

ZÁPADOČESKÁ UNIVERZITA V PLZNI

FAKULTA EKONOMICKÁ

Diplomová práce

Využití umělé inteligence v marketingových aktivitách

Use of artificial intelligence in marketing activities

Bc. Petr Černý

Plzeň 2024

Čestné prohlášení

Prohlašuji, že jsem diplomovou práci na téma

„Využití umělé inteligence v marketingových aktivitách“

vypracoval samostatně pod odborným dohledem vedoucího diplomové práce za použití pramenů uvedených v příložené bibliografii.

Plzeň dne 21.4.2024

v. r. *Bc. Petr Černý*

Zásady pro vypracování práce

1. Charakterizujte umělou inteligenci (AI), strojové učení a její vliv na marketingové aktivity organizací.
2. Zvolte rámec výzkumu.
3. Zpracujte teoretický vstup do řešené problematiky v kontextu zvoleného rámce výzkumu.
4. Připravte a realizujte vlastní výzkum.
5. Na základě zjištěných informací formulujte závěry.

Studijní program

Podniková ekonomika a management

Poděkování

Na tomto místě bych chtěl vyjádřit své upřímné díky vedoucímu mé diplomové práce, panu Ing. Janu Tluchořovi, Ph.D. za jeho cenné rady, odborné vedení a trpělivost, které mi byly po celou dobu studia neocenitelnou oporou.

Poděkování také patří mé rodině a blízkým, kteří mi byli oporou během celého studia.

Obsah

Úvod.....	7
Cíle a metodika	8
1 Umělá inteligence.....	10
1.1 Definice umělé inteligence.....	10
1.2 Rozvoj Umělé inteligence	12
1.3 Koncepty umělé inteligence	14
2 Strojové učení	16
3 Marketing jako disciplína.....	19
3.1 Rozvoj marketingu	19
3.2 Marketing 5.0	21
4 Umělá inteligence a personalizace obsahu	23
5 Zpracování přirozeného jazyka v marketingu	26
5.1 Chatboti a osobní asistenti.....	27
5.2 Velké jazykové modely	30
5.3 Odpovídání na otázky (QA)	32
6 Etické aspekty využití umělé inteligence a strojového učení v marketingu.....	34
6.1 Soukromí a ochrana osobních údajů	34
6.1.1 GDPR	34
6.1.2 Regulace týkající se Umělé inteligence	36
6.2 The New York Times žaloba na společnost OpenAI.....	37
6.3 Environmentální dopad umělé inteligence	38
7 Praktická část	41
7.1 Definování problematiky.....	41

7.2	Představení Fakulty ekonomické Západočeské univerzity v Plzni.....	42
7.3	Postup vývoje chatbota	43
8	Zjišťování preferencí a postojů	45
8.1	Vyhodnocení dotazníku	46
8.1.1	Popis skupiny respondentů	46
8.1.2	Vyhodnocení specifických otázek	50
8.2	Shrnutí dotazníkového šetření	58
9	Návrh a vývoj chatbota	60
9.1	Analýza požadavků na daného chatbota	60
9.2	Návrh	60
9.3	Výběr platformy.....	62
9.4	Použité dokumenty a jejich úprava.....	63
9.5	Vektorová reprezentace slov (Embedding).....	65
9.6	Vektorová databáze.....	66
9.7	Agent.....	67
9.8	Finální systém.....	69
9.9	Uživatelské rozhraní	69
10	Testování navrženého chatbota konverzační agent.....	71
10.1	Porovnání s velkým jazykovým modelem.....	71
10.2	Testování s dalšími prototypy.....	72
10.3	Testování pomocí LangSmith.....	74
10.4	Hodnocení odpovědí	76
10.5	Testování za pomoci Ragas	77
10.6	Shrnutí testování a volba.....	78
11	Testování chatbota studujícími.....	81

11.1	Postup testování.....	81
11.2	Vyhodnocení testování.....	83
11.3	Shrnutí testování.....	84
12	Implementace.....	85
13	Kontrola	88
	Závěr.....	90
	Seznam použitých zkratk.....	92
	Seznam použitých zdrojů.....	93
	Seznam tabulek.....	106
	Seznam obrázků	107
	Seznam příloh	108
	Přílohy	
	Abstrakt	
	Abstract	

Úvod

Umělá inteligence se v současné době stává pilířem inovací napříč mnoha odvětvími včetně marketingu. Využití umělé inteligence v marketingových strategiích a aktivitách umožňuje organizacím analyzovat obrovské množství dat a nalézt tak nové vzorce, které vedou k lepšímu porozumění chování zákazníků a jejich preferencím. Umělá inteligence otevírá dveře řadě nových forem personalizace, umožňuje vytvářet předpovědi na základě historických dat a přináší nové druhy automatizace, jejíž aplikace lze nalézt u řady rutinních a opakujících se úkolů. K vývoji dochází napříč jednotlivými koncepty umělé inteligence jako je například zpracování přirozeného jazyka, které umožňuje lépe identifikovat sentiment, a utvářet tak zprávy, které jsou upravené na základě určitého kontextu.

V rámci tohoto vývoje zpracování přirozeného jazyka se chatboti jeví jako klíčový inovativní prvek, který ukazuje možnosti využití umělé inteligence k posílení vztahů mezi značkami a jejich zákazníky. Zapojení prvků umělé inteligence vylepšuje schopnosti chatbotů natolik, že nyní dokáží udržet lidskou konverzaci na místo výběru určitých předem připravených odpovědí. Tento pokrok významně zlepšuje zákaznickou zkušenost a v kontextu zákaznické podpory se zákazníci cítí více vyslyšeni a oceněni.

Téma této diplomové práce je „Využití umělé inteligence v marketingových aktivitách“. Po domluvě s vedoucím práce je praktická část zaměřena na vývoj a testování chatbota pro potřeby Fakulty ekonomické. Chatbot bude navržen tak, aby podporoval činnosti v oblasti studijní administrativy a akademické podpory, což zahrnuje interakci se studenty a zjednodušení administrativních procesů na fakultě.

Cíle a metodika

Hlavním cílem této práce je vyvinout chatbota, který bude sloužit jako podpora pro studenty ve studijní administrativě a akademickém poradenství na fakultě. Pro splnění hlavního cíle je potřeba stanovit následující dílčí cíle:

1. Vymežit koncept chatbotů a spojení s umělou inteligencí,
2. Ověřit, jaký by byl zájem studentů o trvalou implementaci chatbota a zjistit jaká by byla úroveň důvěry studentů v odpovědi daného chatbota.
3. Vytvořit fakultního chatbota.
4. Analyzovat frekvenci halucinací chatbota během dotazování.
5. Vyčíslit, jaké jsou finanční náklady spojené provozem chatbota pro školní účely.
6. Posoudit, jak vnímají studenti přesnost a užitečnost informací poskytovaných chatbotem.
7. Zhodnotit míru použitelnosti daného fakultního chatbota.
8. Sestavit předběžný plán pro implementaci a následnou kontrolu.

Tato práce je strukturovaná do dvou hlavních částí: teoretické a praktické. První část je vypracována na základě odborné literatury a pokrývá základní koncepty, které jsou kritické pro tuto práci. Kapitoly v této části se zabývají konceptem umělé inteligence, strojovým učením, obecným představením marketingu, zpracováním přirozeného jazyka, chatboty společně s asistenty a v poslední řadě s etickými aspekty, které jsou spojené s využíváním umělé inteligence.

Praktická část je rozdělena na několik klíčových částí. Nejdříve je definována samotná problematika, která představuje důvody proč je následná praktická část sestavena. Následuje samotné dotazníkové šetření, návrh a vývoj chatbota, testování navrženého chatbota a následné testování chatbota se studenty fakulty. V poslední řadě je představen plán implementace a prvky, které je po implementaci potřeba kontrolovat.

Dotazníkové šetření v této práci je využito ke sběru dat od studentů a k zjištění postojů a preferencím k umělé inteligenci a chatbotovi fakulty. Ve vytvořeném dotazníku jsou využity zejména uzavřené a polouzavřené otázky. Využitá metoda dotazníkového šetření je detailněji rozepsána v kapitole osm.

K získání potřebných dat při testování chatbota se studenty je využito speciálního dotazníku použitelnosti CUQ. Metodický popis sběru dat a postup testování chatbota studujícími, včetně použitých nástrojů, je podrobně rozpracován v kapitole jedenáct této práce.

1 Umělá inteligence

V rámci této kapitoly je popsána potřebná teorie týkající se umělé inteligence, často nazývané také jako artificial intelligence (AI). Bude probírána jak různorodost definic, rozvoj umělé inteligence, tak i uvedení konceptů umělé inteligence, které se hojně využívají ve světě.

1.1 Definice umělé inteligence

Definice pojmu umělé inteligence jsou různorodé, některé se opírají spíše filozofickou stránkou tohoto konceptu, jiné zase o stránku technickou. Definice se také měnily v průběhu času, a to zejména kvůli rapidnímu vývoji v této oblasti.

Rozšířená učebnice na téma umělé inteligence s názvem *Artificial Intelligence: A modern Approach* při definici toho konceptu uvádí 4 skupiny, do kterých rozřazuje definice od různých autorů. Skupiny nesou následující názvy:

- Myslet lidsky,
- Myslet racionálně,
- Jednat lidsky,
- Jednat racionálně. (Russell & Norvig, 2014)

Tyto skupiny jsou rozděleny do dvou dimenzí, kdy první dvě skupiny zastřešují definice týkající se myšlenkovými procesy a uvažováním. Spodní dvě skupiny na tomto listu se zabývají jednáním/chováním.

Lze také uvést definici od jednoho z „otců“ umělé inteligence Johna McCarthyho: „Umělá inteligence je věda a technika o vytváření inteligentních strojů, zejména inteligentních počítačových programů.“¹. (McCarthy, 2007, Přeloženo autorem)

V rámci institucionálního zapojení do této problematiky je možné také představit modernější definici orgánu Evropské unie. Evropský parlament uvedl v roce 2020 následující definici pro umělou inteligenci: „Umělá inteligence – Artificial intelligence

¹ It is the science and engineering of making intelligent machines, especially intelligent computer programs.

(AI) - je schopnost strojů napodobovat lidské schopnosti, jako je uvažování, učení se, plánování nebo kreativita“ (Evropský parlament, 2020)

Typy Umělé inteligence

Umělou inteligenci lze rozdělit do 3 skupin, a to dle úrovně propracovanosti, schopností a použití. Zní následovně:

- Artificial narrow intelligence (ANI) (Úzká umělá inteligence)
- Artificial general intelligence (AGI) (Obecná umělá inteligence)
- Artificial super intelligence (ASI) (Umělá superinteligence) (Nancholas, 2023)

Úzká umělá inteligence, často také nazývaná jako slabá umělá inteligence, je speciálně navržena k provádění předem definovaných úkolů, a to bez schopnosti rozšířit své učení nad rámec naprogramovaných funkcí.

Příkladů v této oblasti existuje řada, v rámci virtuálních osobních asistentů lze zmínit Alexu od společnosti Amazon, nebo Siri od společnosti Apple. Další populární příklady existují v překladatelských službách, a to například Deepl. Zmínit lze také společnost Netflix a další poskytovatele obdobných služeb, kteří využívají umělou inteligenci k doporučení relevantního obsahu. Velké jazykové modely jako například Chat GPT – 3 od společnosti OpenAI, nebo Bard od společnosti Google také spadají do této skupiny umělé inteligence. Existuje řada dalších oblastí, ve kterých se systémy slabé umělé inteligence využívají. (Nancholas, 2023)

Obecná umělá inteligence – je teoretická podoba umělé inteligence „schopná řešit libovolný počet hypotetických úloh pomocí zobecněných lidských kognitivních schopností.“ (Bell, 2023)

Jak již bylo uvedeno výše, odvětví obecné umělé inteligence je stále pouze teoretický koncept, existuje zde tedy určitý nesoulad mezi názory odborníků v této oblasti o tom, jak by tento koncept mohl fungovat, dokonce o tom, zdali je vůbec možné takovýto systém sestavit. I přesto lze zjednodušeně AGI popsat jako systémy umělé inteligence, které disponují přiměřenou úrovní sebeporozumění a samostatné sebekontroly, schopností řešit problémy v různých kontextech a schopností naučit se řešení nových problémů, které nebyly při jejich vývoji předpokládány. (Bell, 2023)

Umělá superintelligence – Pojem umělé superintelligence je představován jako systém, který přesahuje lidskou inteligenci a překonává lidské schopnosti, a to na poli většiny oblastí. Ačkoliv se stejně jako předchozí koncept jedná pouze o teoretický koncept, zůstává v oblasti umělé inteligence jakožto předmětem rozsáhlých debat, úvah a sporů. Jelikož by se mělo jednat o skutku převratný systém, mluví se potencionální změně celého světa, jak ho doposud známe. Naproti této logice se staví řada odpůrců, kteří s tímto konceptem mají spojená veliká rizika například ztráta kontroly lidstva nad stroji. (Nancholas, 2023)

1.2 Rozvoj Umělé inteligence

Počátky konceptu umělé inteligence

Jak již bylo nastíněno v předchozí kapitole, umělá inteligence byla poprvé definována v první polovině 20. století, a to vědci a matematiky, ale také filozofy. Alan Turing, britský vědec, je považován za jednoho z prvních, který se touto problematikou zabýval a zkoumal možnost existence tohoto konceptu. Turing polemizoval nad faktem, že lidé dostupné informace určitým způsobem zpracovávají a na základě toho řeší určité problémy a proč by tedy to samé nemohli dělat i stroje. V roce 1950 vydal na toto téma vědeckou práci s názvem „Computing Machinery and Intelligence“, ve které zmíněnou problematiku řeší a diskutuje o tom, jak lze inteligentní stroje vytvořit a následně i otestovat. Turing neměl možnost koncept, o kterém teoretizoval, prakticky vyzkoušet, jelikož k tomu byla potřeba razantní změna v počítačové oblasti. Existovali totiž dva významné problémy, prvním problémem byla neexistence možnosti ukládání příkazů, pouze samotné vykonání. Druhý problém se týkal nákladů, jelikož pronájem na potřebnou techniku dosahoval částek až 200 000 dolarů na měsíc, jež si mohla dovolit jen hrst prestižních institucí. (Anyoha, 2017)

První program a umělá inteligence v období od 70. let do roku 2000

Pět let po Turingovo práci začala iniciativa tento koncept ověřit. Allen Newell, Clifford Shaw a Herbert Simon vytvořili první program, který napodoboval způsob, jakým lidé řeší problémy. Tento program byl prezentován na konferenci o umělé inteligenci na Dartmouthské univerzitě, kterou pořádal již zmíněný John McCarthy a Marvin Minsky a je považován jako první program umělé inteligence. (Buchanan, 2005)

V 70. letech minulého století docházelo k rozkvětu konceptu umělé inteligence, a to díky stále lepší a levnější technologii. Počítače již dokázaly ukládat mnohem více informací a pracovat o mnohem rychleji. Stejně tak docházelo k výraznému zdokonalení algoritmů strojového učení. První modely jako například výše zmíněný univerzální stroj na řešení problémů a také program ELIZA, který dokázal s uživatelem vést jednoduchou komunikaci. Tyto programy a samotní vědci v tomto oboru přiměly americké vládní agentury k financování výzkumu umělé inteligence. Optimismus byl veliký, ale po prolomení počáteční mlhy došlo k zjištění několika významných problémů. Hlavním problémem byl nedostatečný výpočetní výkon, který by umožnil cokoli významného. Jelikož výzkum pokračoval z tohoto důvodu pomalu, docházelo k úbytku finančních prostředků a celý výzkum se zpomalil. (Anyoha, 2017)

V osmdesátých letech se umělá inteligence znovu dostala do popředí, a to ze dvou hlavních důvodů. Prvním důvodem bylo rozšíření konceptu deep learning, který umožňuje počítačům se učit na základě zkušeností. Druhým konceptem, který byl představen v tomto období jsou takzvané expertní systémy. Tyto systémy dokáží napodobit chování lidského experta v dané oblasti. Druhým důvodem bylo zvýšení financí v odvětví umělé inteligence, a to zejména v Japonsku. Japonská vláda se rozhodla výrazně investovat především do expertních systémů, a to cílem dosáhnout revoluce ve zpracování dat a celkově zdokonalit umělou inteligenci. Navzdory tomuto vývoji se znovu umělá inteligence dostala do ústraní, protože většina velmi ambiciózních cílů nebyl splněna. (Anyoha, 2017)

I přes nedostatek finančních prostředků ke konci 80. let se podařilo dostat koncept umělé inteligence v 90. letech znovu do popředí. Významným triumfem v této oblasti bylo vytvoření AI, která dokázala porazit světového šampiona Gary Kasparova v šachách. Deep Blue, takzvaný superpočítač navržený společností IBM, dokázal pomocí algoritmu zpracování stromu vyhodnotit až 200 miliónů pohybů za sekundu a stát se tak prvním programem, který dokázal porazit člověka. V roce 1997 vydal Windows software pro rozpoznávání řeči. (Tableau, n.d.)

1.3 Koncepty umělé inteligence

Z řady různých definic, je zřejmé, že umělou inteligenci lze rozdělit do řady konceptů a kategorií. V následující podkapitole budou vyjmenovány jednotlivé koncepty umělé inteligence a ve zkratce vysvětleny.

1. **Strojové učení (Machine learning)** - Prvním a jedním z nejznámějších konceptů umělé inteligence je strojové učení. Umělá inteligence se častokrát vysvětluje společně se strojovým učení. Strojové učení je podobor informatiky, ale také se řadí jako podřád umělé inteligence. Strojové učení se podstatě zabývá vytvářením určitých algoritmů, které se zakládají na určitém souboru příkladů nějakého jevu. (Burkov, 2019)
2. **Neuronové sítě/hluboké učení** – Neuronovou síť lze popsat jako sérii algoritmů, které se snaží rozpoznat určité vztahy v daném datovém souboru, a to skrze procesy, které napodobují chování člověka. Výhodou neuronových sítí je fakt, že se dokáží přizpůsobit vstupním datům, což znamená, že zde nevychází požadavek přepracování vstupních kritérií. (Mahesh, 2020)
3. **Zpracování přirozeného jazyka** – Dalším odvětvím umělé inteligence je Zpracování přirozeného jazyka v angličtině zvaném Natural Language Processing (zkráceně NLP). Zpracování přirozeného jazyka je v podstatě souhrn výpočetních technik, které slouží k automatické analýze a zobrazení různých lidských jazyků. Existuje řada aplikací tohoto odvětví jako například: kontrola pravopisu, kategorizace informací, odpovídání na otázky, překlad textu a další. (Kavlakoglu, 2020)
4. **Robotika** – V oblasti robotiky lze sledovat takzvané strojové smysly neboli například sluch, zrak a hmat. Tyto smysly se následně mimo jiné techniky implementují v rámci oblasti robotiky. Umělá inteligence je v oblasti robotiky klíčová především, protože umožňuje strojům využívat zmíněné smysly a tím pádem tedy vnímat, učit se, rozhodovat se a provádět určité úkony. (Chowdhary, 2020)
5. **Počítačové vidění** – Počítačové vidění neboli strojové vidění je dalším oborem umělé inteligence, který se velmi rapidně rozvíjí. Cílem tohoto odvětví AI je porozumět k čemu dochází před danou kamerou a na základě toho použít toto

porozumění například k ovládnání počítače či robotické systému. Technologie počítačového vidění je schopná vytvářet nové vizuální materiály z existujících obrázků a přetvářet je do forem, které jsou buď informativnější, nebo esteticky atraktivnější. (Chowdhary, 2020)

- 6. Expertní systémy** – Expertní systém je počítačový program, který bývá určen jako komponent člověka či organizace, který má určité zkušenosti v dané oblasti. Expertní systémy se snaží za pomoci využívání technologie umělé inteligence k simulaci úsudku a způsobu jednání zmíněného experta či organizace. (Lutkevich, 2023)

2 Strojové učení

Strojové učení je obor umělé inteligence, který se zabývá získáváním znalostí z určitých dat. Metody strojového učení jsou v každodenním životě v posledních letech hojně využívány. Například lze uvést automatická doporučení určitého obsahu, jako jsou filmy, seriály ale i jídlo, které si lze objednat přes určitou mobilní aplikaci. Strojové učení neboli spíše modely strojového učení lze sledovat na spoustě mezinárodně populárních webových stránkách jako je Amazon, Netflix a mnoho dalších. (Müller & Guido, 2016)

Strojové učení je postaveno na řadě vědních disciplín. Stejně tak je naopak strojové učení důležitým nástrojem pro řadu vědních oborů. Vztahy k ostatním oborům lze například sledovat ve spojení **k matematice**, a to přesněji ve spojení s lineární algebrou. Další důležitá vědní disciplína, na které je strojové učení postavené, je **počítačová věda**. Metody strojového učení využívají sérii výpočetních operací na vstupní data. Výsledek těchto operací je určitá predikce neboli předpověď. Interpretace strojového učení jakožto výpočetních systémů umožňuje využít teoretické znalosti a nástroje právě z počítačové vědy. Další dvě důležitá odvětví jsou **statistika** společně s **teorií pravděpodobnosti**. Statistika je pro koncept strojového učení silný základ při sestavení řady algoritmů. Teorie pravděpodobnosti nabízí určité metody pro odhad pravděpodobnostního rozdělení z daného souboru. Teorie pravděpodobnosti tedy v podstatě napomáhá k porozumění nejistoty a náhodnosti v daném souboru dat. (Jung, 2022)

Teorie informací zkoumá efektivní komunikaci zpráv, a to navzdory šumu či jiným nedokonalostem v rámci komunikačního kanálu. V rámci strojového učení je teorie informací užitečná při analýze a návrhu metod strojového učení. Teorie informací v podstatě řeší 2 problémy v rámci strojového učení, a to v případě vytváření predikce a v rámci kroku učení daného modelu, upravením či zlepšením dané hypotézy. V poslední řadě je nutné zmínit **optimalizaci**. Základním požadavkem či principem při návrhu metod strojového učení je definovat tyto problémy jako optimalizační problémy. Optimalizace tedy v kontextu strojového učení má klíčovou roli v procesu trénování modelů a jejich následném efektivním využití. Metody strojového učení jsou využívány na různých úrovních. V rámci nižší úrovně se lze bavit o extrakci dané relevantní informace ze vstupních dat. Metody strojového učení lze také využít ke klasifikaci obrázků do definovaných kategorií. Tyto výstupy se následně využívají pro další analýzu či jiné

funkce na vyšších úrovních. Strojové učení lze nalézt v rámci vyšších úrovní například při zpracování přirozeného jazyka či detekování určitých objektů. (Jung, 2022)

Koncepty strojového učení

Strojové učení se nejčastěji rozděluje na 3 kategorie – učení s učitelem (supervised learning), učení bez učitele (unsupervised learning) a zpětnovazební učení (reinforcement learning) Zmínit lze také méně využívanou kombinaci učení s učitelem a učení bez učitele.

Učení s učitelem (supervised learning)

Kategorie učení s učitelem je postavená na určité trénovací množině, která je složena z anotovaných datových bodů. Při učení s učitelem je daným algoritmem zajištěné dostatečné množství datových bodů, které jsou anotované a následně jsou určeny proměnné pro posouzení korelací. (Tucci, n.d.)

Strojové učení s učitelem se v podstatě snaží nalézt takovou hypotézu, která je schopna napodobit chování lidského anotátora neboli člověka, který například spolupracoval na přípravě trénovacího datového souboru. Tento koncept se snaží předpovědět určité označení/anotaci podle vlastností daného datového bodu. (Jung, 2022)

Algoritmy učení s učitelem se používají pro řadu úloh včetně těchto základních:

- Binární klasifikace – tato klasifikace rozděluje data do 2 kategorií,
- Klasifikace více tříd – při této klasifikaci dochází k výběru z více než dvou odpovědí,
- Sestavování – zde dochází ke kombinaci predikcí z řady modelů strojového učení a následně k vypracování přesnější předpovědi,
- Regresní modelování – jedná se o předpověď spojité veličiny na základě vztahů mezi daty. (Tucci, n.d.)

Učení bez učitele (unsupervised learning)

U učení bez učitele je daný data set souborem neanotovaných příkladů neboli výše zmíněných datových bodů. Metody učení bez učitele se tedy zcela spoléhají na vlastní strukturu daného souboru datových bodů pro určení kvalitní hypotézy. (Burkov, 2019)

Mimo jiné existují, mimo jiné, dvě hlavní metody učení bez učitele – shlukování a redukce dimenzionality. V rámci shlukování dochází k rozdělení datových bodů, které se nacházejí ve využívaném datovém souboru, do menšího počtu podskupin, a to na základě určité podobnosti mezi nimi. Redukce dimenzionality spočívá ve snížení počtu proměnných v daném datovém souboru. (Jung, 2022),

Učení s částečným dohledem

V případě učení s částečným dohledem neboli výše zmíněná kombinace předchozích dvou kategorií, pracuje s datovou sadou, která obsahuje jak anotované, tak neoznačené datové body. Ve většině případů tato datová sada obsahuje větší množství neoznačených datových bodů. Cíl u této metody je v podstatě stejný jako u metody učení s učitelem, ale s tím rozdílem, že za použití velkého množství neoznačených datových bodů, je zde snaha pomoci učicímu algoritmu najít lepší model. (Burkov, 2019)

Zpětnovazební učení (reinforcement learning)

Poslední zmíněná metoda strojového učení je zpětnovazební učení. Jedná se v podstatě o techniku, která se určitým způsobem liší od předchozích metod tím, že pracuje na způsobu pokusů a omylů. Zpětnou vazbu následně získává z vlastních akcí a zkušeností. Rozdíl mezi touto metodou a metodou učení s učitelem je takový, že zpětná vazba u metody učení s učitelem je poskytnuta ve formě správných kroků pro provedení daného úkolu. U metody zpětnovazebního učení je využíváno odměn a trestů jako signálů pro příznivé a nepříznivé chování. Při porovnání s učením bez učitele je zde rozdíl v rámci cílů. U učení bez učitele je cílem identifikovat podobnosti a rozdíly mezi datovými body, u metody zpětnovazební je cílem maximalizace kumulativní odměny neboli nalézt takový sled akcí, za které dostane nejvyšší celkovou odměnu. (Burkov, 2019)

3 Marketing jako disciplína

V této části bude představena disciplína marketingu, rozvoj marketingu s největším důrazem na etapu marketing 5.0. Rozvoj marketingu bude představen na etapách, které zmiňuje Philip Kotler ve svých publikacích. Hlavní zaměření na etapu marketingu 5.0 je z toho důvodu, že v této etapě je nejvíce zmiňována právě umělá inteligence společně s další novou technologií. I přes vysoký dopad těchto technologií na marketing zde stále figuruje pozice lidského marketéra jako kritický pilíř pro interpretaci chování zákazníka.

Existuje řada způsobů, jak definovat marketing, buďto lze využít některou z řad akademických publikací od odborníků v tomto odvětví, či od významných institucí. Pro účely této práce byla využita definice z české publikace profesora Světlíka a zní následovně: „Marketing je proces řízení, jehož výsledkem je poznání, předvídání, ovlivňování a v konečné fázi uspokojení potřeb a přání zákazníka efektivním a výhodným způsobem zajišťujícím splnění cílů organizace.“ Rozdíly mezi dalšími definicemi mohou být různorodé, vždy mají ale společného jednoho účastníka, jímž je zákazník a určité uspokojování potřeb a přání všem zainteresovaným osobám v dané transakci. (Světlík, 2018, s. 7)

3.1 Rozvoj marketingu

Rozvoj marketingu lze popsat různými způsoby, v této práci bude využito rozdělení do pěti fází, které zmiňuje jeden z nejznámějších autorů v oblasti marketingu Philip Kotler. Společně s rozvojem marketingu lze uvést 5 generací od Baby Boomers až po generaci narozenou po roce 2010, která nese název Alfa.

Marketing 1.0 neboli první etapa marketingu, kterou lze nalézt v řadě publikací je převážně produktově zaměřena. Tento produktově orientovaný marketing vznikl v 50. letech minulého století ve Spojených státech amerických a byl hlavně zaměřen na zmíněnou generaci Baby Boomers (Lidé narození o roku 1945-1965). Slouží hlavně bohatým lidem této generace a hlavním cílem je produkovat perfektní produkty a služby, které přinášejí maximální hodnotu pro zákazníka. Společnosti se snaží nabídnout svým zákazníkům optimální výhody, což nevyhnutelně vede k požadavku na vyšší ceny v delším časovém období. Z tohoto důvodu základní marketingové koncepce, které se

v této éře objevily, kladly důraz na vývoj produktu, řízení životního cyklu a vytváření čtyř P neboli produkt (product), cena (price), místo (place) a propagace (promotion). Hlavním cílem se stala spokojenost zákazníka. (Kotler a kol., 2021)

Marketing 2.0 – Společenské a obchodní prostředí se změnilo zejména v důsledku technologického pokroku, procesu globalizace a privatizace, což následně vedlo k nástupu marketingu 2.0. Marketing 2.0 je často charakterizován určitým přechodem k zákazníkem orientovanému přístupu. Marketing 2.0 bývá také charakterizován či spojován s využitím nástrojů webu 2.0. Web 2.0 přináší nové interaktivní komunikační kanály, a to mezi zákazníky a podniky. Facebook, YouTube a další jsou v této fázi využívány pro sběr a uspořádání dat. Následně jsou vyhodnoceny demografické informace o spotřebitelích, společně s určením určitých trendů na trhu, pro stanovení strategií. V tomto přístupu tedy dochází k upřednostnění cesty k pochopení zákazníka a jeho potřeb. (Mehta, 2022)

Marketing 3.0 – Nástup generace Y, společně s globální recesí na konci roku 2000 vyvolalo zásadní proměnu v oblasti marketingu. Podniky, které přednostně nadřazovaly zisk před společenskou či environmentální odpovědností či přínosem, pocíťovaly sníženou důvěru od nově nastupující generace Y. V souladu s tím, aby se obnovila důvěra této generace, musí společnosti vytvořit takové nabídky a kulturu společnosti, které budou klást důraz na pozitivní společenské a environmentální výsledky. Fáze marketingu 3.0 neboli orientovaného na člověka začala v momentě, kdy společnosti začaly vyvíjet a integrovat společensky odpovědné marketingové postupy a strategie do svých business modelů. (Kotler a kol., 2021)

Marketing 4.0 – Marketing 4.0 je častokrát označován jako revoluční. Nejčastěji se hovoří o integraci či spojení sféry online a offline interakcí, a to mezi podniky a zákazníky, plynulou integraci stylu a obsahu při vývoji značky. Jedná se o kombinaci tradičního a digitálního marketingu, tak, aby jejich koexistence získala větší přízeň zákazníka. (Kotler a kol., 2016)

Razantně větší sběr dat je následně využíván k lepšímu uspokojování zákazníků. Rozhodování založené na datech je možné provádět na základě důkladných analýz dat k pochopení preferencí, či trendů. Následně dochází k lepšímu uspokojování zákazníků. Zohledněn je v této fázi také vliv sociálních sítí na značku společnosti neboli způsob,

jakým zákazník využívá sociální média a jestli produkt podpoří či kritizuje. Samotná nákupní cesta je komplexnější. Z tohoto důvodu se využívá řada digitálních způsobů, jak přesvědčit zákazníka, jako například pomocí obsahového marketingu, či pomocí UGC (User Generated Content) obsahu vytvořeným uživateli/zákazníky společnosti. (EDC Paris Business School, 2022)

3.2 Marketing 5.0

Poslední etapu, kterou Philip Kotler zmiňuje ve své nové knize a nese tak i stejný název, je Marketing 5.0. Marketing 5.0 lze uvést pomocí následující definice: „představuje použití technologií napodobujících člověka k vytváření, komunikaci, poskytování a zvyšování hodnoty v rámci celé cesty k zákazníkovi.“² (Kotler a kol., 2021) (Přeloženo autorem)

Etapa Marketing 5.0 je tedy silně spojená s novými nastupujícími technologiemi mezi něž spadá již zmíněná umělá inteligence ale také koncept NLP (zpracování přirozeného jazyka), robotika, AR (rozšířená realita), VR (virtuální realita), IoT (internet věcí) a blockchain. Zmíněná umělá inteligence hraje významnou roli při učení se ze zákaznických dat a zejména tedy u nestrukturovaných dat. Výstupy, které vycházejí z těchto dat, jsou následně využívány k vytváření personalizovaných nabídek správným zákazníkům. (Kotler a kol., 2021)

Společně s tímto přístupem využívání umělé inteligence je stále více v oblibě takzvaný marketing segmentů jednoho. Tento koncept existuje již řadu let a jedná se o spojení konceptů vyhledávání informací a poskytování služeb. Využívá se zde vlastní databáze preferencí a nákupního chování zákazníků a na druhé straně přístup řízeného poskytování služeb. Následně dochází k poskytnutí přizpůsobeného balíčku služeb jednotlivým zákazníkům. (Winger & Edelman, 1989)

Existuje řada reálných aplikací Marketingu 5.0 jako například využití umělé inteligence pro personalizaci emailových zpráv. Dále lze zmínit analýzu nákupních zvyklostí

² Marketing 5.0, by definition, is the application of human-mimicking technologies to create, communicate, deliver, and enhance value across the customer journey.

zákazníků anebo dynamickou segmentaci zákazníků. Umělá inteligence má již také významný vliv na tvorbu obsahu, a významně také na celý koncept chatbotů, u kterých se významně využívá umělá inteligence, specificky NLP pro generaci zpráv. (Haleem a kol., 2022)

Je nutné podotknout, že ačkoliv je v etapě Marketing 5.0 významný důraz na technologie, v ústřední pozici je stále lidský faktor a lidská inteligence. Využití umělé inteligence v marketingu umožňuje objevit nové modely chování zákazníků z obrovského množství dat. Samotná interpretace těchto modelů nebo vzorců je ale stále v rukou marketérů, kteří dokážou pochopit a následně využít motivy při kterých zákazníci jednají. Role lidského marketéra je stále kritická v etapě Marketingu 5.0. a v rámci využití technologie a lidského faktoru v této etapě je nejdůležitějším krokem správná kombinace, kam by stroje a lidé přinesli nejlepší hodnotu pro zákazníka. (Kotler a kol., 2021)

4 Umělá inteligence a personalizace obsahu

V marketingu se personalizací rozumí využití dat k přizpůsobení zpráv podle konkrétních preferencí jednotlivých uživatelů. Personalizovaný marketing má pro firmy jednoznačné výhody, ať už se to týká snížení nákladů na získání a udržení zákazníka, nebo zvýšení celkových příjmů a v neposlední řadě zvýšení návratnosti investic do marketingu (ROI). Existuje řada výhod, které tento koncept přináší a naléhavost na implementaci toho konceptu se jenom zvýšila při pandemické krizi COVID-19. Personalizace má ale také své stinné stránky, jelikož zde může dojít k poskytnutí příliš relevantního obsahu, jež může být pro uživatele znepokojující. (McKinsey & Company, 2023)

Je také nutné zmínit rozdíl mezi problematikou personalizace (personalization) a přizpůsobením (customization). Personalizace, jak již bylo zmíněno na začátku této podkapitoly, nastává, když se firma rozhoduje, jaký marketingový mix je nejvhodnější pro daného uživatele či potenciálního zákazníka, a to na základě již získaných dat o zákaznících. Přizpůsobení se liší v tom smyslu, že zde daný zákazník určitým způsobem aktivně specifikuje alespoň jeden z prvků marketingového mixu. (Kumar et al, 2019)

V digitální sféře v rámci personalizace lze zmínit řadu zástupců, kteří využívají různé komplexní techniky tohoto konceptu jako například Amazon při doporučování produktů. Sekce „doporučení pro vás“, které ukazují již výsledky personalizace, které využívá zmíněná společnost Amazon také využívá řada dalších společností. Jako další významné příklady lze zmínit například Netflix, Spotify nebo Pandora. Jako příklad v offline prostoru se dá uvést firma Sprint, která aplikuje prediktivní analytiku pro vytváření personalizovaných strategií pro udržení klientů. Dochází tedy k personalizaci nabídek, a to z toho důvodu, že zde hrozí odchod daného zákazníka. (Kumar et al, 2019)

Umělá inteligence se využívá v konceptu personalizace z důvodu vysoké míry personalizace. Implementace umělé inteligence do konceptu personalizace změnila tradiční způsob řešení personalizace takzvaného přístupu založeného na pravidlech, které spadají do expertních systémů k přístupu hlubokého učení. AI dokáže pracovat tak diskrétně, že si uživatel častokrát ani neuvědomí, že určitým způsobem komunikuje s technologií. Technologie dokáží pracovat na osobní úrovni, což podporuje silný vztah. V marketingu lze takovéto spojení využít pro vytváření hodnoty pro zákazníky a její

potenciál je obrovský. Nicméně úspěch personalizačních snah je limitován množstvím a kvalitou zákaznických dat, schopností společnosti získat z těchto dat poznatky a efektivní implementací těchto poznatků. K překonání této problematiky a k pozdvižení personalizované nabídky na lepší úroveň se nyní podniky obracejí na řešení, které jsou poháněné AI. (Kumar et al, 2019)

Systémy pro doporučení

Systémy doporučení je řada algoritmů nebo AI, který je většinou spojován se strojovým učením, a to z toho důvodu, že zde dochází k využití velkého souboru dat k tomu, aby doporučoval další produkty či jiné elementy v souladu s daným odvětví využití. Kritéria, pomocí kterých dochází k doporučení, mohou být různorodá, například demografické údaje či předchozí nákupy. (Nvidia, n.d.a)

Systémy doporučení lze rozdělit do 3 obsáhlých skupin, a to i za předpokladu, že existuje celá řada algoritmů, které pracují s problematikou doporučení. Tyto skupiny se nazývají následovně:

- Kolaborativní filtrování,
- Obsahové filtrování,
- Kontextuální filtrování.

Kolaborativní filtrování jsou algoritmy, které fungují tak, že navrhnou položky na základě preferencí více uživatelů, přičemž využívají podobnosti preferenčního chování, které je podloženo jejich předchozí interakcí s danými položkami. Tato skupina systému doporučování se vyznačuje tím, že předpovídá budoucí interakci, a to na základě historických dat. Základní koncepce spočívá v tom, že pokud jednotlivci dříve učinili podobné volby a nákupy, například v případě filmových preferencí, existuje velká pravděpodobnost, že se budou v budoucnu při výběru shodovat. (Nvidia, n.d.a)

Tyto systémy mohou fungovat na základě explicitní nebo implicitní zpětné vazby, kdy explicitní zpětná vazba častokrát představuje kvalitnější cestu, jelikož se jedná o typ zpětné vazby jako například hodnocení uživatele daného produktu pomocí hvězdičkového či obdobného hodnocení. Implicitní zpětná vazba tedy znamená, že je potřeba názor získat pozorováním chování uživatelů, častokrát je nutné využít tuto cestu, jelikož explicitní vazba není vždy k dispozici. (Koren et al, 2022).

Obsahové filtrování vychází z vlastností daného produktu (obsahu) a následně navrhuje podobné položky, které odpovídají preferencím daného uživatele. Vychází se tedy z podobnosti mezi vlastnostmi daného obsahu a zohledňuje informace o daném uživateli, ale také jeho interakce s daným obsahem. Lze uvést příklad, kdy daný uživatel sledoval v minulosti dva filmy, může daný systém doporučit další film, který spadá do stejného či podobného žánru či je zde podobné obsazení. (Nvidia, n.d.a)

Kontextuální filtrování, často také nazývané jako doporučení s ohledem na kontext, řeší problematiku dvou předchozích skupin i jejich kombinace. Předchozí skupiny nezohledňují kontextuální informace, například časový kontext. Různé typy kontextového filtrování využívají různé světově známé společnosti jako je Netflix či Spotify. Netflix například využívá následující kontextuální informace – lokace, typ zařízení, které je využíváno ke sledování obsahu, počasí, eventy v okolí, ale také určitý denní režim daného uživatele. Ve společnosti Spotify se pro doporučování hudby využívají kontextové informace jako nálada, dny v týdnu, lokace nebo region a podobně jako v předchozím případě také typ použitého zařízení. (Adomavicius e al, 2022)

Další využití umělé inteligence v marketingových nástrojích lze nalézt v **příloze G**.

5 Zpracování přirozeného jazyka v marketingu

Zpracování přirozeného jazyka (v angličtině Natural language processing – NLP, v této práci bude využívána tato zkratka), jak již bylo nastíněno v jedné z předešlých kapitol, využívá techniky strojového a hlubokého učení pro následné plnění určitého úkolu, jako například zodpovězení otázek či překlad textu. Je nutné také zmínit, že často používaný termín generování přirozeného jazyka (Natural language generation – NLG) je podmnožinou zpracování přirozeného jazyka a umožňuje v podstatě vytvářet samotné textové odpovědi, a to v lidském jazyce. Druhou podmnožinou NLP, která využívá syntaktickou a sémantickou analýzu k porozumění významu daného textu se nazývá porozumění přirozeného jazyka (Natural language understanding – NLU). (Kavlakoglu, 2020)

Jak již bylo několikrát zmíněno v této práci, množství nestrukturovaných dat neustále narůstá a umělá inteligence je jedním z nástrojů či konceptů, který dokáže nalézt určité vzorce, jenž byly dříve neznámé. Nárůst na popularitě této skutečnosti způsobil, že v dnešní době existuje řada aplikací, které využívají nestrukturovaná data v marketingu. Text bývá jednou ze základních a také nejvíce využívaných forem data typu a je to zřejmě dáno tím, že je řádově jednodušší získat a následně zpracovat tuto formu. (Hartmann & Netzer, 2023)

Metody NLP získaly také na významu s prudkým nárůstem výpočetního výkonu a rychlosti zpracování. Milníky, které se podařilo splnit lze uvést od jednoduchých statistických modelů textů pomocí četností slov až po jazykové modely Generative Pretrained Transformer – 3 (GPT - 3), nebo nejnovější multimodální velký jazykový model Generative Pretrained Transformer – 4 (GPT - 4). (Shankar & Parsana, 2022)

Aplikace NLP v marketingu se neustále rozšiřuje, NLP má v marketingu různorodé využití, včetně vyhodnocování kreditu zákazníků, sledování aktivity na sociálních sítích, zavádění chatbotů a vytváření osobních asistentů. (Shankar & Parsana, 2022) Analýza sentimentu je jedna z dalších aplikací, při které se využívá konceptu NLP.

5.1 Chatboti a osobní asistenti

Na úvod je nejprve důležité uvést samotnou definici chatbota, která může znít následovně: „Chatboti jsou inteligentní konverzační počítačové programy, které napodobují lidskou konverzaci v její přirozené podobě.“. (Caldarini e al, 2022, s. 1) V kontextu umělé inteligence lze zmínit ještě následující definici: „Chatboti jsou programy, které napodobují lidskou konverzaci pomocí umělé inteligence.“. (Ranoliya et al., 2017)

Tyto konverzační počítačové programy, využívají zpracování přirozeného jazyka, ale také koncept analýzy sentimentu, a to, aby mohl komunikovat podobně jako lidé prostřednictvím buďto psaného nebo mluveného jazyka. Komunikovat může jak s lidmi, tak i s jinými chatboty. Interaktivní agenti, digitální či osobní asistenti a další umělé konverzační entity se také zahrnují do skupiny chatbotů. Aplikace chatbotů existuje napříč mnoha odvětvími od využití ve vzdělání, businessu, ecommerci až po zdravotnický segment a zábavu. V podnikatelském sektoru došlo k velkému rozšíření chatbotů, a to zejména díky svým schopnostem spravovat více zákazníků ve stejný okamžik a být tedy tak nákladově úspornější. Ve srovnání s FAQ sekci, kde uživatelé mohou pouze vyhledávat informace, je chatbot uživatelsky přívětivější a, jak bylo již zmíněno, díky využití NLP (zpracování přirozeného jazyka) působí více lidsky. (Adamopoulou & Moussiades, 2020)

Existuje několik způsobů implementace chatbotů, lze je ale rozdělit do dvou hlavních skupin, které zní následovně:

- Chatbot založený na pravidlech,
- Chatbot založený na umělé inteligenci.

Chatboti založené na pravidlech byly první snahou implementace v této problematice, systémy tohoto typu totiž bývají jednodušší na realizaci a uvedení do provozu. Tyto systémy, jak již vychází z názvu, pracují na základě jednoduchých, předem stanovených pravidlech. Když uživatel zadá dotaz, chatbot porovná tento dotaz s uloženými pravidly a pokusí se najít shodu. Problémem zde nastává jejich omezenost, co se týče odpovídání na složité dotazy. Další problém, který tato skupina s sebou nese, je složitost nebo spíše nemožnost přenést daného chatbota nebo systém na řešení jiného problému, jelikož je

silně vázaný na hledání shody pouze v určité oblasti. Existují 3 základní nebo spíše nejběžněji používané jazyky implementace chatbotů založených na pravidlech. Chronologicky za sebou zni následovně: AIML (Artificial Intelligence Markup Language), Riverscript a Chacript. (Caldarini et al, 2022) Je nutné zmínit, že i s příchodem umělé inteligence má stále tento způsob své využití i v praxi. (Pradeep a kol, 2018)

Chatbot založený na umělé inteligenci – Modernější způsob implementace chatbotů je s využitím umělé inteligence. Chatboti, kteří upřednostňují strojové učení před párováním vzorců neboli způsob, kterým pracují již zmíněné systémy, využívají NLP neboli zpracování přirozeného jazyka k extrakci obsahu vstupu, který zadal daný uživatel, a dokonce mají schopnost se učit z daných konverzací. Významnou výhodou tohoto typu systémů je, že dokáží brát v úvahu celkový kontext a nefixují se tedy pouze na aktuální dotaz/obrat a nemají potřebu mít již předem stanovenou odpověď na každou možnou otázku, kterou uživatel může zadat. Jednou z obtíží zde bývá potřeba robustního trénovacího souboru, jehož nalezení nebo vytvoření může být komplikované. (Adamopoulou & Moussiades, 2020)

Chatboti založené na umělé inteligenci mohou využívat různé přístupy a technologie. Významné metody, jakými mohou být chatboti vytvořeny lze rozdělit do skupin: modely pro vyhledávání informací, generativní modely a transformery.

Modely pro vyhledávání informací – tyto modely získávají informace, které jsou relevantní pro dotaz daného uživatele ze souboru textových informací. Běžně se využívá algoritmů mělkého učení, ale lze také nalézt aplikace hlubokého učení. Výhodou tohoto systému je zajištění kvality odpovědí, jelikož jsou předem definovány, zároveň to lze považovat za nevýhodu tohoto systému, jelikož jsou pro konverzační nebo sociální chatboty méně efektivní, z toho důvodu, že negenerují odpovědi, ale pouze je načítají z připraveného souboru. (Caldarini et al, 2022)

Generativní modely – u těchto modelů dochází k jinému postupu vytváření odpovědi, tyto modely totiž vytvářejí odpovědi slovo po slově, a to na základě uživatelského vstupu/dotazu. U těchto modelů tedy může docházet k tvorbě zcela nových vět, je zde ale požadavek na trénování a testování. Většina modelů tohoto typu využívá techniky hlubokého učení, a to zejména neuronové sítě. Využívá se architektury kodér-dekodér

a rekurentní neuronové sítě (RNN), které často bývají doplněné o prvek dlouhodobé a krátkodobé paměti (LSTM). LSTM umožňují modelu si zapamatovat či zapomenout informace v dlouhých sekvencích, což napomáhá pro pochopení kontextu v dané konverzaci. (Caldarini et al, 2022)

Transformer – První zmínky o tomto modelu (česky také transformátor) lze sledovat již v roce 2017 a jedná se o jednu z posledních nejzajímavějších inovací v tomto odvětví. Tento model je v podstatě neuronová síť, která se dokáže učit z kontextu a také najít význam díky sledování souvislostí v rámci sekvenčních dat, jako jsou například slova ve větě. Transformery jsou nyní hojně využívané modely zejména s aplikací jako jsou velké jazykové modely. Dochází také k velkému nahrazení již zmíněných rekurentních a konvolučních neuronových sítí za Transformátory pro výzvy v oblasti zpracování přirozeného jazyka. (Merritt, 2022)

Transformátory umožňují diferencovaně vážit relevanci každé vstupní části dat, a dokonce umožňují paralelní trénování, což znamená, že je zde umožněn trénink na větších souborech dat. Systémy jako BERT (Bidirectional Encoder Representations from transformers) nebo GPT (Generative Pre-trained Transformer), které jsou vytrénovány na obrovských jazykových datech, využívají transformační architektury. Původní aplikace se měla převážně zabývat strojovým překladem, využívají se i k řadě jiných úkolů po určité adaptaci. (Caldarini et al, 2022)

Na závěr tohoto rozdělení je nutné podotknout, že v praxi bývá zvykem určitá kombinace různé techniky pro optimalizaci výkonu daného systému.

Obecná architektura chatbotů

Existuje řada způsobů, jak vytvořit určitého chatbota a zároveň již zmíněné dvě hlavní skupiny. Každý chatbot může být postaven z různých komponent. Přesnou architekturu tedy nelze přímo stanovit, můžeme popsat určitou obecnou architekturu. V první řadě je nutné zmínit určité **uživatelské rozhraní** neboli určité prostředí, ve které daný uživatel může vložit svoji zprávu a zároveň sledovat vytvořenou odpověď. (Adamopoulou & Moussiades, 2020)

Komponenta analýzy uživatelské zprávy je další obecnou částí, která se zabývá zpracováním uživatelské zprávy. V této části je například využíváno NLU k překladu

zprávy do jazyka chatbota, nebo dochází k analýze sentimentu u daného vstupu. Správné formátování určitého vstupu pro jádro chatbota je klíčovou částí správné funkčnosti. (Achtaich et al, 2021)

Jádro chatbota je komponentou, která se napříč různými chatboty může významně lišit, jelikož zde může využita řada modulů. V této části se řeší problematika nejednoznačnosti otázek, práce s daty a určité zpracování případných chyb. Také dochází k vyhledávání potřebných informací k vytvoření odpovědi. Tato operace vyhledávání je realizována s využitím specifické znalostní databáze, jejíž struktura se odvíjí od charakteristik konkrétního typu chatbota. (Adamopoulou & Moussiades, 2020)

Komponenta generování odpovědi lze znovu rozdělit podle toho, o jaký typ chatbota se jedná. Jak již z názvu této komponenty vychází stará se o výslednou generaci zprávy či jiného výstupu. Systémy založené na pravidlech v tento moment vybírají nejvhodnější předefinovanou odpověď. Modely založené na umělé inteligenci využívají NLG k vytvoření daného výstupu, tak aby byly formulované v přirozeném jazyce a upravené na základě daného vstupu uživatele. (Achtaich et al, 2021)

5.2 Velké jazykové modely

Velké jazykové modely (v angličtině se využívá zkratka LLM) jako jsou například GPT (GPT 3.5, GPT 4) dosáhly velmi významného pokroku, zejména díky společnosti OpenAI, která se významně prosadila svým systémem ChatGPT. Tento název nesou hlavně z toho důvodu, že jsou tyto modely trénované na obrovském množství textových či jazykových dat a následně jsou schopné generovat text, který se podobá lidskému. Tyto modely dokáží odpovídat na otázky, plnit určité úkoly, které souvisejí s jazykem, a to s vysokou přesností. (Kasneci et al., 2023)

Velké jazykové modely jsou založeny na transformační architektuře a využívají základní mechanismus pozornosti, jak bylo popsáno v předchozí kapitole. Tyto schopnosti výrazně napomohly velkým jazykovým modelům zvládat dlouhodobé závislosti v textech přirozeného jazyka. (Kasneci et al., 2023) Příchod transformátorů také umožnil trénování na významně větším množství dat, což posunulo celý koncept o krok napřed. (Caldarini et al, 2022)

Nejvýznamnější velké jazykové modely jsou již zmíněné modely GPT od společnosti OpenAI, kdy je spekulováno, že nejnovější model tedy GPT-4 se skládá z 8 modelů, kdy má každý 220 miliard parametrů, dohromady by se tedy mělo jednat o zhruba 1,76 bilionů parametrů. Předchůdce tohoto modelu neboli GPT – 3 byl trénován na 175 miliardách parametrů. ³ (Bastian, 2023)

Využití velkých jazykových modelů je spousta, s neustálým vývojem v tomto odvětví se toto využití rozšiřuje do dalších sektorů. První způsob využití můžeme pojmenovat jako generativní, který lze například využít v marketingu k vytvoření určitého obsahu. Překlad a sumarizace jsou další dvě hlavní funkce velkých jazykových modelů, s rozsáhlou možností využití. (Nvidia, n.d.b) Klasifikace jako analýza sentimentu zákazníků dosáhla velkého pokroku společně s vývojem těchto modelů. A v neposlední řadě samotná implementace v rámci chatbotů. Odpovídání na otázky na základě znalostní báze a vytváření kódu jsou další významné odvětví aplikace těchto modelů. (AWS, n.d.)

Ačkoliv jsou velké jazykové modely hojně využívány a úspěšné při plnění různorodých úkolů, jejich hlavní limitace je takzvané kontextové okno neboli spíše jeho velikost. Mnoho velkých jazykových modelů je trénována tak, že je velikost jejich kontextového okna 4 096 – 16 385 tokenů.⁴ Při překročení tohoto okna dochází ke zhoršení kvality odpovědí, což ovlivňuje praktické využití těchto modelů. Je ale nutné zmínit, že nejnovější verze modelu GPT-4 má již kontextové okno o maximální velikosti 128 000 tokenů což značně mění situaci u této limitace. (Ding et al., 2024)

Halucinace

Ačkoliv je přínos těchto modelů obrovský, čelí komplikaci, které se říká halucinace. Halucinace u velkých jazykových modelů vzniká v případech, kdy daný model vygeneruje odpověď, která je buďto nesprávná, nesmyslná nebo neodpovídá na danou otázku či zadání. Nejhorší případy tohoto problému jsou ty, kdy daný model vytvoří odpověď, která logicky dává smysl, ale není postavená na faktickém kontextu. Tento

³ Parametry jsou v podstatě váhy v neuronové síti, které se využívají během trénování modelu a určitým způsobem upravují model, aby lépe porozuměl a následně generoval text.

⁴ Jeden token je zhruba 4 textové znaky v anglické jazyce, tedy zhruba 100 tokenů odpovídá 75 slovům v angličtině. (OpenAI, n.d.a)

případ může být totiž pro uživatele těžko zjistitelný, jestliže neprovede určitou kontrolu za pomoci jiných zdrojů. (Lee, n.d.)

5.3 Odpovídání na otázky (QA)

Jednou z nejdůležitějších úloh zpracování přirozeného jazyka je odpovídání na otázky. Takzvané Question Answering (QA) má za cíl využít NLP k vytvoření správné či odpovídající odpovědi na vyslovenou otázku, a to za pomoci nestrukturovaného korpusu. S rozvojem velkých jazykových modelů a dalších odvětví v rámci umělé inteligence dochází k tvorbě stále nových metod pro řešení této problematiky. (Wang, 2022)

Generování odpovědí s rozšířeným vyhledáváním (RAG)

Retrieval-augmented Generation (RAG) je jednou z nejnovějších metod konceptu odpovídání na otázky, která nově také zapojuje velké jazykové modely do procesu vytváření odpovědí na dotazy. Tento způsob spojuje externí specifické zdroje s jakýmkoliv velkým jazykovým modelem za účelem vylepšit schopnosti systémů odpovídání na otázky. Tímto způsobem může daný systém citovat zdroje, ze kterých čerpal, a zároveň dochází ke snížení již zmíněné halucinace. Využití této metody může být velmi široké, jelikož umožňuje uživatelům vést konverzaci s určitými soubory. Ve své podstatě může jakýkoliv podnik využít tuto metodu a proměnit tak určitou dokumentaci do konverzační podoby pro účely zákaznické podpory, školení zaměstnanců či jiné aplikace. (Merritt, 2023)

Návrh systému RAG

Při návrhu systému RAG je potřeba zmínit několik hlavních stavebních bloků, ze kterých se tento systém skládá. V první řadě je proces přípravy dokumentů pro tento systém spočívající v převodu dokumentů na textový formát. Existuje několik nástrojů, které tento převod mohou provést, ale mají své limity, jako je práce s tabulkami nebo diagramy. Následný krok v procesu přípravy je segmentace textů na určité menší části. U této části se již ve většině případů využívá určitých nástrojů, který tento proces provádějí, je ale potřeba určit jakým způsobem budou text segmentovat. Text segmentovat na základě počtu znaků, určitých nadpisů či speciálních znaků. Další část je takzvaná **vektorová reprezentace slov a vektorová databáze**. V této fázi dochází za pomoci speciálních modelů (Embedding models) k převodu slov na vektory. Díky tomuto převodu lze

následně lépe vyhledávat potřebné informace za pomoci vektorové databáze. Na základě například kosinové podobnosti, která je nejvíce využívána ve vektorových databázích k vyhledávání, lze určit úhel mezi jednotlivými vektory a určit tak obsahovou podobnost mezi jednotlivými texty (segmenty). Tento postup je zejména účinný pro sémantické vyhledávání. (Ke et al., 2024)

Poslední důležitou součástí je použitý **vyhledávací agent**. Tento agent využívá stejného modelu jako předchozí část k převodu vyvolané otázky uživatelem na vektorovou podobu a následně vyhledává nejbližší vektory k této otázce. Vyhledané vektory slouží jako kontext či znalost, ze které daný agent vytváří relevantní odpověď. Aby tento agent mohl vytvořit smysluplnou odpověď, je potřeba zapojit určitý již zmíněný **velký jazykový model**. (Ke et al., 2024)

Evaluace systémů RAG

Jako i u jiných systémů QA je možné provést určité evaluace systémů RAG. Existuje řada metrik, které lze vyhodnotit a analyzovat v rámci systému RAG. Jelikož se stále jedná o velmi novou metodu řešení problematiky odpovídání na otázky nejsou přímo stanovené metriky, které je potřeba při evaluaci měřit. Jelikož má ale RAG dvě nejvýznamnější části, jimiž je proces vyhledávání dat a proces generace odpovědí jsou nejčastěji evaluovány tyto části. V tabulce níže lze sledovat některé z využívaných metrik k evaluaci systémů RAG.

Tab. 1 Využívané metriky při evaluaci RAG

Práce	Metriky
Lyu a kol. (2024)	RAGQuestEval, Přesnost, Úplnost
Balaguer a kol. (2024)	Správnost, Úplnost, Stručnost,
Stelmakh a kol. (2022)	Úplnost, Komplexnost, ROUGE-L
Es a kol. (2023)	Věrnost, Relevance odpovědi, Relevance kontextu
Falcon a kol. (2023)	Relevance kontextu, Věrnost odpovědi, Relevance odpovědi

Zdroj: vlastní zpracování, 2024

6 Etické aspekty využití umělé inteligence a strojového učení v marketingu

V této kapitole se soustředíme na klíčové etické aspekty, které souvisí s využitím umělé inteligence v oblasti marketingu. Předmětem pozornosti budou otázky soukromí a ochrany osobních údajů, dopady na pracovní trh a zodpovědné použití dat. Dále se zaměříme na GDPR, ostatní regulace či nově připravované regulace v této oblasti. Tato kapitola tedy poskytne určitý pohled na výzvy a překážky, které mohou vzniknout při implementaci umělé inteligence v marketingových aktivitách.

6.1 Soukromí a ochrana osobních údajů

V současné době je soukromí a ochrana osobních údajů neustále diskutovaným tématem a také klíčovým aspektem u využití umělé inteligence v oblasti marketingu ale i v jiných odvětvích. Rapidním vývojem technologií a čím dál tím větším růstem cloudových možností, roste také schopnost sbírat data z různých zdrojů. Následná analýza na obrovském množství dat může být prováděna mimo jiné i právě díky umělé inteligenci. S tímto rozvojem ale narůstají i obavy ohledně soukromí a ochrany osobních údajů a také etického využívání těchto údajů a informací. Určité nařízení a právní regulace již existují, co se týče soukromí a ochrany osobních údajů, v Evropské unii je to například již zmíněné GDPR neboli obecné nařízení o ochraně osobních údajů. Další regulace na úrovni umělé inteligence jsou ale stále ještě v procesu přípravy.

6.1.1 GDPR

Již v polovině 90. let poznala Evropská Unie potřebu určité legislativy a právního rámce pro dosažení určité ochrany soukromí. Tento první pokus, který se nazýval směrnice Evropského parlamentu a Rady o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů, popisoval podrobně 7 konkrétních cílů ohledně ochrany soukromí a ochrany dat. V roce 2016 bylo obecně přijato členy EU nové nařízení o ochraně osobních údajů, které se začalo vyvíjet již v roce 2012, zkráceně GDPR. V roce 2018 následně vstoupilo v platnost. (Alfayad, 2020)

GDPR je považováno za nejprísnejší zákon o ochrane osobných údajů a bezpečnosti na světě. I když byl tento zákon vytvořen a schválen EU, určuje povinnosti v této problematice i společnostem, které sídlí kdekoli, pokud shromažďují údaje o osobách v EU. Pokuty, které jsou spojené při porušení GDPR jsou veliké, rozdělují se na 2 úrovně, první úroveň neboli méně závažná porušení mohou mít za následek pokutu 10 milionů Eur nebo 2 % celosvětových příjmů firmy, podle toho, která částka je vyšší. U druhé úrovně neboli u případů, kdy dochází k vážnějšímu porušení, mohou dosahovat pokuty maximálně 20 milionů eur nebo 4 % celosvětových příjmů (znovu pak záleží, která částka je vyšší). (Wolford, n.d.a)

Jestliže daný subjekt zpracovává data, musí dodržovat několik základních klíčových principů ochrany a zodpovědnosti, které jsou popsány v článku 5 v GDPR. Osobní data jsou:

1. Zpracována zákonně, korektně a transparentně (zákonnost, korektnost a transparentnost),
2. zpracována pro legitimní účely, které byly subjektům výslovně uvedené při shromažďování (omezení účelem),
3. shromažďována a zpracována pouze v takovém množství, kolik je nezbytné pro dané účely (minimalizace dat),
4. přesné a aktualizované (přesnost),
5. uchovávány pouze po dobu, která je nezbytná pro stanové účely (omezení uložení),
6. zpracovány způsobem, který zajišťuje potřebné zabezpečení, integritu a důvěrnost osobních údajů (integrita a důvěrnost). (Wolford, n.d.b)

Posledním principem je (zodpovědný přístup k prokázání souladu), kdy správce odpovídá za dodržení všech zmíněných principů.

Pojem umělá inteligence, či AI není v rámci GDPR výslovně zmíněna ani žádné jiné termíny podobné tomuto konceptu, je ale nutné zmínit, že řada ustanovení, která se nachází v GDPR je pro AI velmi relevantní. Vazbu mezi GDPR a umělou inteligencí lze uvést na následujícím příkladu osobních údajů specificky identifikace, identifikovatelnosti a opětovné identifikace. Metody umělé inteligence totiž zvyšují riziko opětovné identifikace a spojení s jednotlivci na datech, u kterých byla údajná

anonymita. Existuje řada příkladů údajných anonymních datových sad, které byly znovu opětovně identifikovány za pomoci umělé inteligenci či algoritmů. (Sartor, 2020)

6.1.2 Regulace týkající se Umělé inteligence

V rámci Evropské Unie jeden z prvních kroků vpřed při regulaci umělé inteligence je takzvaný „Artificial Intelligence Act“ (zkráceně AI Act, Česky – Akt o umělé inteligenci). Již první návrh na toto téma byl vydán v roce 2021 a jednalo se o text Evropské komise. Tento návrh stanovoval harmonizovaná pravidla pro umělou inteligenci a měnil určité legislativní akty unie. V době, kdy byl tento návrh poprvé představen, měl za cíl bojovat proti novým rizikům a negativní důsledkům pro společnost v rámci užívání umělé inteligence. Zaměření tohoto aktu bylo převážně na nástroje a odvětví, ve kterých došlo k určité implementaci umělé inteligence. Jeden z hlavních problémů, který nastal u tohoto návrhu je, že došlo k razantním změnám v této oblasti. Pravidla představená v roce 2021 nebyla dostatečně vybavena k regulaci systémů jako například ChatGPT od společnosti OpenAI, který byl spuštěn pro veřejnost v listopadu roku 2022. (Weatherbed, 2023)

Ke konci roku 2023, po měsících intenzivních diskusí a jednání trialogu (tedy jednání Evropského parlamentu, Rady Evropské unie a Evropské komise) bylo dosaženo dohody o Aktu o umělé inteligenci. Akt EU o AI je charakteristický svým přístupem, který je založený na hodnocení rizik a rozděluje jednotlivé systémy umělé inteligence do čtyř rizikových tříd, přičemž každá tato třída pokrývá různé typy využití umělé inteligence. (Hainsdord et al., 2023)

První třída jsou **nepřijatelné rizikové** systémy AI. Do této skupiny spadají systémy, které klasifikují lidi na základě jejich chování či osobních charakteristik, systémy biometrické identifikace a kategorizace osob, systémy, které manipulují s kognitivním chováním lidí nebo jiných citlivých skupin jako jsou například děti a další... Tyto systémy jsou považovány za hrozbu a ve většině případů budou zakázány, až na určité výjimky, jako například využití biometrické identifikace se značným zpožděním pro stíhání závažných trestných činů. (European Parliament, 2023a)

Vysoce rizikové systémy jsou další skupinou v tomto rozdělení, kdy se jedná o systémy, které mohou významně poškodit zdraví, životní prostředí, demokracii a právní stát. Jedná

se o systémy používané v oblasti vymáhání práva, výkonu spravedlnosti, systémy pro určování přístupu do vzdělávacích institucí, zdravotnické prostředky a systémy v různých kritických infrastrukturách. Na systémy v této skupině se vztahuje řada povinností, které se týkají snižování rizika, lidského dozoru, správy dat, transparentnosti, přesnosti ale také kybernetické bezpečnosti. (Hainsdord et al., 2023)

Třetí skupina se týká **obecných a generativních systémů** AI. K zohlednění široké škály úkolů, které tyto systémy mohou plnit a také k zohlednění rychlého vývoje v této oblasti, bylo dohodnuto, že tyto systémy budou muset dodržovat požadavky na transparentnost. Což znamená vypracování technické dokumentace, dodržování autorského práva pro EU a zveřejnění shrnutí dat, na kterých byl daný systém trénován. (European Parliament, 2023b)

Do poslední skupiny systémů AI spadají systémy s **omezeným či minimálním rizikem**, které by měly splňovat minimální požadavky na transparentnost, které by měli uživatelům po interakci s těmito systémy pomoci se rozhodnout, zdali chtějí tyto systémy nadále využívat. Převážná většina AI systémů spadá do této kategorie. (European Commission, 2023)

Stejně jako u GDPR zde existují určité sankce a pokuty za porušení tohoto aktu. Pokuty se pohybují od 35 milionů eur nebo 7 % celosvětového ročního obrátu (znovu se to řídí podle toho, která částka je vyšší) za porušení zakázaných aplikací AI, 3 % nebo 15 milionů eur za porušení povinností, které vyplývají z aktu o AI a 7,5 milionů eur nebo 1,5 % za poskytnutí nesprávných informací. (European Commission, 2023)

Další kroky v momentální chvíli jsou formální schválení Evropskou radou a Evropským parlamentem a následně vstoupí v platnost 20 dní po schválení. Akt se bude moci začít uplatňovat dva roky po jeho vstupu v platnost. (Evropská rada, 2023)

6.2 The New York Times žaloba na společnost OpenAI

První problémy s hojně využívanými LLM neboli velkými jazykovými modely začínají přicházet na světlo a jedná se o častokrát vážné problémy týkající se porušení autorských práv. Nejnovější příklad této události je žaloba od jednoho z největších deníků ve Spojených státech amerických se sídlem v New Yorku, The New York Times na

společnost OpenAI, která vytvořila a provozuje systémy jako ChatGPT, což je jeden z nejvyužívanějších chatbotů na světě. (Grynbaum, & Mac, 2023)

Jedná se v tomto ohledu o první žalobu od významné mediální společnosti směrem na tvůrce modelů GPT. Tato žaloba se nevztahuje jen na společnost OpenAI, ale také na významnou nadnárodní společnost Microsoft. Je to z toho důvodu, že nově nasazený chatbot Copilot od společnosti Microsoft je postavený na velkém jazykovém modelu GPT – 4 od společnosti OpenAI. Deník The New York Times ve své žalobě tvrdí, že jmenované společnosti vytvořily své modely AI kopírováním a následným využíváním značného množství článků a nyní je deník považuje za potenciální konkurenci v oblasti zpravodajství. (Roth, 2023)

Ve zmíněné žalobě je také uvedeno, že při otázkách týkajících se aktuálních událostí či jiných zpravodajských témat, mohou tyto modely generovat odpovědi, které jsou postavené na článcích deníku The New York Times. V žalobě lze nalézt řadu příkladů, kdy daný chatbot poskytl uživatelům identickou část článku, který je jinak přístupný pouze s placeným předplatným. Obavy deníku v tomto ohledu jsou, že se daný čtenář spokojí s výstupem chatbota a již nebude mít potřebu navštívit deník, tím by mohlo dojít ke snížení příjmů z reklamy, která se odvíjí od návštěvnosti webu. (Grynbaum, & Mac, 2023)

Mimo žalobu se také deník The New York Times a další mediální společnosti rozhodli k blokaci web crawlera společnosti OpenAI, který slouží k shromažďování dat a obsahu z webových stránek a následně se využívají k trénování daných modelů. Tento krok učinil deník již v průběhu srpna 2023. (Roth, 2023)

6.3 Environmentální dopad umělé inteligence

S umělou inteligencí a jejím rozšířením je spojeno mnoho environmentálních problémů. Dva hlavní problémy v této oblasti jsou nadměrná spotřeba elektřiny a spotřeba vody pro chlazení potřebných komponent.

Spotřeba elektřiny

Velké jazykové modely mají zřejmě největší dopad na rostoucí hodnoty spotřeby elektřiny spojené s umělou inteligencí. Jelikož je určitým trendem neustálý růst těchto modelů, roste s tím také potřeba milionů GPU hodin (což je v podstatě míra výpočetní

techniky), která je potřeba pro trénování těchto velkých jazykových modelů s miliardy parametrů. Následné nasazení modelu a jeho provoz je také vysoce nákladné. Trénování modelu s názvem BLOOM, který pochází z projektu BigScience a má 176 miliard parametrů, trvalo přes 118 dní a za tuto dobu bylo spotřebováno 433 196 kWh elektřiny. (Luccioni et. al., 2022)

Jelikož společnost OpenAI přímo nezveřejnila hodnoty spotřebované elektřiny při trénování jejich nejnovějšího modelu GPT-4, je možné pouze využít odhadované rozmezí, které experti určili z hodnot elektrické spotřeby předchůdce GPT-3. Toto rozmezí pro model GPT-4 s 170 biliony parametrů je odhadováno na 51 733 MWh – 62 319 MWh. (Lai, 2023)

Jak již bylo zmíněno, trénování není jediná fáze, při které je spotřebováno velké množství elektřiny. Po fázi trénování jsou modely nasazeny do daného prostředí a přechází tak do fáze, kdy generují určité výstupy. Pro tuto problematiku můžeme využít nejrozšířenější AI systém ChatGPT. Je odhadováno, že fáze užívání spotřebovává 10x více elektřiny než proces trénování. Některé uváděné hodnoty na systém ChatGPT poukazují, že energie, která je spotřebována pro zpracování milionů dotazů, by se mohla rovnat přibližně 1 gigaawatt hodině na den, což je častokrát přirovnáváno k spotřebě elektřiny 33 000 domácností v USA. (Cohan, 2023) Jiné zdroje uvádějí, že spotřeba energie při zpracování jednoho dotazu se pohybuje v rozmezí od 0,001-0,01 kWh modelu GPT-4 kdy záleží na počtu tokenů, které byly zpracovány a také na daném typu modelu. K porovnání lze uvést, že jedno vyhledání na vyhledávači Google spotřebuje zhruba 0,0003 kWh. (Sreedhar, 2023)

Spotřeba vody

Další částí této podkapitoly týkající se environmentální problematiky spojené s AI je spotřeba vody. Všechna výpočetní technika, která je potřebná pro plynulý chod velkých jazykových modelů jako je například GPT-4, vyžaduje, spoustu elektřiny. Dochází ale také ke tvorbě značného tepla, k udržení požadované teploty nebo tedy k snížení teploty zejména v horkých dnech, je potřeba pumpovat značné množství vody, které slouží k chlazení. Velké společnosti jako Microsoft, Meta, Google, a další se nyní potýkají se zvýšenými hodnotami ve spotřebě vody z důvodu rostoucí poptávky po jejich nástrojích generativní umělé inteligence. (O'Brien & Fingerhut, 2023)

Tomuto nákladu se někdy říká skrytý náklad, jelikož oproti spotřebě elektřiny, se nejedná o tak veřejně známý problém. Microsoft ve své poslední zprávě týkající se environmentální udržitelnosti zmínil, že se její spotřeba vody v období od roku 2021 do 2022 zvýšila o více než třetinu jejich dosavadní hodnoty. Tyto hodnoty byly zveřejněné před spuštěním vlastního systému Copilot, který tyto hodnoty může významně změnit. Obdobně je na tom také společnost Google, která zveřejnila, že se její celková spotřeba vody zvýšila o 21 % za rok 2022 oproti předchozímu roku. Stejně jako u společnosti Microsoft se jedná o hodnoty, které nezahrnují spuštění systému Google Bard, nebo dokonce nejnovější model Gemini. (Meredith. 2023)

Výzkumníci z Kalifornské univerzity, kteří se snaží určit dopad umělé inteligence, a tedy především generativní umělé inteligence, ve své publikaci uvádějí, že při trénování modelu GPT-3 v data centru společnosti Microsoft dokáže spotřebovat až 5,4 milionů litrů čisté vody. Stejně jako u spotřeby elektřiny dochází ke spotřebě vody i při samotném chodu tohoto modelu. Uvádí se hodnota 500 ml vody pro zhruba 10-50 odpovědí, podle toho, kdy a kde je nasazen. Tyto hodnoty byly vypočítané pro model GPT-3 a mohou se zvýšit pro nové modely jako je GPT-4, které dokáží pohltnout více kontextu a vytvářet obrázky. (Li et al., 2023)

7 Praktická část

V této části diplomové práce je vytvořen a následně testován chatbot, který je vytvořen pro potřeby studentů a uchazečů Fakulty ekonomické Západočeské univerzity v Plzni. Cílem tohoto chatbota je usnadnit studentům fakulty přístup k důležitým a relevantním informacím a odpovídat na dotazy administrativního typu, které se týkají studia na fakultě. Současně by tato implementace mohla uvolnit pracovní kapacitu některým pracovníkům fakulty pro jiné úkoly.

7.1 Definování problematiky

V této části diplomové práce se zaměříme na obtížnosti spojené s nalezením potřebných informací či dokumentů studenty nebo potencialními studenty, které se týkají studia na Fakultě ekonomické Západočeské univerzity v Plzni.

Zejména nastupující studenti na vysokou školu mohou být zaskočení významnými změnami prostředí a nově se potřebují orientovat ve velkém množství jak studijních materiálů, tak i v důležitých dokumentech fakulty či univerzity. Pro některé z nich to může být v podstatě první setkání s různými typy vyhlášek, řádů, norem a dalších důležitých dokumentů. Pro tyto studenty tedy může být obtížné číst a interpretovat dokumenty, ale také identifikovat a získat relevantní informace pro jejich konkrétní potřeby.

Následně může dojít k přehlcení informacemi, což může mít negativní následek na motivaci těchto studentů pokračovat ve svém studiu z důvodu pocitů ztracení a frustrace v počátcích své akademické cesty. Špatná interpretace dat, nebo vyhledání starších harmonogramů, norem či jiných dokumentů může také vést k šíření nesprávných či zastaralých informací.

Studenti dalších ročníků, a hlavně studenti závěrečných ročníků, si musí kontrolovat řadu důležitých termínů, spojených s kvalifikačními pracemi, státními zkouškami, podáním přihlášek a další důležité informace spojené s jejich povinnostmi. Stejně jako u předchozího příkladu může dojít k přehlcení informacemi.

Implementace chatbota v této souvislosti představuje nové řešení, které by mohlo studentům ulehčit práci s dokumenty fakulty a univerzity. Chatbot by primárně sloužil

k personalizaci informací, čímž by umožnil studentům efektivnější vyhledávání relevantních údajů a zároveň by přispěl k redukci přehlcení informacemi. Díky tomuto řešení by mohlo dojít k rychlejšímu procesu orientace nových studentů, snížení potencionálního poklesu motivace a zvýšení obeznámenosti studentů s důležitými dokumenty.

Pozitivní dopady by se mohly objevit i na straně pracovníků, především tedy u pracovníků studijního oddělení. Uvolnění tohoto personálu od rutinních dotazů, by jim totiž umožnilo věnovat více času komplexnějším či jiným důležitým úkolům.

7.2 Představení Fakulty ekonomické Západočeské univerzity v Plzni

Fakulta ekonomická (FEK) byla založena v roce 1990 a je jednou z 9 fakult Západočeské univerzity v Plzni, která má své působení ve dvou krajích a městech ČR, v Plzeňském kraji ve městě Plzeň a v Karlovarském kraji ve městě Cheb. Fakulta nabízí široké spektrum oborů v oblasti ekonomie, a to na úrovni bakalářského stupně studia, navazujícího magisterského stupně studia, a dokonce doktorského stupně studia. Fakulta má za své doby existence řadu absolventů, kteří se následně stávají odborníky ve vybrané oblasti. Díky intenzivní spolupráci s více než 60 zahraničními partnerskými univerzitami poskytuje fakulta ekonomická bohaté možnosti pro mezinárodní výměny a stáže, jenž podporují vědeckou činnost, ale především připravuje studenty pro úspěšný začátek své profesní kariéry v dynamickém trhu práce. (Fakulta ekonomická Západočeské univerzity v Plzni, 2023).

Obr. 1 Logo Fakulty ekonomické



Zdroj: Fakulta ekonomická Západočeské univerzity v Plzni (n.d.)

Fakulta se rozděluje na 5 kateder, Katedra ekonomie a kvantitativních metod, Katedra financí a účetnictví, Katedra geografie, Katedra marketingu, obchodu a služeb a v poslední řadě katedra podnikové ekonomiky a managementu, které jsou zaměřené na

určitou specifickou oblast. Dále existují na FEK jiná centra a střediska jako je například centrum celoživotního vzdělávání, Centrum podnikání a udržitelnosti a další. (Fakulta ekonomická Západočeské univerzity v Plzni, 2023).

7.3 Postup vývoje chatbota

V rámci této práce dojde k několika krokům, které na sebe budou chronologicky navazovat. V první řadě je potřeba realizovat průzkum postojů studentů Fakulty ekonomické k umělé inteligenci a chatbotům. Tento průzkum bude vyhotoven za pomoci dotazníkového šetření. Dotazník jako nástroj ke sběru dat v této části byl vybrán z toho důvodu, že umožňuje získat rozsáhlý objem dat od velkého souboru respondentů za kratší časový úsek. Větší množství dat v tomto šetření umožňuje lepší následnou konstrukci daného chatbota, jelikož může být zpracováno více požadavků či názorů respondentů napříč fakultou. Tímto šetřením bude také zkoumána důvěra v chatboty společně s tím, zda mají studenti zájem o takový technologický nástroj. (Eger & Egerová, 2017)

Po dotazníkovém šetření bude následovat samotná konstrukce chatbota. U této části bude využito několik výstupů z dotazníkové šetření, tak aby byl daný chatbot zkonstruován na základě požadavků studentů. Mimo tyto požadavky bude také potřeba zahrnout určité požadavky ze strany fakulty.

Další fází je testování daného chatbota pro fakultní účely. V této fázi budou zkoumány různé metriky, které jsou charakteristické pro funkčnost a provoz daného chatbota, jako je například rychlost vytvoření odpovědi, finanční náklady, chybovost a další.

Následující fází bude testování chatbota s jednotlivými studenty za účelem získání zpětné vazby k použitelnosti daného chatbota společně určitými návrhy na zlepšení či rozšíření chatbota. Zpětná vazba bude shromažďována za pomoci speciálního dotazníku použitelnosti chatbotů (The Chatbot Usability Questionnaire (CUQ)). Tento dotazník, se tedy využívá po následném testování vybranými studenty. Výsledky z této části lze porovnávat s určitými standardy v dané oblasti, či s jinými chatboty.

V poslední části budou představeny procesy případné implementace a kontroly daného chatbota.

Časový harmonogram

V následující tabulce lze sledovat harmonogram práce, který se skládal ze zmíněného dotazníkového šetření, vlastního testování a testování se studenty. V tomto harmonogramu lze také nalézt období, ve kterém docházelo k návrhu a vývoji chatbota.

Tab. 2 Časový harmonogram vývoje

Fáze	Aktivita	od	do
Dotazníkové šetření	Tvorba dotazníku	29.1.24	7.2.24
	Pilotáž dotazníku	16.2.24	17.2.24
	Sběr dat	21.2.24	7.3.24
	Zpracování dat	8.3.24	8.3.24
	Analýza a vyhodnocení	9.3.24	13.3.24
Návrh a vývoj	Tvorba chatbota	22.1.24	16.3.24
Vlastní testování	Testování s LangSmith	16.3.24	18.3.24
	Hodnocení odpovědí	18.3.24	21.3.24
	Testování s Ragas	21.3.24	22.3.24
Testování se studenty	Úprava dotazníku	23.3.24	26.3.24
	Testování se studenty	26.3.24	9.4.24
	Analýza a vyhodnocení	9.4.24	12.4.24

Zdroj: vlastní zpracování, 2024

V přiložené tabulce je možné sledovat určitý paralelní průběh některých částí. Výběr platformy, identifikace vhodné technologie a testování funkčnosti představují klíčové komponenty návrhu a vývoje, které byly realizovány paralelně s dotazníkovým šetřením, neboť nebyla vyžadována data z dotazníku. Tento přístup umožnil efektivní alokaci času během realizace této práce.

8 Zjišťování preferencí a postojů

První částí bylo dotazníkové šetření mezi studenty, které mělo za cíl zjistit určité postoje k využívání fakultního chatbota mezi studenty FEK. Toto dotazníkové šetření bylo realizováno před dokončením vývoje daného chatbota. Důvodem bylo, že strukturovaný dotazník zahrnoval otázky zaměřené na to, jaké účely by respondenti chtěli chatbotem pokrýt a jaký typ komunikace by od něj očekávali, což bylo klíčové pro jeho finální nastavení.

Cílem této části je ověřit jaký by byl zájem studentů o trvalou implementaci chatbota a zjistit jaká by byla úroveň důvěry studentů v odpovědi daného chatbota.

Předpoklady:

1. Studenti vykáží značný zájem o implementaci chatbota na webu fakulty.
2. Více jak 50 % studentů plně důvěřuje v odpovědi chatbota.

Statistická hypotéza pro 1. předpoklad:

Předpokládané rozdělení odpovědí je určeno na základě vědecké práce od Nguyen a kol. (2021), ve které je popsána úspěšná implementace chatbota s významným procentem spokojenosti jednotlivých uživatelů a s velkým množstvím vytvořených dotazů na chatbota za určité období.

Tab. 3 Předpokládané rozdělení

Určitě ano	Pravděpodobně ano	Nevím	Pravděpodobně ne	Určitě ne
35,00%	40,00%	10,00%	10,00%	5,00%

Zdroj: vlastní zpracování, 2024

H0: Neexistuje statisticky významný rozdíl mezi zjištěným zájmem respondentů o používání chatbota na webu fakulty a předpokládanými hodnotami stanovenými na základě podobného výzkumu.

H1: Existuje statisticky významný rozdíl mezi zjištěným zájmem respondentů o používání chatbota na webu fakulty a předem předpokládanými hodnotami.

Sběr dat

Sběr dat u tohoto dotazníku byl uskutečněn pomocí dotazníkového šetření, které bylo uskutečněno na platformě Google Forms. Dotazníkové šetření tedy probíhalo v online prostředí. Dotazník byl rozeslán studentům pomocí využívaných online kanálů a sociálních sítí. Dotazník byl dostupný 16 dní a po tomto období byl následně uzavřen.

Pro zajištění adekvátní kontroly a standardu výzkumného procesu byl proveden pilotní výzkum se zapojením deseti účastníků. Tito účastníci přinesli důležité poznatky o funkčnosti dotazníku. Na základě těchto informací byly provedeny úpravy a detailní ladění vytvořeného dotazníku.

Výzkumný soubor tvořili pouze studenti Fakulty ekonomické, tj. základní soubor tvoří ca. 1 600 studentů.

8.1 Vyhodnocení dotazníku

V této kapitole budou zpracována a vyhodnocena data, která byla získána v rámci dotazníkového šetření. Tento dotazník byl rozeslán zejména pomocí sociálních sítí a komunikačních platforem. Platforma, která byla využita k zprostředkování dotazníků byla Google Forms. Následně byla data stažena, zkontrolována a zpracována v Excelu. Finální počet respondentů vytvořeného dotazníku byl 101 respondentů, což je pro účely tohoto výzkumu dostačující.

8.1.1 Popis skupiny respondentů

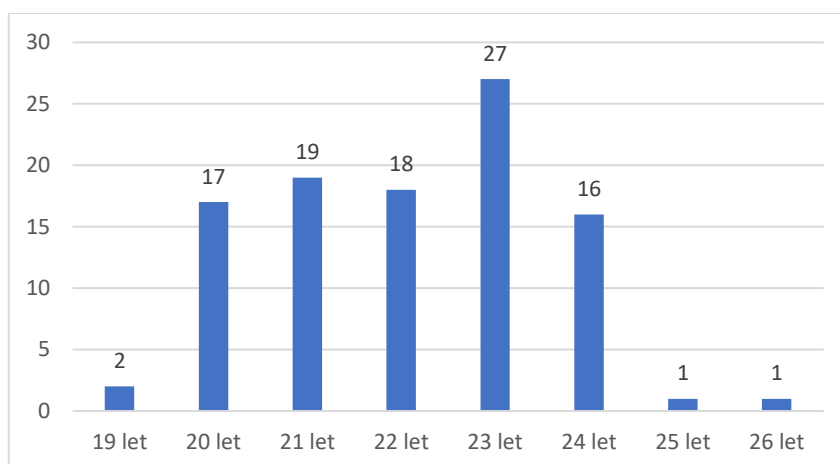
Výběrový soubor pro dotazníkové šetření tvořili studenti Fakulty ekonomické Západočeské univerzity v Plzni bez omezení na pohlaví, specifický studijní program, rok studia nebo jiné demografické faktory.

Jak již bylo zmíněno, finální počet respondentů činil 101 studentů. Převaha respondentů u dotazníkového šetření byla ze strany mužů, jelikož finální počet mužů, kteří se zúčastnili dotazníkového šetření činil 59 (58,4 %), finální počet žen, kteří se zúčastnili dotazníkového šetření činil 42 (41,6 %). Ani v jednom případě nebyla vybrána možnost jiné.

Otázka týkající se **věku** v tomto případě byla otevřeného typu, Nejčtenější skupinou jsou respondenti ve věku 23 let, kterých je 27. Následují respondenti ve věku 24 let, kterých

je 16, a poté těsně respondenti ve věku 22 let, jejichž počet je 18. Ve věku 21 let je 19 respondentů a ve věku 20 let jich je 17. Na druhém konci spektra jsou respondenti ve věku 19, 25 a 26 let, kde každé z těchto věkových kategorií je zastoupeno velmi malým počtem – 19letí jsou zastoupeni dvěma respondenty a 25 a 26letí po jednom respondentovi. Graf poskytuje přehled o věkové struktuře účastníků šetření a ukazuje, že většina respondentů je ve věku od 20 do 24 let.

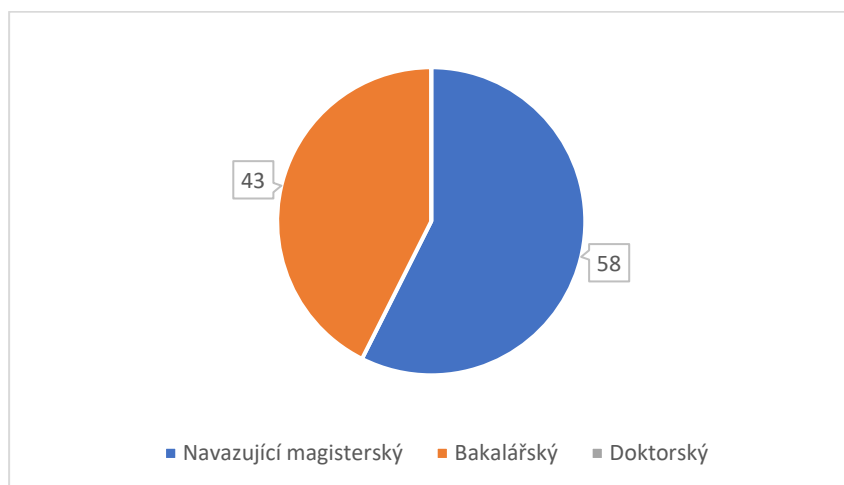
Obr. 2 Věk respondentů



Zdroj: vlastní zpracování, 2024

V další otázce týkající se daného **stupně studia**, vybírali respondenti, jestli studují bakalářský, navazující magisterský či doktorský studijní program. Větší část respondentů 58 (57,4 %) vybrala navazující studijní program, zbytek respondentů 43 (42,6 %) momentálně studuje některý z bakalářských studijních programů. Nikdo z respondentů nevybral možnost doktorský studijní program, což je zřejmě dáno tím, že počet studentů u tohoto stupně studia je několika násobně menší než u předchozích dvou. Je důležité zdůraznit, že výběr respondentů bez zahrnutí doktorských programů byl záměrný, neboť dotazník byl šířen převážně na sociálních sítích a ve školních skupinách, kde jsou zastoupeni primárně studenti bakalářských a magisterských programů

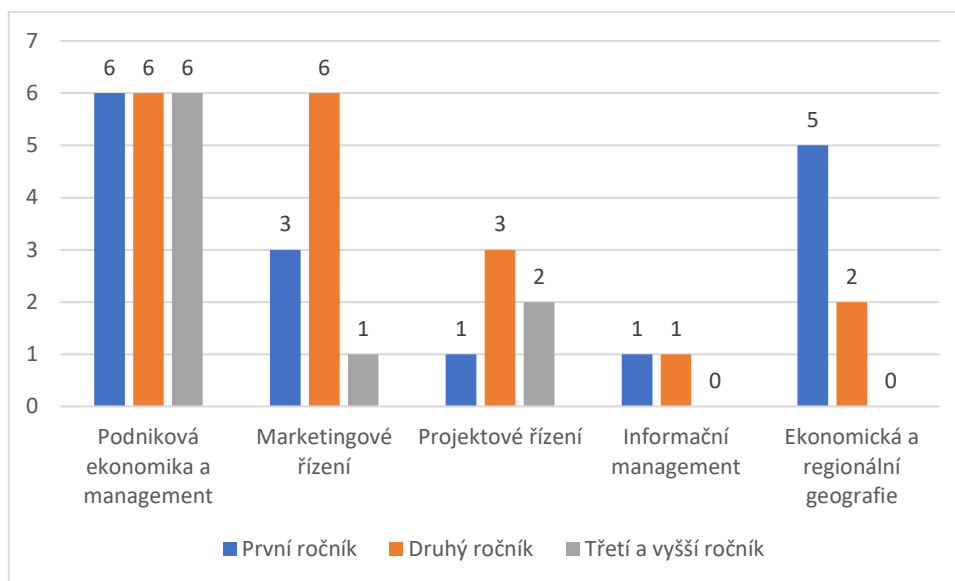
Obr. 3 Stupeň studia



Zdroj: vlastní zpracování, 2024

Poslední dvě otázky se týkají **konkrétního studovaného programu** u daného stupně studia a **ročníku studia**, ve kterém se daný respondent nachází. Tyto dvě otázky lze spojit do jednoho grafického zobrazení, které lze sledovat níže pro jednotlivé stupně studia. U bakalářského stupně studia bylo největší zastoupení respondentů ze studijního programu podniková ekonomika a management s 18 respondenty rozděleny rovnoměrně mezi tři ročníky studia. Následuje skupina respondentů ze studijního programu marketingové řízení, kde byl celkový počet respondentů 10. Dále pokračuje program projektové řízení a Ekonomická a regionální geografie. V poslední řadě 2 respondenti ze studijního programu Informační management.

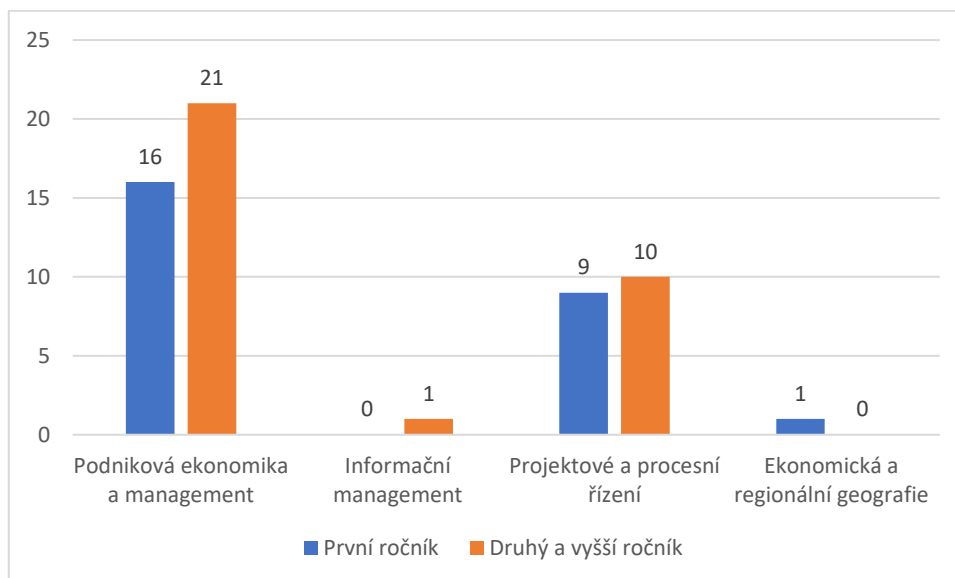
Obr. 4 Zastoupení u bakalářských studijních programů



Zdroj: vlastní zpracování, 2024

U navazujícího magisterského studia se respondenti dle jednotlivých studijních programů dělí na významnější dvě skupiny. Větší část respondentů (37) pochází ze studijního programu podniková ekonomika a management, druhá, menší část (19) pochází ze studijního programu projektové a procesní řízení.

Obr. 5 Zastoupení u navazujících magisterských studijních programů



Zdroj: vlastní zpracování, 2024

Tato část představovala popis respondentů neboli tedy výběrové souboru, a to z demografického hlediska. Následuje již vyhodnocení hlavní části dotazníku.

8.1.2 Vyhodnocení specifických otázek

Na začátku dotazníkového šetření bylo společně s krátkým představením dotazníku také stručný teoretický vstup, který představil pojmy jako je chatbot a zároveň upřesnil rozdíl mezi chatbotem a živým chatem s operátorem. Pro tento dotazník bylo totiž klíčové, aby nedocházelo k záměně a aby bylo respondentům jasné, že se při dotazování jedná o webového chatbota.

Otázka 1. Už jste někdy viděli chatbota na nějaké webové stránce?

U této otázky odpověděli všichni respondenti pozitivně. Tento výsledek naznačuje, že studenti mají povědomí o chatbotech a zároveň poukazuje na široké rozšíření chatbotů na internetu. Například velká většina českých bank již využívá určitý typ chatbota pro své účely.

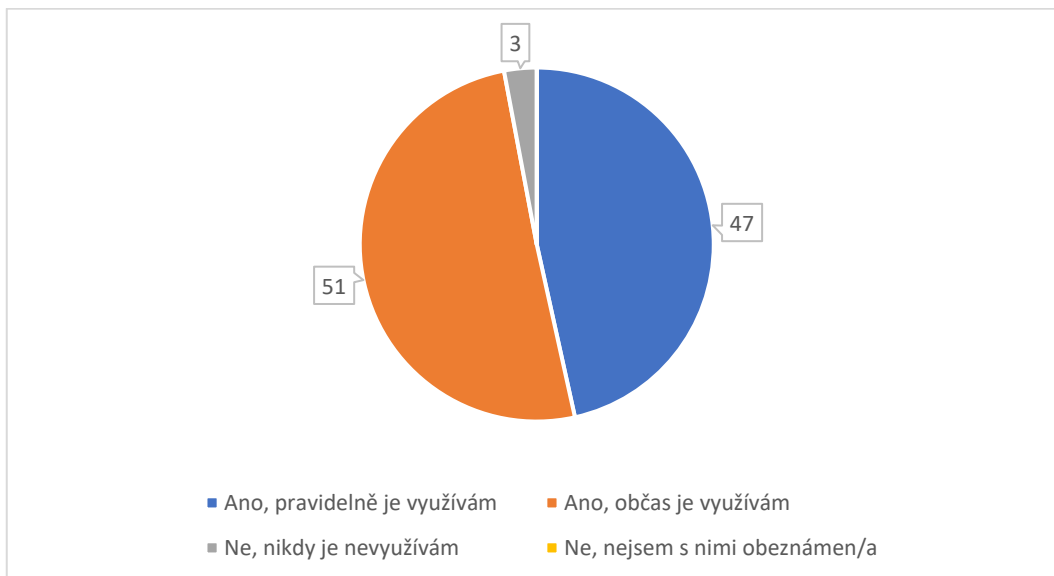
Otázka 2. Už jste někdy využili chatbota pro získání informací nebo asistence online?

Většina respondentů 81 (80,2 %) u této otázky odpověděla, že už využili chatbota za účelem získat určité informace či asistence v online prostředí. Ačkoliv 100 % respondentů odpovědělo, že už někdy viděli nějakého chatbota, existuje zde menší skupina respondentů 20 (19,8 %), kteří nikdy žádného chatbota nevyužili pro získání informací či asistence.

Otázka 3. Využíváte AI jazykové modely jako je GPT, Bard či jiné?

Při dotazování studentů na velké jazykové modely byla snaha zjistit to, zdali respondenti využívají dané modely, jelikož výsledný chatbot bude využívat určitý model ke generaci odpovědi. 51 respondentů u této otázky odpovědělo, že občas využívají jazykové modely, druhá o trochu menší skupina respondentů (47) odpověděla, že tyto modely využívají pravidelně. Pouze 3 respondenti odpověděli, že tyto modely nevyužívají. Všichni ze skupiny respondentů, jsou již určitým způsobem obeznámeni s těmito modely.

Obr. 6 Využití jazykových modelů

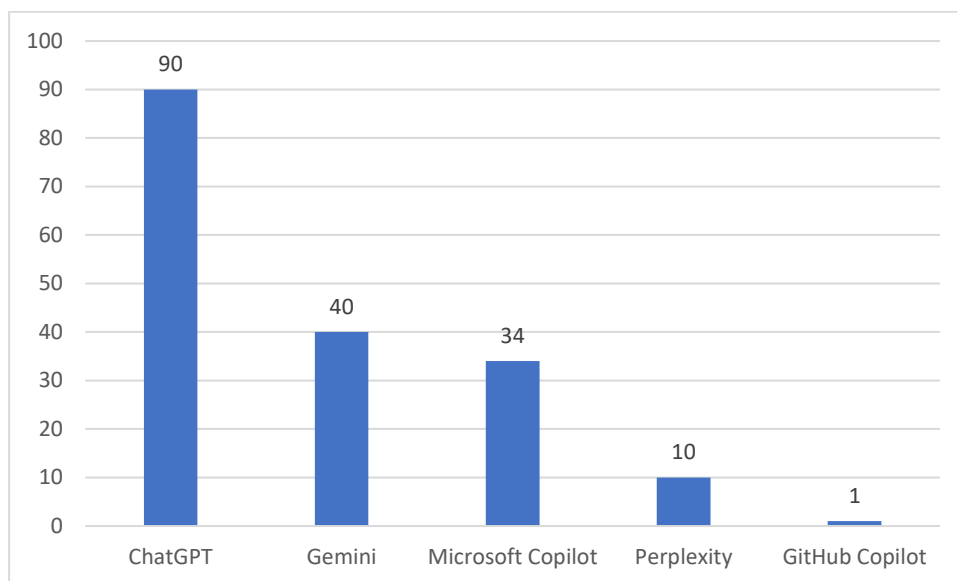


Zdroj: vlastní zpracování, 2024

Otázka 4. *Které jazykové modely využíváte konkrétně?*

Tato otázka se týkala upřesnění, jaké jazykové modely daný respondent využívá. U této otázky mohli respondenti vybrat více možností, a přidat i vlastní odpověď. Nejvíce využívaný systém v tomto případě byl očekávaný výsledek (90), ChatGPT. Tento systém, jak již bylo zmíněno, byl významným krokem vpřed při využívání AI v marketingu, ale i jiných odvětví. Další modely, které byly vybrány respondenty jsou následující – Google Gemini (40), Microsoft Copilot (34), Perplexity (10) a v poslední řadě GitHub Copilot, který vybral pouze jeden respondent.

Obr. 7 Specifické velké jazykové modely a jejich využívání



Zdroj: vlastní zpracování, 2024

Otázka 5. *Využili byste AI chatbota, pokud by byl dostupný na webových stránkách fakulty ekonomické?*

Tato otázka se již týkala nově vytvořeného chatbota, který by byl dostupný na stránkách FEK a sloužil by studentům pro různé potřeby. Největší skupina 50 (49 %) respondentů odpověděla, že by pravděpodobně využili fakultního chatbota, 29 (29 %) respondentů odpovědělo, že by rozhodně využilo daného chatbota, poslední významná skupina 10 (10 %) respondentů vybralo odpověď „nevím“. Počet respondentů s negativním postojem je relativně nízký, a tedy odráží určitou otevřenost studentů vůči technologickým nástrojům jako je právě chatbot využívající AI ve fakultním prostředí. Je ale potřeba zmínit, že při případné následné integraci daného chatbota do provozu je potřeba udržovat existující komunikační kanály, minimálně z toho důvodu, že existuje určitá skupina respondentů, kteří nechtějí využívat tento inovativní způsob.

Tab. 4 Porovnání očekávaných hodnot se zjištěnými

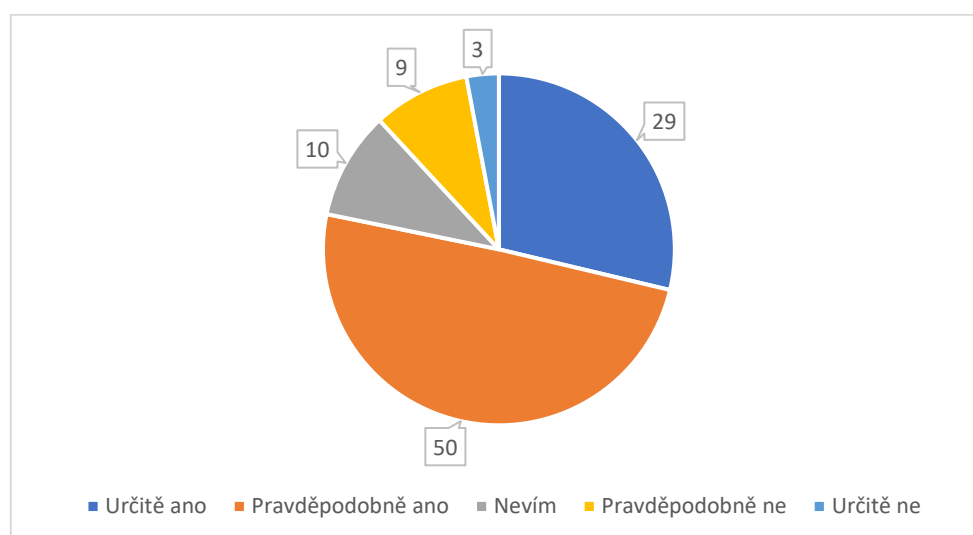
	Určitě ano	Pravděpodobně ano	Nevím	Pravděpodobně ne	Určitě ne
Očekávané	35	41	10	10	5
Zjištěné	29	50	10	9	3

Zdroj: vlastní zpracování, 2024

V tabulce výše lze sledovat porovnání očekávaných a zjištěných hodnot s tím, že procentuální zastoupení, které bylo stanoveno u očekávaných hodnot bylo přepočtené na celá čísla z celkového počtu respondentů. Na základě těchto dat lze vytvořit chí-kvadrát test dobré shody a porovnat tak zmíněné předpokládané četnosti s opravdovými hodnotami z tohoto dotazníku. Chí-kvadrát χ^2 v tomto případě vyšel 3,904. P-hodnota byla vypočítaná za pomoci výsledné hodnoty chí-kvadrátu společně se 4 stupni volnosti (n-1 tedy v tomto případě 5-1). Výsledná p-hodnota vyšla 0,419 nelze zamítnout nulovou hypotézu.

Tato otázka také sloužila jako rozcestník, kdy respondenti, kteří odpověděli, že nemají zájem o využití chatbota pro fakultní účely pokračovali na demografické otázky, ostatní respondenti odpovídali na další otázky týkající se chatbota.

Obr. 8 Postoj respondentů k využití AI chatbota na webových stránkách FEK



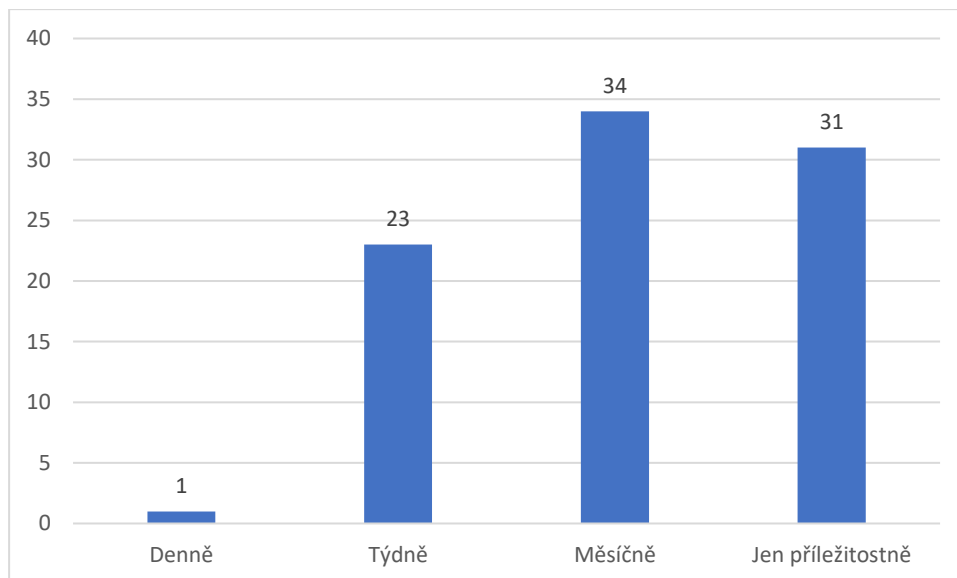
Zdroj: vlastní zpracování, 2024

Otázka 6. *Jak často si myslíte, že byste využívali chatbota na webové stránce fakulty?*

Respondenti odpověděli následovně: 1 (1,12 %) respondent uvedl, že by chatbota využíval denně, 23 (25,84 %) respondentů by ho využívalo týdně, 34 (38,20 %) respondentů měsíčně, a 31 (34,83 %) respondentů by chatbota využívalo jen příležitostně. Nejčastější očekávaná frekvence využití chatbota je měsíčně, těsně následována skupinou respondentů, která by využívala chatbota je příležitostně. Týdenní využití je také poměrně vysoce vybraná kategorie, pouze o denní využití nemá téměř nikdo zájem. Tyto výsledky

jsou zejména důležité při případném plánování určitého rozpočtu pro provoz daného chatbota.

Obr. 9 Frekvence využívání chatbota



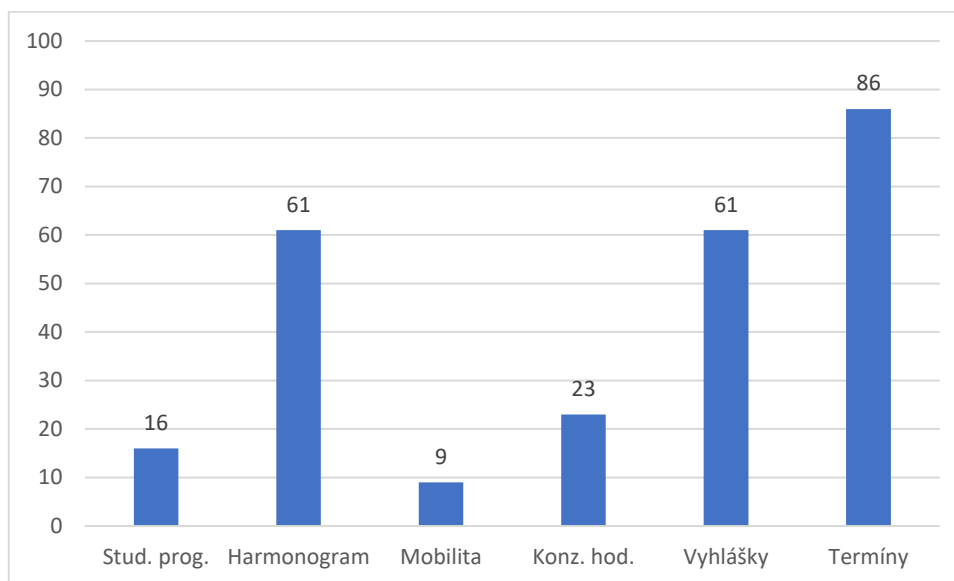
Zdroj: vlastní zpracování, 2024

Otázka 7. *Za jakým účelem byste využili AI chatbota na webových stránkách FEK?*

Další otázka řešila to, za jakým účelem by respondenti využívali chatbota. Tato otázka je významná pro následnou konstrukci chatbota, jelikož poukazuje na specifické využití chatbota studenty. Respondenti u této otázky měli možnost vybrat několik odpovědí, nebo vytvořit jiný. V grafu níže lze sledovat několik sloupců, kategorie Termíny (v dotazníku odpověď: Dotazy na důležité termíny - např. odevzdání KP, termíny SZZ a další), byla nejvíce vybranou kategorií napříč respondenty. Následují dotazy na harmonogram akademického roku a dotazy na vyhlášky a dokumenty typu Studijní zkušební řád a další (oba s počtem 61). Poté další odpovědi razantně klesají dotazy na konzultační hodiny a informace o akademických pracovnících již vybralo pouze 23 respondentů. Předposlední nejčastěji vybranou kategorií je kategorie týkající se informací ohledně studijních programů. Nejnižší zájem, pouze 9 respondentů, vybralo kategorii informace týkající se mobility na FEK a Erasmu.

Lze vyvodit, že studenti mají zájem využívat chatbota k získání široké škály informací, přičemž největší důraz je kladen na praktické aspekty studia, jako jsou termíny a pravidla.

Obr. 10 Graf preferovaného využití chatbota na webových stránkách FEK



Zdroj: vlastní zpracování, 2024

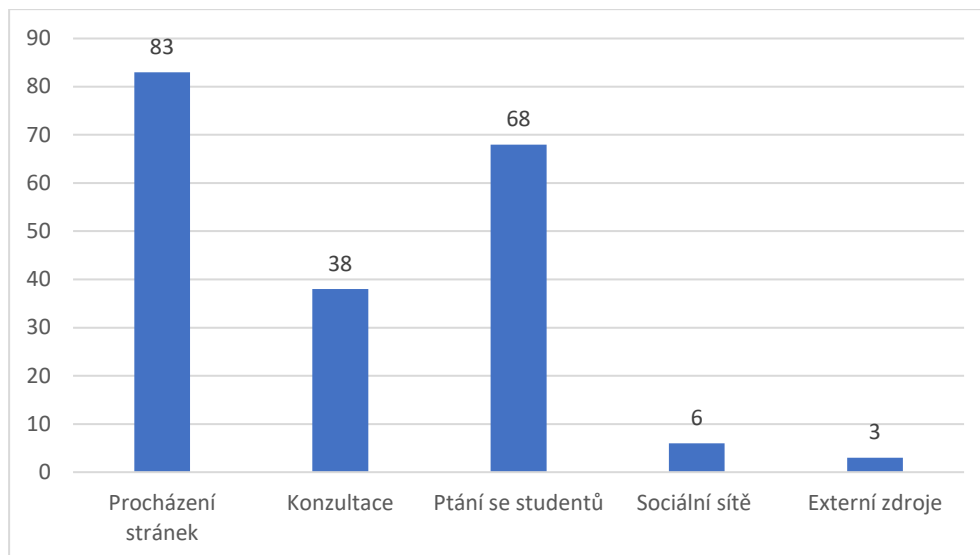
Otázka 8. *Jakým způsobem nejčastěji získáváte informace o tématech zmíněných ve vašem předchozím výběru?*

Další otázka navazuje na předchozí a zaměřuje se na způsoby, jakými respondenti získávají informace, které byly vybrány v předchozí otázce. Nejčastější metodou získávání informací mezi respondenty (83) je procházení oficiálních stránek fakulty. Tento způsob dominuje oproti dalším kategoriím a ukazuje na důvěru v tento zdroj, společně s jednoduchým přístupem. Druhým nejčastějším způsobem mezi respondenty (68) je dotazování se spolužáků či starších studentů na určité informace. Konzultace s akademickými pracovníky je třetím nejvíce vybraným způsobem získávání informací mezi studenty, kteří odpovídali na dotazník (38), přímý kontakt se studijním oddělením a dalšími pracovníky fakulty je tedy cenný zdroj pro získávání určitých specifických informací.

Respondenti na druhou stranu takřka nevyužívají jako zdroje k získávání informací sociální síť fakulty (6), nebo jiné externí zdroje informací mimo webovou stránku fakulty (3). Je to zřejmě částečně dáno tím, že informace jako jsou termíny státních závěrečných zkoušek, vyhlášky a normy nebo harmonogram akademického roku (tedy nejvíce vybrané kategorie u předchozí otázky), nejsou jediné informace, které jsou zveřejňovány

na sociálních sítích fakulty. Mimo jiné jsou na sociálních sítích fakulty zveřejňovány informace týkající se Erasmu, Free movers a dalších zajímavých událostí.

Obr. 11 Preferované metody získávání informací o vybraných tématech



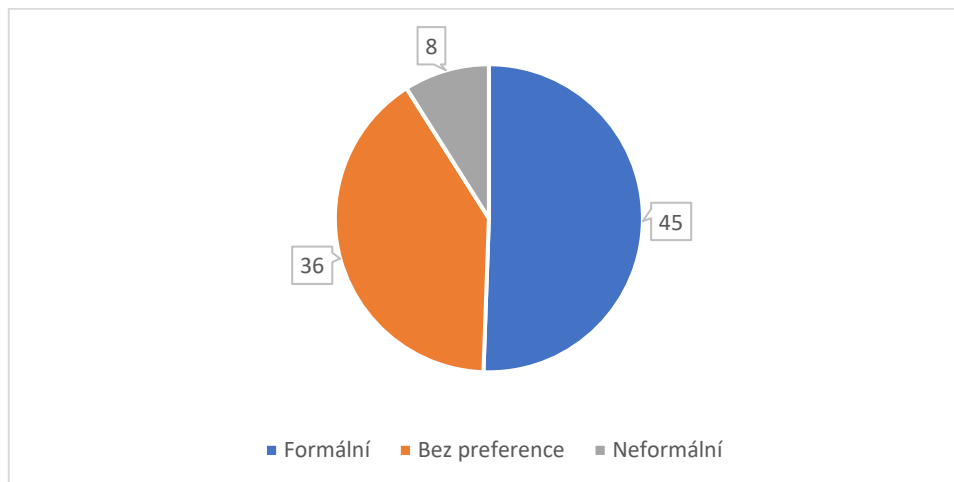
Zdroj: vlastní zpracování, 2024

Otázka 9. *Preferovali byste spíše formální nebo neformální tón chatbota?*

Další otázka, která je klíčová pro následnou konstrukci chatbota se týkala preference formálního či neformálního tónu. U této otázky odpovědělo 45 (50,56 %) respondentů, že preferují spíše formální tón, a 36 (40,45 %) respondentů, tedy druhá významná skupina respondentů, odpovědělo, že nemají žádnou preferenci. Pouze malá skupina respondentů 8 (8,99 %) vyslovila preferenci neformálního tónu pro chatbota. Velká skupina respondentů tedy očekává spíše profesionální zpracování jejich dotazů. Významná část respondentů nepovažuje tón komunikace chatbota za klíčový faktor.

Z těchto výsledků bude tedy vyvozeno, že při návrhu a samotné konstrukci chatbota pro fakultní webové stránky bude kladen důraz na určitý formálnější tón komunikace.

Obr. 12 Preference tónu chatbota



Zdroj: vlastní zpracování, 2024

Otázka 10. *Jaký máte postoj k integraci AI jazykových modelů do vzdělávacích nástrojů a webů vzdělávacích institucí?*

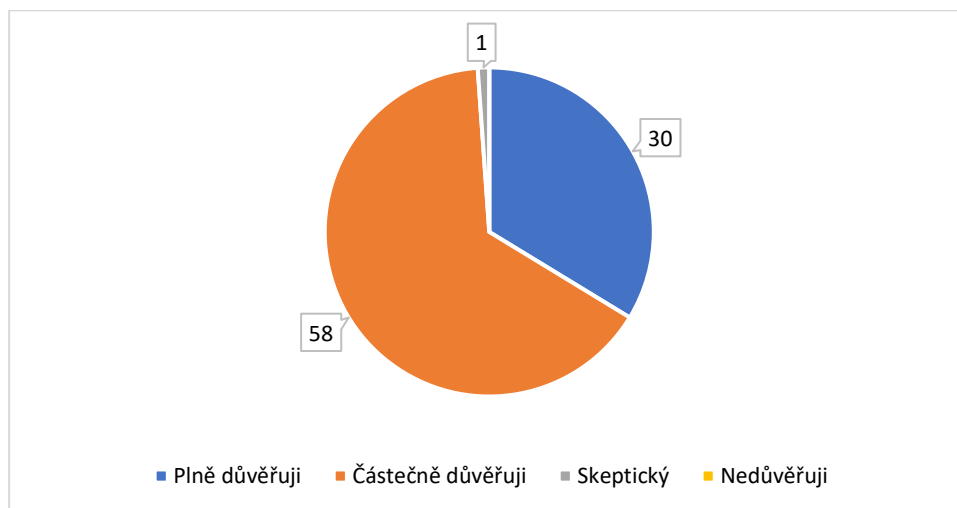
Doplňující otázka, která se týkala postojů vůči integraci jazykových modelů do vzdělávacích nástrojů a jiných prostředků získala pozitivní výsledky. 46 (51,7 %) respondentů odpovědělo, že je jejich postoj k této integraci poměrně pozitivní, druhá významná skupina respondentů 40 (44,9 %) odpověděla, že je jejich postoj velmi pozitivní a pouze 3 (3,4 %) respondenti odpověděli, že je jejich postoj k této integraci neutrální.

Otázka 11. *Jakou míru důvěry byste přikládali odpovědím poskytnutým chatbotem ohledně vašich studijních záležitostí?*

Klíčovou otázkou pro vybrané respondenty je, zda mají důvěru v daného chatbota jako zdroj pro získávání informací. Respondenti 30 (33,71 %), kteří by plně důvěřovali danému chatbotovi představují jednu ze dvou významných skupin u této otázky. Částečná důvěra ve výsledky chatbota je nejvíce vybranou odpovědí u této otázky, 58 (65,17 %) respondentů vybralo tuto odpověď. U této odpovědi byl také doplněk, že by respondent sice důvěřoval dané odpovědi, ale rád by si ji zkontroloval s jiným zdrojem, tento problém lze vyřešit tím, že společně s vytvořenou odpovědí poskytne chatbot také zdroj, ze kterého čerpal neboli tedy určité úryvky textu. Pouze jeden respondent vybral odpověď „Skeptický“, která představovala situaci, kdy by daný respondent konzultovat odpovědi od chatbota se studijním oddělením. Nikdo nevybral možnost plné nedůvěry, což může

být z toho důvodu, že respondenti, kteří danému chatbotovi nedůvěřují, zároveň odpověděli, že by fakultního chatbota nevyužili a tím pádem nepostoupili do podrobného dotazování.

Obr. 13 Míra důvěry v chatbota



Zdroj: vlastní zpracování, 2024

8.2 Shrnutí dotazníkového šetření

Dotazníkové šetření odhalilo několik klíčových informací o postojích k jazykovým modelům a potenciálnímu fakultnímu chatbotu. Odpovědi, které se týkají fakultního chatbota budou následně využity při finální konstrukci.

Studenti vykáží značný zájem o implementaci chatbota na webu fakulty.

Vypočítaná p-hodnota u otázky 5., která se týkala využití fakultního chatbota vyšla 0,419 při hladině významnosti 0,05, nelze zamítnout nulovou hypotézu. Neexistuje statisticky významný rozdíl mezi zjištěným zájmem respondentů o používání chatbota na webu fakulty a předpokládanými hodnotami stanovenými na základě podobného výzkumu.

Ačkoliv se některé hodnoty liší od předpokládaných, zájem o fakultního chatbota významně převyšuje negativní odpovědi, tak jak tomu bylo i u očekávaných hodnot. Zájem o fakultního chatbota mezi studenty je výrazný.

Více jak 50 % studentů plně důvěřuje v odpovědi chatbota.

Plná důvěra v chatbota nebyla nejvíce zastoupenou kategorií při dotazování na důvěru v odpovědi, ale druhou nejvíce zastoupenou. Pouze 33,71 % respondentů vybralo

možnost plné důvěry, v tomto případě se tedy předpoklad nepotvrdil. Nejvíce zastoupená kategorie vyjadřovala částečnou důvěru v odpovědi, kdy by si daný respondent některé informace kontroloval i s externími zdroji. Toto zastoupení poukazuje na to, že dokumenty a informace, které jsou uchovávány na oficiálních stránkách fakulty, by i při implementaci chatbota bylo potřeba neustále aktualizovat a udržovat ve stejné kvalitě.

9 Návrh a vývoj chatbota

Tato kapitola se bude zaměřovat na samotný návrh a vývoj chatbota. Jednotlivé kroky jsou rozděleny do dalších podkapitol pro přehlednost a oddělení jednotlivých kroků tohoto procesu.

9.1 Analýza požadavků na daného chatbota

Na základě dotazníkového šetření vzešly hlavní požadavky na to, aby chatbot uměl odpovídat na dotazy, které se týkaly určitých důležitých termínů, jako jsou například státní závěrečné zkoušky a další. Společně s tímto požadavkem byly také vysoké nároky na informace pocházející ze studijního zkušebního řádu, harmonogramu akademického roku a další...

Aby bylo zabráněno úniku interních či citlivých dat, bylo možné využít pouze informace a data, které jsou veřejně dostupná na stránkách fakulty ekonomické.

9.2 Návrh

Nejjednodušší způsob tvorby spolehlivého chatbota je chatbot, který je založený na pravidlech. Tento typ chatbota, jak již bylo nastíněno v předchozí části této práce, reaguje na uživatelské vstupy předem definovanými pravidly. Tento typ chatbota dokáže velmi dobře reagovat na často kladených otázek, což částečně řeší zmíněnou problematiku. (Albuger, 2023)

Problém nastává v případě složitých dotazů či konverzací, kdy tento typ chatbota nemusí správně fungovat. Tento typ také nejlépe pracuje s předem připravenými odpověďmi. Při využívání většího množství dokumentů je potřeba vytvořit velké množství těchto odpovědí, které následně mohou být irelevantní při aktualizaci určitých dokumentů. Tento typ tedy není optimálním řešením daného problému.

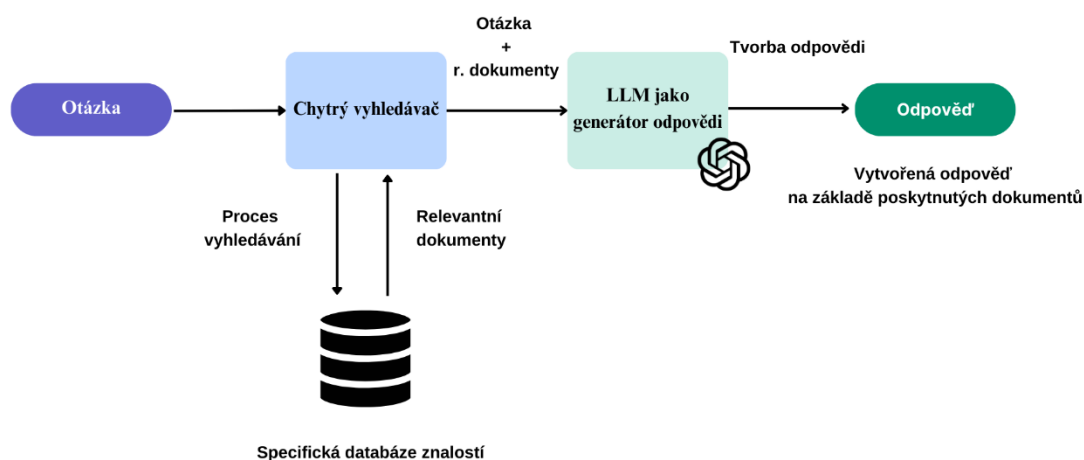
Velké jazykové modely, a hlavně tedy systémy, které jsou na nich následně postavené jako například ChatGPT, Gemini a další, jsou nové inovativní způsoby řešení problematiky chatbotů. Dokáží generovat kontextově relevantní odpovědi a vytvářet tak plynulou konverzaci s uživatelem. Samotný model od společnosti OpenAI lze implementovat a modifikovat pomocí API do vlastního systému. Problém při

implementaci samotného modelu GPT-4 je, že daný model s velkou pravděpodobností nebude studentovi podávat správné informace i když se tak daná struktura odpovědi bude zdát, tomuto termínu se říká již zmíněná halucinace. Je to dáno tím, že velké jazykové modely jsou trénované na obrovském množství dat, ale ne na specifických datech fakulty ekonomické Západočeské univerzity v Plzni. Zároveň nejsou učené, aby byli více obezřetné, protože by to mohlo vést k odmítnutí odpovědět i když zrovna daný model zná správnou odpověď. (OpenAI, 2022)

RAG

Řešení tohoto problému a zároveň architektura chatbota pro studijní účely funguje na rámci RAG. RAG (retrieval-augmented generation) neboli generování s rozšířeným vyhledáváním je nový rámec v této oblasti, který částečně řeší problém halucinace u velkých jazykových modelů a zároveň tedy velmi specifické dotazy. Společně s velmi specifickými dotazy řeší tento způsob také časově citlivé informace jako je například v případě fakulty harmonogram studia pro daný akademický rok. V zásadě odbourává problémy, které by vznikaly při samotné implementaci holého velkého jazykového modelu. RAG zapojuje externí získávání dat do procesu tvorby odpovědi a tím pádem zlepšuje schopnosti modelu správně odpovídat. Dochází v podstatě k tomu, že se daný model dotazuje jiného vedlejšího zdroje dat k získání relevantních informací pro odpovězení dané otázky. Tento systém tedy zajišťuje to, aby byly odpovědi daného modelu vytvářeny na základě získaných informací z externích dat a zvýšila se přesnost výstupu. (Gao et al., 2023)

Obr. 14 Jednoduchá architektura RAG



Zdroj: Blumenfeld. & Htet (2023), vlastní zpracování

Všechny prvky z výše uvedené architektury RAG budou vysvětleny v následujících kapitolách.

9.3 Výběr platformy

Systém, který byl využit pro tvorbu chatbota, se nazývá FlowiseAI. Tento nástroj spadá do skupiny otevřený systém (open-source). Přesněji tento software spadá do takzvané licence Apache, která požaduje po uživateli, aby zachoval autorství. Dále zde existují takzvané zřeknutí odpovědnosti za jakékoliv škody spojené s tímto softwarem. Tento typ licence umožňuje uživateli komerční využití, modifikace a distribuci. (Apache, n.d.)

Nástroj FlowiseAI umožňuje tvorbu velkých jazykových modelů a také AI agentů za využití takzvaného low-code vizuálního přístupu. Tento přístup umožňuje rychlý způsob prototypování a zkoušení různých způsobů a přístupů konstrukce AI agentů. Tento nástroj umožňuje propojit řadu nástrojů jako jsou paměti, nástroje načítání dat, rozdělovače textů, velké jazykové modely a další nové rámce v oblasti velkých jazykových modelů jako například rámec od společnosti LangChain, který byl využit pro tvorbu chatbota. Tento nástroj také umožňuje další rozšíření a integraci do vlastních aplikací za pomoci API (aplikační programové rozhraní). Zajímavou funkcí, který tento systém nabízí, je částečně

předpřipravené uživatelské rozhraní pro chatbota, které lze následně upravit a nastavit do vlastního systému. (FlowiseAI, n.d.)

9.4 Použité dokumenty a jejich úprava

Dokumenty, které byly využity pro vytvoření chatbota, jsou ve většině případů ve formátu PDF. Vybrané dokumenty zahrnují vyhlášky k přijímacímu řízení, studijní a zkušební řád, harmonogram pro daný akademický rok, dokumenty týkající se stipendií, směrnice týkající se kvalifikačních prací, vyhláška týkající se celoživotního studia, motivační vyhláška a další vyhlášky, směrnice a jiné důležité dokumenty.

Aby byl chatbot schopný odpovídat na otázky týkajících se informací, které nejsou obsažené v jednotlivých dokumentech FEK, ale lze je nalézt na stránkách fakulty, bylo potřeba vytvořit řadu vlastních souborů. Informace týkající se zahraničních studijních pobytů, informace o jednotlivých programech na fakultě, informace o akademických pracovnících jsou některé oblasti, na které bylo potřeba vytvořit soubory, které následně může chatbot využívat. Pro informace, které se mění v kratších časových intervalech, byl vytvořen program za pomoci programovacího jazyka Python, který tyto informace ze stránek fakulty zkopíruje a naformátuje dle daných potřeb. Těmto programům se také říká v angličtině web scraper⁵. V **příloze E** lze nalézt python kód, který zprostředkovává tuto funkcionalitu, v **příloze F** je vytvořené jednoduché uživatelské rozhraní pro tento program.

K optimálnímu fungování chatbota je potřeba vybrané dokumenty náležitě upravit. Přestože již momentálně existují způsoby načítání dat z formátu PDF, který se na fakultě používá. Systémy, které to umožňují nejsou perfektní, zejména při transformaci souborů v češtině. Aby byl tento problém vyřešen, bylo potřeba změnit formát veškerých souborů a přeformátovat tabulky do textového formátu. Tento způsob přepracování také umožnil přidat do souboru prvky markdown značkovacího jazyka, a tedy přidat určitou strukturu souborů jako několik úrovní nadpisů. Pro lepší přehlednost, ale hlavně pro další proces

⁵ „Web scraping je proces získávání dat z určité webové stránky.“ (Rouse, 2023)

přípravy dokumentů bylo využito tři úrovní nadpisů za pomoci tohoto značkovacího jazyka.

Dokumenty bylo také potřeba řádně očistit o určité duplicity nebo části, které již nejsou aktuální, jako například podpis děkana u vyhlášky, která je sice platná, ale tato osoba již není na pozici děkana. Tato informace totiž může následně vyvolat špatnou odpověď při dotazu „Kdo je děkan na fakultě ekonomické?“.

Segmentace textu (Text splitter)

V rámci procesu vyhledávání relevantních dokumentů často není nutné a zároveň je finančně nákladné vracet celý dokument velkému jazykovému modelu k vygenerování odpovědi. Stejně tak existuje výše zmíněné kontextové okno, které má své určité omezení. Z tohoto důvodu je využit proces segmentace textu na menší části (v angličtině chunk).

Existuje řada strategií, jakými je možné text segmentovat na menší části. Chunking neboli již zmíněná segmentace je další částí procesu přípravy dokumentů pro velké jazykové modely. Při určení strategie segmentace, je potřeba zvážit řadu proměnných, které hrají významnou roli při tomto rozdělování. V první řadě je potřeba určit povahu dokumentů, které jsou využívány, tedy jestli se jedná o dlouhé dokumenty nebo kratší zprávy. V případě fakultního chatbota se tato proměnná významně liší, jelikož některé dokumenty jako například studijní a zkušební řád jsou rozsáhlé, ale v celkovém souboru dokumentů je také řada jiných, které jsou spíše kratší povahy. Dále je potřeba vybrat takzvaný model embedování, o této problematice pojednává následující kapitola. Další proměnná se týká zohlednění, pro jaké využití budou jednotlivé segmenty využívány, kdy v tomto případě zejména k odpovídání na dotazy. (Cohen, 2023a)

Fixovaná nebo rekurzivní segmentace se v podstatě snaží udržet části, které mají stejný nebo podobný kontextuální význam spolu. Jelikož je spousta dokumentů, kterých bylo využito, již rozdělena na určité články/hlavy bylo možné toto rozdělení využít a označit je již zmíněnými nadpisy markdown jazyka. U částí, které neměli toto rozdělení, bylo potřeba jednotlivý text rozdělit vlastnoručně, tak aby nedošlo ke kontextuálnímu porušení. Za těchto podmínek bylo tedy využito specializovaného rozdělovače markdown, který rozpoznává syntaxi markdown jazyka a rozděluje text na jeho základě. Z vlastního testování vzešel požadavek přepracování souboru akademických pracovníků

na formát csv a využití takzvaného rekurzivního rozdělování⁶, jelikož u tohoto souboru nebylo dosahováno požadovaných výsledků. Oba tyto nástroje pocházejí z aplikačního rámce LangChain (LangChain, n.d.a)

9.5 Vektorová reprezentace slov (Embedding)

Důležitou částí vyhledávacího (retrieval) procesu je takzvaná vektorová reprezentace slov, kterou lze také nazvat vnoření slov.⁷ Tento proces v podstatě slouží k převodu textu či obrázků, neboli tedy vysoko dimenzionálních dat, na typ méně dimenzionálních dat za předpokladu, že dojde k zachování důležitých vztahů a struktury. Vnořená slova jsou vektory neboli v podstatě určitý seznam čísel, který může být ukládán v různých databázích. Generování a ukládání těchto vektorů zejména slouží k zefektivnění procesu vyhledávání. (Yu, 2023)

Vektory mají ideální datovou strukturu pro algoritmy strojového učení, a proto se tedy pro tyto účely využívají. Vektory lze popsat v prostoru a lze říci, že některé jsou si blíže a zároveň některé jsou od sebe vzdálené. Při převodu textu na vektory je důležité, aby texty, které mají podobný sémantický význam byly v prostorovém uspořádání blízko sebe. (Cohen, 2023b)

Pro převod textu na vektory lze využít řadu nástrojů, které se touto problematikou zabývají, v případě této práce bylo využito nejnovějšího modelu společnosti OpenAI text-embedding-3-small. Tento model převýšil svého předchůdce (ada v2) v srovnávacím testu (MTEB – Massive Text Embedding Benchmark) o 1,3 % procentního bodu, jeho významná výhoda je ale ve snížení ceny z \$0.00010 na tisíc tokenů na \$0.00002 za tisíc tokenů. I když tento model neleží na samotné špici srovnávacího testu, bylo potřeba také zohlednit právě zmíněnou cenu a také možnosti, které vybraná platforma nabízela. V rámci těchto modelů neboli tedy v rámci procesu vnoření slov, se objevuje pojem dimenzionalita, která v tomto případě značí délku daného vektoru. U nejnovějších modelů vnoření slov od společnosti OpenAI lze tuto dimenzionalitu konfigurovat, ale je

⁶ Tento rozdělovač segmentuje text na základě předem definovaných znaků, tím, že byl soubor přepracován do tabulkového formátu, byl rozdělovač nastaven na „n“ neboli rozděluje soubor na jednotlivé řádky kdy na každém řádku je právě jeden pracovník. (LangChain, n.d.a)

⁷ Pochází z anglického názvu Embedding.

nutné zohlednit, že větší dimenze mohou sice nést více informací, ale jsou tím pádem více finančně nákladné. V rámci vybraného modelu bylo použito základní nastavení 1536 dimenzí. (OpenAI, n.d.b)

Tento převod textu se využívá jak pro využití dokumenty, ze kterých se vytvořila znalostní databáze, tak i pro jednotlivé dotazy od uživatelů. Takto převedené dotazy se totiž následně využívají pro vyhledávání ve vektorové databázi, o které je následující kapitola.

Příklad embeddování:

„## čl. 7 – Opakování státní závěrečné zkoušky

1. Opakování státní závěrečné zkoušky nebo jejího předmětu se řídí čl. 49 Studijního a zkušebního řádu.

4. ... má právo státní závěrečnou zkoušku nebo její předmět opakovat v průběhu studia nejvýše dvakrát.“

Ve vektorové podobě tento text vypadá pro představu následovně:

[0.007572209, 0.0202694498, 0.0159868896, -0.0324953087, -0.0119650932, 0.0582909659, 0.0084397532, -0.00806866493, 0.0109420931... dalších 1527 dimenzí]

9.6 Vektorová databáze

Komponenta, která je jádrem systému RAG, je právě vektorová databáze. Vytvoření vektory z fáze vnoření slov je potřeba na určitém místě uložit, aby je bylo možné následně využít k získání relevantních dokumentů. Vektorové databáze jsou v podstatě databáze, které mají obdobné funkce jako regulární databáze s tím, že jejich specializace je právě na vytvořené vektory. Důležitou součástí vektorové databáze je následné vyhledávání v této databázi, ve které nedochází k hledání exaktní shody dat (tak jak tomu bývá u normálních databází.), ale k vyhledávání nejbližší shody. Poté tato databáze vrátí relevantní dokumenty k danému dotazu, které mají podobný sémantický nebo kontextový význam. Některé databáze také nabízejí typ vyhledávání s názvem maximální mezní relevance a další způsoby. (Microsoft, 2023)

V případě vytvořeného chatbota bylo využito vektorové databáze od společnosti Pinecone, která poskytuje všechny potřebné funkce k plynulému fungování vyhledávání

a vracení relevantních dat. Jelikož jsou v celkovém souboru dokumentů určité podobné dokumenty jako například vyhlášky týkající se přijímacího řízení pro jednotlivé stupně vysokoškolského vzdělání, bylo potřeba vytvořit více vektorových databází a rozdělit je na určité skupiny. K popsání jednotlivých databází, tak aby daný chatbot věděl, kterou využít, bylo využito specifického nástroje pro vyhledávač. Při tvorbě databáze bylo důležité nastavit specifické parametry jako dimenzionalitu, tak aby souhlasila s dimenzionalitou vybraného modelu pro vnoření slov, tedy 1536. V rámci nastavení databází je také možné nastavit počet nejlepších výsledků, které by měly být načtené (hodnota k nejlepším výsledkům). Tato hodnota byla nastavena u všech databází na 4. Při vyšším počtu dochází k zaplnění kontextového okna, a tedy k vyšší finanční zátěži, jelikož je spotřebováno více tokenů.

Dále bylo potřeba vybrat typ měření podobnosti mezi vektory. Tyto matematické metody slouží k porovnání vektorů v prostoru neboli tedy jak si jsou podobné. Existuje několik způsobů, společnost OpenAI ale u svých modelů vnoření slov doporučuje využít metodu kosinové podobnosti, která tedy následně byla použita u vytvořených databází. U kosinové podobnosti dochází k měření kosinu úhlů, které svírají dva vektory v prostoru, kdy se jejich hodnoty pohybují od -1 a 1. Hodnota -1 značí, že jsou vektory přesně opačné, 0 představuje ortogonální vektory a hodnota 1 identické vektory. (Cohen, 2023c)

Vyhledávací nástroj

Tato komponenta slouží v programu Flowise, ale zejména v rámci LangChain k pojmenování a popisu jednotlivých vektorových databází, tak aby daný agent/chatbot věděl, za jaké situace má danou databázi využít. Zároveň to slouží jako propojení mezi využitým agentem a databázemi. Každá z těchto komponent obsahuje stručný popis dokumentů, které jsou obsažené v dané vektorové databázi.

9.7 Agent

V poslední řadě je důležité vybrat takového agenta/asistenta, který je optimální pro daný problém. V případě fakultního chatbota bylo využito takzvaného konverzačního vyhledávacího agenta. Tento agent využívá velký jazykový model ke generování odpovědi, vytvořené nástroje vyhledávání k získání potřebných informací, a zároveň má

zabudovanou paměť, díky které dokáže s uživatelem udržet plynulou konverzaci. Společně s těmito komponentami lze chatbotovi nastavit určitou systémovou zprávu, která by měla sloužit jako určité instrukce pro daného agenta, jak by měl pracovat.

Vybraný agent je optimalizovaný pro proces vyhledávání a vracení dokumentů v určité konverzaci, pomocí volání funkcí od společnosti OpenAI, kdy v podstatě daný model dokáže rozpoznat, kdy má zavolat danou funkci. V případě fakulního chatbota se jedná o zavolání funkcí vyhledávání správných informací z jednotlivých vektorových databází. (LangChain, n.d.b)

Tento agent byl vybrán z důvodu, že dokáže využít řadu vektorových databází díky zmíněné možnosti volání funkcí, a zároveň, že daný agent využívá také komponentu paměti.

Generativní část (openai)

Jak již bylo zmíněno, agent využívá velký jazykový model, a to zejména ke generování výsledné odpovědi. Společně s vyhledanými dokumenty/segmenty textu a dotazem uživatele je také přidávána určitá historie konverzace v případě, že se jedná o navazující dotaz, jako vstup pro daný jazykový model. Zapojení paměti do vstupu tohoto modelu umožňuje opravdovou konverzaci s chatbotem, jelikož může reagovat na zmíněné dodatečné dotazy. Jazykový model dokáže pracovat s vybranými segmenty textu a zpracovat z nich smysluplnou odpověď na danou otázkou. (Gao et al., 2023)

V rámci vytvořeného chatbota bylo využito velkého jazykového modelu od společnosti OpenAI. Tento chatbot využívá nejnovější model z řady GPT-4 Turbo, gpt-4-0125-preview. Tento model disponuje velkým kontextovým oknem o velikosti 128 tisíc tokenů, a zároveň má pozoruhodné skóre ve srovnávacím testu MMLU (Massive Multitask Language Understanding) v mnoha jazycích. (Achiam et al., 2024)

Paměť

Další komponenta důležitá pro vybraného agenta je paměť. Vybraná paměť, kterou výsledný chatbot využívá se nazývá Buffer memory a jedná se o dočasnou paměť kam se daná data z konverzací ukládají. Uložení trvá do té doby, dokud není daný chatbot restartován, nebo pokud není manuálně vyčištěna. Tento způsob lze nahradit cloudovou databází, kde mohou být data uchována napořád.

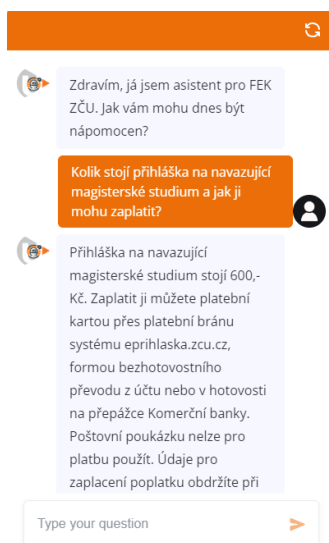
9.8 Finální systém

Finální celkový systém je vcelku komplexní. Tak aby byla minimalizována situace, kdy dochází k halucinaci či špatnému vrácení informací, bylo potřeba systém RAG rozdělit na několik skupin. Specificky bylo potřeba rozdělit část RA, při které dochází k vyhledávání souborů a následnému vrácení relevantních textů společně s uživatelským dotazem, tak aby mohl daný jazykový model přesně odpovědět. Rozdělení bylo provedeno do 8 skupin, které reprezentují vždy určitou společnou oblast, například skupina navazujícího magisterského studia obsahuje soubory týkající se přijímací řízení, odpovídající studijní programy a další směrnice. Všechny tyto skupiny jsou následně propojeny s vybraným agentem, který za pomoci zmíněného procesu volání funkcí vybírá právě danou hodící se skupinu. V **příloze B** a v **příloze C** lze vidět reprezentaci jedné skupiny a celého systému v platformě Flowise.

9.9 Uživatelské rozhraní

V této kapitole je nastíněno uživatelské rozhraní daného chatbota. Pro účely následného testování chatbota se studenty, ale také v podstatě k simulaci toho, jak by daný chatbot mohl vypadat ve výsledné implementaci, bylo vytvořeno uživatelské rozhraní na webové stránce fakulty ekonomické. Uživatelské rozhraní chatbota bylo navrženo s cílem maximalizovat jeho přívětivost a uživatelský komfort pro studenty zapojené do testování. Hlavní funkce tohoto chatbota byly převzaty z platformy Flowise, která nabízí již částečně připravené uživatelské rozhraní. Toto rozhraní bylo následně upravené, tak aby odpovídalo webové stránce fakulty ekonomické.

Obr. 15 Uživatelské rozhraní chatbota



Zdroj: vlastní zpracování, 2024

V **příloze A** lze sledovat uživatelské rozhraní chatbota na webové stránce fakulty ekonomické.

Logo chatbota a název

Pro vytvoření chatbota bylo také potřeba vytvořit logo a název, které by ho představovalo. Při tvorbě loga byl kladen důraz na to, aby bylo logo sladěno se vzhledem webové stránky fakulty, ale také s logem fakulty. Proto bylo nejdříve vytvořeno samotné logo chatbota s barvami FEK, a následně bylo do pozadí přidáno logo fakulty ekonomické. (Airey, 2014)

Obr. 16 Logo chatbota



Zdroj: vlastní zpracování, 2024

Dále bylo potřeba vytvořit název pro daného chatbota, tak aby se mohl představit v úvodní zprávě a zároveň při ukončení zprávy jako podpis. Název, který byl pro daného chatbota vytvořen je, FEK ŽČU asistent.

10 Testování navrženého chatbota konverzační agent

V rámci testování bylo porovnáno několik modelů, které byly zhotovené jako prototypy. U některých prototypů docházelo k příliš velké halucinaci či k jiným chybám od samého počátku, a tedy k podrobnějšímu testování již nebyly zařazeny.

Cíle:

1. Analyzovat frekvenci halucinací chatbota během dotazování.
2. Vyčíslit jaké jsou finanční náklady spojené provozem chatbota pro školní účely.

V prvním stupni testování šlo o porovnání vytvořeného systému RAG s velkým jazykovým modelem. Tento jednoduchý test byl zhotoven na základě poskytnuté otázky pro oba systémy a následné analýzy jednotlivých odpovědí.

Testování s dalšími prototypy je významnou částí tohoto testování, jedná se porovnání několika nejlepších prototypů, které vznikly při tvorbě optimálního chatbota. V této části bylo využito několik nástrojů tak, aby bylo možno co nejpřesněji zachytit nejdůležitější metriky, které jsou klíčové pro následný provoz chatbota. Nejdůležitější aspekty, které lze u chatbota testovat, je rychlost (latence), cena, halucinace a celková tvorba odpovědi. Tyto aspekty v rámci následujícího testování byly rozděleny do určitých specifických a měřitelných metrik, tak aby bylo možné jednotlivé prototypy mezi sebou porovnávat.

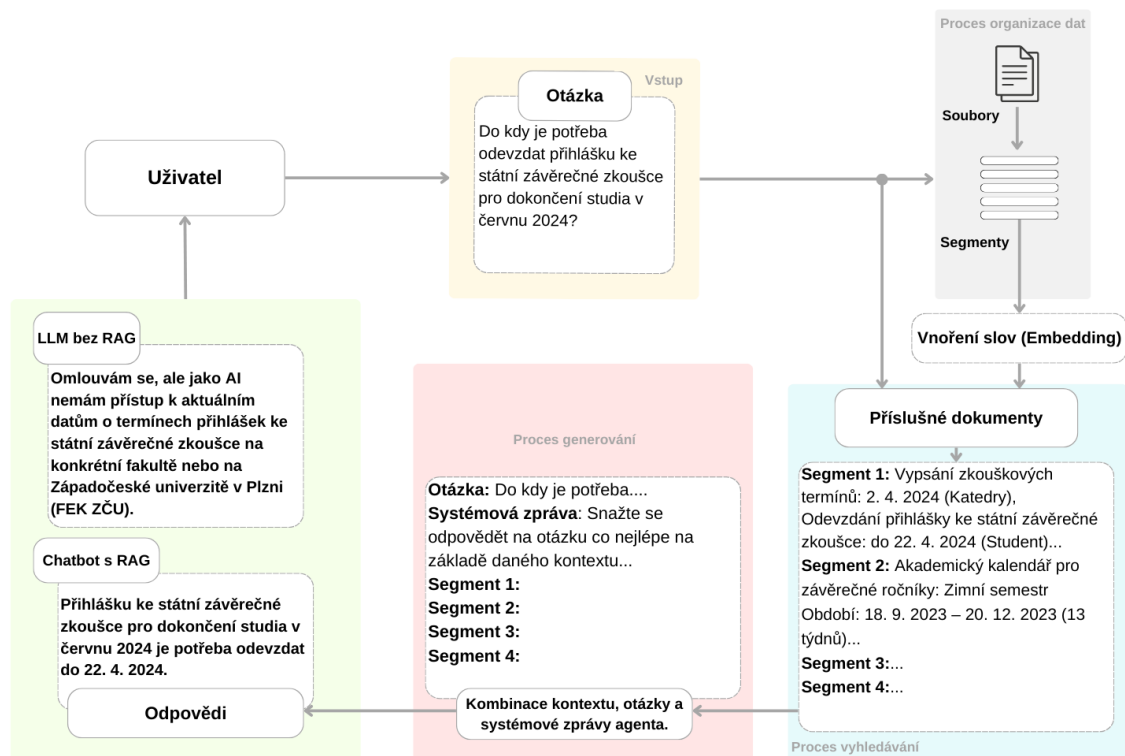
Postup testování

Testování je rozděleno do 3 částí, kdy v každé části byla potřebná data sbírána jiným způsobem. V první fázi testování byl využit u většiny případů nástroj LangSmith, u některých výjimek bylo potřeba využít také platformy OpenAI. V druhé fázi testování byla potřebná data přebírána z jednotlivých modelů a následně zpracovávána v programu Microsoft Excel. V poslední fázi bylo využito nástroje Ragas k získání hodnot jednotlivých důležitých metrik.

10.1 Porovnání s velkým jazykovým modelem

V první řadě lze jednoduše porovnat vytvořeného chatbota oproti velkému jazykovému modelu (GPT-4-1106-preview), který nevyužívá vytvořený systém RAG a ukázat tak tedy důležitost tohoto systému pro některé aplikace.

Obr. 17 Porovnání vytvořeného chatbota s velkým jazykovým modelem



Zdroj: Gao et al. (2023), vlastní zpracování

V tomto případě byla dotázána jednoduchá otázka týkající se termínu podání přihlášky na státní závěrečnou zkoušku, kdy chatbot využívající RAG odpověděl správně díky získanému kontextu. Samotný velký jazykový model v tomto případě uvedl, že nemá přístup k aktuálním datům a není tedy schopen danému studentovi s touto otázkou pomoci. Tento model lze mimo jiné vylepšit funkcí, při které model prohledává informace na internetu tak, jak je tomu v systému ChatGPT. I při této funkci ale model neodpověděl správně a na místo správného datumu uvedl termín pro přijímací řízení na fakultu. Je to dáno tím, že správná informace se nachází v příloženém souboru na stránce fakulty. Samotný velký jazykový model, v tomto případě GPT-4-1106-preview nelze využít pro účely fakultního chatbota.

10.2 Testování s dalšími prototypy

V rámci přípravování chatbota bylo vytvořeno několik prototypů za účelem najít nejlepší řešení pro tuto problematiku fakultního chatbota. S postupným přidáváním souborů totiž určité systémy přestávaly být dostatečně validní, a bylo tedy potřeba najít lepší řešení.

V rámci podrobnějšího testování prototypů bylo využito 5 různých asistentů, agentů či jiných „řetězců“⁸, které slouží jako chatbot pro studentské záležitosti. V rámci těchto prototypů je také výsledný využitý chatbot s názvem Konverzační agent pro porovnání s ostatními. V následující kapitole jsou zjednodušeně popsány tyto prototypy.

Jednotlivé prototypy

Assistant API je rozhraní, které umožňuje vytvořit si vlastního asistenta pro určité účely ve vlastní aplikaci. Toto rozhraní je vytvořené společností OpenAI a představuje jeden z nejjednodušších způsobů, jak vytvořit funkčního chatbota pro fakultní účely. Toto rozhraní nabízí všechny potřebné nástroje jako je prohlížení webových stránek a vyhledávání v znalostní databázi. Společně s těmito nástroji také umožňuje tvůrci nastavit asistentovi určité instrukce, kterými by se měl řídit. Je nutné zmínit, že znalostní databáze, a hlavně tedy proces vyhledávání nelze nikterak upravovat či měnit což představuje určité problémy. Stejně tak má znalostní databáze omezení na 20 souborů, které lze využít. Ke generaci odpovědi bylo využito modelu (GPT-4-1106-preview). (OpenAI, n.d.c)

Ethan (2023) ve svém testování tohoto typu asistenta uvádí, že nejlepší výsledky vycházejí při využití pouze jednoho souboru v znalostní databázi. Tento názor se potvrdil i při vlastním testování, a proto finální verze tohoto prototypu využívá pouze 1 PDF soubor.

Vectara je další platforma, která umožňuje spojit generativní umělou inteligenci s různými aplikacemi. V rámci portfolia svých služeb nabízí pro účely fakultního chatbota důležité vyhledávání v dokumentech. Další prototyp byl zhotoven za využití vektorové databáze od společnosti Vectara a řetězce k odpovídání dotazů⁹ od stejné společnosti. Ke generaci odpovědi bylo využito modelu GPT 3.5. (Vectara, n.d.)

Konverzační QA je řetězec, který pochází z rámce LangChain, který standartně vyhledává potřebné dokumenty a následně z nich vrací danou odpověď. Tento řetězec je

⁸ Řetězec je pojem, který využívá rámec Langchain a jedná se o spojení určitých komponent v daném pořadí a slouží k provádění určitých úkolů, v tomto případě k odpovídání na dané otázky podle vyhledaného kontextu. (Sahota, 2023)

⁹ vectara-experimental-summary-ext-2023-10-23-small (gpt-3.5-turbo)

zároveň doplněn funkcí paměti, stejně jako tomu je u konverzačního agenta a dokáže tedy pracovat s navazujícími dotazy. Rozdíl oproti konverzačnímu agentovi je, že tento systém dokáže pracovat pouze s jednou vektorovou databází. Ke generaci odpovědi bylo využito modelu (GPT-4-1106-preview). (Langchain, n.d.c)

Multi Retrieval QA je další řetězec z rámce LangChain, který funguje na trochu jiné bázi než předchozí. Tento řetězec totiž dynamicky vybírá systém vyhledávání, který by měl být využit pro vyhledání potřebných dokumentů a poté pomocí něj odpoví na danou otázku. V tomto případě je tedy využito několik vektorových databází a systémů vyhledávání. Oproti předchozímu řetězci tento řetězec nemá komponentu paměti. Ke generaci odpovědi bylo využito modelu (GPT-4-1106-preview). (Langchain, n.d.d)

10.3 Testování pomocí LangSmith

První způsob testování těchto prototypů byl vyhotoven za pomoci platformy LangSmith. LangSmith umožňuje monitorování aplikací využívající velké jazykové modely. Při tvorbě či užívání velkých jazykových modelů může dojít k řadě neočekávaných událostí. Pomocí této platformy lze sledovat jednotlivé kroky, ke kterým při průběhu došlo a najít tak určitou chybu či problém. LangSmith lze také využít k monitorování latence, spotřebovaných tokenů pro daný dotaz a k určení ceny za daný dotaz a odpověď. Mimo tyto tři hlavní metriky byl LangSmith také využit při návrhu finálního chatbota k monitorování jednotlivých kroků a následné optimalizaci. (LangChain, 2023)

Následující tabulka ukazuje jednotlivé hodnoty vybraných metrik pro všechny prototypy kromě prototypu Vectara, který se nezdařilo integrovat s platformou LangSmith. Hodnoty, které jsou v této tabulce, představují údaje pro 30 vyvolaných otázek a odpovědí.

Tab. 5 Testování za pomoci LangSmith

		LangSmith		
Typ	Znalostní databáze	Latence	Tokeny	Cena
Assistant API GPT- 4	PDF 1 soubor	16,32s	8105,7	0,0926
Konverzační agent	26 txt souborů	20,24s	3595,519	0,0465
Vectara (GPT-3.5)	26 txt souborů	x	x	x
Konverzační QA	26 txt souborů	21,53s	2289,3	0,0278
Multi Retrieval QA	26 txt souborů	14,96s	2569,556	0,0308

Zdroj: vlastní zpracování, 2024

Latence v kontextu chatbotů značí, jak dlouho trvá, než daný chatbot odpoví studentovi na jeho otázku. V této době je zahrnuto zpracování vstupu (otázky) a následná generace odpovědi, přičemž je nutné zmínit, že se tato latence počítá až po dokončení generativní části. „Nejrychlejší“ u této metriky je Multi Retrieval QA se P50 14,96s. Latence P50 značí 50. percentil latence odkazuje tedy na medián pro změřenou latenci. Všechny hodnoty latence jsou uvedeny v těchto mediánových hodnotách. Výsledný chatbot (konverzační agent) má v tomto případě hodnotu latence P50 20,24s

Počet tokenů, které byly spotřebovány při zpracování jedné otázky a vytvoření jedné odpovědi lze sledovat v sloupci tokeny. U této metriky bylo využito průměru pro reprezentaci této hodnoty. Nejšetrnější na spotřebu tokenů byl Konverzační QA s 2289 tokeny. Tokeny lze rozdělit na vstupní a výstupní, pro zjednodušení byly ale spojené dohromady. Významné navýšení spotřeby tokenů lze sledovat u Assistant API. Jelikož se jedná o uzavřený systém, který v platformě LangSmith není rozdělen na jednotlivé kroky a k nim přiřazené tokeny není možné sledovat, jak jsou dané tokeny spotřebované. Hodnoty spotřebovaných tokenů bylo potřeba sledovat na oficiálních stránkách společnosti OpenAI. Konverzační agent v tomto případě stojí na předposledním místě, jeho navýšená spotřeba tokenů je částečně vyvolána tím, že při výběru správné funkce musí přečíst popis všech funkcí, které jsou k dispozici.

Při určení **ceny** je potřeba rozdělit vstupní tokeny a výstupní tokeny, jelikož se cena za tyto dva druhy liší. Cena za výstupní tokeny neboli generativní část bývá vyšší. Hodnoty v tabulkách jsou v Amerických dolarech (USD). Tento postup bylo potřeba provést u prototypu Assistant API, jelikož stejně jako správný počet tokenů není v platformě LangSmith určená cena pro tohoto asistenta. U ostatních hodnot byl vypočítán průměr z jednotlivých hodnot, které jsou již v platformě správně vypočítané. Jelikož se cena odvíjí od spotřebovaných tokenů, nemění se pořadí jednotlivých chatbotů oproti metrice tokenů. U prototypu Vectary lze zmínit, že tato společnost nabízí zajímavou cenovou nabídku, 15 000 otázek a odpovědí zdarma na měsíc, a poté 1,25 USD za každých 1000 otázek a odpovědí.

10.4 Hodnocení odpovědí

V této části testování bylo využito stejných 30 vyvolaných otázek a odpovědí a následně u nich bylo anotováno hodnocení jednotlivých odpovědí. Hodnocení bylo rozděleno na 3 skupiny, kdy první skupina „Korektní“ představuje odpovědi, které byly jak kontextově, tak smysluplně a detailně zpracované. Skupina „Halucinace“ představuje takové odpovědi, u kterých došlo buďto k úplné halucinaci systému a odpověď tedy vůbec neodpovídala dané otázce, nebo odpovědi, u kterých byla zaměněna důležitá část jako je například datum. Skupina „Neúplné“ představuje takové odpovědi, jak už z názvu napovídá, u kterých chyběla určitá část na to, aby byla odpověď zcela kompletní.

Tab. 6 Vlastní hodnocení odpovědí

		Vlastní testování				
Typ	Znalostní databáze	Korektní	Halucinace	Neúplné	% Halucinace	Celkem
Assistant API GPT- 4	PDF 1 soubor	18	6	6	20,00%	30
Konverzační agent	26 txt souborů	26	1	3	3,33%	30
Vectara (GPT-3.5)	26 txt souborů	15	3	12	10,00%	30
Konverzační QA	26 txt souborů	18	9	3	30,00%	30
Multi Retrieval QA	26 txt souborů	25	1	4	3,33%	30

Zdroj: vlastní zpracování, 2024

Assistant API měl největší problémy se záměnou důležitých datumů, v některých případech ale došlo k úplné halucinaci, kdy tento asistent odpovídal za využití irelevantního souboru. V rámci neúplných odpovědí docházelo zejména k rozpůlení určitého seznamu informací, které měli být v jedné odpovědi, jako například seznam studijních programů pro daný stupeň studia.

Konverzační agent v tomto hodnocení uspěl nejlépe, jeho jediný nedostatek spočívá s větších listech informací, jako například vyjmenování všech partnerských vysokých škol v dané zemi, kde jich je více jak 5. V tomto případě je potřeba se zeptat na dodatečný dotaz, který díky komponentě paměti dokáže správně zpracovat.

Vectara při tomto testování neměla příliš velký počet halucinací, ale oproti jiným prototypům, častokrát docházelo k neúplným odpovědím. Těžko lze usoudit z jakého důvodu k této situaci docházelo, protože se jedná o uzavřený systém, částečně na to ale může mít vliv, že tento prototyp využívá model GPT 3.5 ke generaci odpovědí.

Konverzační QA – tento prototyp měl podobné potíže jako Assistant API, ale v některých případech dokonce kombinoval špatné dokumenty, tak, že při otázkách na přijímací řízení pro určitý stupeň studia utvářel odpověď z dokumentů pro různé stupně studia.

Multi Retrieval QA je druhý v pořadí v rámci tohoto hodnocení. Úroveň odpovědí odpovídá kvalitě Konverzačního agenta v podstatě skoro ve všech aspektech. Problém u tohoto prototypu přichází v situacích, kdy je daná odpověď neúplná a není zde možnost se doptat na další informace, jelikož tento řetězec, na kterém je prototyp postavený, neumožňuje zapojení komponenty paměti.

10.5 Testování za pomoci Ragas

V poslední části testování byl použit rámec Ragas, který slouží k hodnocení systémů RAG. Ragas nabízí nový způsob, jak hodnotit a systémy RAG, a to za pomoci velkých jazykových modelů, které v pozadí vyhodnocují určitá data.

V rámci tohoto testování bylo potřeba vytvořit speciální datovou sadu, která obsahovala otázky a správné odpovědi, kdy následně byl testován vyhledaný kontext (segmenty textu) a vygenerovaná odpověď u jednotlivých prototypů u těchto otázek. Je nutné zmínit, že Assistant API zde nebylo možné testovat z důvodu uzavřenosti jeho procesu vyhledávání.

Tab. 7 Testování s Ragas

		Ragas			
Typ	Znalostní databáze	Věrnost	Relevantnost od.	Vybavení kontextu	Relevance kont.
Assistant API GPT- 4	PDF 1 soubor	x	x	x	x
Konverzační agent	26 txt souborů	0,9870	0,8938	0,8704	0,3061
Vectara (GPT-3.5)	26 txt souborů	0,9563	0,8721	0,8678	0,3462
Konverzační QA	26 txt souborů	0,9667	0,6136	0,9167	0,1931
Multi Retrieval QA	26 txt souborů	0,9750	0,8945	0,8721	0,2964

Zdroj: vlastní zpracování, 2024

První metrika **věrnost** ukazuje na to, do jaké míry se vygenerovaná odpověď shoduje s vyhledaným kontextem. Počítá se za pomoci analýzy odpovědi a daného kontextu a její hodnota je v intervalu od 0 do 1, kdy 1 znamená nejlepší hodnotu/shodu. Ačkoliv jsou výsledky u této metriky velmi podobné, nejlépe se umístil výsledný chatbot konverzační

agent. Lze konstatovat, že většina prototypů generovala danou odpověď na základě poskytnutého kontextu. (Ragas, n.d.a)

Relevantnost odpovědi je metrika, která se soustředí na určení, jak moc je daná odpověď relevantní vzhledem k dané otázce a kontextu. Tato metrika nezohledňuje pravdivost odpovědi jako spíše to, zdali je odpověď neúplná či přeplněná nepotřebnými informacemi. U této metriky se v podstatě o první místo dělí prototyp Multi Retrieval QA společně s výsledným chatbotem z toho důvodu, že je jejich proces získávání jak kontextu, tak následné generace odpovědi velmi podobný. (Ragas, n.d.b)

Následující dvě metriky se týkají zejména vyhledaného kontextu, a tedy hodnotí vyhledávací proces daného systému.

Vybavení kontextu poukazuje na to, jak se daný vyhledaný kontext shoduje s připravenou správnou odpovědí. Určuje se na základě porovnání této správné odpovědi a vyhledaného kontextu, kdy výsledné hodnoty jsou v rozmezí 0 až 1, 1 označuje lepší hodnotu. Překvapivě je na prvním místě u této metriky Konverzační QA, ale po následné manuální analýze kontextu bylo zjištěno, že tento prototyp v určitých případech vracel stejný segment textu několikrát a tím pádem výsledek u této metriky je spíše zkreslený. (Ragas, n.d.c)

Relevance kontextu je jedna z nejdůležitějších metrik, co se týče vyhledávání kontextu, u této metriky dochází k porovnání dané otázky a vyhledaného kontextu, přičemž se počítá jako podíl relevantních vět k dané otázce v kontextu s celkovým počtem vět v daném kontextu. U této metriky lze sledovat významný pokles oproti předchozím a je to zřejmě dáno tím, že jednotlivé prototypy (kromě Vectara) vracejí vždy 4 nejlepší výsledky, přičemž některé z těchto výsledků již nejsou zjevně důležité pro danou otázku. Tento problém lze řešit snížením počtu této veličiny nebo přidáním komponenty filtr (tuto komponentu nelze využít u všech prototypů). Tato veličina nebyla snížena u finálního chatbota z toho důvodu, že u komplexnějších otázek je potřeba vrátit více segmentů textu. (Ragas, n.d.d)

10.6 Shrnutí testování a volba

Ačkoliv již byla představena struktura daného chatbota, je potřeba představit finální shrnutí, které vyhodnotí důvody pro výběr tohoto chatbota. Bylo tedy potřeba přepočítat

nejdůležitější metriky neboli normalizovat metriky a převést hodnoty z různých metrik na určitou srovnatelnou škálu. U metrik jako je například cena bylo využito takzvané obrácené normalizace, kdy Assistant API, jakožto nejdražší z prototypů má normalizovanou hodnotu 1 a ostatní vyšší, protože je jejich cena nižší.

Příklad výpočtu normalizované hodnoty ceny Assistant API: $\frac{0,09257}{0,09257} = 1$

Příklad výpočtu normalizované hodnoty ceny Konverzační agent: $\frac{0,09257}{0,0465} = 1,99$

Výpočet přímé normalizace byl využit u metrik Relevantnost odpovědi a Relevance kontextu. Neboli tedy nejvyšší hodnota (Multi Retrieval QA) má normalizovanou hodnotu 1 a ostatní prototypy mají nižší.

V poslední řadě bylo také potřeba vytvořit metriku paměť a následné váhy pro jednotlivé metriky. Paměť představuje důležitou komponentu u jednotlivých prototypů a řeší například problematiku u některých neúplných odpovědí, ale také schopnosti udržet konverzaci. Prototypům s integrovanou danou komponentou byla přiřazena hodnota 2, zatímco prototypům bez této komponenty byla přiřazena hodnota 0. Váhy byly určeny na základě konzultace s vedoucím práce, kde vzešlo, že nejdůležitějším aspektem v tomto případě je nejnižší počet výskytů halucinace. Dále je významná samotná cena, z důvodu finanční zátěže pro fakultu, stejnou váhu má i již zmíněná paměť. Poté následuje metrika latence, relevance odpovědi a relevance kontextu.

Normalizované hodnoty byly následně vynásobeny určenými vahami. Výsledné hodnoty, společně s finální celkovou hodnotou pro jednotlivé prototypy, lze sledovat v tabulce níže.

Tab. 8 Souhrnná tabulka testování

Metriky	Assistant API GPT- 4	Konverzační agent	Vectara (GPT-3.5)	Konverzační QA	Multi Retrieval QA	Váhy
Cena	0,20	0,40	0,40	0,67	0,60	20,00%
Latence	0,13	0,11	0,00	0,10	0,14	10,00%
Halucinace	0,60	3,60	1,20	0,40	3,60	40,00%
Paměť	0,40	0,40	0,00	0,40	0,00	20,00%
Relevantnost od.	0,00	0,05	0,05	0,03	0,05	5,00%
Relevance kont.	0,00	0,04	0,05	0,03	0,04	5,00%
Celkem	1,33	4,60	1,70	1,63	4,44	100,00%

Zdroj: vlastní zpracování, 2024

Na základě testování byl vybrán již zmíněný a představený **konverzační agent**. Navzdory tomu, že prototyp Multi Retrieval QA prokázal srovnatelné výsledky s nižšími náklady a menší latencí ve srovnání s konverzačním agentem, absence možnosti integrace komponenty paměti znemožnila výběr tohoto chatbota pro testování se studenty. Konverzační agent byl následně využit při testování se studenty.

Míra halucinace u finálního chatbota byla vypočítána pro 30 vytvořených otázek a odpovědí. Výsledná procentuální hodnota halucinace činila 3,33 %. Tato hodnota je přijatelná, je ale nutné zmínit, že tento chatbot měl 3 neúplné odpovědi.

Finanční náklady spojené s provozem chatbota se u stejného počtu testování rovnala 0,0465 USD za jeden zpracovaný dotaz a následnou odpověď. Tedy s momentálním kurzem 23,4167 USD/CZK vychází jeden zpracovaný dotaz s odpovědí na 1,09 Kč.

11 Testování chatbota studujícími

V této kapitole bude rozebrána metoda testování se studenty a následné vyhodnocení použitelnosti chatbota. V úvodu praktické části již bylo zmíněno, že vyhodnocení bude provedeno pomocí speciálně vytvořeného dotazníku použitelnosti chatbotů (CUQ).

Cíle:

Posoudit, jak vnímají studenti přesnost a užitečnost informací poskytovaných chatbotem.

Zhodnotit míru použitelnosti daného fakulního chatbota.

11.1 Postup testování

Dotazník použitelnosti CUQ pochází z práce od Holmes a kol. (2019), která se zabývá testování chatbota ve zdravotnictví. V rámci této práce byl tento dotazník poprvé otestován a validován. Jedná se v podstatě o specializovaný dotazník směřovaný na konverzační chatboty, který se, ačkoliv je na něm částečně postaven, odklání od obecnějšího přístupu stupnice použitelnosti systému (SUS – System Usability Scale). Tento způsob testování výhradně zkoumá a hodnotí osobnost, porozumění, odpovědi, zpracování chyb a inteligenci daného chatbota.

CUQ je postaven tak, aby bylo možné jednotlivé systémy porovnávat i s výsledky s více využívaným způsobem SUS. Tento dotazník obsahuje 16 otázek, které jsou zaměřené na zmíněné charakteristiky chatbota, na které odpovídá respondent za pomoci vytvořené Likertovy škály s 5 stupni. Škála začíná od „Silně nesouhlasím“ až po „Silně souhlasím“. Originální otázky z tohoto dotazníku bylo potřeba přeložit a upravit do českého jazyka. (Holmes a kol., 2019)

Otázky z CUQ dotazníku lze v podstatě rozdělit do 8 skupin, každá z těchto skupin má jednu otázku, která se dotazuje na pozitivní aspekty a jednu, která se dotazuje na negativní aspekty. Skupiny zní následovně: Osobnost chatbota (otázky 1 a 2), úvodní zprávy (otázky 3 a 4), účel (otázky 5 a 6), udržení toku konverzace (otázky 7 a 8), porozumění (otázky 9 a 10), odpovědi (otázky 11 a 12), zpracování chyb vstupu (otázky 13 a 14), v poslední řadě jednoduchost použití (otázky 15 a 16). (Boian et al, 2024)

Sběr dat

Sběr dat v tomto případě probíhal tak, že nejprve daný student testoval vytvořeného chatbota. Jelikož se chatbot v této fázi nachází v uzavřeném prostředí, byl autor přítomen u všech studentů, kteří testovali vytvořeného chatbota. Aby výsledky testování co nejpřesněji reflektovaly reálné podmínky a užívání chatbota, nebylo autorem zasahováno do jednotlivých testování. Každému účastníkovi byl daný chatbot představen společně s následným dotazníkem a poté byl testování nechán volný průběh. Jednotlivé konverzace lze následně analyzovat za pomoci nástroje LangSmith a optimalizovat tak chatbota.

Dotazník použitelnosti byl vytvořen v papírové formě k výslednému sběru dat a následně byla data přepsána do program Excel.

Výzkumný soubor

Výzkumný soubor pro testování chatbota byl vybrán ze studentů, kteří se přihlásili k testování tohoto chatbota. Výběr probíhal zejména dvěma způsoby, studenti se mohli přihlásit pomocí dotazníkového šetření anebo přímým kontaktem s autorem práce. Zejména díky první možnosti bylo umožněno získat širokou a rozmanitou skupinu respondentů.

Zpracování dat

Data, která jsou shromážděná za pomoci zmíněného formuláře jsou následně zpracována za pomoci programu Excel. Kalkulační nástroj, který je poskytován výzkumným týmem z Ulsterovi University byl využit k vypočítání finálního skóre. Vzorec, který tento systém využívá lze sledovat níže.

$$CUQ = \left(\left(\sum_{n=1}^m 2n - 1 \right) - 5 \right) + \left(25 - \left(\sum_{n=1}^m 2n \right) \right) * 1.6 \quad (1)$$

Kde: CUQ... Dotazník použitelnosti,

m... počet otázek v tomto případě se rovná 16,

n... skóre jednotlivých otázek na účastníka.

11.2 Vyhodnocení testování

Výsledný počet respondentů CUQ dotazníků, a tedy zejména počet studentů, kteří se zapojili do testování byl 22. Dotazník CUQ nemá stanovený minimální počet respondentů ani doporučenou hodnotu. Ve studiích, kde se používal tento dotazník použitelnosti, se počet účastníků většinou pohyboval mezi 22 a 30. Tedy výsledný počet respondentů by měl být dostačující.

Výsledné skóre CUQ dotazníku použitelnosti bylo 79,9 (Směrodatná odchylka = 8,8). Medián u dotazníku použitelnosti vyšel 78,9. Rozpětí skóre jednotlivých respondentů bylo od 65,6 až do 93,8. Na stupnici SUS by tento finální výsledek spadal do přijatelných systémů s označením dobrého systému. V **příloze H** lze nalézt škálu systému SUS, ze které byly vyvozeny tyto výsledky. V **příloze I** lze také nalézt průměrné výsledky a směrodatnou odchylku u jednotlivých otázek.

Tyto výsledné hodnoty lze porovnat s hodnotami z ostatních prací. Pro toto porovnání byly vybrány tři vědecké práce, které využily dotazník CUQ k vyhodnocení použitelnosti určitých systémů. Aloqayli a Abdelhafez (2023) se zabývali tvorbou a evaluací chatbota pro přijímací řízení na vysokou školu. Boian et al. (2024) se zabývali tvorbou a evaluací konverzačního agenta pro screening duševního zdraví. V poslední řadě Larbi et al. (2022) se zabývali návrhem, evaluací a následnou implementací chatbota na sociálních sítích pro zvýšení fyzické aktivity.

Tab. 9 Porovnání CUQ skóre

(zpracování, 2024)	Výsledky
Vytvořený chatbot	Výsledné CUQ skóre bylo 79,9 se směrodatnou odchylkou 8,8.
Aloqayli & Abdelhafez, (2023)	Výsledné CUQ skóre bylo 76,6 se směrodatnou odchylkou 11,5.
Boian et al. (2024)	Výsledné CUQ skóre bylo 75,82 se směrodatnou odchylkou 10,68.
Larbi et al. (2022)	Výsledné CUQ skóre bylo 59,5, směrodatná odchylka nebyla v této práci uvedena.

Zdroj: vlastní zpracování, 2024

Z tabulky lze konstatovat, že výsledek vytvořeného chatbota se nachází na podobné úrovni jako je tomu u dvou zmíněných prací, je ale nutné zmínit, že faktory jako různý počet účastníků testování či různý účel chatbota mohou mít určitý vliv na výsledky u tohoto dotazníku.

11.3 Shrnutí testování

Z výsledku testování lze konstatovat, že se vytvořený chatbot pohybuje na akceptovatelné části škály dle SUS. Dále má také ve srovnání s ostatními vybranými chatboty velmi dobré skóre použitelnosti CUQ.

Ačkoliv to nebylo součástí zkoumání této části, je potřeba zmínit, že v některých málo případech při testování došlo k použití nesprávné funkce agentem, který následně využil nesprávné dokumenty pro vytvoření odpovědi na otázku. Těmto problémům se lze jen obtížně vyhnout, jelikož klíčovým aspektem je způsob, jakým přebírá vstup pro proces volání funkcí agent od společnosti OpenAI. Aktualizace tohoto modulu by mohla snížit počet výskytů této chyby.

Na základě průměrného skóre z otázky 11 a 12, „Odpovědi chatbota byly užitečné, přiměřené a informativní“ a „Odpovědi chatbota byly irelevantní“ lze posoudit, jak studenti vnímají **přesnost a užitečnost poskytovaných informací**. Otázka 11 měla průměrné skóre 4,5 a smodch. 0,5, otázka 12 měla průměrné skóre 1,5 a smodch. 0,7. Lze tedy konstatovat, že studenti vnímali odpovědi jako velmi relevantní užitečné, jelikož u otázky 11 se průměrná hodnota již pohybuje na úrovni odpovědi "silně souhlasím". Opačně u otázky 12 se průměrná hodnota pohybuje na pomezí mezi „Silně nesouhlasím“ a „nesouhlasím“.

Výsledná míra použitelnosti dle dotazníku CUQ byla 79,9 se směrodatnou odchylkou 8,8. Tato hodnota je na přijatelné úrovni a lze označit systém za dobrý.

12 Implementace

Postup nasazení

Při samotné implementaci tohoto chatbota je potřeba převést celý systém z testovacího prostředí na určitý server, který bude nepřetržitě dostupný jednotlivým uživatelům. Platforma FlowiseAI umožňuje vcelku jednoduchý přesun vytvořeného chatbota na určitý server. Systém Flowise lze přesunout v podstatě na jakoukoliv cloudovou platformu, jak je například Render, Railway, Sealos a další.

Nejjednodušší a také nejčastěji se využívá Render při využívání platformy Flowise. Render umožňuje hostovat jakékoliv webové aplikace jako je například právě platforma Flowise, a tedy i výsledného chatbota.

V první řadě je potřeba zkopírovat repositář platformy Flowise z webové služby GitHub a propojit tento repositář s vybranou hostovací platformou (v tomto případě zmíněný Render). Následně je potřeba adekvátně nastavit parametry jako jsou region, jméno, prostředí. Dále je potřeba vybrat dostačující typ instance, které se liší poskytovanou výpočetní technikou a také tedy následnou cenou, v tomto případě jsou optimální verze „Starter“ nebo „Standard“. Standard verze nabízí následující výpočetní techniku daného serveru, 2GB RAM a 1 CPU. Poté je také důležité nastavit ochranné přístupy pro administrátory tohoto nástroje, a to za pomoci proměnných „FLOWISE_NAME“ a „FLOWISE_PASSWORD“, které představují přístupové jméno a heslo do systému správce. (Render, n.d.a)

Poté je již platforma Flowise nasazena na daném serveru a lze se k ní připojit odkudkoliv. Flowise nabízí možnost kopírovat a vložit jednotlivé prototypy, tato možnost významně ulehčuje celý proces přesunu vytvořeného chatbota v uzavřeném prostředí na daný server. Staženou kopii chatbota, který byl vytvořen v rámci této práce, lze tedy jednoduše vložit do prostředí daného serveru a následně pouze přidat API klíče a jednotlivé soubory. Po těchto úpravách by daný chatbot měl představovat plnohodnotnou kopii chatbota, který byl vytvořen v této práci. V posledním kroku je pouze potřeba tohoto chatbota vložit na webovou stránku fakulty, třeba za pomoci upraveného uživatelského prostředí, které přímo nabízí Flowise. V tuto chvíli by daný chatbot měl být připraven odpovídat na jednotlivé dotazy nepřetržitě.

Časová náročnost nasazení

Celý zmíněný proces nasazení je také potřeba vyčíslit jako počet hodin práce při samotné implementaci. V tabulce níže lze sledovat jednotlivé akce, které je potřeba provést pro úspěšnou implementaci chatbota na webovou stránku fakulty. Některé z těchto zmíněných akcí nemusí přesně trvat stanovený čas, záleží na kompetencích daného pracovníka, který provede proces implementaci a případných chybách, které by mohly při nastavení nastat.

Tab. 10 Časová náročnost nasazení

Akce	Doba trvání (h)
Přípravné procesy	3
Konfigurace serveru a spuštění	2
Převod vytvořeného chatbota a konfigurace	4
Vložení na webovou stránku	2
Kontrola a testování funkčnosti	2
Celkem	13

Zdroj: vlastní zpracování, 2024

Náklady

V následující části je prezentována prognóza nákladů, které jsou spojené s využíváním velkých jazykových modelů.

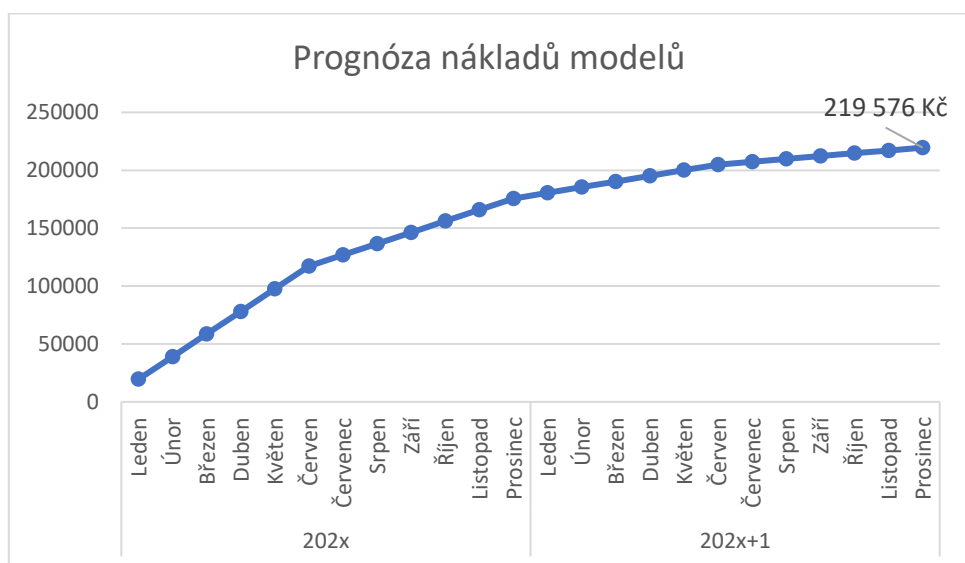
Graf níže představuje kumulativní náklady za využívání chatbota během dvouletého období, a zobrazuje určitý nárůst celkových výdajů na tento nástroj. U této prognózy bylo potřeba určit jednotlivé proměnné, jako jsou počet studentů, frekvence využívání, počet dotazů na danou konverzaci a očekávané snížení ceny využívaných modelů.

Počet studentů byl stanoven na 1600, frekvence využívání byla rozřazena dle procentuálního zastoupení z dotazníkového šetření, podle kategorií denní, týdenní, měsíční a čtvrtletní (převod z kategorie „jen příležitostně“, tak aby bylo možné vypočítat přesné hodnoty). Počet dotazů na jednu konverzaci byl stanoven na hodnotu 6 a to na základě analýzy výsledků testování chatbota studenty. Očekávané snížení ceny využívaných modelů je v podstatě předpokládané snížení ceny za využívání velkých jazykových modelů, kdy na základě historických informací docházelo ke snížení ceny jednotlivých modelů o průměrných 50% každý půlrok. Technologické společnosti, které

se zabývají vývojem velkých jazykových modelů, je neustále optimalizují, což jim umožňuje také snižovat ceny za využívané modely.

Nejprve bylo potřeba vypočítat počet použití za měsíc pro každou kategorii. Neboli tedy například u kategorie „denní“ vypadal přepočet následovně: Počet studentů (1600) * procentuální zastoupení z dotazníkového šetření (1,12%) * počet dní v měsíci (30). Následně se vypočítaly náklady pro jednotlivé měsíce s tím, že každý půlrok došlo ke zmíněnému snížení ceny. Vypočítané náklady bylo potřeba také převést na CZK z USD dle aktuálního kurzu, který byl 23,41 CZK/USD.

Obr. 18 Graf prognózy nákladů za využívané modely na dva roky



Zdroj: vlastní zpracování, 2024

Na vytvořeném grafu lze sledovat již zmíněný růst nákladů za dvouleté období. Na začátku neboli tedy prvních 6 měsíců jsou zmíněné náklady nejvyšší a postupné očekávané snížení ceny využívaných modelů snižuje i nárůst těchto nákladů v dalším období. V některých měsících lze očekávat zvýšené využívání chatbota, v jiných měsících zase menší (například letní prázdniny), u této prognózy je využívání rozděleno rovnoměrně v rámci jednotlivých měsíců.

Náklady spojené s Render

V případě využití platformy Render je také nutné počítat s měsíčním poplatkem 25 USD neboli tedy 585 Kč při kurzu 23,41 CZK/USD. Tento poplatek je spojený s provozem vytvořeného chatbota na daném serveru.

13 Kontrola

Po úspěšné implementaci daného chatbota je potřeba také provádět pravidelné kontroly. Určité kontroly funkčnosti systému lze provádět pomocí již zmíněného nástroje LangSmith, který umožňuje sledovat jednotlivé konverzace, a tedy hlavně vytvořené odpovědi chatbota. Tento nástroj také umožňuje sledovat přesné náklady za využívání velkého jazykového modelu.

V rámci procesu kontroly je potřeba naplánovat, a hlavně provádět určité aktualizace využívaného systému a dokumentů. Jelikož využívané dokumenty slouží jako znalostní databáze pro chatbota, je nezbytné ji udržovat stejně aktuální jako informace na webové stránce.

Aktualizace souborů

Při implementaci vytvořeného chatbota je potřeba zajistit, aby data, využívaná jako znalostní databáze, byla udržována v aktuálním a přesném stavu, což je nezbytné pro zajištění efektivní komunikace, a hlavně k prevenci vytvoření potencionálních nedorozumění. Různé soubory je potřeba aktualizovat v různých intervalech, vyhlášky týkající se přijímacího řízení je například potřeba aktualizovat minimálně na roční bázi. Stejný postup je potřeba udělat například u akademického harmonogramu. U jiných důležitých termínů může být potřeba aktualizace i několikrát za daný semestr.

Vytvořený systém je rozdělený na zmíněných několik částí, které zaobírají určité společné téma. V rámci dané části lze upravovat, měnit či přidávat vybrané soubory, je ale potřeba dodržet určitá pravidla. Jelikož většina těchto částí využívá nadpisů markdown jazyka k segmentaci textu, je potřeba při úpravě či přidání dokumentů dodržet tento styl úpravy dokumentů. Nejedná se o příliš zdlouhavý či složitý postup, není však možné pouze nahrát originální PDF soubor do tohoto systému.

Aktualizace Flowise

Za poslední roky intenzivního rozvoje v odvětví umělé inteligence dochází také k častým aktualizacím platformy Flowise. Jednotlivé komponenty, které jsou využívány i ve vytvořeném chatbotovi, s různými aktualizacemi dostávají určitá vylepšení či modifikace. Je na místě tento systém pravidelně aktualizovat, tak aby bylo udrženo

nejlepší kvality a zároveň sníženo procento halucinace. Render má funkci automatické aktualizace při změně zdrojového kódu ve zkopírovaném repositáři. Je tedy potřeba kontrolovat pouze zdali je zkopírovaný repositář synchronizovaný s repositářem Flowise. (Render, n.d.b)

Aktualizace velkých jazykových modelů

Podle prognózy nákladů na využívání velkých jazykových modelů dochází přibližně každých šest měsíců k významné optimalizaci těchto modelů, což vede ke snížení cen za jejich používání. Je nutné sledovat vývoj těchto modelů a v případě, že dojde k vydání nové verze používaného modelu, provést aktualizaci tohoto modelu také v systému Flowise.

Závěr

Hlavním cílem této práce bylo vyvinout chatbota, který bude sloužit jako podpora pro studenty ve studijní administrativě a akademickém poradenství na fakultě.

První část diplomové práce se věnovala teoretickému zakotvení základů souvisejících s tématem práce. Nejprve byl představen koncept umělé inteligence a strojového učení a disciplína marketingu. Následně byl charakterizován chatbot a jeho další rozdělení. Nakonec byly představeny etické problémy, které přináší umělá inteligence.

Prostřednictvím dotazníkového šetření byla získána data od studentů týkající se jejich preferencí a postojů k umělé inteligenci a fakultnímu chatbotovi. Z analýzy výsledků dotazníkového šetření vzešla informace, že studenti mají zájem o využívání nových nástrojů jako jsou právě AI chatboti. Z dotazníkového šetření také vyplynulo, že by respondenti využívali chatbota spíše od týdenní báze a méně, a to zejména za účelem dohledávání důležitých termínů. Důvěra ve výsledné odpovědi byla mezi respondenty rozložena na dvě skupiny, jedna skupina plně důvěřuje v odpovědi, druhá větší skupina potřebuje určité ověření z jiného zdroje.

Všechny zmíněné poznatky byly přeneseny do procesu vývoje chatbota, tak aby tento chatbot odpovídal požadavkům studentů, tak i Fakulty ekonomické. Požadavek na chatbota ze strany fakulty byl využívat zdroje, které jsou pouze dostupné na veřejných stránkách fakulty, tak aby nedošlo k úniku citlivých dat.

Vyprototypováno bylo 5 nejlepších systémů, které byly následně porovnány na nejběžnějších metrikách. Při testování byly změřeny finanční náklady spojené s provozem a počet halucinací. Na základě těchto metrik byl vybrán nejvhodnější prototyp. Dále bylo také provedené testování se studenty, tak aby bylo možné určit použitelnost chatbota, a to za pomoci odpovědí koncových uživatelů tohoto chatbota. V práci byly popsány jednotlivé specifikace a komponenty vybraného finálního chatbota.

V poslední řadě byl představen plán implementace a určité kontroly, kterou je potřeba provádět při případné implementaci. V rámci implementace byla představená určitá prognóza nákladů.

V závěru lze konstatovat, že implementace AI chatbota na ekonomické fakultě je nejen technologicky proveditelná, ale také výrazně přispívá ke zlepšení studijního prostředí tím, že poskytuje studentům snadný přístup k důležitým informacím a podporuje tak celý proces vyhledávání administrativních informací. Vzhledem k zájmu studentů a pozitivním výsledkům testování lze doporučit přistoupení k implementaci chatbota, přičemž je důležité věnovat pozornost kontinuálnímu zlepšování a aktualizaci systému na základě zpětné vazby od uživatelů a změn v akademickém prostředí. Je nutné ale počítat se zmíněnými náklady, které mohou být v prvních měsících ještě vyšší, než bylo predikováno z toho důvodu, že se jedná o nový nástroj. I přes pozitivní výsledky z testování je potřeba počítat s určitým nižším procentem halucinace využitého velkého jazykového modelu. Neschopnost předem určit pevné náklady a nestabilita nákladů jsou klíčové překážky pro sestavení rozpočtu. Problematika rozpočtu a nižší procento halucinace mohou představovat hlavní důvody, proč by nebyl tento systém implementován.

Seznam použitých zkratek

AGI	Artificial general intelligence (Obecná umělá inteligence)
AI	Artificial intelligence (Umělá inteligence)
ANI	Artificial narrow intelligence (Úzká umělá inteligence)
API	Aplikační programové rozhraní
ASI	Artificial super intelligence (Umělá superinteligence)
CPU	Centrální procesorová jednotka
CUQ	The Chatbot Usability Questionnaire (Dotazník použitelnosti)
EU	Evropská unie
FEK	Fakulta ekonomická
GDPR	Obecné nařízení o ochraně osobních údajů
GPU	Grafický procesor
GPT	Generative Pretrained Transformer (Generativní předtrénovaný transformátor)
LLM	Large language model (Velký jazykový model)
LSTM	Dlouhodobá krátkodobá paměť
NLG	Natural language generation (Generování přirozeného jazyka)
NLP	Natural Language Processing (Zpracování přirozeného jazyka)
NLU	Natural language understanding (Porozumění přirozeného jazyka)
QA	Question Answering (Odpovídání na otázky)
RAG	Retrieval-augmented Generation (Generování s využitím vyhledávání)
RNN	Rekurentní neuronové síť
SUS	System Usability Scale (Stupnice použitelnosti systému)
ZČU	Západočeská univerzita v Plzni
FAQ	Často kladené otázky
Smoch.	Směrodatná odchylka

Seznam použitých zdrojů

- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, 100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- Adomavicius, G., Bauman, K., Tuzhilin, A., & Unger, M. (2022). Context-aware recommender systems: From foundations to recent developments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (s. 211-250). Springer. https://doi.org/10.1007/978-1-0716-2197-4_6
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Alteschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., ... Zoph, B. (2024). *GPT-4 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2303.08774>
- Achtaich, K., Fagroud, F. Z., & Achtaich, N. (2021). AI-Powered Health Chatbots: Toward a general architecture. *Procedia Computer Science*, 191, 355–360. <https://doi.org/10.1016/j.procs.2021.07.048>
- Airey, D. (2014). *Logo design love: A guide to creating iconic brand identities*. Peachpit Press.
- Albuger, J. (2023). *Rule-Based Chatbots vs. AI Chatbots: Key Differences*. Dostupné (15.2.2024) z: <https://www.hubtype.com/blog/rule-based-chatbots-vs-ai-chatbots>
- Alfayad, F. (2020). The European Union's GDPR and its effect on Data-Driven Marketing Strategies. *International Journal of Marketing Studies*, 12(1), 39. <https://doi.org/10.5539/ijms.v12n1p39>
- Allal-Chérif, O., Puertas, R., & Carracedo, P. (2024). Intelligent influencer marketing: How AI-powered virtual influencers outperform human influencers. *Technological Forecasting and Social Change*, 200, 123113. <https://doi.org/10.1016/j.techfore.2023.123113>
- Aloqayli, A., & Abdelhafez, H. (2023). Intelligent chatbot for admission in higher education. *International Journal of Information and Education Technology*, 13(9), 1348–1357. <https://doi.org/10.18178/ijiet.2023.13.9.1937>

- Anyoha, R. (2017). *The History of Artificial Intelligence*. Dostupné 9.11.2023 z: [The History of Artificial Intelligence - Science in the News \(harvard.edu\)](https://www.harvard.edu/history-of-artificial-intelligence-science-in-the-news)
- Apache (n.d.). *APACHE LICENSE, VERSION 2.0*. Dostupné (15.2.2024) z: <https://www.apache.org/licenses/LICENSE-2.0>
- AWS (n.d.). *What are Large Language Models (LLM)?* Dostupné (3.1.2024) z: <https://aws.amazon.com/what-is/large-language-model/>
- Balaguer, A., Benara, V., de Freitas Cunha, R. L., Estevão Filho, R. M., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv <https://doi.org/10.48550/arXiv.2401.08406>
- Bastian, M. (2023). *GPT-4 has more than a trillion parameters – Report*. Dostupné (3.1.2024) z: <https://the-decoder.com/gpt-4-has-a-trillion-parameters/>
- Bell, E. (2023) *Artificial General Intelligence (AGI): Definition, How It Works, and Examples*. Dostupné 18.10.2023 z: <https://www.investopedia.com/artificial-general-intelligence-7563858>
- Bhatt, S. (2018). *Reinforcement Learning 101*. Dostupné 14.11.2023 z: <https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292>
- Blumenfeld, Z. & Htet, E. (2023). *What Is Retrieval-Augmented Generation (RAG)?* Dostupné (16.2.2024) z: <https://neo4j.com/blog/what-is-rag/>
- Boian, R., Bucur, A., Todea, D., Luca, A. R., Rebedea, T., & Podină, I. R. (2024). A conversational agent framework for mental health screening: design, implementation, and usability. *Behaviour & Information Technology*, 1–15. <https://doi.org/10.1080/0144929x.2024.2332934>
- Buchanan, B. (2005). Brief History of Artificial Intelligence. *AI Magazine*, 26(4), 53. <https://doi.org/10.1609/aimag.v26i4.1848>
- Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, 13(1), 41. <https://doi.org/10.3390/info13010041>

- Cohan, P. (2023). *As ChatGPT And Other AI Tools Increase Energy Demand, Here's What Investors Need To Know*. Dostupné (28.1.2024) z: <https://www.forbes.com/sites/petercohan/2023/11/09/equinix-and-vertiv-stock-prices-could-rise-on-generative-ais-energy-use/>
- Cohen, R. (2023a). *Chunking Strategies for LLM Applications*. Dostupné (25.2.2024) z: <https://www.pinecone.io/learn/chunking-strategies/>
- Cohen, R. (2023b). *Vector Embeddings for Developers: The Basics*. Dostupné (25.2.2024) z: <https://www.pinecone.io/learn/vector-embeddings-for-developers/>
- Cohen, R. (2023c). *What is a Vector Database & How Does it Work? Use Cases + Examples*. Dostupné (26.2.2024) z: <https://www.pinecone.io/learn/vector-database/#What-is-a-Vector-Database>
- Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., & Yang, M. (2024). *LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens*. *arXiv:2402.13753 [cs.CL]*. <https://doi.org/10.48550/arXiv.2402.13753>
- EDC Paris Business School. (2022). *Marketing 4.0 : Definition and features*. Dostupné 22.11.2023 z: <https://www.edcparis.edu/en/blog/marketing-4-0-definition-and-features>
- Eger, L., & Egerová, D. (2014). *Základy metodologie výzkumu*. Západočeská univerzita v Plzni.
- Eger, L., & Egerová, D. (2017). *Základy metodologie výzkumu*. Západočeská univerzita v Plzni.
- Errante, D. (2023). *AI Keyword Research: The Essential Tools You Need*. Dostupné (15.1.2024) z: <https://typli.ai/blog/ai-keyword-research-tools>
- Es, S., James, J., Anke, L., & Schockaert S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Aletras, N., & Clerq, O. (Eds.) *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. (s. 150-158). Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-demo.16>

Ethan, P. (2023). *RAG Evaluation Series: Validating OpenAI Assistant's RAG Performance*. Dostupné z: <https://www.tonic.ai/blog/rag-evaluation-series-validating-openai-assistants-rag-performance>

European Commission (2023). *Commission welcomes political agreement on Artificial Intelligence Act**. Dostupné (28.1.2024) z: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473

European Parliament (2023a). *EU AI Act: first regulation on artificial intelligence*. Dostupné (19.1.2024) z: <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

European Parliament (2023b). *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI*. Dostupné (24.1.2024) z: <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai>

Evropská rada (2023). *Akt o umělé inteligenci: Rada a Parlament dosáhly dohody o prvních pravidlech pro umělou inteligenci na světě*. Dostupné (28.1.2024) z: <https://www.consilium.europa.eu/cs/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>

Evropský parlament (2020). *Co je umělá inteligence a jak ji využíváme?* Dostupné 17.10.2023 z: <https://www.europarl.europa.eu/news/cs/headlines/society/20200827STO85804/umela-intelligence-definice-a-vyuziti>

Fakulta ekonomická Západočeské univerzity v Plzni (2023). *Výroční zpráva*

Fakulta ekonomická Západočeské univerzity v Plzni (n.d.). *Logo* [Logo]. <https://macbeth.fek.zcu.cz/>

FlowiseAI (n.d.). *Build LLM Apps Easily*. Dostupné (15.2.2024) z: <https://flowiseai.com/>

Frankel, A. (2023). *Types of SEO: What Does Your Site Need to Succeed?* Dostupné (7.1.2024) z: <https://www.semrush.com/blog/types-of-seo/>

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- Geysler, W. (2023). *The State of Influencer Marketing 2023: Benchmark Report*. Dostupné (5.1.2024) z: <https://influencermarketinghub.com/influencer-marketing-benchmark-report/#toc-2>
- Goodwin, D. (n.d.). *What Is SEO – Search Engine Optimization?* Dostupné (7.1.2024) z: <https://searchengineland.com/guide/what-is-seo>
- Google for Developers (2022) *What creators should know about Google's August 2022 helpful content update*. Dostupné (7.1.2024) z: <https://developers.google.com/search/blog/2022/08/helpful-content-update>
- Grynbaum, M., & Mac, R. (2023). *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. Dostupné (28.1.2024) z: <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>
- Hainsdorf, C., Hickman, T., Lorenz, S., & Rennie, J. (2023). *Dawn of the EU's AI Act: political agreement reached on world's first comprehensive horizontal AI regulation*. Dostupné (19.1.2024) z: <https://www.whitecase.com/insight-alert/dawn-eus-ai-act-political-agreement-reached-worlds-first-comprehensive-horizontal-ai>
- Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3, 119–132. <https://doi.org/10.1016/j.ijin.2022.08.005>
- Hartmann, J., & Netzer, O. (2023). Natural Language Processing in Marketing. In K. Sudhir, & O. Toubia, (Eds.) *Artificial Intelligence in Marketing (Review of Marketing Research, Vol. 20)*, (s. 191-215). Emerald Publishing Limited, Leeds, <https://doi.org/10.1108/S1548-643520230000020011>
- Holmes, W., Moorhead, A., Bond, R. R., Zheng, H., Coates, V., & McTear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In M. Mulvenna, & R. Bond (Eds.), *Proceedings of the*

- 31st European Conference on Cognitive Ergonomics: "Design for Cognition" (s. 207-214). Association for Computing Machinery. <https://doi.org/10.1145/3335082.3335094>
- Chinn, A. (2022). *What's the System Usability Scale (SUS) & How Can You Use It?* Dostupné z: <https://blog.hubspot.com/service/system-usability-scale-sus>
- Chowdhary, K. (2020). *Fundamentals of artificial intelligence*. Springer.
- Joshi, Y., Lim, W., Jagani, K., & Kumar S. (2023). Social media influencer marketing: foundations, trends, and ways forward. *Electron Commer Res* <https://doi.org/10.1007/s10660-023-09719-z>
- Jung, A. (2022). *Machine learning: The Basics*. Springer Nature.
- Junia (n.d.). *Keyword Research Done by AI*. Dostupné (15.1.2024) z: <https://www.junia.ai/tools/ai-keyword-research>
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kavlakoglu E., (2020). *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts*. Dostupné (19.12.2023) z: <https://www.ibm.com/blog/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/>
- Ke, Y., Jin, L., Elangovan, K., Abdullah, H., Liu, N., Sia, A., Soh, C., Tung, J., Lingong, J., Ting, D. (2024). *Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report*. arXiv. <https://doi.org/10.48550/arXiv.2402.01733>
- Keller, K. (2007). *Strategické řízení značky*. Grada
- Koren, Y., Rendle, S., & Bell, R. (2022). Advances in collaborative filtering. In Ricci, F., Rokach, L., & Shapira, B. (Eds.), *Recommender Systems Handbook* (s. 91-142). Springer. https://doi.org/10.1007/978-1-0716-2197-4_3

- Kotler, P., Kartajaya, H., & Setiawan, I. (2016). *Marketing 4.0: Moving from Traditional to Digital*. John Wiley & Sons.
- Kotler, P., Kartajaya, H., & Setiawan, I. (2021). *Marketing 5.0: Technology for humanity*. Wiley
- Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the Role of Artificial Intelligence in Personalized Engagement Marketing. *California Management Review*, 61(4), 135-155. <https://doi.org/10.1177/0008125619859317>
- Lai, C. (2023). *AI is harming our planet: addressing AI's staggering energy cost (2023 update)*. Dostupné (28.1.2024) z: <https://www.numenta.com/blog/2023/08/10/ai-is-harming-our-planet-2023/>
- LangChain (2023). *Announcing LangSmith, a unified platform for debugging, testing, evaluating, and monitoring your LLM applications*. Dostupné z: <https://blog.langchain.dev/announcing-langsmith/>
- LangChain (n.d.a). *Text Splitters*. Dostupné (25.2.2024) z: https://python.langchain.com/docs/modules/data_connection/document_transformers/
- LangChain (n.d.b). *OpenAI tools*. Dostupné (26.2.2024) z: https://python.langchain.com/docs/modules/agents/agent_types/openai_tools
- LangChain (n.d.c). *Conversational Retrieval QA*. Dostupné z: https://js.langchain.com/docs/modules/chains/popular/chat_vector_db
- LangChain (n.d.d). *Dynamically selecting from multiple retrievers*. Dostupné z: https://js.langchain.com/docs/modules/chains/additional/multi_retrieval_qa_router
- Larbi, D., Denecke, K., & Gabarrón, E. (2022). Usability testing of a social media chatbot for increasing physical activity behavior. *Journal of Personalized Medicine*, 12(5), 828. <https://doi.org/10.3390/jpm12050828>
- Lee, K. (n.d.). *Understanding LLM Hallucinations and how to mitigate them*. Dostupné z: <https://kili-technology.com/large-language-models-llms/understanding-llm-hallucinations-and-how-to-mitigate-them#1>
- Leist, R. (2022). *How to Do Keyword Research for SEO: A Beginner's Guide*. Dostupné (15.1.2024) z: <https://blog.hubspot.com/marketing/how-to-do-keyword-research-ht>

- Leung, F., Gu, F., & Palmatier, R. (2022). Online influencer marketing. *J. of the Acad. Mark. Sci.* 50, 226–251. <https://doi.org/10.1007/s11747-021-00829-4>
- Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models.* *arXiv*. <https://doi.org/10.48550/arXiv.2304.03271>
- Luccioni, A., Viguiet, S., & Ligozat, A. (2022). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.* *arXiv*. <https://doi.org/10.48550/arXiv.2211.02001>
- Lutkevich, B. (2023). *expert systém.* Dostupné 9.11.2023 z: <https://www.techtarget.com/searchenterpriseai/definition/expert-system>
- Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., & Chen, E. (2024). CRUD-RAG: A comprehensive Chinese benchmark for Retrieval-Augmented Generation of Large Language Models. *arXiv* <https://doi.org/10.48550/arxiv.2401.17043>
- Mahesh, B. (2020). Machine Learning Algorithms – A Review. *International Journal of Science and Research*, 9(1), 381-386, <https://www.ijsr.net/getabstract.php?paperid=ART20203995>
- Marin (2024). *Harnessing the Power of AI for SEO with Automated Keyword Research.* Dostupné (15.1.2024) z: <https://www.marinsoftware.com/blog/harnessing-the-power-of-ai-for-seo-with-automated-keyword-research>
- McKinsey & Company (2023). What is personalization? Dostupné (12.12.2023) z: <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-personalization>
- Mehta, S. (2022) The Evolution of Marketing 1.0 to Marketing 5.0. *International Journal of Law Management & Humanities*, 5(4), 469-485. <https://doi.org/10.10000/IJLMH.113373>
- Meredith, S. (2023). *A 'thirsty' generative AI boom poses a growing problem for Big Tech.* Dostupné (29.1.2024) z: <https://www.cnbc.com/2023/12/06/water-why-a-thirsty-generative-ai-boom-poses-a-problem-for-big-tech.html>

- Merritt, R. (2022). *What Is a Transformer Model?* Dostupné (5.1.2024) z: <https://blogs.nvidia.com/blog/what-is-a-transformer-model/>
- Merritt, R. (2023). *What Is Retrieval-Augmented Generation, aka RAG?* Dostupné (7.1.2024) z: <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>
- Microsoft (2023). *What is a vector database?* Dostupné (26.2.2024) z: <https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db>
- Molenaar K. (2023). *Discover The Top 12 Virtual Influencers for 2024 – Listed and Ranked!* Dostupné (6.1.2024) z: <https://influencermarketinghub.com/virtual-influencers/#toc-3>
- Monteiro, V. (2023) *AI KEYWORD RESEARCH: HOW TO USE IT TO DOMINATE THE COMPETITION.* Dostupné (15.1.2024): <https://vettted.com/blog/ai-powered-keyword-research-future-seo-strategy/>
- Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* “O’Reilly Media, Inc.
- McCarthy, J. (2007). *WHAT IS ARTIFICIAL INTELLIGENCE?* Dostupné 17.10.2023 z: <https://www-formal.stanford.edu/jmc/whatisai/node1.html>
- Burkov, A. (2019) *The hundred-page machine learning book.*
- Nancholas, B. (2023). *What are the different types of artificial intelligence?* Dostupné 18.10.2023 z: <https://online.wlv.ac.uk/what-are-the-different-types-of-artificial-intelligence/>
- Nguyen, M. (2023). *Virtual influencers: meet the AI-generated figures posing as your new online friends – as they try to sell you stuff.* Dostupné (6.1.2024) z: <https://theconversation.com/virtual-influencers-meet-the-ai-generated-figures-posing-as-your-new-online-friends-as-they-try-to-sell-you-stuff-212001>
- Nguyen, T., Le, A. D., Hoang, H., & Nguyen, T. T. (2021). NEU-chatbot: Chatbot for admission of National Economics University. *Computers & Education: Artificial Intelligence*, 2, 100036. <https://doi.org/10.1016/j.caeai.2021.100036>
- Nvidia (n.d.a). *Recommendation System.* Dostupné (17.12.2023) z: <https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/>

Nvidia (n.d.b). *Large Language Models Explained*. Dostupné (3.1.2024) z: <https://www.nvidia.com/en-us/glossary/large-language-models/>

o činnosti FEK ZČU 2022. <https://drive.google.com/file/d/1RybiuchWJsEixTGltSnnL-uh0ZFhjcD6/view>

O'Brien, M., & Fingerhut H. (2023). *A.I. tools fueled a 34% spike in Microsoft's water consumption, and one city with its data centers is concerned about the effect on residential supply* Dostupné (29.1.2024) z: <https://fortune.com/2023/09/09/ai-chatgpt-usage-fuels-spike-in-microsoft-water-consumption/>

OpenAI (2022). *Introducing ChatGPT*, Dostupné (15.2.2024) z: <https://openai.com/blog/chatgpt>

OpenAI (n.d.a). *Tokenizer*. Dostupné (15.1.2024) z: <https://platform.openai.com/tokenizer>

OpenAI (n.d.b). *Embedding*. Dostupné (25.2.2024) z: <https://platform.openai.com/docs/guides/embeddings>

OpenAI (n.d.c). *Assistants API*. Dostupné z: <https://platform.openai.com/docs/assistants/overview>

Pradeep, A. K., Appel, A., & Sthanunathan, S. (2018). *AI for Marketing and Product Innovation: Powerful New Tools for Predicting Trends, Connecting with Customers, and Closing Sales*. John Wiley & Sons.

Ragas (n.d.a). *Faithfulness*. Dostupné z: <https://docs.ragas.io/en/latest/concepts/metrics/faithfulness.html>

Ragas (n.d.b). *Answer Relevance*. Dostupné z: https://docs.ragas.io/en/latest/concepts/metrics/answer_relevance.html

Ragas (n.d.c). *Context Recall*. Dostupné z: https://docs.ragas.io/en/latest/concepts/metrics/context_recall.html

Ragas (n.d.d). *Context Relevancy*. Dostupné z: https://docs.ragas.io/en/latest/concepts/metrics/context_relevancy.html

Ranoliya, B., Raghuwanshi, N., & Singh, S. (2017). *Chatbot for university related FAQs*. *2017 International Conference on Advances in Computing, Communications and*

- Informatics (ICACCI)*, (s. 1525-1530). IEEE
<https://doi.org/10.1109/ICACCI.2017.8126057>
- Render (n.d.a). *Web Services*. Dostupné (5.4.2024) z: <https://docs.render.com/web-services>
- Render (n.d.b). *How Deploys Work*. Dostupné (5.4.2024) z: <https://docs.render.com/deploys#skipping-an-auto-deploy>
- Roth, E. (2023). *The New York Times is suing OpenAI and Microsoft for copyright infringement / A lawsuit claims OpenAI copied millions of Times' articles to train the language models that power ChatGPT and Microsoft Copilot*. Dostupné (28.1.2024) z: <https://www.theverge.com/2023/12/27/24016212/new-york-times-openai-microsoft-lawsuit-copyright-infringement>
- Rouse, M. (2023). *Web Scraping*. Dostupné z: <https://www.techopedia.com/definition/5212/web-scraping>
- Russell, S., & Norvig, P. (2014). *Artificial intelligence: A Modern Approach, Global Edition*. Pearson
- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2023). ARES: an Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv*. <https://doi.org/10.48550/arxiv.2311.09476>
- Sahota, H. (2023). *Chaining the Future: An In-depth Dive into LangChain*. Dostupné z: <https://www.comet.com/site/blog/chaining-the-future-an-in-depth-dive-into-langchain/>
- Sands, S., Campbell, C., Plangger, K., & Ferraro, C. (2022). Unreal influence: leveraging AI in influencer marketing. *European Journal of Marketing*, 56(6), 1721–1747. <https://doi.org/10.1108/ejm-12-2019-0949>
- Sartor, G. et. al. (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*.
- Shankar, V., & Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *Journal of the Academy of Marketing Science*, 50(6), 1324–1350. <https://doi.org/10.1007/s11747-022-00840-3>

- Silva, C. (2022). *What Is Off-Page SEO? A Guide to Off-Page SEO Strategy*. Dostupné (15.1.2024) z: <https://www.semrush.com/blog/off-page-seo/>
- Sreedhar, N. (2023). *AI and its carbon footprint: How much water does ChatGPT consume?* Dostupné (29.1.2024) z: <https://lifestyle.livemint.com/news/big-story/ai-carbon-footprint-openai-chatgpt-water-google-microsoft-111697802189371.html>
- Stelmakh, I., Luan, Y., Dhingra, B., & Chang, M. (2022). ASQA: Factoid Questions Meet Long-Form Answers. In Goldberg, Y., Kozareva, Z., & Zhang, Y. (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. (s. 8273–8288). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.566>
- Světlík, J. (2018). *Marketing – cesta k trhu*. VŠPP. Dostupné z: <http://www.marketingsvetlik.cz/knihy/Marketing%20-%20cesta%20k%20trhu%20-%20Jaroslav%20Svetlik.pdf>
- Tableau. (n.d.). *What is the history of artificial intelligence (AI)?* Dostupné 10.11.2023 z: <https://www.tableau.com/data-insights/ai/history>
- Tucci, L. (n.d.). *What is machine learning and how does it work? In-depth guide*. Dostupné 13.11.2023 z: <https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML>
- Upwork (2023). *How To Use AI Tools for Keyword Research*. Dostupné (15.1.2024) z: <https://www.upwork.com/resources/ai-tools-for-keyword-research>
- Vectara (n.d.). *What is Vectara?* Dostupné z: <https://vectara.com/platform/>
- Wang, Z. (2022). *Modern Question Answering Datasets and Benchmarks: A Survey*. arXiv. <https://doi.org/10.48550/arXiv.2206.15030>
- Weatherbed, J. (2023). *Why the AI Act was so hard to pass*. Dostupné (19.1.2024) z: <https://www.theverge.com/2023/12/13/23999849/eu-ai-act-artificial-intelligence-regulations-complicated-delays>
- Winger R., & Edelman, D. (1989). *Segment-of-One Marketing*. Dostupné (17.12.2023) z: <https://www.bcg.com/publications/1989/strategy-segment-of-one-marketing>

Wolford, B. (n.d.a). *What are the GDPR Fines?* Dostupné (16.1.2024) z: <https://gdpr.eu/fines/>

Wolford, B. (n.d.b). *What is GDPR, the EU's new data protection law?* Dostupné (16.1.2024) z: <https://gdpr.eu/what-is-gdpr/>

Yu, F. (2023). *Text Embedding — What, Why and How?* Dostupné (25.2.2024) z: <https://medium.com/@yu-joshua/text-embedding-what-why-and-how-13227e983ba7>

Seznam tabulek

Tab. 1 Využívané metriky při evaluaci RAG.....	33
Tab. 2 Časový harmonogram vývoje	44
Tab. 3 Předpokládané rozdělení	45
Tab. 4 Porovnání očekávaných hodnot se zjištěnými	52
Tab. 5 Testování za pomoci LangSmith.....	74
Tab. 6 Vlastní hodnocení odpovědí.....	76
Tab. 7 Testování s Ragas.....	77
Tab. 8 Souhrnná tabulka testování	79
Tab. 9 Porovnání CUQ skóre	83
Tab. 10 Časová náročnost nasazení.....	86

Seznam obrázků

Obr. 1 Logo Fakulty ekonomické	42
Obr. 2 Věk respondentů	47
Obr. 3 Stupeň studia.....	48
Obr. 4 Zastoupení u bakalářských studijních programů	49
Obr. 5 Zastoupení u navazujících magisterských studijních programů.....	49
Obr. 6 Využití jazykových modelů.....	51
Obr. 7 Specifické velké jazykové modely a jejich využívání.....	52
Obr. 8 Postoj respondentů k využití AI chatbota na webových stránkách FEK.....	53
Obr. 9 Frekvence využívání chatbota	54
Obr. 10 Graf preferovaného využití chatbota na webových stránkách FEK	55
Obr. 11 Preferované metody získávání informací o vybraných tématech	56
Obr. 12 Preference tónu chatbota	57
Obr. 13 Míra důvěry v chatbota.....	58
Obr. 14 Jednoduchá architektura RAG.....	62
Obr. 15 Uživatelské rozhraní chatbota	70
Obr. 16 Logo chatbota	70
Obr. 17 Porovnání vytvořeného chatbota s velkým jazykovým modelem.....	72
Obr. 18 Graf prognózy nákladů za využívané modely na dva roky	87

Seznam příloh

Příloha A Uživatelské rozhraní na stránce FEK

Příloha B Jedna vektorová databáze společně s využitým agentem v platformě Flowise

Příloha C Celý systém v platformě Flowise

Příloha D CUQ otázky

Příloha E Web scraper pro akademické pracovníky

Příloha F Uživatelské rozhraní pro scraper pro akademické pracovníky

Příloha G Další příklady využití umělé inteligence v marketingu

Příloha H System usability Scale skóre

Příloha I Průměrné skóre pro jednotlivé otázky a směrodatná odchylka

Příloha J CUQ skóre u jednotlivých účastníků

Příloha K Příklad konverzace z testování

Příloha L Dotazník postojů k AI

Příloha A Uživatelské rozhraní na stránce FEK

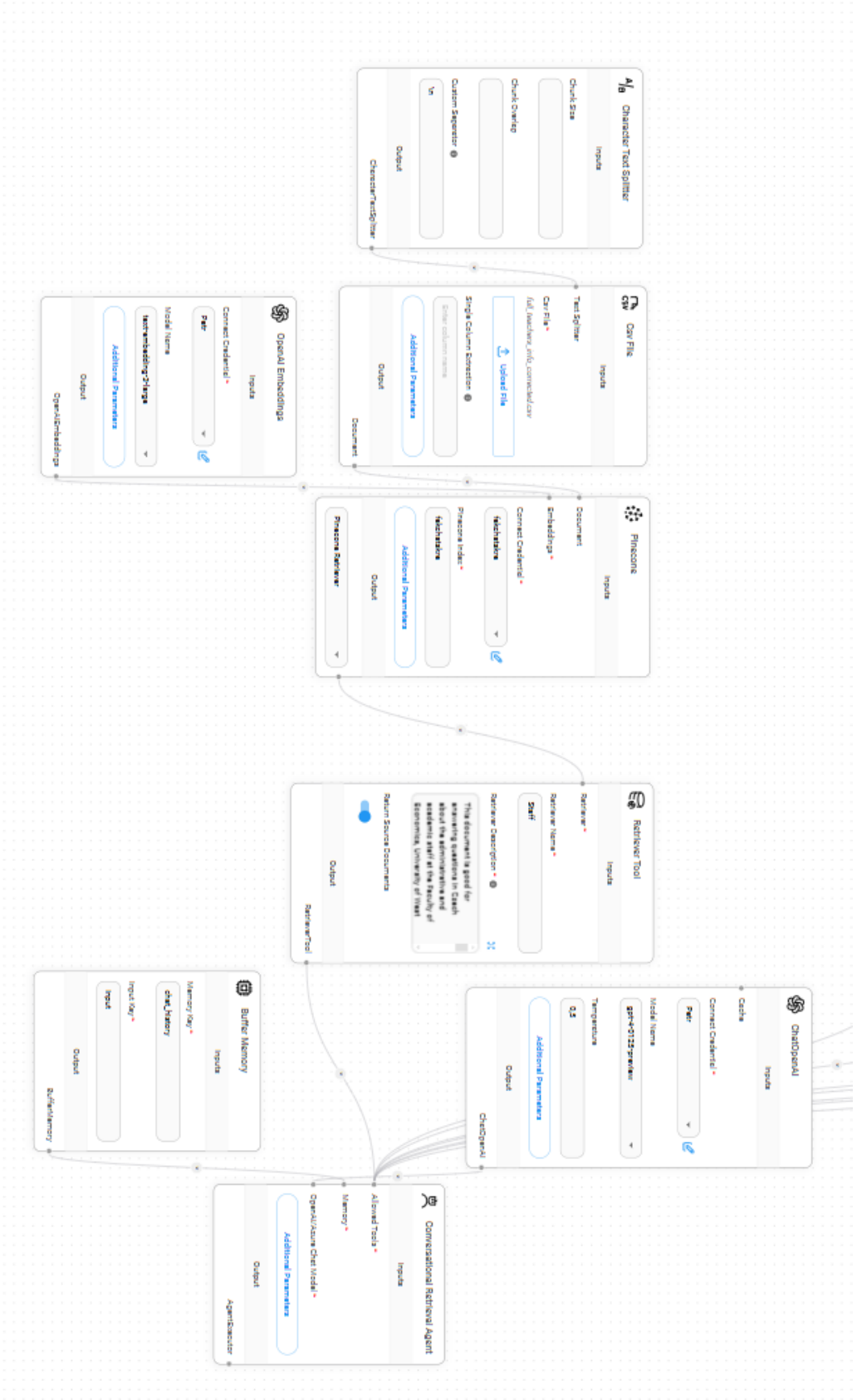
The screenshot displays a website interface for the Faculty of Economics and Business Administration (FEK) at Palacký University in Olomouc. The background image shows a group of students sitting on a lawn. The navigation menu at the top includes: **FAKULTA EKONOMICKÁ Z PRÁVEČESKÉ UNIVERZITY V OLOMOUCI**, **STUDIUM**, **FAKULTA**, **VÝZKUM**, **SPOLUPRÁCE**, **Studenti**, **Zaměstnanci**, **Absolventi**, and a search icon.

The main content area features a large orange banner with the text: **NÁŠÍ VIZÍ JE VÁŠ ÚSPĚCH!** Below this, a white box contains the text: **Nášim studentům nabízíme nejen moderní výuku v přátelském prostředí, ale také možnost absolovovat stáže ve firmách a na univerzitách po celém světě.** To the right of this text is a button labeled **NABÍDKA STUDIA**.

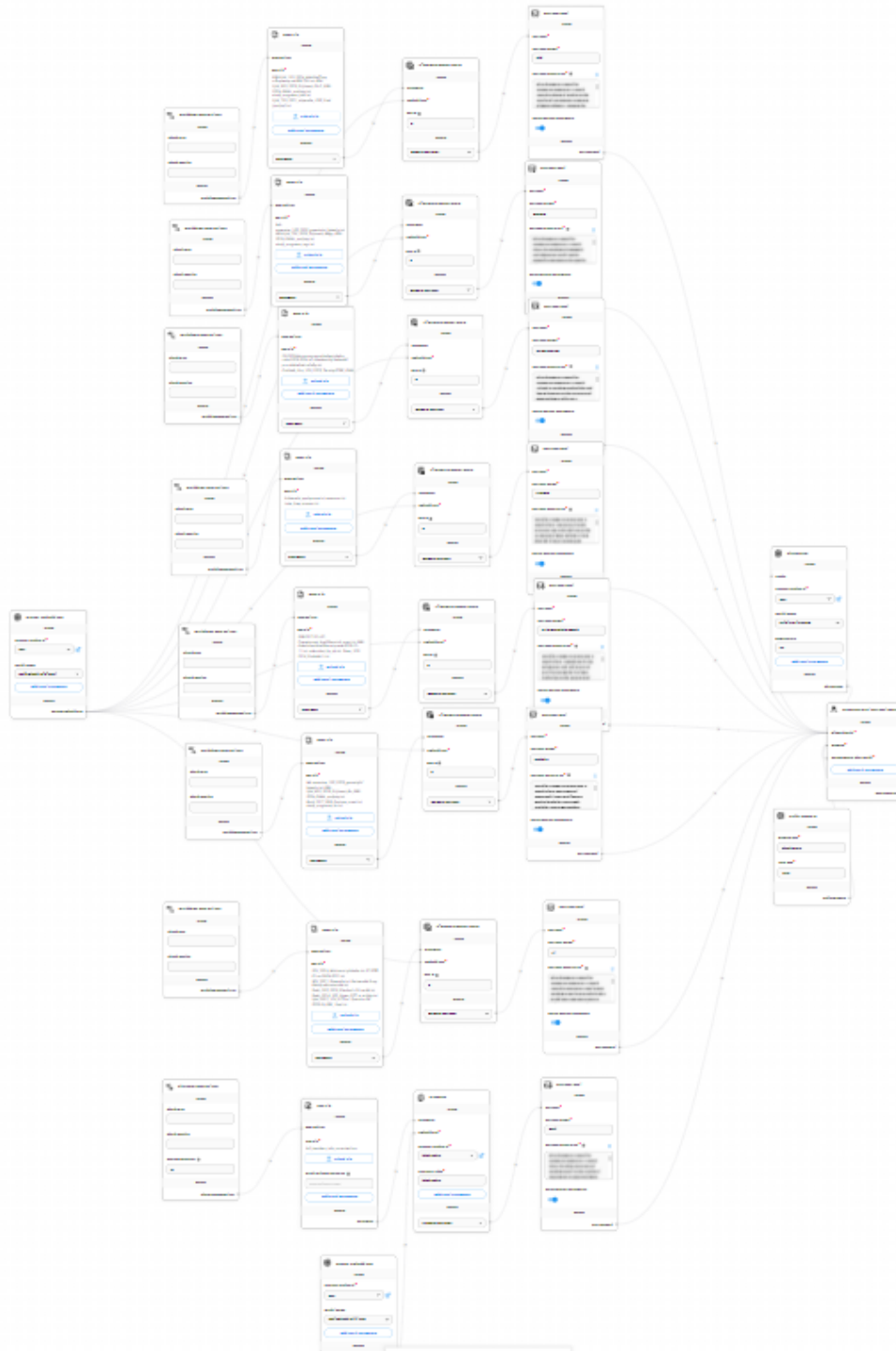
Below the banner is a chatbot interface. It includes a search icon, a question input field with the placeholder text "Type your question", and a "Powered by Flowise" logo. The chatbot's response is displayed in a white box: **Příhláška na navazující magisterské studium stojí 600,- Kč. Zaplatit ji můžete platební kartou přes platební bránu systému eprhlaska.zcu.cz, formou bezhotovostního převodu z účtu nebo v hotovosti na přepážce Komerční banky. Poštovní poukážku nelze pro platbu použít. Uďále pro zaplacení poplatku obdržíte při**

At the bottom of the page, there are two buttons: **DEN OTEVŘENÝCH DVEŘÍ** and **STUDIUM V CHEBU**.

Příloha B Jedna vektorová databáze společně s využitým agentem v platformě Flowise



Příloha C Celý systém v platformě Flowise



Příloha D CUQ otázky

1.	Osobnost chatbota byla realistická a poutavá.
2.	Přišel(a) mi chatbot příliš robotický.
3.	Chatbot byl přívětivý během počátečního nastavení.
4.	Chatbot působil velmi nepřátelsky.
5.	Chatbot dobře vysvětlil svůj účel a rozsah možností.
6.	Chatbot nedal žádnou indikaci o svém účelu.
7.	Bylo snadné se v chatbotu orientovat.
8.	Používání chatbotu by mohlo být matoucí.
9.	Chatbot mě dobře pochopil.
10.	Chatbot nerozpoznal mnoho mých vstupů.
11.	Odpovědi chatbota byly užitečné, přiměřené a informativní.
12.	Odpovědi chatbota byly irelevantní.
13.	Chatbot dobře zvládal jakékoli chyby nebo omyly.
14.	Chatbot se zdál být neschopný zvládnout jakékoli chyby.
15.	Používání chatbota bylo velmi snadné.
16.	Chatbot byl velmi složitý na používání.

Příloha E Web scraper pro akademické pracovníky

```
import tkinter as tk
from tkinter import filedialog, messagebox, scrolledtext
import threading
import requests
from bs4 import BeautifulSoup
import re
import time
import random

def scrape_teacher_info(url):
    try:
        response = requests.get(url)
        if response.status_code != 200:
            return "neexistuje"

        soup = BeautifulSoup(response.text, 'html.parser')
        page_text = soup.get_text(strip=True)
        if "Pracovník již není zaměstnancem" in page_text:
            return f"Information not available for URL: {url}\n"

        # Extrakce jména and titulu
        name_title_section = soup.find('h2', {'class': 'nadpisosobni'})
        name_title = name_title_section.get_text(strip=True) if name_title_section else "neexistuje"

        # Extrakce kontaktní info s regex a kontrola &nbsp;
        contact_info = soup.find('div', {'class': 'kontakty'})
        email_regex = re.compile('mailto:')
        phone_regex = re.compile('telefon\s?(&nbsp;)?')
        office_hours_regex = re.compile('konzultační/úřední hodiny\s?(&nbsp;)?')
        office_regex = re.compile('kancelář\s?(&nbsp;)?')

        email = contact_info.find('a', {'href': email_regex}).get_text(strip=True) \
            if contact_info and contact_info.find('a', {'href': email_regex}) else "neexistuje"
        phone = contact_info.find('td', string=phone_regex).find_next_sibling('td').get_text(strip=True) \
            if contact_info and contact_info.find('td', string=phone_regex) else "neexistuje"
        office_hours = contact_info.find('td', string=office_hours_regex).find_next_sibling('td').get_text(strip=True) \
            if contact_info and contact_info.find('td', string=office_hours_regex) else "neexistuje"
        office = contact_info.find('td', string=office_regex).find_next_sibling('td').get_text(strip=True) \
            if contact_info and contact_info.find('td', string=office_regex) else "neexistuje"

        # Extrakce Pozice
        positions_section = soup.find('td', {'colspan': '2'}).find('i')
        positions = positions_section.get_text('; ', strip=True).split('; ') if positions_section else "neexistuje"
```

```

# Extrakce predmetu
courses_section = soup.find('h2', {'id': 'predmety'})
courses = []
if courses_section:
    course_rows = courses_section.find_all_next('tr', valign="top")
    for row in course_rows:
        course_code = row.find('td', {'class': 'osobni'}).get_text(strip=True)
        course_name = row.find('b').get_text(strip=True)
        courses.append(f"{course_code}: {course_name}")

# Formátování získaných dat
output_str = f"Jméno a tituly: {name_title}\nPozice: {'; '.join(positions)}\nEmail: {email}\nPhone: {phone}\n \
f"Konzultační hodiny / Úřední hodiny: {office_hours}\nKancelář: {office}\nVyučované předměty: {'; '.join(courses)}\n\n"

return output_str

except Exception as e:
    return f"Error processing URL {url}: {str(e)}\n"

def append_to_file(data, file_path):
    with open(file_path, 'a', encoding='utf-8') as file:
        file.write(data)

def process_multiple_urls(url_list, output_file_path):
    for url in url_list:
        print("Processing URL:", url) # Debugging
        teacher_info = scrape_teacher_info(url)
        if teacher_info != "neexistuje":
            append_to_file(teacher_info, output_file_path)
            update_progress(f"Scraped successfully: {url}\n")
        else:
            update_progress(f"No information available or teacher no longer works there: {url}\n")
            delay_seconds = random.randint(2, 7) #delay abych nebyl nahlášen plus aby nedošlo k DDOS
            time.sleep(delay_seconds)
            update_progress(f"Scraped {url} (waited {delay_seconds} seconds)\n")
    on_all_scraping_complete()

def update_progress(message):
    progress_text.insert(tk.END, message)
    progress_text.see(tk.END) # Auto-scroll to the bottom

```

```

def update_progress(message):
    progress_text.insert(tk.END, message)
    progress_text.see(tk.END) # Auto-scroll to the bottom

def generate_urls(start_id, end_id):
    base_url = "https://fek.zcu.cz/osobni.php?IDWorker="
    return [base_url + str(i) for i in range(start_id, end_id + 1)]

def start_scraping():
    start_id = int(start_id_entry.get())
    end_id = int(end_id_entry.get())
    urls = generate_urls(start_id, end_id)
    output_file_path = filedialog.asksaveasfilename(defaultextension=".txt")
    if output_file_path:
        status_label.config(text="Scraping in progress..")
        threading.Thread(target=lambda: process_multiple_urls(urls, output_file_path), daemon=True).start()

def on_all_scraping_complete():
    status_label.config(text="Sběr dokončen.")
    messagebox.showinfo("Úspěch, všechna data byla úspěšně zkopírována!")

```

Příloha F Uživatelské rozhraní pro scraper pro akademické pracovníky

```
# Tkinter GUI setup
root = tk.TK()
root.title("Web Scraper")
root.geometry("600x400") # nastavení velikosti okna

# URL vstup
tk.Label(root, text="Enter start ID:").pack(pady=5)
start_id_entry = tk.Entry(root, width=50)
start_id_entry.pack(pady=5)

tk.Label(root, text="Enter end ID:").pack(pady=5)
end_id_entry = tk.Entry(root, width=50)
end_id_entry.pack(pady=5)

# Scrape tlačitko
scrape_button = tk.Button(root, text="Scrape Data", command=start_scraping, width=20)
scrape_button.pack(pady=10)

progress_text = scrolledtext.ScrolledText(root, height=10, width=70)
progress_text.pack(pady=10)

# Status značka
status_label = tk.Label(root, text="")
status_label.pack(pady=10)

root.mainloop()
```

Příloha G Další příklady využití umělé inteligence v marketingu

Influencer marketing AI influenceři

Marketing prostřednictvím online influencerů (anglicky Online influencer marketing) zahrnuje výběr a motivování influencerů společnostmi, aby interagovali s jejich publikem na sociálních médiích, s cílem využít jedinečných výhod influencerů k podpoře produktů společnosti, s konečným cílem zlepšit výsledky společnosti. Web 3.0, který představil mobilní zařízení a aplikace také razantně zrychlil popularitu sociálních médií. Toto umožnilo v podstatě jakémukoliv uživateli si vytvořit online sledující. Je nutné také zmínit změny v chování zákazníků, zákazníci či spotřebitelé totiž vyjadřují určitý rostoucí negativní postoj vůči přímému marketingu firem. Firmy také začínají být více nákladově citlivé či uvědomělé v rámci marketingových výdajů. Tyto faktory napomáhají konceptu influencer marketingu, jelikož pro firmy může představovat levnější alternativní variantu oproti například využívání známých celebrit k propagování svých produktů či služeb. (Leung a kol, 2022)

Podle softwarové společnosti Influencer marketing hub byly výdaje v tomto odvětví v roce 2022 16,4 miliardy dolarů a její odhad pro rok 2023 je růst na hodnotu 21,1 miliardy dolarů neboli 29 %. (Geysler, 2023)

Influencery lze rozdělit či segmentovat podle počtu sledujících, kdy existují 3 hlavní kategorie mikro – influenceři, mezo – influenceři a makro – influenceři. Příspěvky mohou mít dva hlavní účely z hlediska marketingu: prvním cílem je zvýšit zájem o nákup u jejich sledujících a druhým cílem je zlepšit atraktivitu a produktové znalosti jejich sledujících. Influenceři často vytvářejí příspěvky obsahující informace a doporučení týkající se vlastností produktů, které propagují. To vede k vyšší úrovni informativního obsahu o daném produktu a také tedy i povědomí o produktu. Pro úspěch influencera je kritický vztah mezi influencerem a sledujícími, na kterém totiž závisí i budoucí chování sledujících neboli tedy interakce či úmysl k nákupu. Při spolupráci influencera se značkou je rolí influencera působit jako ambasador společnosti, a tedy vytvářet takový sponzorovaný obsah, který vede ke zlepšení image a pověsti dané značky. Tento obsah by měl také zvyšovat nebo podporovat věrnost zákazníků vůči značce. (Joshi et al., 2023)

Influencerům bývá poskytnuta určitá kreativní svoboda, aby reklamní obsah byl v souladu s jejich osobní značkou a obsahem, který normálně přidávají na své sociální platformy. Určité směrnice a klíčové zprávy ale mohou být také poskytnuty, kterými by

se měli infleunceři řídit. Tato možnost umožňuje firmám využívat jedinečné zdroje influencerů, zejména jejich sledovanost a důvěru, kterou si vybudovali. Existuje zde samozřejmě určité riziko ztráty kontroly na propagačním sdělením. (Leung a kol, 2022)

Virtuální AI influenceři

I když stále více podniků využívá influencersy k propagaci svých produktů či služeb, a to nejen na úrovni mezinárodních podniků ale také i na úrovni malých a středních podniků vzniká zde určité riziko spolupráce s těmito osobami. Jestli je určitý influencer považován za vcelku bez rizikový partnera i tak je nutné počítat s určitou pravděpodobností, že se daná osoba může ztratit svou přitažlivost v podstatě přes noc. Je to dáno tím, že influenceři v širokém slova smyslu často dostávají do určitých skandálů. Společnosti spolupracující s těmito influencersy musejí následně zřídit určitou krizovou komunikaci. Často je také potřeba ukončit určité smlouvy, které mají s touto osobou uzavřené. (Allal-Chérif a kol., 2024)

Nahrazení lidského influencera za influencera vytvořeného pomocí umělé inteligence představuje revoluční metodu zapojení zákazníků, při které se společnost vyhýbá potencionálním problémům spojených s lidskými influencersy. Trend využívání virtuálních influencerů neustále roste na popularitě, přednost jím dávají například v Brazílie či Jižní Korea. Nejčastěji lze virtuální influencersy nalézt u luxusních značek zabývajících se například výrobou oblečení, doplňků či kosmetiky. (Allal-Chérif a kol., 2024)

Virutální influencer může mít také velkou spoustu sledujících na různých platformách, a tak může být považován za důvěryhodného poradce či „trend settera“ v nejedné oblasti. S tím, že se dá virtuální influencer naprogramovat a vytvořit tak aby se choval vypadal jako reálný lidský influencer, značí to, že by AI influenceři mohli mít stejný efekt jako je tomu u lidských influencerů. Je ale nutné zmínit, že neexistuje stále ještě dostatečné množství informací o tom, jak lidé reagují na tento koncept. (Sands et al., 2022)

Častokrát se jedná o virtuální influencersy, kteří mají představovat ženy mladého věku, existuje ale také řada animovaných postav, nebo postav z počítačových her, kteří vystupují na sociálních sítích jako influenceři.

Následující list představuje některé z nejznámějších virtuálních influencerů, kteří již spolupracovali se světovými značkami, a jejich počet sledujících na sociální síti Instagram.

- Lu do Magalu – 6,7 milionu sledujících
- Lil Miquela – 2,6 milionu sledujících
- Imma – 392 tisíc sledujících
- Shudu – 241 tisíc sledujících
- Bermuda – 234 tisíc sledujících (Molenaar, 2023)

Tento list obsahuje virtuální influencers, které lze zařadit do skupiny repliky člověka. Tato skupina bývá nejčastěji využívána k propagaci určitých produktů či služeb. Existují také skupiny jako nelidský influenceri a animovaní influenceri. Lil miquela již pózovala v reklamě pro Calvin Klein v roce 2019, a na jejím Instagramu lze také nalézt i novější spolupráce například společně s automobilkou BMW při představení elektrického vozu BMW iX2. (Nguyen, 2023)

Je ale také možné rozeznat rozdíl mezi maskotem společnosti a influencerem. Společnosti jako například Geico využívají svého maskota, malou ještěrku, jako jeden ze způsobů, kterým komunikují na sociálních sítích.

SEO a analýza klíčových slov

SEO neboli optimalizace pro vyhledávače (v originále search engine optimization) lze zjednodušeně popsat jako proces při, kterém dochází k vylepšování webových stránek, a to za účelem je více zviditelnit ve vyhledávačích jako je Google, Mozilla Firefox, Microsoft Bing, Safari a další. Je zde tedy snaha, aby stránka dané společnosti byla na prvních pozicích ve vyhledávačích při vyhledávání určitého produktu, služby nebo informací o dané problematice, kterou se společnost zabývá. (Goodwin, n.d.)

Vyhledávač od společnost Google neustále prochází různými aktualizacemi a vylepšeními. Už jen z této reality je nutné chápat, že proces SEO je neustálý proces optimalizace dané stránky. V roce 2022 vydal Google aktualizaci s názvem „Focus on people-first content“, která pojednává o obsahu zaměřený na lidi. Tato aktualizace byla zejména navržena, aby odměňoval obsah, který poskytuje uspokojivý zážitek uživatelům. Lze zmínit určitá kritéria, které by měl úspěšný obsah splňovat. Prvním kritériem je, že by daný obsah měl být zaměřený na dané publikum a neměla by zde být v první řadě snaha vytvořit obsah, který by uspokojil vyhledávač. Obsah by měl mít jasný účel, vzdělávat uživatele, uchovat pocit uspokojení po přečtení. (Google for Developers, 2022)

SEO lze rozdělit do určitých typů nebo skupin. Existuje jich řada lze zmínit například 2 hlavní skupiny:

- On-page SEO,
- Off-page SEO,

On-page SEO se soustředí na obsah, který je daném webu a je klíčový při snaze dostat se na vyšší umístění ve vyhledávačích. Aby bylo dosaženo určitého úspěchu u této problematiky je potřeba disponovat takovým obsahem, který lze snadno nalézt a je srozumitelný. Google častokrát upřednostňuje obsah, který přináší uživateli určitý užitek a je něčím jedinečný. Do činností, které spadají do on-page skupiny lze například zařadit výběr klíčových slov, optimalizace názvů stránek, aby obsahovali daná klíčová slova, optimalizace a aktualizace popisů jednotlivých položek na e-shopech, přidávání relevantního a kvalitního obsahu, využívání HTML tagů k správnému strukturování stránek například nadpisy H1, H2 a další. (Frankel, 2023)

Off-page SEO – někdy také off site SEO je další prvek celkového SEO, který je klíčový pro celkový výkon SEO dané stránky. Vztahuje se na všechny aktivity, které jsou prováděné mimo daný web, které pomáhají zvyšovat jeho důvěryhodnost a autoritu. Hlavní roli v tomto případě zde hraje takzvaný link building, nejedná se ale pouze o tuto strategii v rámci off page SEO. Link building je takzvané budování sítě zpětných odkazů a je to další aktivita SEO, která se snaží zvýšit návštěvnost daného webu. U link buildingu se řeší, kolik unikátních domén obsahuje link zpět na požadovanou stránku. Další metrika řeší, jak kvalitní/autoritativní jsou stránky, ze kterých pocházejí linky zpět na požadovano stránku. Další techniky v této oblasti může být, obsahový marketing, recenze a další. (Silva, 2022)

SEO nástroje založené na umělé inteligenci – vyhledávání klíčových slov

Vyhledávání klíčových slov nebo průzkum klíčových slov je proces, který identifikuje slova, slovní spojení či fráze, které využívá daná cílová skupina a existuje pravděpodobnost, že je znovu využije při vyhledávání produktů, služeb nebo dalších informací. V této fázi nastupuje firma, která tyto produkty či služby nabízí a využívá daná klíčová slova. Začlenění využívaných klíčových slov do obsahu pomáhá vyhledávačům jako je například již zmíněný Google přivést zákazníky na daný web. (Upwork, 2023)

Vyhledávání klíčových slov lze provádět manuálně, ale jedná se o činnost, která je časově náročná a rozsáhlá. Populární vyhledávače také neustále mění či upravují využívané

algoritmy za účelem zlepšení své funkčnosti, což vytváří určitý nátlak na tvorbu kvalitního obsahu, který zaujme potřebné publikum. Nástroje, které jsou založené na umělé inteligenci poskytují komplexní řešení pro vyhledávání klíčových slov. (Upwork, 2023)

Při vyhledávání klíčových slov existují 3 hlavní aspekty: relevance, důvěryhodnost a frekvence vyhledávání. Relevance a důvěryhodnost se týkají toho, jestliže daný obsah splňuje podmínky vyhledávačů a obsahuje užitečné informace. Aspekt frekvence vyhledávání v tomto případě znamená, že se daná stránka může umístit na první stránce s výsledky vyhledávání pro dané klíčové slovo, je ale nutné také zjistit, jestli je dané klíčové slovo často vyhledáváno, nebo se takto stránka umístila, protože se dané klíčové slovo téměř nevyhledává. Záměr uživatele je také nutno sledovat a analyzovat při vyhledávání klíčových slov. Klíčová slova totiž mohou nést více významů, je tedy nutné pro daná klíčová slova analyzovat typy výsledků. (Leist, 2022)

Nástroje, které využívají umělou inteligenci nebo upravené modely jako například některý z modelů GPT, přináší řadu výhod při procesu vyhledávání klíčových slov. Následující seznam představuje list výhod využití AI pro vyhledávání klíčových slov:

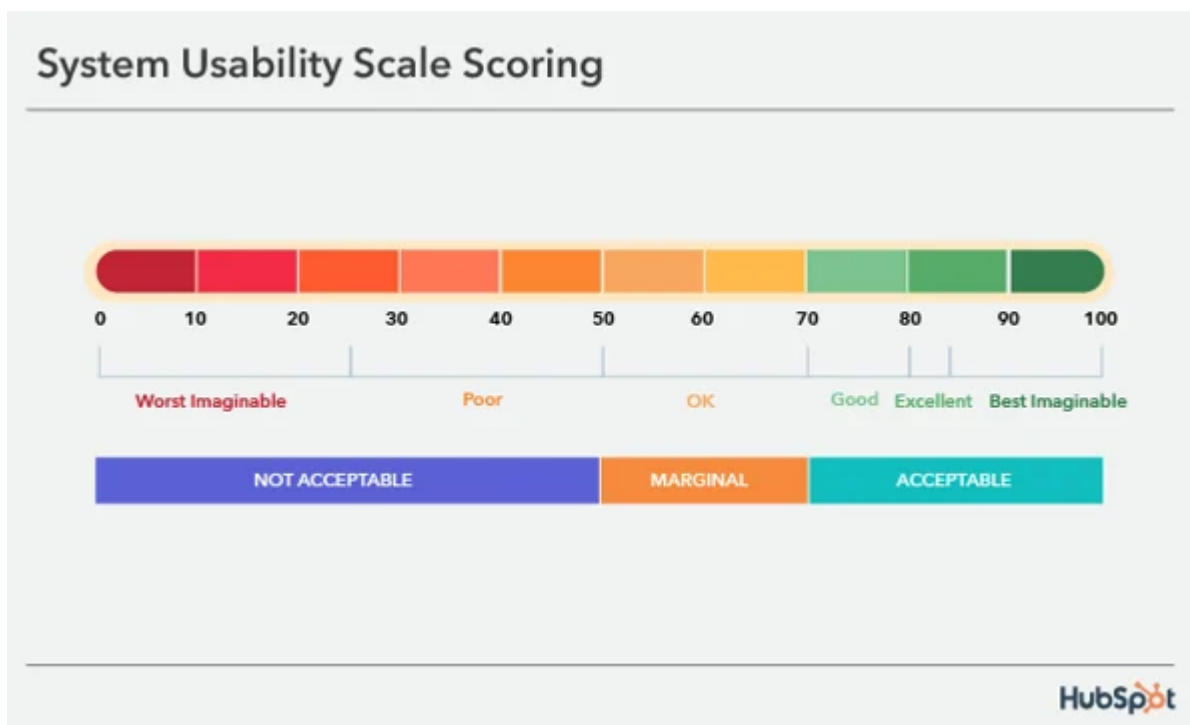
1. **Seznam potencionálních klíčových slov** – K vytvoření seznamu potencionálních klíčových slov lze například využít modelů GPT a zadat výchozí klíčové slovo, ze kterého následně daný model vytvoří sadu dalších potencionálních. Tento proces může ušetřit čas při brainstormování potencionálních klíčových slov. (Monteiro, 2023)
2. **Analýza počtu vyhledávání a konkurence** – Nástroje využívající AI jsou schopné analyzovat údaje o počtu vyhledávání identifikovat tak klíčová slova, která jsou výkonná. Dále také dokáží vyhodnotit úroveň konkurence z dat od Google Ads a určit tak, jak obtížné by bylo získat vysokou pozici pro určité slovo. (Junia, n.d.)
3. **Analýza záměru vyhledávání** – AI dokáže také vyřešit často složitý úkol analýzy záměru vyhledávání a identifikovat tak pravé záměry uživatelských dotazů. Toto zajišťuje, že je daný obsah sladěný s potřebami uživatelů a je relevantní. (Marin, 2024)
4. **Generování dlouhých klíčových slov a identifikace souvisejících a relevantních klíčových slov** – Významnou výhodou AI modelů je schopnost procházet obrovská množství dat a odhalit tak určité vzory a následně například

dlouhá klíčová slova. Kromě toho mají také schopnost identifikovat další relevantní klíčová slova, a to na základě pochopení záměru a preferencí uživatelů. (Monteiro, 2023)

5. **Předpovídání trendů klíčových slov** – Nezanedbatelnou výhodou AI nástrojů je schopnost pracovat s historickými daty a vytvářet určité predikce. AI nástroje využívají data z keyword planner od Google Ads k identifikaci určitých vzorů z historických trendů v rámci klíčových slov. To znamená, že určitá klíčová slova, která byla populární před 5 lety se mohou znovu stát populárními, kdy AI může identifikovat podobná vznikající klíčová slova, než získají popularitu ve velkém měřítku. (Errante, 2023)

Výhod, které umělá inteligence přináší pro vyhledávání klíčových slov je řada, je nutné ale podotknout, že stále nese určité limity. Jednotlivé nástroje postavené na AI jsou stále pořád jen nástroje a nelze tedy nahradit lidský dohled nebo kritické myšlení. I když AI dosáhla významného pokroku v rámci porozumění textu, stále se nejedná o perfektní proces, znovu je tedy potřeba lidský dohled marketéra a správná interpretace. Přílišná závislost na těchto nástrojích také limituje určitou kreativitu. (Monteiro, 2023)

Příloha H System usability Scale skóre



Zdroj: Chinn, 2022

Příloha I Průměrné skóre pro jednotlivé otázky a směrodatná odchylka

Otázka	Skóre	Smoch.
1	3,5	0,7
2	1,7	0,5
3	4,1	0,5
4	1,1	0,3
5	3,9	1,0
6	2,0	0,9
7	4,0	0,7
8	1,6	0,5
9	4,5	0,5
10	1,9	1,1
11	4,5	0,5
12	1,5	0,7
13	3,4	0,6
14	2,2	0,9
15	4,6	0,5
16	1,5	0,6

Příloha J CUQ skóre u jednotlivých účastníků

Účastník	CUQ skóre
8	93,8
1	92,2
3	92,2
5	89,1
19	89,1
16	87,5
6	85,9
12	84,4
2	82,8
18	82,8
11	79,7
15	78,1
20	78,1
21	78,1
9	76,6
4	73,4
13	73,4
10	70,3
14	70,3
7	67,2
17	67,2
22	65,6

Příloha K Příklad konverzace z testování

Otázka 1: "S čím mi můžeš pomoci?"

Odpověď 1: "Mohu vám pomoci s širokou škálou dotazů týkajících se Fakulty ekonomické Západočeské univerzity v Plzni. Zde je několik příkladů, jak vám mohu být nápomocen:

- Informace o doktorských studijních programech – Příjímací kritéria, popisy programů, stipendijní předpisy a další.
- Informace o magisterských studijních programech – Garantované obory, postupy přijímacího řízení, popisy jednotlivých programů.
- Akademický kalendář a klíčové milníky – Detailní rozvrhy pro všechny studenty, termíny pro podání diplomových prací a státní zkoušky.
- Mezinárodní studijní příležitosti – Programy bilaterální spolupráce, nabídky dvojího diplomu, Erasmus+ a další.
- Postupy pro podávání a publikaci kvalifikačních prací – Elektronické podávání, odklady publikace, směrnice pro bakalářské a magisterské práce.
- Informace o bakalářských programech – Principy přidělování studentů, proces přijímacího řízení, studijní plány.
- Předpisy a směrnice – Motivační vyhlášky, akademické předpisy, stipendijní předpisy, přechody mezi studijními programy.
- Akademický a administrativní personál – Informace o profesorech, jejich pozicích, oblastech výuky a kontaktních údajích.

Pokud máte konkrétní dotaz nebo potřebujete informace o něčem, co není uvedeno výše, neváhejte se zeptat, a já se pokusím vám co nejlépe pomoci."

Otázka 2. "Do kdy je potřeba odeslat přihlášku na bakalářský studijní program?"

Odpověď 2. "Termín pro podání přihlášky na bakalářský studijní program na Fakultě ekonomické Západočeské univerzity je do 30. dubna 2024. Přihlášku je potřeba vyplnit v elektronické formě na webové stránce. Uchazeč může podat více přihlášek na různé bakalářské studijní programy a na různá místa studia (Plzeň a Cheb). Každé přihlášce je potřeba zadat prioritu, která udává pořadí, v jakém má uchazeč zájem o studium programů se zohledněním místa studia. Dále je potřeba do 10. června 2024 nahrát sken maturitního vysvědčení v systému E-Přihláška ZČU. Pokud máte další dotazy nebo potřebujete pomoc, můžete kontaktovat fakultu na e-mailu."

Otázka 3. "Kdy je konec výuky v letního semestru?"

Odpověď 3. "Konec výuky v letním semestru pro akademický rok 2023/2024 je 17. května 2024."

Otázka 4. "Na jaké univerzity ve Španělsku lze vycestovat přes program Erasmus?"

Odpověď 4 "Přes program Erasmus+ můžete vycestovat na následující univerzity ve Španělsku:

- Universidade de Vigo, Facultade de Ciencias Economicas e Empresariais – Kód: E VIGO01, Adresa: Campus Universitario Lagoas Marcosende Vigo (Pontevedra), Webové stránky, Počet míst: 2 místa na 9 měsíců pro každého studenta (nebo 3 místa na 6 měsíců), Stupeň studia: Bc. Mgr., Obor studia: Business Studies and Management Science, Jazyk studia: ŠJ (B1) nebo AJ (B2).

- Universitat Rovira I Virgili – Kód: E TARRAGO01, Adresa: Avinguda Països Catalans, 17-19 43007 Tarragona, Webové stránky, Počet míst: 10 míst (Bc. stupeň 5 míst na 5 měsíců; NMgr. stupeň 3 místa na 5 měsíců; PhD. stupeň 2 místa na 10 měsíců) pro každého studenta

- Stupeň studia: Bc. Mgr. PhD.

- Obor studia: Business and Administration and Economics

- Jazyk studia: AJ (B1) nebo ŠJ (B1)

Pro více informací o podmínkách programu Erasmus+ a postupu přihlášení navštivte oficiální stránky programu."

Otázka 5: "Jsou i nějaké další možnosti?"

Odpověď 5. "Ano, kromě výše uvedených univerzit ve Španělsku jsou další možnosti pro studium prostřednictvím programu Erasmus+. Další španělské univerzity, kam lze vycestovat, zahrnují:

- Universidad de Granada – Kód: E GRANADA01, Adresa: Av. del Hospicio, s/n, 18071 Granada, Webové stránky, Počet míst: 2 místa na 5 měsíců pro každého studenta, Stupeň studia: Bc. Mgr., Obor studia: Business Studies and Management Science, Economics, Jazyk studia: ŠJ (B1) nebo AJ (B2).

- Universidad de Sevilla – Kód: E SEVILLA01, Adresa: C. San Fernando, 4, 41004 Sevilla, Webové stránky, Počet míst: 2 místa na 5 měsíců pro každého studenta, Stupeň studia: Bc. Mgr., Obor studia: Business Studies and Management Science, Economics, Jazyk studia: ŠJ (B1) nebo AJ (B2).

- Universitat de València – Kód: E VALENCI01, Adresa: Av. de Blasco Ibáñez, 13, 46010 València, Valencia, Webové stránky, Počet míst: 2 místa na 5 měsíců pro každého studenta, Stupeň studia: Bc. Mgr., Obor studia: Business Studies and Management Science, Economics, Jazyk studia: ŠJ (B1) nebo AJ (B2).

Toto jsou pouze příklady dalších univerzit ve Španělsku, které jsou dostupné pro studenty prostřednictvím programu Erasmus+. Nabídka univerzit a počet dostupných míst se mohou rok od roku měnit, proto je důležité sledovat aktuální informace na webových stránkách vaší univerzity nebo se obrátit na koordinátora programu Erasmus+ na vaší fakultě pro nejnovější informace a podrobnosti o přihlašovacím procesu."

Příloha L Dotazník postojů k AI



Dotazník o využívání a postojích k AI mezi studenty FEK

Vážení kolegové,

doufám, že se máte dobře. Chtěl bych vás oslovit s prosbou o spolupráci na diplomové práci, kterou momentálně připravuji a která se zaměřuje na využití umělé inteligence a chatbotů na naší Fakultě ekonomické. Měl bych zájem o získání vašich názorů a zkušeností s touto technologií, stejně jako o váš pohled na potenciální implementaci chatbotu pro naše akademické prostředí.

Pro tento účel jsem vytvořil krátký dotazník, který by mi měl poskytnout cenné informace pro další vývoj práce. Vyplnění by vám nemělo trvat déle než 10 minut.

Vaše odpovědi jsou zcela anonymní a budou použity především pro účely této diplomové práce. Vaše spolupráce mi bude velmi cenná.


Také bych vás chtěl pozvat k účasti na testování chatbota, který by mohl být užitečný pro naši fakultu. Na konci dotazníku je možnost vyplnit vaše kontaktní údaje, pokud máte zájem o bližší informace o testování.

Děkuji vám za vaši pozornost a čas věnovaný vyplnění tohoto dotazníku.

Petr Černý

cernype@gapps.zcu.cz [Přepnout účet](#)



 Není sdíleno

* Označuje povinnou otázku

Chatboti jsou inteligentní konverzační počítačové programy, které napodobují lidskou konverzaci v její přirozené podobě.

Je nutné rozlišit mezi "webovým chatbotem", automatizovaným nástrojem pro komunikaci, a "živým chatem s operátorem", který umožňuje přímý dialog s člověkem.

1. Webový chatbot:

Automatizovaný nástroj na webové stránce pro okamžité odpovědi na dotazy návštěvníků 24/7.

2. Živý Chat s Operátorem:

Interaktivní chatovací okno na webové stránce, poskytující přímou komunikaci s lidským operátorem pro osobní asistenci a odpovědi na dotazy.

Už jste někdy viděli chatbota na nějaké webové stránce? *

Ano

Ne

Už jste někdy využili chatbota pro získání informací nebo asistence online? *

Ano

Ne

Využíváte AI jazykové modely jako je GPT, Bard či jiné? *

- Ano, pravidelně je využívám
- Ano, občas je využívám
- Ne, nikdy je nevyžívám
- Ne, nejsem s nimi obeznámen/a
- Ne, nemám zájem je využívat

Které jazykové modely využíváte konkrétně?

- Gemini
- ChatGPT
- Microsoft Copilot
- Grok
- Jiné: _____

Využili byste AI chatbota, pokud by byl dostupný na webových stránkách fakulty ekonomické? *

- Určitě ano
- Pravděpodobně ano
- Nevím
- Pravděpodobně ne
- Určitě ne

Jak často si myslíte, že byste využívali chatbota na webové stránce fakulty? *

- Denně
- Týdně
- Měsíčně
- Jen příležitostně

Za jakým účelem byste využili AI chatbota na webových stránkách FEK? (Vyberte *
vše, co se vztahuje)

- Informace o studijních programech
- Dotazy na harmonogram akademického roku
- Informace o Erasmu a jiné mobility na FEK ZČU
- Informace o konzultačních hodinách a učitelích na FEK ZČU
- Dotazy na studijní a zkušební řád (kreditní hranice, opakovaný zápis předmětu, zápisová propustka a další)
- Dotazy na důležité termíny (odevzdání KP, termíny SZZ a další)
- Jiné: _____

Jakým způsobem nejčastěji získáváte informace o tématech zmíněných ve vašem předchozím výběru?

- Procházením oficiálních stránek fakulty.
- Konzultací s akademickými pracovníky nebo studijním oddělením.
- Hledáním informací na sociálních sítích nebo ve skupinách spojených s fakultou.
- Ptáním se starších studentů nebo spolužáků.
- Použitím externích zdrojů informací mimo oficiální webové stránky fakulty.
- Jiné: _____

Preferovali byste spíše formální nebo neformální tón chatbota? *

- Formální
- Neformální
- Bez preference

Jaký máte postoj k integraci AI jazykových modelů do vzdělávacích nástrojů a webů vzdělávacích institucí? *

- Velmi pozitivní
- Poměrně pozitivní
- Neutrální
- Spíše negativní
- Velmi negativní

Jakou míru důvěry byste přikládali odpovědím poskytnutým chatbotem ohledně vašich studijních záležitostí? *

- Plně důvěřuji - považuji informace za stejně spolehlivé jako od akademického personálu.
- Částečně důvěřuji - informace považuji za užitečné, ale potřebuji je ověřit s dalšími zdroji.
- Skeptický - informace beru s rezervou a vždy bych je konzultoval s někým ze studijního oddělení nebo s učitelem.
- Nedůvěřuji - spoléhám pouze na informace z oficiálních akademických zdrojů nebo přímo ze studijního oddělení.

Jakého jste pohlaví? *

- Muž
- Žena
- Jiné

Kolik vám je let?

Vaše odpověď _____

Který studijní program aktuálně studujete? *

- Bakalářský
- Navazující magisterský
- Doktorský

Jaký bakalářský studijní program studujete? *

- Podniková ekonomika a management
- Marketingové řízení
- Projektové řízení
- Informační management
- Ekonomická a regionální geografie

Ve kterém roce studia se nacházíte v rámci tohoto programu? *

- Jsem v prvním roce.
- Jsem ve druhém roce.
- Jsem ve třetím, nebo vyšším roce

Jaký navazující magisterský studijní program studujete? *

- Podniková ekonomika a management
- Informační management
- Projektové a procesní řízení
- Ekonomická a regionální geografie

Ve kterém roce studia se nacházíte v rámci tohoto programu? *

- Jsem v prvním roce.
- Jsem ve druhém, nebo vyšším roce

Ve kterém roce studia se nacházíte v rámci tohoto programu? *

- Jsem v prvním roce.
- Jsem ve druhém roce.
- Jsem ve třetím roce.
- Jsem ve čtvrtém, nebo vyšším roce.

Abstrakt

Černý, P. (2024). *Využití umělé inteligence v marketingových aktivitách*. [Diplomová práce, Západočeská univerzita v Plzni].

Klíčová slova: umělá inteligence, chatbot, marketing, RAG

Tématem této diplomové práce je využití umělé inteligence v marketingových aktivitách. Cílem této diplomové práce je vyvinout chatbota, který bude sloužit jako podpora pro studenty ve studijní administrativě a akademickém poradenství na fakultě. Teoretická část obsahuje vymezení pojmů umělé inteligence, a zejména koncept chatbotů a zapojení umělé inteligence do chatbotů. V závěru této části jsou představené etické aspekty spojené s umělou inteligencí. V praktické části jsou nejdříve zkoumány preference a postoje studentů k umělé inteligenci a fakulnímu chatbotu. Další kroky zahrnují vývoj chatbota, jeho testování a srovnání s dalšími prototypy, následované testováním chatbota s vybranými studenty. V poslední části je představen plán implementace do provozu a plán aktivit, které je potřeba provádět v rámci následné kontroly.

Abstract

Černý, P. (2024). *Use of artificial intelligence in marketing activities*. [Master's Thesis, University of West Bohemia].

Key words: artificial intelligence, chatbot, marketing, RAG

The topic of this master's thesis is the use of artificial intelligence in marketing activities. The main objective of this thesis is to develop a chatbot that will serve as a support for students in student administration and academic advising at the faculty. The theoretical part includes the definition of artificial intelligence, and in particular the concept of chatbots and the involvement of artificial intelligence in chatbots. This section concludes with an introduction to the ethical considerations associated with artificial intelligence. In the practical part, students' preferences and attitudes towards AI and faculty chatbots are first explored. The next steps include the development of the chatbot, its testing and comparison with other prototypes, followed by testing the chatbot with selected students. In the last section, a plan for the implementation into operation and a plan for the activities to be carried out in the follow-up are presented.