



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

**DIALOGOVÝ SYSTÉM
JAKO UŽIVATELSKÝ POMOCNÍK
PRO SOFTWARE ZPRACOVÁNÍ FOTOGRAFIÍ**

DIALOG SYSTEMS AS USER SUPPORT FOR PHOTOGRAPHY PROCESSING SOFTWARE

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

LUKÁŠ FOUKAL

VEDOUCÍ PRÁCE

SUPERVISOR

prof. Dr. Ing. PAVEL ZEMČÍK, dr. h. c.

BRNO 2025

Zadání bakalářské práce



164240

Ústav: Ústav počítačové grafiky a multimédií (UPGM)
Student: **Foukal Lukáš**
Program: Informační technologie
Název: **Dialogový systém jako uživatelský pomocník pro software zpracování fotografií**
Kategorie: Zpracování obrazu
Akademický rok: 2024/25

Zadání:

1. Seznamte se se současným stavem v dialogových systémech se zaměřením na použití jako uživatelský návod či podpora v práci s fotografiemi a fotografickým softwarem.
2. Definujte sadu testů či generátor testů pro dialogové systémy tak, aby demonstrovaly použitelnost dialogových systémů v této oblasti.
3. Navrhněte způsoby a postupy vedoucí ke zvýšení použitelnosti dialogových systémů v testované oblasti.
4. Implementujte systém dle bodů výše a zhodnoťte jeho možnosti a výsledky.
5. Diskutujte dosažené výsledky a možnou budoucí práci.

Literatura:

- Dle doporučení vedoucího práce

Při obhajobě semestrální části projektu je požadováno:

- Body 1, 2 a 3 zadání

Podrobné závazné pokyny pro vypracování práce viz <https://www.fit.vut.cz/study/theses/>

Vedoucí práce: **Zemčík Pavel, prof. Dr. Ing., dr. h. c.**

Vedoucí ústavu: Černocký Jan, prof. Dr. Ing.

Datum zadání: 1.11.2024

Termín pro odevzdání: 14.5.2025

Datum schválení: 12.11.2024

Abstrakt

Cílem této práce bylo vytvořit dialogový systém, který bude schopen v českém jazyce pomoci lidem obsluhovat pokročilý fotografický editor. Navrhnuté řešení je retrieval-augmented generation systém, který využívá sémantického vyhledávání s modelem simcse-dist-mpnet-paracrawl-cs-en a tvorbu odpovědi pomocí Llama 3.3. Výsledek práce, založený na open-source technologiích, dosahuje na sadě otázek z praxe stejné úspěšnosti zodpovídání dotazů jako produkt GPTs od OpenAI.

Abstract

The goal of this work was to develop a chatbot capable of assisting users in operating an advanced photo editor in the Czech language. The presented solution is a retrieval-augmented generation system that uses semantic search with the simcse-dist-mpnet-paracrawl-cs-en model and response generation via Llama 3.3. The proposed system, built on open-source technologies, achieves the same success rate in answering real-world queries as OpenAI's GPTs product.

Klíčová slova

retrieval-augmented generation, dialogový systém, velké jazykové modely, sémantické vyhledávání, český jazyk, uživatelská podpora

Keywords

retrieval-augmented generation, dialogue system, large language models, semantic search, Czech language, user support

Citace

FOUKAL, Lukáš. *Dialogový systém*

jako uživatelský pomocník

pro software zpracování fotografií. Brno, 2025. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce prof. Dr. Ing. Pavel Zemčík, dr. h. c.

Dialogový systém jako uživatelský pomocník pro software zpracování fotografií

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením profesora Pavla Zemčíka. Další informace mi poskytli Milan Behro, Michal Prouza a Pavel Fridrich ze společnosti ZONER a.s. Při komplikacích s ovladači pro platformu AMD ROCm byly velmi nápomocné rady od Milana Hrabánka, též ze společnosti ZONER a.s. Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Lukáš Foukal
11. května 2025

Poděkování

Chtěl bych poděkovat profesoru Zemčíkovi za jeho čas, odborné vedení a ochotu. Také chci poděkovat Milanu Behrovi za jeho podporu.

Obsah

1 Úvod	3
2 Velké jazykové modely	4
2.1 Vznik velkých jazykových modelů	4
2.2 Terminologie jazykových modelů	5
2.3 Kroky učení velkých jazykových modelů	7
2.4 Schopnosti jazykových modelů	7
2.5 Evoluce velkých jazykových modelů v OpenAI	8
2.6 Otevřené velké jazykové modely	12
2.7 Nasazení jazykových modelů	15
3 Získávání informací a jazykové modely	16
3.1 Text se znalostmi a jeho dělení	17
3.2 Sémantické vyhledávání	18
3.3 Aktuální produkční systémy	20
4 Zhodnocení současného stavu a specifikace řešení	21
4.1 Zhodnocení současného stavu	21
4.2 Implementační cíl práce	22
4.3 Specifikace požadovaného výstupu	23
5 Implementace systému	24
5.1 Implementace vytěžení znalostí	25
5.2 Implementace znalostní báze	26
5.3 Výběr modelu pro generaci odpovědi a jeho zapojení	30
5.4 Ostatní detaily implementace	33
5.5 Vyhodnocení vlastností implementace	34
5.6 Srovnání systému s alternativami	34
6 Závěr	37
Literatura	38
A Otázky z praxe	42

Seznam obrázků

2.1	Ilustrace text-to-text architektury modelu T5	5
2.2	Llama a některé modely na ní založené	13
3.1	Obrázek triviálního „retrieval-augmented generation“ systému	17
5.1	Rozdělení systému do tří základních částí	24
5.2	Obecná „retrieval-augmented generation“ architektura	27
5.3	Konkrétní „retrieval-augmented generation“ architektura	31

Kapitola 1

Úvod

„Každý produkt, který potřebuje návod, je z principu rozbitý.“, to světu před deseti lety sdělil Elon Musk. S postupem vývoje jsou ale počítače používány na stále složitější úkoly a vznikají programy, s čím dál více funkcemi. Ne vždy je rozhraní vhodné pro zkušeného uživatele intuitivní i pro začátečníka. Tato práce navrhuje podpůrný systém, který usnadní novým či občasným uživatelům orientaci v pokročilém editoru fotografií.

V průběhu historie se toto autoři komplexních softwarů pokoušeli řešit různými způsoby, od menu nápovědy, přes takzvané „tool tips“ i pana Sponku, kterého Microsoft uvedl na přelomu tisíciletí.

Novým paradigmatem poslední doby jsou „chatboti“, popularizováni produktem Chat-GPT, kteří využívají velké jazykové modely. V rámci populárního trendu můžeme pozorovat jejich integraci do široké škály aplikací, pro tento úkol ale mohou být opravdu užitečné.

Cílem této práce je za pomoci moderních dialogových systémů vytvořit lepší návod, tedy systém na odpovídání uživatelských dotazů o programu Zoner Photo Studio, za pomoci znalostní báze, sémantického vyhledávání a velkého jazykového modelu – to celé v češtině.

Této oblasti jsem se rozhodl věnovat, protože vidím velký potenciál v rozšíření velkých jazykových modelů znalostmi specifických oblastí.

Následující kapitola je věnována velkým jazykovým modelům, otázce, co je liší od modelů dřívějších, a vysvětlení, jak je možno je využívat pro širokou škálu úloh. V kapitole 3 jsou pak rozebrány aplikace velkých jazykových modelů v oblasti práce se znalostmi. Kapitola 4 vyhodnocuje současný stav předcházejících kapitol a prezentuje očekávané specifikace systému. Navazující kapitola 5 prezentuje provedené experimenty, výsledný systém a jeho výsledky. Provedená práce a její závěry jsou pak shrnuty v kapitole 6.

Kapitola 2

Velké jazykové modely

Tato kapitola se věnuje velkým jazykovým modelům a obsahuje informace bezprostředně relevantní k práci. Vzhledem k povaze a maximálnímu rozsahu této práce se nejedná o encyklopedický přehled.

Jak vlastně vypadá velký jazykový model, který vidíme v ChatGPT? Jaká byla k němu technická cesta? Jaká proběhla evoluce za poslední 2 roky?

Vzhledem k vysokým nákladům na tento výzkum [18] se mu převážně věnují velké technologické společnosti ze Silicon Valley. Některé z nich neposkytují detailní technické informace o svých systémech.

2.1 Vznik velkých jazykových modelů

Jazykové modelování se obecně zabývá výskyty sekvencí slov a jejich předpovídáním [32]. Díky inovacím posledních let lze tímto způsobem vykonávat všemožné úlohy [21].

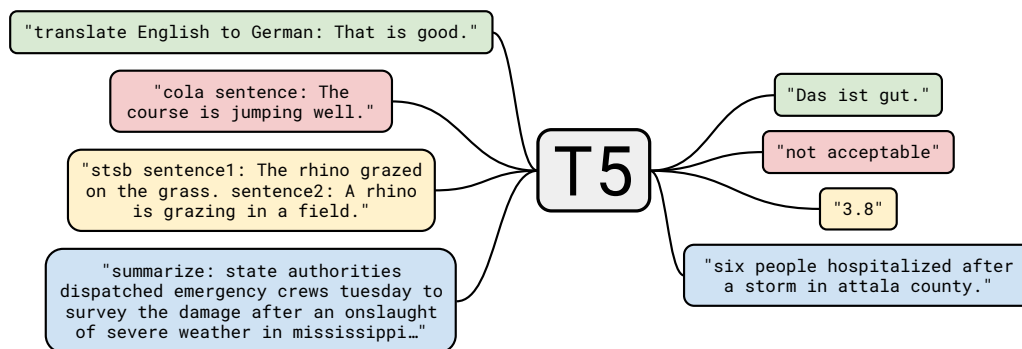
Zlomovým bodem pro jazykové modely se stala architektura transformerů [21], z publikace [28]. Skupina výzkumníků, převážně z Google Brain, zde prezentuje architekturu na základě tzv. pozornosti (attention). Touto novou architekturou, kde plně nahrazují rekurentní neuronové sítě mechanismem pozornosti, dosahují rychlejšího učení a lepší paralelizace učení.

Dalším faktorem v rozmachu jazykových modelů je dostupnost velkého množství dat a výpočetního výkonu pro trénování stále větších modelů, dnes řádově v desítkách až stovkách miliard parametrů [21]. Tímto se z Pretrained Language Models (PLM, předtrénovaných jazykových modelů) stávají Large Language Models (LLM, velké jazykové modely).

V roce 2018 vznikl v OpenAI model [23], dnes označovaný jako GPT-1 [32], který využívá architekturu transformeru z publikace [28] a učení s částečným dohledem (semi-supervised sequence learning) ze stejnojmenné publikace [5], kde bylo aplikováno na rekurentní neuronové sítě. Tato metoda učení spočívá v předučení (pre-training), které spadá do učení bez učitele, a doladění (fine-tuning), které klasifikujeme jako učení s učitelem [23].

Rané modely, jako Google T5 ($\leq 11B$ parametrů) [24], se soustředili na transfer learning. Model je prvně před-trénován na velkém objemu textu a následně doladěn pro konečný počet tradičních úloh zpracování textu, jako je překlad či sumarizace [24]. T5 zavádí sjednocený přístup, kde všechny úkoly jsou převedeny na text-to-text formát, viz obrázek 2.1. Publikace k T5 dokonce ve vlastním názvu uvádí, že prohledává limity tohoto postupu [24].

¹převzato z literatury [24]



Obrázek 2.1: Ilustrace text-to-text architektury modelu T5¹

Proti tomu stojí publikace [3] a model GPT-3, kde autoři ukazují, že významně větší (175 miliard parametrů, 10× větší než největší dosavadní husté modely, 16× větší než největší varianta T5) model dokáže podobně jako člověk fungovat pomocí tzv. few-shot pochopení (viz sekci 2.4), kde je v rámci zadání úlohy uveden příklad a pro každý typ úlohy není nutno tradiční trénování pro daný úkol. Učení GPT-3 spočívá pouze v učení bez učitele na velkém objemu textu. V některých úkolech je tento přístup GPT-3 srovnatelný s předchozími postupy, kde byl model doladěn s pomocí specifických anotovaných dat. Hlavní předností je samozřejmě možnost adaptace na širokou škálu úkolů, pro které ani nemusí existovat trénovací sada dat [3].

Na základě předučeného GPT-3 vyvinulo OpenAI model InstructGPT [22]. Jedná se o model dotrénovaný na lidské zpětné vazbě metodou reinforcement learning from human feedback (RLHF). Tento model výrazně lépe následuje instrukce a odpovídá na otázky, protože není naučen pouze předpovídat další slovo, ale i bezpečně vykonávat úkoly od uživatele. Jedná se o aplikaci dlouholetého výzkumu do tzv. alignmentu, viz sekci 2.3.

V listopadu roku 2022 vznikl systém blízce příbuzný k InstructGPT, ale založený na konverzaci – ChatGPT². Zde vidíme možnost ptát se na doplňující otázky nebo blíže usměrnit žádanou odpověď. Základem pro prvotní ChatGPT byl model z řady GPT-3.5. K jeho vzniku vedla stejná metoda učení z lidské zpětné vazby (RLHF) jako pro InstructGPT.

Uvedení ChatGPT na trh nelze popsat jinak než jako virální³. Konverzační systém byl k dispozici zdarma a již v lednu (měsíc po spuštění) jej denně použilo průměrně 13 milionů uživatelů. Během pouhých 2 měsíců získal 100 milionů měsíčních aktivních uživatelů, jednalo se o nejrychlejší dosažení této laťky pro spotřebitelskou aplikaci v historii internetu. A nezapomeňme, že chatgpt.com nebylo dostupné na čínském trhu [33].

2.2 Terminologie jazykových modelů

Tato sekce blíže vysvětluje některé pojmy, které jsou používány v kontextu velkých jazykových modelů.

Počet parametrů Výrazem velké jazykové modely (large language models, LLMs) se dnes běžně myslí modely architektury transformer, které mohou obsahovat až stovky miliard

²<https://openai.com/index/chatgpt/>

³<https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>

parametrů [32]. Dnes dokonce i více⁴. Jejich schopnosti jsou provázány s počtem parametrů modelu pomocí tzv. scaling laws (škálovací zákony) [32]. S velikostí modelů logicky roste i výpočetní náročnost.

Kvantizace Vzhledem k obrovskému počtu parametrů kladou velké jazykové modely vysoké nároky na operační paměť. Jedním z technických řešení tohoto problému je kvantizace (quantization) [32], kdy jsou hodnoty s plovoucí desetinou čárkou typicky převedeny na menší celočíselné reprezentace. Tento proces je možností provádět pro váhy i aktivace modelu. Vliv na výstupy modelu může být relativně malý při reprezentaci na 8 bitech, pro 4 bity je logicky větší. Také je možno využít různých počet bitů pro různé vrstvy.

Žetony Jazykové modely běžně nepracují s jazykem ve smyslu ASCII nebo Unicode znaků, jako textový editor [21]. Jazyk lze modelovat na mnoha úrovních jemnosti [20], včetně jednotlivých znaků, nejpůvodnější jsou ale tzv. tokeny (žetony), které velikostně zapadají někde mezi slovo a znak. Pro účely této práce není nutné přesně chápat proces tokenizace a význam tokenů. Jen je nutné mít na vědomí, že tyto jednotky například pro OpenAI GPT-4 odpovídají zhruba čtyřem znakům a 100 tokenů bude asi 75 anglických slov⁵.

Kontextové okno Modely architektury transformer dokáží do generace následujícího tokenu zvážit pouze omezený počet tokenů předcházejících [32]. Počet těchto tokenů je někdy označován jako délka kontextového okna.

„Mixture-of-experts“ Původní transformer modely (například T5) stály na encoder-decoder architektuře [32]. Většina dnes relevantních (například série GPT, Llama) spíše užívá causal decoder.

Diskuze konkrétních architektur, jejich výhod a detailů není předmětem této práce. Je ale vhodné vysvětlit rozšíření mixture-of-experts (směs expertů, MoE), které využívají některé nejnovější modely [32]. To má velký vliv na výkonnost, výpočetní cenu a tedy i odezvu konverzačního systému. Tyto odlišné výpočetní vlastnosti ovlivňují nasazení velkého jazykového modelu, což bude diskutováno později.

Princip MoE architektury spočívá v rozdělení grafu modelu na podgrafy (jednotlivé experty), které jsou řídicí aktivovány [32]. To nabízí možnost zvyšovat celkový počet parametrů při zachování výpočetní náročnosti. V danou chvíli je aktivováno pouze několik málo expertů [21].

Halucinace Termín halucinace pochází již z dřívějšího výzkumu do zpracování přirozeného jazyka [11]. Popisuje fenomén, kdy model generuje věrohodný text, který ale není založený na informacích v textu zdrojovém. Halucinace lze dále dělit podle toho, zda je výsledný text přímo v rozporu (intrinsic hallucination) nebo jen výstup nelze vyvodit ze vstupu (extrinsic content). V následující kapitole této práce jsou diskutovány techniky, které se snaží halucinace ve velkých jazykových modelech omezit.

⁴<https://ai.meta.com/blog/llama-4-multimodal-intelligence/>

⁵<https://platform.openai.com/tokenizer>

2.3 Kroky učení velkých jazykových modelů

Jak je zmíněno začátkem kapitoly, jednou z odlišujících vlastností dnešních velkých jazykových modelů je víceřadový proces učení, původně uvedený v publikaci [5]. Tato sekce je věnována jednotlivým krokům dnešních postupů.

„Pre-training“

V procesu předučení (pre-training) získávají velké jazykové modely znalosti a pochopení světa [32]. V tomto procesu efektivně komprimují informace do svých parametrů [6], proto některé publikace [15] označují tyto znalosti jako parametrické. Schopnost velkých jazykových modelů věrohodně přistupovat k uvedeným informacím je omezená a stále neexistují jednoduché způsoby aktualizace těchto informací [15]. Tyto informace také nemohou být osobní pro daného uživatele, nýbrž musí být zahrnuty v učící sadě při tomto prvním kroku procesu⁶.

„Fine-tuning“

Tento krok, někdy překládaný jako doladění, může probíhat dvěma způsoby. Tradiční způsob transfer-learning pro jeden úkol jako u T5 [24] nebo ladění pro následování instrukcí, což umožňuje modelu odpovídat na dotazy a následovat obecné instrukce [21]. Daty pro učení jsou v tomto případě páry vstupu a výstupu. Toto vede ke zlepšení tzv. zero-shot výsledků (viz sekci 2.4).

„Alignment-tuning“

Velké jazykové modely mají tendenci generovat nepravdivé, ovlivněné nebo jinak škodlivé texty [21]. Pro dosažení pravdivějších a užitečnějších výsledků jsou jazykové modely dále trénovány podle lidské zpětné vazby metodou zvanou reinforcement learning with human feedback (RLHF, posilované učení s lidskou zpětnou vazbou) [21]. Model je považován za aligned, pokud je ochotný, upřímný a neškodný.

2.4 Schopnosti jazykových modelů

Za emergentní schopnosti (emergent abilities, vynořující se schopnosti) považujeme ty, které nejsou přítomny u malých modelů, ale náhle se objevují u modelů velkých [32]. Toto chování lze přirovnat ke změně skupenství ve fyzice. Tento efekt je neočekávaný a nelze jej předpovídat extrapolací [29]. Existence takových schopností vzbuzuje otázky, jestli další zvětšování jazykových modelů povede k novým schopnostem.

Tato sekce se věnuje těmto vlastnostem a jejich využívání v rámci promptingu. Kolem promptů vznikla celá disciplína přezdívaná prompt engineering, která se zabývá tím, jak pomocí přirozeného jazyka na vstupu získat co nejlepší výsledky z velkého jazykové modelu [32].

„In-context learning“ Volně přeloženo jako učení uvnitř kontextu, tato vlastnost popisuje schopnost velkých jazykových modelů vykonávat úkoly, na které nebyli předem tré-

⁶<https://openai.com/index/chatgpt-plugins/>

novány a jejich popis či příklady se objevují až uvnitř kontextu [32]. Tato schopnost byla prvotně objevena v GPT-3.

V praxi se promptu, který využívá několik málo příkladů dané úlohy, přezdívá **few-shot** [29].

„Instruction following“ V překladu následování instrukcí, tato schopnost umožňuje vykonávat nové úlohy, na které nebyly trénovány, z jejich popisu [32]. Vzniká pomocí doladění modelu na různorodé sadě popisů úloh. Toto je využíváno v **zero-shot** promptech bez příkladu.

„Chain-of-thought reasoning“ Jednoduše uvažování proudem myšlenek, tato schopnost, na kterou poukázali vědci ze skupiny Google Brain [30], popisuje stav, kdy model „přemýšlí nahlas“. Toto lze vyvolávat v běžných, neupravených modelech, včetně GPT-3, pomocí vhodného promptu. Autoři zároveň zmiňují, že tento přístup má v mnoha úkolech rostoucí škálovací křivku, kde běžné dotazování stagnuje. Jinými slovy, nový typ dotazování vede ke zlepšování výsledků pro větší a větší modely, i v úkolech, kde se toto dříve neprojevovalo. Nejen že úspěšnost na testech roste rychleji s počtem parametrů, ale výsledky jsou v předložených testech lepší než s tradičními metodami. Publikace [30] tedy předkládá teorii, že dosavadní výsledky velkých jazykových modelů jsou pouze spodní hranicí možného pro daný počet parametrů.

Dotazování

Mimo to, že dnešní dotazovací (prompting) strategie využívají vynořujících se vlastností pomocí instrukcí, příkladů a chain-of-thought, existují také obecné pravidla pro správné dotazování.

Důležitý je kontext, ve kterém se úkol odehrává. Dotaz „*Jaký je rekord pro milovou trať?*“ je velmi jednoznačný, pokud sedíte na atletickém oválu. To ale model neví a proto může vytvořit odpověď například z kontextu motorsportu [11]. Jako kontext bychom zařadili například výstupy z vyhledávací fáze retrieval-augmented generation systému (generování rozšířené o načtení, viz kapitolu 3) [32], který je předmětem této práce.

Je důležité, aby dotaz byl ve vhodném stylu [32]. Obecně dobré jsou jednoznačné dotazy a detailní instrukce. Do stylu patří i další triky, například prefix „*Uvažujme krok po kroku*“, často vede k lepší dekompozici úkolu. Úvody jako „*Jsi expertem na daný úkol*“ mohou zároveň posílit přesvědčivost modelu. Pro konverzační systémy také může pomoci rozdělení zadání do více zpráv. Dalším trikem je rozdělení samotného úkolu na více kroků, ideálně i očíslovaných.

Příklad vhodného promptu je zde zařazen jako tabulka 2.1.

2.5 Evoluce velkých jazykových modelů v OpenAI

Pro lepší pochopení původní mise a dnešní uzavřenosti OpenAI je zde zařazena krátká vsuvka o historii této organizace.

⁸Převzato z publikace [32], zkráceno, přeloženo.

Tabulka 2.1: Příklad správného promptu. Text **modře** označuje popis úkolu, text **červeně** označuje kontextové informace, text **zeleně** označuje ukázky a text **žlutě** označuje styl výzvy.⁸

Použij poskytnuté články ohraničené trojitými uvozovkami k zodpovězení otázek. Pokud odpověď v článcích nenajdeš, napiš "Odpověď jsem nenašel."

Články: ""Joao Moutinho je portugalský fotbalista, který naposledy hrál jako střední záložník za klub anglické Premier League Wolverhampton Wanderers a portugalskou reprezentaci.""

Otázka: Je následující věta věrohodná? 'Joao Moutinho byl vyautován na třetí metě.'

Odpověď: Pojď me na to krok za krokem. Joao Moutinho je fotbalista. Být vyautován na třetí metě je součástí baseballu, ne fotbalu. Takže odpověď je Ne.

...
<Ukázky>

Články: <vlož články, každý ohraničený trojitými uvozovkami>

Otázka: <vlož otázku>

Odpověď:

Historie OpenAI

OpenAI vzniklo jako nezisková organizace s cílem posouvat hranice umělé inteligence ve prospěch celého lidstva, bez nutnosti generovat finanční zisk⁹. Výzkum nové organizace vedl Ilya Sutskever, do čela společnosti zasedli společně Elon Musk a Sam Altman. Celkem zakladatelé a jiní partneři (Peter Thiel, Y Combinator, Amazon Web Services a další) přislíbili společnosti miliardu dolarů.

Následující kapitola historie je předmětem soudních sporů. Faktem je, že v roce 2016 společnost oznámila spolupráci s Microsoftem¹⁰, která byla dále posílena v letech 2019, 2021 a 2023¹¹. V roce 2019 vznikla společnost s omezeným ziskem, v níž je Microsoft investorem a kterou ovládá neziskovka OpenAI¹². V posledních měsících se OpenAI snaží dále změnit svou korporátní strukturu.

Spor spočívá v tom, že Elon Musk považuje plánované změny za přeměnu OpenAI na ziskovou společnost a zánik neziskové organizace, kterou částečně založil¹³. Společnost naopak argumentuje, že dobročinná větev bude silnější než dřív, díky prostředkům z komerčních aktivit. Zároveň poukazuje na Elonův neúspěšný pokus přivlastnit si OpenAI pod křídla Tesly v roce 2017¹⁴. Je také nutno podotknout, že Musk je kvůli xAI v přímé konkurenci s OpenAI.

Otevřenost

Co se otevřenosti z názvu týče, OpenAI naposledy publikovalo parametry modelu GPT-2 v roce 2019¹⁵. To stejné podle svého ředitele¹⁶ plánuje učinit v průběhu letošního roku pro nový, blíže neurčený model. Toto stále není plně open-source model, jako můžeme vidět u startupu DeepSeek (viz sekci 2.6). Novější modely společnosti z posledních 6 let

⁹<https://openai.com/index/introducing-openai/>

¹⁰<https://openai.com/index/openai-and-microsoft/>

¹¹<https://openai.com/index/openai-and-microsoft-extend-partnership/>

¹²<https://openai.com/index/why-our-structure-must-evolve-to-advance-our-mission/>

¹³<https://openai.com/index/court-rejects-elon/>

¹⁴<https://openai.com/index/elon-musk-wanted-an-openai-for-profit/>

¹⁵<https://github.com/openai/gpt-2-output-dataset>

¹⁶<https://x.com/sama/status/1906793591944646898>

tedy nejsou k dispozici k provozu na vlastním vybavení nebo k bližšímu nahlédnutí. Proto v následující sekci bude diskutována cena spíše finanční než výpočetní, protože skutečné technické informace nejsou veřejně dostupné. I přes to nelze díky rekordním výsledkům [12] jejich modelů OpenAI vynechat z této práce.

Modely

První modely série GPT a vývoj ChatGPT byly zmíněny v sekci 2.1, tato sekce tedy navazuje tam, kde předchozí končí.

GPT-4 Příchod ChatGPT na veřejnou scénu posunul světem, a na místě nestáli ani jeho tvůrci. Necelé 3 měsíce od prvotního spuštění přibyla možnost využívat nový model GPT-4¹⁷. Ten je schopen projít americkou právníkou zkouškou lépe než 90% lidských uchazečů, má výrazně bezpečnější chování a zároveň podporuje i obrazové vstupy, i když tato funkcionalita byla uveřejněna až později. Stejně jako každý jiný transformer se GPT-4 potýká s halucinacemi, ale po dotrénování na lidském vstupu dosahuje o 40% lepších výsledků na testu pravdivosti výstupu, než stejně upravený GPT-3.5.

Konkrétní technické změny, které vedly k těmto lepším výsledkům nejsou v technické zprávě popsány. Toto utajení je v anglickém originálu [1] zdůvodněno soutěžním prostředím i bezpečností:

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Existují rozšířené spekulace [32], že GPT-4 je založen na architektuře mixture-of-experts (viz sekci 2.2), ale k této teorii neexistuje oficiální potvrzení.

GPT-4 Turbo V rámci akce DevDay v listopadu 2023 bylo GPT-4 vylepšeno na GPT-4 Turbo¹⁸. Místo skutečného turbodmychadla byl model opatřen novějšími znalostmi, delším kontextem (až 128 tisíc tokenů, ekvivalent 300 stran textu) a vyšší efektivitou, což se projevilo na až trojnásobném poklesu ceny pro vývojáře používající API. Přes API bylo také možné konečně využívat obrazové vstupy GPT-4 nebo provádět vlastní doladění OpenAI modelů.

GPT-4o Dalším krůčkem pro OpenAI je GPT-4o, kde „o“ značí „omni“¹⁹. Mimo text a obraz, který dokáže zpracovávat GPT-4, si poradí i se zvukem a videem, zároveň na výstupu, kde přechází modely ze série GPT produkovali text, dokáže vytvářet zvuk i obrázky. V minulosti bylo možné hlasově používat GPT-4 zřetězený s modely pro přepis do textu a následné čtení (tzv. Voice Mode), to ale vedlo k průměrnému zpoždění přes 5 sekund. Nový model toto dokáže od začátku po konec a s průměrným zpožděním pouhých 320 milisekund. Díky této nativní podpoře dokáže model zachytit emoce, intonaci a ruch, zároveň to umožňuje GPT-4o i zpívat.

I z textové pohledu této práce je GPT-4o velký skok vpřed. Nový model je dvakrát rychlejší a k dispozici za poloviční cenu. Změny v tokenizaci, nutné pro nové formáty vstupu

¹⁷<https://openai.com/index/gpt-4-research/>

¹⁸<https://openai.com/index/new-models-and-developer-products-announced-at-devday/>

¹⁹<https://openai.com/index/hello-gpt-4o/>

a výstupu, také vylepšili efektivitu pro psaný text. V angličtině je stejný vstup zhruba o 10% méně tokenů, pro ruštinu (jediný uvedený příklad slovanského jazyka) je to dokonce o 40% méně. Tento model byl uveřejněn v květnu 2024, od prvního ChatGPT ho tedy dělí 18 měsíců.

o1 V září 2024 společnost odhalila o1²⁰, velký jazykový model, který byl v rámci fáze učení s učitelem přizpůsoben na uvažování řetězcem myšlenek (tzv. chain-of-thought, viz sekci 2.4).

OpenAI testovalo model o1 na zkoušce pro nejnadanější středoškoláky. V nejuvážnějším zapojení, kde model řeší každý úkol tisíckrát a následně sám vybírá finální odpověď, dosáhl skóre 93% a umístil by se mezi 500 nejlepších studentů ve Spojených státech.

o3 Na konci roku 2024 naznačila společnost budoucnost modelů v sérii o. Sam Altman v rámci projevu²¹ prozradil, že model navazující na o1 byl pojmenován o3, sám odůvodnil, že tak činí z respektu k přátelům ze společnosti Telefónica²² a navazují na významnou tradici špatných jmen v OpenAI.

Předností o3 je hlavně vyšší výpočetní efektivita oproti předchozí generaci²³. Ve vybraných úkolech dokáže i menší o3-mini s nejvyšším nastaveným úsilím (délce řetězce uvažování) předběhnout větší o1.

GPT-4.5 V únoru 2025 OpenAI uvedlo GPT-4.5. Je to logický postup série GPT, větší model, více učení bez učitele, více znalostí o světě, dle společnosti také pocitově vyšší EQ²⁴. Dle vyjádření ředitele společnosti na síti X²⁵ se jedná o poslední model bez vnitřního chain-of-thought. Výsledkem tohoto škálování jsou také větší náklady. V dubnu 2025 je model v ceníku OpenAI API²⁶ 30× dražší pro vstupní tokeny a 15× dražší za tokeny na výstupu než GPT-4o.

Tato vysoká cena ale přináší velké výsledky, podle publikace [12] z března 2025 dokáže GPT-4.5 projít Touring testem. Ten spočívá ve hře, kde jeden vyšetřovatel textově komunikuje se dvěma svědky – člověkem a strojem. Cílem obou svědků je přesvědčit vyšetřovatele, že právě oni jsou člověkem. GPT-4.5 dokázalo takto obelstít 73% zúčastněných.

Pro úplnost je uvedeno, že testem zároveň prošel systém Llama 3.1 405B (viz sekci 2.6), ale s úspěšností pouhých 56% [12].

GPT-5 Podle původního plánu publikovaného na síti X²⁷ z února 2025 je cílem společnosti sjednotit všechnu technologii pod střechem GPT-5. To znamená zkombinovat sérii GPT a „o“ do jednoho produktu v API i ChatGPT. V dubnu Altman na svém profilu na síti dříve známé jako Twitter uvedl, že toto rozhodnutí otáčí. Příspěvek²⁸ informuje, že vytvořit kombinaci chain-of-thought („o“) a GPT modelu bylo těžší, než se zdálo.

²⁰<https://openai.com/index/learning-to-reason-with-llms/>

²¹<https://www.youtube.com/watch?v=SKBG1sqdyIU>

²²Telefónica je mateřskou společností značky O2, viz <https://www.telefonica.com/en/about-us/brands/>.

²³<https://openai.com/index/openai-o3-mini/>

²⁴<https://openai.com/index/introducing-gpt-4-5/>

²⁵<https://x.com/sama/status/1889755723078443244>

²⁶<https://openai.com/api/pricing/>

²⁷<https://x.com/sama/status/1889755723078443244>

²⁸<https://x.com/sama/status/1908167621624856998>

2.6 Otevřené velké jazykové modely

Ve vývoji velkých jazykových modelů lze pozorovat trend k open-source modelům [21].

Meta AI

Společnost Meta²⁹ vyvíjí své modely poněkud odlišně od OpenAI. Divize Meta AI, která se soustředí na rozvoj umělé inteligence, uvádí na svém webu³⁰ jako své klíčové hodnoty otevřenost, spolupráci, excelenci a škálovatelnost.

Tyto hodnoty nejsou jen marketingovým cvičením. V oblasti velkých jazykových modelů vytváří Meta AI modely série Llama. Ty se staly velmi oblíbenými díky jejich otevřenosti a efektivitě [32]. Otevřenost zároveň vedla ke vzniku mnoha upravených variant, viz obrázek 2.2. Na sklonku roku 2024 měla Llama a její varianty přes 650 milionů stažení³¹.

Llama 1 První model Llama byl představen v únoru 2023 (3 měsíce po senzaci ChatGPT) v publikaci [26]. Sada pro učení obsahuje $1,4 \times 10^{12}$ tokenů, učení pro největší variantu trvalo 21 dní na 2048 GPU Nvidia A100. Výsledkem práce jsou tzv. foundational (základní) modely s počty 7, 13, 33 a 65 miliard parametrů. U těchto modelů neproběhlo doučení na následování instrukcí.

Llama 2 Llama 2 [27] navazuje na předchozí práci a obsahuje modely od 7 do 70 miliard parametrů. Objem dat použitý pro učení je o cca 40% větší. Tentokrát se nejedná pouze o předtrénované modely, ale Meta publikuje i Llama 2-Chat, model přizpůsobený pro konverzační použití. Podle publikace Llama 2 přesahuje výsledky ostatní otevřených modelů ve většině testů a v některých úkolech je zaměnitelná i za proprietární ChatGPT.

Llama 3/3.1 Třetí instance série [9] je rozsáhlejší skupina modelů, sahající od 8 po 405 miliard parametrů.

Základem je více dat, konkrétně 15×10^{12} tokenů, to je $8\times$ více než pro Llama 2. Učení největšího modelu taky použilo $50\times$ více výpočetních operací než největší Llama 2 a proběhlo na 16 000 GPU Nvidia H100 (80GB VRAM, 700W TDP).

Model podporuje kontext až 128 tisíc tokenů (zhruba 300 stran textu, stejně tokenů jako GPT-4 Turbo).

Ve své architektuře se významně neliší od svých předchůdců, ale používá nové optimalizační triky okolo mechanismu pozornosti [9].

Jedním z cílů pro Llama 3 je vícejazyčnost, proto je v učící sadě krom převládající angličtiny oficiálně i 7 dalších jazyků a tomuto byl přizpůsoben i tokenizátor. Vývojáři přiznávají, že model může být schopen generovat text ve více jazycích³³, ale silně nedoporučují konverzovat v jiných než oficiálně podporovaných.

Llama 3.2 Navazující verze 3.2 se soustředí převážně na multimodalitu³⁴ a odlehčené textové verze pro mobilní zařízení. Výsledkem jsou 11 a 90 miliardové verze zaměřené na

²⁹dříve Facebook, viz <https://about.fb.com/news/2021/10/facebook-company-is-now-meta/>

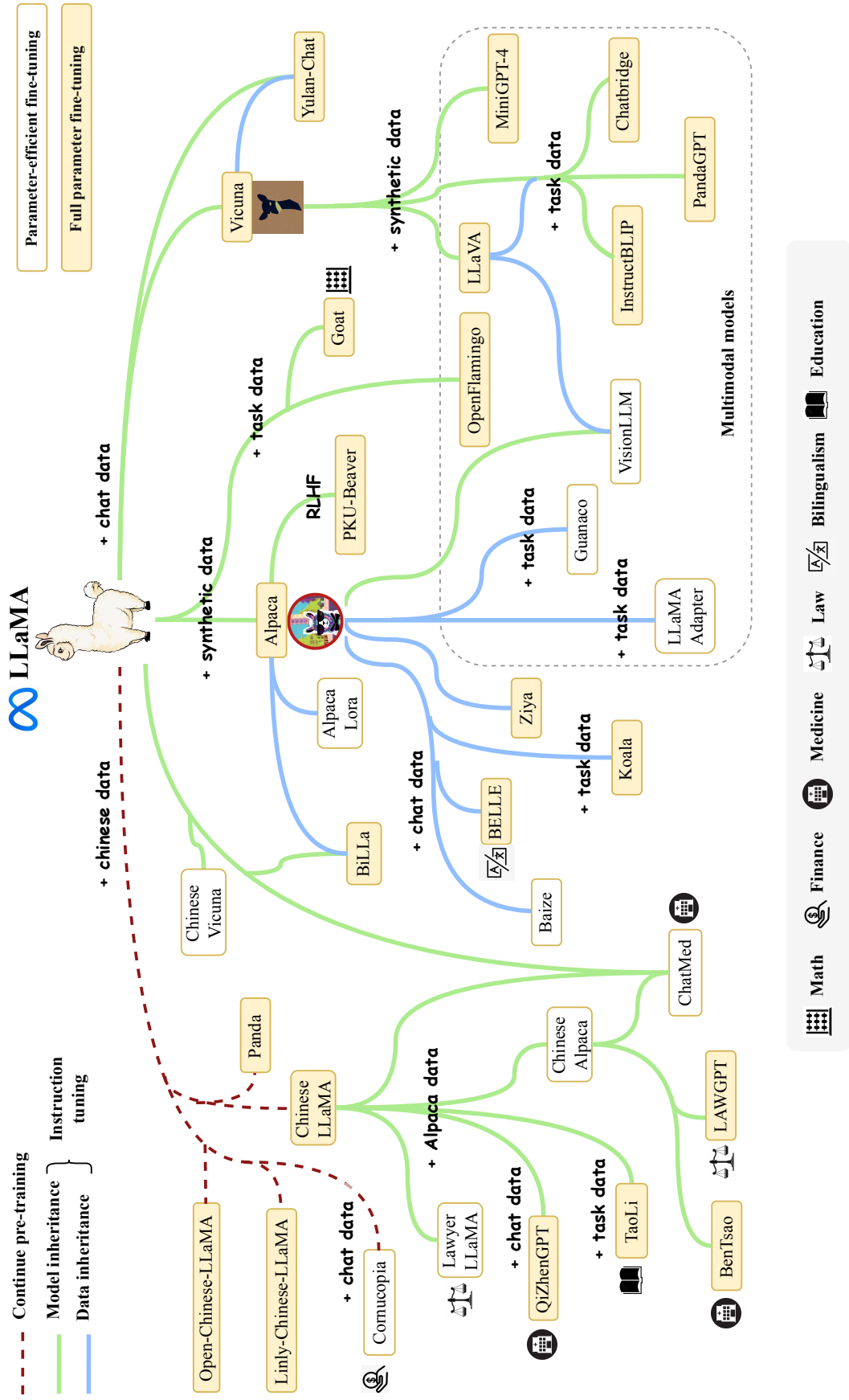
³⁰<https://ai.meta.com/about/>

³¹<https://ai.meta.com/blog/future-of-ai-built-with-llama/>

³²převzato z publikace [32]

³³https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

³⁴více typů vstupů a výstupů - typicky například zvuk a obraz



Obrázek 2.2: Llama a některé modely na ní založené³²

uvažování nad obrázky³⁵. Ve druhé větvi 1B a 3B modely pracující pouze s textem³⁶. Licence k multimodálním modelům z tohoto vydání není udělena osobám a organizacím, které sídlí v Evropské unii³⁷. Jedním z možných vysvětlení jsou regulace jako tzv. AI Act³⁸, ale to není oficiálně potvrzeno.

Llama 3.3 Vydání 3.3 je velmi zaměřené. Součástí je jediný, čistě textový, model o 70 miliardách parametrů laděný pro následování instrukcí³⁹. V některých úkolech přesahuje i výsledky významně většího 3.1 405B.

Llama 4 Poslední vydání⁴⁰, z dubna 2025, přináší spoustu inovací do rodiny Llama. Jedná se o první modely série, které jsou založeny na architektuře mixture-of-experts (viz sekci 2.2). Aktuálně dostupné jsou Llama 4 Scout, složený z 16 expertů a celkem 109B parametrů, a Llama 4 Maverick o 128 expertech a 400 miliardách parametrů. V obou modelech je v jednu chvíli aktivních pouze 17 miliard parametrů. Větší z nich překonává podle autorů GPT-4o a disponuje impresivním kontextem milion tokenů. Menší Scout zvládá až 10 milionů tokenů, což je téměř o dva řády více, než Llama 3.

Tyto modely jsou vytvořeny destilací z většího Llama 4 Behemoth, ten obsahuje celkem 2T parametrů, 288B z nich aktivních a podle autorů překonává GPT-4.5. K jeho uveřejnění ale ještě nedošlo.

Celá skupina disponuje předtrénováním pro 200 jazyků, 100 z nich o více než 1 miliardě učicích tokenů. Seznam oficiálně podporovaných jazyků se rozrostl na 12⁴¹.

Podmínky užití zakazují používání těchto modelů pro škodlivé účely a také pro osoby a společnosti v Evropské unii⁴².

Llama v českém průmyslu Otevřenost modelů Llama jim přinesla popularitu ve výzkumném i obchodním prostředí [32]. Vzhledem k jejich vícejazyčnosti není divu, že se Llamy chytil i český průmysl.

Česká internetová jednička Seznam.cz na konci roku 2024 představila vlastní model SeLLMa⁴³. Ten v té době vycházel z Llama 3.1 70B⁴⁴. Pro účely dalšího učení pro ještě lepší češtinu investovala společnost do více než stovky akceleratorů Nvidia H100. Do celého projektu ladění AI a integrace do produktů jsou zapojeny desítky lidí s celkovou investicí kolem 100 milionů korun.

O zapojení umělé inteligence se snaží i Rohlík.cz. Společnost integruje umělou inteligenci do každodenního rozhodování a nabírá přes 50 expertů, například na tzv. prompt enginee-

³⁵https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md

³⁶https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md

³⁷https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/USE_POLICY.md

³⁸<https://medium.com/techwisdom-io/why-does-meta-restrict-the-usage-of-llama3-2-in-the-eu-4079946abb07>

³⁹https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

⁴⁰<https://ai.meta.com/blog/llama-4-multimodal-intelligence/>

⁴¹https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md

⁴²https://github.com/meta-llama/llama-models/blob/main/models/llama4/USE_POLICY.md

⁴³<https://blog.seznam.cz/2024/10/diana-hlavacova-sellma-aneb-jak-v-seznamu-krotime-drave-jazykove-modely/>

⁴⁴<https://blog.seznam.cz/2025/01/seznam-cz-nabizi-na-sve-domovske-strance-shrnuti-clanku-s-vyuzitim-vlastnich-ai-technologii/>

ring⁴⁵. Na čem stojí Rohlík dnes oficiálně nevíme, v minulosti ale společnost využívala modely ze série Llama⁴⁶.

DeepSeek

V lednu 2025 čínský start-up DeepSeek otřásl světem umělé inteligence. Uvedl R1 – chain-of-thought model (viz sekci 2.4) – založený na dřívějším V3, který dosahuje srovnatelných výsledků s OpenAI o1 [10]. Přitom učení podle autorů stálo méně než 6 milionů dolarů, zhruba 1/50 běžného nákladu pro srovnatelné modely [18]. Model je optimalizován pouze pro čínštinu a angličtinu [10]. Výsledný model (671B) i destilované verze (1.5B až 70B) byly publikovány pod MIT licencí [10]. Efektivita modelu vzbudila otázky ohledně tradičního přístupu hrubou silou od společností jako OpenAI a vedla k poklesu tržní hodnoty společnosti Nvidia o 17% [18].

Google

Společnost Google sehrála významnou roli v raném vývoji velkých jazykových modelů, jak je popsáno v úvodu této kapitoly. Od roku 2024 publikuje otevřené modely série Gemma, které jsou založeny na proprietárním Gemini [19]. V březnu 2025 společnost publikovala verzi Gemma 3 [13]. Ta je zde relevantní, protože slibuje out-of-the-box podporu pro více než 35 jazyků⁴⁷, i když není publikováno, o které se jedná.

2.7 Nasazení jazykových modelů

Open-source či open-weight modely lze provozovat na vlastním hardwaru. Vývojáři ale nedodávají .exe soubor ani nic podobného.

Pro vývoj a ladění všemožných systémů strojového učení existují populární knihovny jako PyTorch a Tensorflow [21]. Pro používání předtrénovaných velkých jazykových modelů jsou populární knihovny s vyšší mírou abstrakce, jako Hugging Face Transformers⁴⁸. Tato knihovna stojí nad technologiemi jako PyTorch, Tensorflow a poskytuje uživatelsky přívětivější rozhraní pro využívání předtrénovaných modelů. Další populární knihovnou je llama.cpp⁴⁹, která se soustředí na efektivní inferenci v C++.

Dalším krokem k vyšší abstrakci jsou softwary jako Ollama⁵⁰, který stojí na llama.cpp. Toto už je hotová aplikace, kterou uživatel nainstaluje jako binární soubor. Následně má k dispozici širokou škálu nejlepších modelů z předpřipravené knihovny přes CLI nebo REST API. S Ollama a jejím API jsou pak kompatibilní desítky webových i nativních grafických uživatelských rozhraní, které poskytují uživatelský zážitek bližší produktům jako ChatGPT.

Pro provoz malých modelů v řádu jednotek miliard parametrů je minimálním požadavkem libovolný počítač s 8 GB RAM. Pro větší modely diskutované v této práci je ale standardem provoz na grafických kartách [21]. Ollama podporuje akceleraci na kartách od společností Nvidia, AMD Radeon i Apple⁵¹.

⁴⁵<https://www.rohlik.fund/aktuality/rohlik-zapoji-naplno-ai-rozhodovat-bude-na-zaklade-algoritmu>

⁴⁶<https://www.lupa.cz/aktuality/rohlik-zacal-nasazovat-generativni-ai-na-trenovani-nepouziva-cloud-ale-server-za-tri-miliony/>

⁴⁷<https://blog.google/technology/developers/gemma-3/>

⁴⁸<https://huggingface.co/>

⁴⁹<https://github.com/ggml-org/llama.cpp>

⁵⁰<https://ollama.com/>

⁵¹<https://github.com/ollama/ollama/blob/main/docs/gpu.md>

Kapitola 3

Získávání informací a jazykové modely

To, že jazykové modely umí produkovat věrohodně působící text, ještě neznamena, že takový text obsahuje skutečně hodnotné informace. Velké jazykové modely jsou pro vyhledávání informací omezené [11]. V procesu předučení získávají znalosti světa do své parametrické paměti (viz kapitolu 2.3), pro některé aplikace je ale vhodné k nim připojit i jinou paměť [15].

Tato kapitola je tedy zaměřená na vyhledávání informací (information retrieval) systémy s využitím jazykových modelů. Vzhledem k povaze a očekávanému rozsahu této práce se nejedná o encyklopedický přehled.

Metoda „retrieval-augmented generation“

Tuto techniku, někdy překládanou jako generování rozšířené o načtení¹, uvádí publikace [15] z dílny Meta AI (v době publikace ještě Facebook AI Research). Cílem této architektury je připojit k paměti parametrické i paměť neparаметrickou. Ta v případě publikace spočívá v indexu Wikipedie a neuronové síť pro sémantické vyhledávání v něm. Po vyhledání relevantních úryvků je odpověď vytvořena modelem BART (400M parametrů). Za pomoci tohoto přístupu a společného dotrénování modelů pro získávání úryvků a tvorbu odpovědi dosahuje publikace state-of-the-art (nejlepších dosažených) výsledků na 3 různých testech pro systémy na odpovídání otevřených dotazů. Zároveň zjišťuje, že model méně halucinuje.

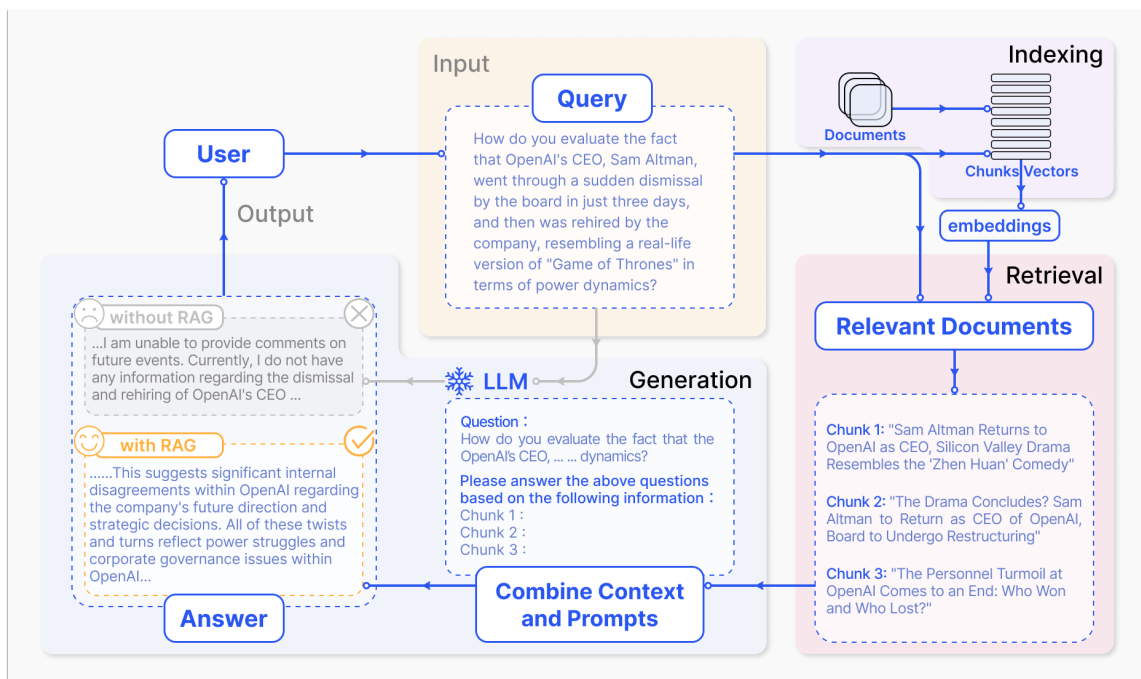
Triviální RAG (retrieval-augmented generation) postup vypadá následovně. Vstupní texty jsou po vhodně velkých kusech zpracovány modelem a vznikají vektory (embeddingy), které reprezentují význam daného úseku textu [8]. Když přijde dotaz, je také vytvořena jeho vektorová reprezentace. Vektory sobě blízké jsou obvykle texty podobného významu, proto se zpravidla vybere K nejbližších textů a ty se použijí jako kontext a spolu s dotazem se stávají vstupem do velkého jazykového modelu, který generuje odpověď. Tento zjednodušený princip je k vidění i v obrázku 3.1.

Z pohledu jazykového modelu je RAG příkladem in-context learning (viz kapitolu 2.4).

Získané dokumenty se stávají součástí promptu pro jazykový model spolu se systémovými instrukcemi, které by měli odkazovat na použití dodaného kontextu, a dotazem od

¹<https://azure.microsoft.com/cs-cz/resources/cloud-computing-dictionary/what-is-retrieval-augmented-generation-rag>

²převzato z publikace [8]



Obrázek 3.1: Obrázek triviálního „retrieval-augmented generation“ systému²

uživatele [32]. V případě protichůdných či irelevantních informací v kontextu může být kvalita odpovědi ovlivněna.

Konverzační „retrieval-augmented generation“

Při použití této architektury u dialogového (chat) modelu je nutno aktualizovat načtený kontext pro nový dotaz. Ten ale zároveň může odkazovat na dotazy a odpovědi předchozí. Proto je v konverzačních RAG systémech vložen krok kontextualizace dotazu, kdy je dotaz velkým jazykovým modelem přepsán na dotaz samostatně pochopitelný, který se pak použije pro vyhledávání nové sady úryvků³. Tento krok může samozřejmě být přeskočen u dotazu prvního.

Alternativou tohoto přístupu je využití samotného velkého jazykového modelu pro rozhodnutí, zda je načtení nového kontextu vůbec nutné.

3.1 Text se znalostmi a jeho dělení

V praxi mohou být dokumenty, ze kterých chceme čerpat znalosti, různých formátů od webů (HTML) po PDF [8]. I když dnes existují multi-modální jazykové modely a RAG systémy (například [31]), nejčastější je reprezentace informací v textu [8]. Základem je tedy převod dokumentů se znalostmi z jejich původních formátů do prostého textu.

Protože dokumenty mohou být rozsáhlé a kontextová okna (viz kapitolu 2.2) jazykových modelů omezené, dělení dokumentů na menší kusy se stalo nedílnou součástí RAG systémů [11]. Nevhodně zvolená velikost segmentů může vést k poškození významové celistvosti i relevance jednotlivých úryvků.

³https://python.langchain.com/v0.2/docs/tutorials/qa_chat_history/

Při výběru velikosti je nutno zvážit nejen potřeby modelu generativního, ale i vyhledávacího, o čemž bude řeč v následující kapitole.

Nejčastější je technika dělení na pevně velké segmenty, například každých 100 znaků [11], s případných překryvem. Tento přístup ale špatně zachovává strukturu původních dokumentů. Vylepšením tohoto přístupu jsou různé rekurzivní algoritmy, které se snaží zachovat celky struktury původního dokumentu.

Nejpokročilejší postupy dělení užívají jazykových modelů pro oddělení významových celků nebo shrnutí delších sekcí pro pochopení na vyšší úrovni [11].

3.2 Sémantické vyhledávání

Jakmile jsou kusy vhodně velké, mohou být zpracovány modelem vyhledávacím. Ten z textu vytvoří vektory (tzv. embeddings), které jsou uloženy ve vektorové databázi [11]. Vektorové databáze nejsou nic nového a dokáží efektivně vyhledávat vektory na základě podobnosti. Pro AI aplikace existují vhodné open-source řešení, jako například Chroma⁴.

Tento model musí být vhodně zvolený pro obsažené informace, aby správně dokázal identifikovat významovou podobnost [11].

Jakmile je zpracováván dotaz, je ve fázi obvykle označované jako retrieval (vyhledání) převeden do vektorové reprezentace pomocí stejného modelu, použitého při indexaci znalostí [8]. Potom jsou nalezeny dokumenty s nejpodobnějšími vektory a jistý počet z nich je vybrán.

V případě zaměření systému na obecně známá témata, je vhodné zvážit nutnost použití externích znalostí pro každý dotaz, protože zavedení dokumentů do kontextu, může být v některých jednoduchých dotazech kontraproduktivní [11]. Toto může být rozhodnuto heuristikou nebo samotným velkým jazykovým modelem.

Embedding modely a sémantické vyhledávání nejsou nic unikátního pro RAG [2]. Sémantické vyhledávání je používáno například v internetových vyhledávačích. Embedding modely mají své užití i v mnoha tradičních úlohách zpracování přirozeného jazyka, jako je analýza sentimentu nebo překlad textu.

České modely pro vektorové reprezentace

Většina výzkumu v těchto modelech je přirozeně zaměřena na angličtinu [2]. O kvalitní embedding model pro češtinu ale má zájem například vyhledávač Seznam.cz, který tyto modely využívá pro vylepšení organického vyhledávání i vybrané úryvky přímo na stránce výsledků.

Pro účely vývoje těchto modelů vytvořili v Seznamu dataset DaReCzech [14]. Ten obsahuje přes milion párů dotaz-dokument a je založen na skutečných uživatelských požadavcích s úpravou překlepů. Průměrný dotaz má o trochu méně než 3 slova a ani kvantil 0.75 nepřesáhne 4. Dokumenty jsou charakterizovány svým URL, názvem, obsahem a relevancí. Publikace [14] nezmiňuje užití v RAG.

Ve stejné publikaci [14] jako dataset DaReCzech je také publikován model Small-E-Czech. Model je založen na Electra-small a disponuje 13 miliony parametrů. V roce 2021 byl zapojen do vyhledávače Seznam.cz [2]. Seznam inferenci provádí pomocí AVX512 na CPU.

V publikaci [2] uvádí čtveřice autorů ze společnosti Seznam několik modelů, které navazují na předchozí Small-E-Czech a překonávají jeho výsledky na datasetu DaReCzech (viz

⁴<https://www.trychroma.com/>

tabulku 3.1). Tyto modely byly následně publikovány pod Creative Commons licencemi. V jejich vlastní evaluaci dosahují na DaReCzech lepších výsledků než proprietární OpenAI Ada.

Model	Spearman's correlation				Accuracy	P@10
	SVOB-IMG	SVOB-HL	CNA	STS-Average	Costra	DaReCzech
Random baseline	2.40	3.85	29.09	11.78	49.54	38.10 ± 0.31
Random-small	64.61	55.30	67.69	62.53	68.73	40.38 ± 0.35
Avg. fastText	54.81	47.52	72.47	58.26	65.75	37.88 ± 0.31
Small-E-Czech	39.67	42.43	62.80	48.30	64.29	37.31 ± 0.33
RetroMAE-Small	78.88	66.21	83.82	76.30	69.66	42.16 ± 0.36
Dist-MPNet-ParaCrawl	90.11	77.66	84.99	84.25	70.42	42.33 ± 0.32
Dist-MPNet-CzEng	90.94	83.89	87.97	87.60	71.22	42.01 ± 0.37
SimCSE-Small-E-Czech	61.70	59.75	77.26	66.24	66.44	39.20 ± 0.38
TSDAE-Small-E-Czech	77.45	66.17	83.16	75.59	69.42	40.54 ± 0.37
SimCSE-RetroMAE	78.88	71.92	85.19	78.66	69.63	42.04 ± 0.37
RankCSE-RetroMAE	79.91	72.03	85.10	79.01	69.79	41.97 ± 0.36
InfoCSE-RetroMAE	79.30	65.58	84.31	76.40	69.89	41.77 ± 0.38
SimCSE-Dist-MPNet-ParaCrawl	90.29	78.80	85.91	85.00	71.12	42.38 ± 0.35
SimCSE-Dist-MPNet-CzEng	90.73	84.22	88.56	87.83	71.77	42.18 ± 0.38
OpenAI Ada Embedding	83.51	78.04	86.21	82.59	69.01	42.21 ± 0.31

Tabulka 3.1: Nulové (zero-shot) vyhodnocení modelů z publikace Bednář a spol. a několika základních modelů na úlohách STS, Costra a DaReCzech. Tabulka je rozdělena do čtyř horizontálních sekcí: základní modely, předtrénované modely, modely doladěné pro embeddingy a externí služby.⁶

OpenAI také vyvíjí embedding modely, lze je ale používat pouze jako službu jejich API⁷. V aktuální nabídce v dubnu 2025 je text-embedding-ada-002, jenž ve vyhodnocení publikace [2] prohrává oproti vlastním modelům Seznamu na českém DaReCzech. K dispozici jsou i novější modely text-embedding-3-small a text-embedding-3-large z ledna loňského roku, pro ty ale není aktuální české srovnání.

Srovnáním embedding modelů na datasetu DaReCzech se zabývá i loňská publikace [25] z FIT VUT. Zde nejlepších výsledků dosahuje BGE-Multilingual-Gemma2⁸, model o 9 miliardách parametrů a tedy o 2 řády větší než největší modely od skupiny v Seznam.cz. Tato publikace je zaměřena na širokou škálu velikostí modelů, zatímco dříve diskutovaná činnost v Seznam.cz [14] [2] se soustředila spíše na malé modely o desítkách až stovkách milionů parametrů. Obě publikace obsahují modely text-embedding-ada-002 a SimCSE-Dist-MPNet-ParaCrawl, ale výsledky nejsou přímo srovnatelné.

„Reranking“

Některé architektury sémantického vyhledávání využívají více úrovní vyhledávání [17] [7]. Ve vícekrokovém systému je úkolem retrieveru získat několik nejlepších výsledků velmi efektivně. Následně je úkolem rerankeru (přeskládavače) je dále seřadit více přesným, ale obvykle více výpočetně náročným procesem [17]. Cílem je umístit nejrelevantnější úryvky na okraje promptu [8], kde mají největší vliv na generovaný text [16].

⁶Převzato z literatury [2].

⁷<https://platform.openai.com/docs/guides/embeddings>

⁸<https://huggingface.co/BAAI/bge-multilingual-gemma2>

3.3 Aktuální produkční systémy

Dialogové systémy jako ChatGPT nejsou jen velké jazykové modely.

Tři měsíce po spuštění ChatGPT OpenAI také uvedlo rozšíření (tzv. plugins)⁹. Podle tiskové zprávy se jedná o „*vznik komunity utvářející paradigma interakce lidí a umělé inteligence*“. Tento systém má sloužit jako „*oči a uši ChatGPT*“ a přinášet informace, které jsou příliš aktuální, osobní nebo specifické na to, aby byly zahrnuty při trénování modelu. Mezi počátečními rozšířeními byly vyhledávače hotelů a restaurací i možnost používat vyhledávač a omezeně surfovat po internetu.

V listopadu stejného roku (2023, 11 měsíců od startu) byla celá logika rozšíření přeformulována do tzv. GPTs, to jsou specializované verze ChatGPT, které může vytvořit kdokoliv, bez žádného programování¹⁰. Stačí přidat instrukce, dodatečné informace a vybrat, které rozšiřující funkce (generace obrázku, hledání na webu, analýza dat) má tento specializovaný nástroj využívat.

Ve stejný den vznikl také systém tzv. Assistants¹¹. Asistenty vytváří vývojáři pro své aplikace užívající OpenAI API. Pro tyto asistenty dokáže OpenAI uchovávat a prohledávat znalosti v RAG systému nazvaném Retrieval. Assistants nejsou jen ChatGPT s RAG znalostmi, mají i jisté vlastnosti agentního systému, jako volání externích funkcí.

O rok později přišel ChatGPT Search¹², ikona glóbusu v okně zadání ChatGPT, která prohledává internet a přináší aktuální informace. Do tohoto režimu je možno se přepnout manuálně nebo může být zvolen automaticky. Odpověď následně obsahuje i odkazy na použité zdroje.

Tato funkce zdaleka nebyla první svého druhu, Bing¹³ nabízel podobné funkce již výrazně dřív.

Otevřené systémy

Populární webové uživatelské rozhraní OpenWebUI¹⁴ také poskytuje plug-and-play RAG systém¹⁵ využívající embedding modely od OpenAI nebo podporované softwarem Ollama. Systém nenabízí příliš rozsáhlou konfiguraci.

Pro větší flexibilitu v tvorbě RAG a agentních systémů vznikají rámce jako LangChain¹⁶. To je open-source knihovna pro Python a JavaScript, která poskytuje rozhraní pro všemožné vektorové databáze a velké jazykové modely a abstrakci nad nimi. Princip spočívá, jak již název napovídá, v tvoření řetězů¹⁷ funkcí a nástrojů, které následně poskytují rozhraní pro jejich běh, ať už jednorázově nebo s nativní podporou streamování tokenů. Tyto objekty (tzv. runnables a chains) pak mohou být nasazeny do produkce pomocí knihovny LangServe¹⁸.

⁹<https://openai.com/index/chatgpt-plugins/>

¹⁰<https://openai.com/index/introducing-gpts/>

¹¹<https://openai.com/index/new-models-and-developer-products-announced-at-devday/>

¹²<https://openai.com/index/introducing-chatgpt-search/>

¹³<https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>

¹⁴<https://openwebui.com/>

¹⁵<https://docs.openwebui.com/features/rag>

¹⁶<https://www.langchain.com/langchain>

¹⁷<https://python.langchain.com/docs/concepts/lcel/>

¹⁸<https://python.langchain.com/docs/langserve/>

Kapitola 4

Zhodnocení současného stavu a specifikace řešení

Cílem této kapitoly je vyhodnotit současný stav ve zmíněných technických oblastech a představit požadavky pro vytvářenou aplikaci dialogových systémů.

4.1 Zhodnocení současného stavu

Vzhledem k povaze diskutovaného systému bude zvlášť vyhodnocena oblast základních modelů a systémů jejich aplikace.

Zhodnocení stavu velkých jazykových modelů

Vývojem základních modelů (foundational models) se zabývá pouze relativně málo subjektů, i když stále více než se vešlo do kapitoly 2 této práce. Zde jsou hlavně zmíněny ty relevantní pro prostředí českého jazyka. Z otevřených modelů na vícejazyčnost cílí Llama 3, což potvrzuje i popularita u českých firem (viz sekci 2.6).

Nový model Gemma 3, který byl publikován až v průběhu této práce, ale slibuje 35 „out-of-the-box“ jazyků, oproti pouhým osmi pro Llama 3. V předtrénování oba modely používají datasety s více než stovkou jazyků, dále se soustředí jen na tyto užší výběry. Oficiálně nevíme, o kterých 35 jazyků se jedná a tedy jestli je čeština mezi nimi. Je otázkou, zda by toto lepší doladění mohlo dohnat náskok 70 miliardové Llama 3.3, i když Gemma 3 je k dispozici jako maximálně 27B model. Detailnější srovnání těchto dvou modelů je uvedeno v tabulce 4.1.

Model	Llama 3.3	Gemma 3	DeepSeek-R1
Společnost	Meta	Google	DeepSeek
Měsíc vydání	prosinec 2024	březen 2025	leden 2025
Počet parametrů	70B	1-27B	1.5-671B
Maximální délka kontextu	128k	128k	128k
Jazyků v předučení	176	140	?
Jazyků v doladění	8	35	2

Tabulka 4.1: Srovnání zvažovaných modelů

Vzhledem k tomu, že oba modely jsou předučeny na výrazně větším počtu jazyků, nabízí se možnost provést vlastní doladění pro jeden z těchto jazyků. Toto je principem projektu SeLLMa společnosti Seznam, jak je zmíněno v kapitole 2.6. I když doladění modelu je obvykle méně náročné než jeho předučení, v Seznamu na tvorbě modelu vyladěného pro perfektní češtinu stále pracují desítky lidí a H100 GPU, což je daleko mimo možnosti této bakalářské práce.

Závěrem je tedy možno říct, že existují velké jazykové modely, na kterých lze dialogový systém založit, ale z publikované literatury nelze jasně určit, který model dosahuje nejlepších výsledků v českém jazyce, nýbrž žádný z nich to explicitně nespecifikuje. Dalšímu výběru jsou věnovány experimenty v následující kapitole.

Zhodnocení stavu „retrieval-augmented generation“ systémů

Rychlý náhled do rejstříku posledního kola Y Combinator¹ ukazuje, že polovina odstartovaných společností pracuje na specializaci AI chatbotů pro nějakou specifickou oblast.

Velké jazykové modely jsou dnes široce dostupné, své znalosti ale získávají v nákladné prvotní fázi učení a tedy disponují pouze zastaralým, obecným přehledem o světě. Zároveň mají tendence halucinovat a vytvářet nepravdy. Technika retrieval-augmented generation perfektně doplňuje tyto mezery.

Retrieval-augmented generation systémy se stávají velmi populární aplikací AI na specifické domény díky jejich jednoduchosti a flexibilitě, to potvrzuje i počet uživatelů knihovny LangChain, který přesáhl 1 milion².

Vzhledem k současnému stavu je metoda retrieval-augmented generation vhodným doplňkem nedostatků velkých jazykových modelů.

Na trhu existují komerční produkty na základě tohoto principu, například od OpenAI, ale jejich využití může být nákladné.

4.2 Implementační cíl práce

Zadáním této práce je vytvořit dialogového pomocníka pro software zpracování fotografií. Vybraným programem je Zoner Photo Studio, fotografický program z Brna s 20 letou tradicí, který se v roce 2002 stal nejpobulárnějším programem svého druhu v České republice a dnes se může pyšnit 100 000 uživateli ve Spojených státech³.

K programu je k dispozici online manuál⁴. Znalosti o programu, tipy a triky také šíří oficiální blog Milujeme fotografii⁵. Společnost pro program také provozuje kontaktní formulář, email a FAQ pro uživatelskou podporu⁶.

Cílem této práce je tedy vytvořit systém, který bude schopen pomoci lidem obsluhovat pokročilý fotografický editor Zoner Photo Studio.

Mezi aplikační cíle projektu patří zpřístupnění softwaru širší klientele a jednodušší osvojení nových funkcí programu pro uživatele stávající.

Technickým cílem je vytvořit řešení na základě open-source technologií, jako alternativu ke komerčním produktům.

¹<https://www.ycombinator.com/companies?batch=X25>

²<https://www.langchain.com/langchain>

³<https://www.zoner.com/en/about-us>

⁴<https://manual.zoner.com>

⁵<https://www.milujemefotografie.cz/>

⁶<https://www.zoner.cz/podpora>

4.3 Specifikace požadovaného výstupu

Celkově ze zadání bakalářské práce, zhodnocení současného stavu a požadavků z praxe vyplývají následující vlastnosti.

- Systém bude věrohodně odpovídat na dotazy.
- Systém bude založený na otevřených modelech.
- Systém bude zaměřený na češtinu.
- Systém by měl být dobře udržovatelný.
- Změny v nových verzích programu se musí co nejjednodušeji projevit i v AI pomocníku.
- Pro inferenci velkých jazykových modelů bude schopen využít standardní infrastrukturu pomocí běžných API.

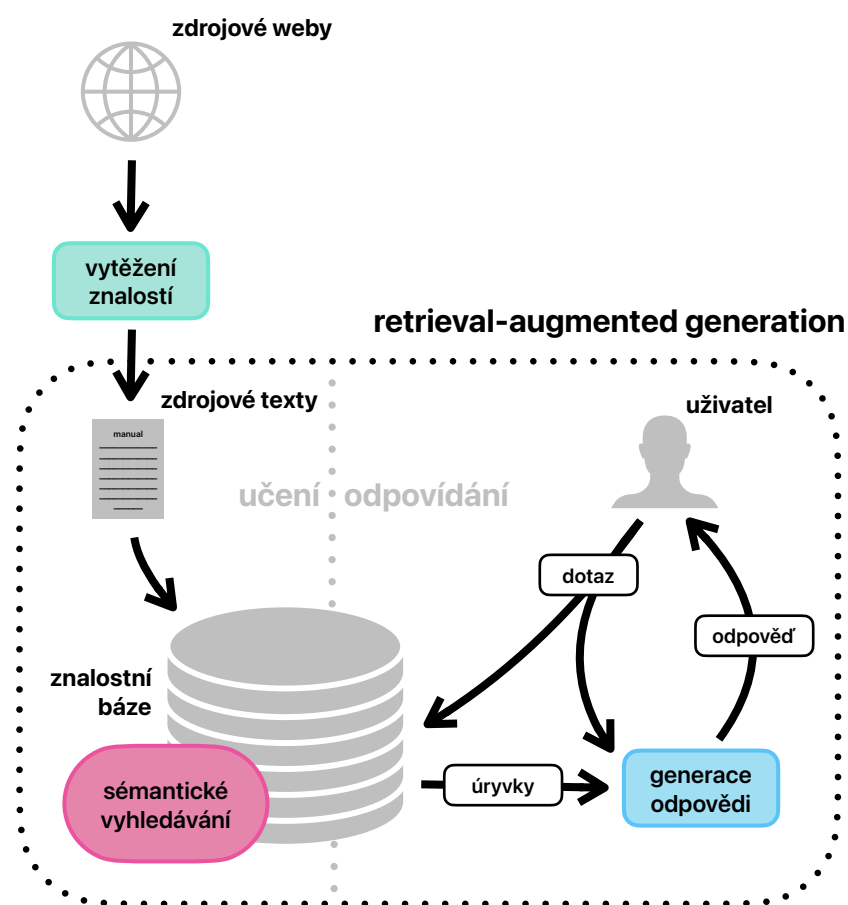
Implementaci podle stanovených cílů se věnuje následující kapitola.

Kapitola 5

Implementace systému

Tato kapitola představuje způsob provedení, technické vlastnosti a dosažené výsledky odevzdaného řešení.

S ohledem na očekávané specifikace se nabízí využití retrieval-augmented generation architektury. Základní pracovní princip systému je rozdělen na tři části představené v obrázku 5.1.



Obrázek 5.1: Rozdělení systému do tří základních částí

Vytěžení znalostí Tento blok zde popisuje prohledání webových zdrojů a vznik strukturovaných Markdown textů pro jednotlivé dokumenty. Tomu se blíže věnuje následující sekce.

Znalostní báze se sémantickým vyhledáváním Tyto texty jsou pak dále děleny na úryvky a zařazeny do znalostní báze organizované pro použité sémantické vyhledávání, tomu se věnuje sekce 5.2.

Generace odpovědi Jakmile jsou pro daný dotaz nalezeny relevantní úryvky, velký jazykový model vytvoří informovanou odpověď, viz sekci 5.3.

Většina tohoto systému, s výjimkou skriptů pro stažení znalostí z webu, běží v rámci rámce (frameworku) LangChain¹ pro Python. Ten poskytuje abstrakce pro API velkých jazykových modelů, vektorových databází a další užitečné funkce. Tyto rozhraní komunikují po síti a tedy je možné, aby například jeden RAG (retrieval-augmented generation) aplikační server rozložil zátěž generování odpovědí na více inferenčních serverů.

5.1 Implementace vytěžení znalostí

Pro získání čistého formátovaného Markdown z online manuálu, blogu a FAQ je použitý Python skript využívající webový crawler (prohledávač) a scraper (extraktor) Crawl4AI². Tento nástroj je velmi populární, repozitář disponuje více než 40 000 hvězdičkami a získal i ocenění nejpopulárnějšího projektu dne na portálu GitHub. Jeho poslední verze si sama poradí i s prozkoumáním celého webu a vyfiltrováním smysluplných informací. V tomto projektu slouží ve své původní funkci a při zavolání na danou URL vrací naformátovaný Markdown výstup vhodný pro velké jazykové modely.

Po krátkém experimentování je pro oba rozsáhlejší weby využitý iterativní průchod. Vzhledem k formátu adres jednotlivých dokumentů se ukázal jako jednodušší, oproti variantě rekurzivní, která by vyžadovala více heuristik určujících, které odkazy má program následovat.

Prvně je načten obsah manuálu, případně archiv blogu, a z každé strany získány všechny odkazy, které Crawl4AI zpřístupňuje ve strukturované podobě. Jakmile je získán seznam všech článků, jsou procházeny jednotlivě a z každého z nich stažen Markdown pouze hlavního těla dokumentu, na základě CSS selektoru této oblasti zdroje. Některé návody na webu milujemefotografii.cz poskytují evergreen informace o základních principech fotografie, jiné ale mají jasnou expiraci. V archivu se tedy nachází i články, které již nejsou aktuální, ty jsou redakcí označeny typizovaným varováním. Pokud je toto varování identifikováno na stránce, bude přeskočena a tyto informace se nedostanou do znalostní báze.

Výsledný Markdown není jen spojen za sebe, výstupem z extrakčních skriptů jsou soubory formátu JSON, které obsahují slovníkovou strukturu, kde klíčem jsou adresy a obsahem řetězec Markdownu. Tento přístup umožňuje snadnější dohledání případných chyb v pozdějších stádiích. Vzhledem k malému množství obrazových informací ve zdrojových webech a přítomnosti řádných popisů těchto obrázků zde není nijak řešen převod obrázků na text.

¹<https://www.langchain.com/langchain>

²<https://github.com/unclecode/crawl4ai>

5.2 Implementace znalostní báze

Pro správné vysvětlení fungování znalostní báze je zde uveden detailnější obrázek fungování retrieval-augmented generation systému 5.2. Zde je možno vidět, jak jsou zdrojové texty rozděleny a pro každý úryvek je vytvořen sémantický vektor. Pomocí těchto vektorů se pak ve fázi odpovídání v pravé části obrázku vyhledávají relevantní úryvky. Ty pak generativní model použije pro zodpovězení dotazu, tomu je věnována sekce 5.3.

Je použita open-source vektorová databáze Chroma³, která je kompatibilní se zvoleným frameworkem a běží jako vlastní Docker kontejner. Díky tomu nebylo nutné implementovat vlastní řešení.

Motivace dělení dokumentů

Po provedení předchozího kroku vzniklo bezmála 1300 tzv. dokumentů o maximální délce 26 316 znaků. Na první pohled se může zdát, že dodání co největšího kontextu povede k lépe informované odpovědi, dokumenty v RAG ale mohou být i příliš dlouhé [11]. Výstupní jazykový model by optimisticky potřeboval 6 000 tokenů pouze na načtení jediného dokumentu. To vyžaduje paměť pro uložení tohoto kontextu, výpočetní čas pro načtení jednotlivých tokenů do modelu a díky literatuře víme, že v takto obsáhlém dokumentu se mohou relevantní informace ztratit [16], i když architektury dnešních modelů technicky umožňují je zpracovat.

Zároveň vytvořit sémantický vektor pro takový dokument není jednoduché. Například modely od Seznamu [14] [2] dokáží načíst pouze 128 až 512 tokenů pro vytvoření embeddingu, což je reprezentativní rozmezí i pro modely testované v publikaci z FIT VUT [25] s výjimkou modelu OpenAI Ada, který zvládne přes 8 000 tokenů. Ne vždy je nutné, aby sémantický vektor byl vypočten z celého dokumentu, testování v publikaci [25] bylo založeno na využití začátku dokumentu a odseknutí všeho, co se nevešlo. Vzhledem k obvyklé logické struktuře webových dokumentů není překvapivé, že tento přístup také funguje.

V případě delšího textu o více tématech se embedding logicky stane kompromisem, což není žádoucí. Řešením může být využití více vektorů k jednomu dokumentu, například pro jeho části a generované shrnutí⁴. Tento přístup umožňuje ve velkých dokumentech efektivně vyhledávat, stále ale zůstávají nežádoucí vlastnosti dlouhých dokumentů z prvního odstavce.

Výběr vhodné velikosti úryvků

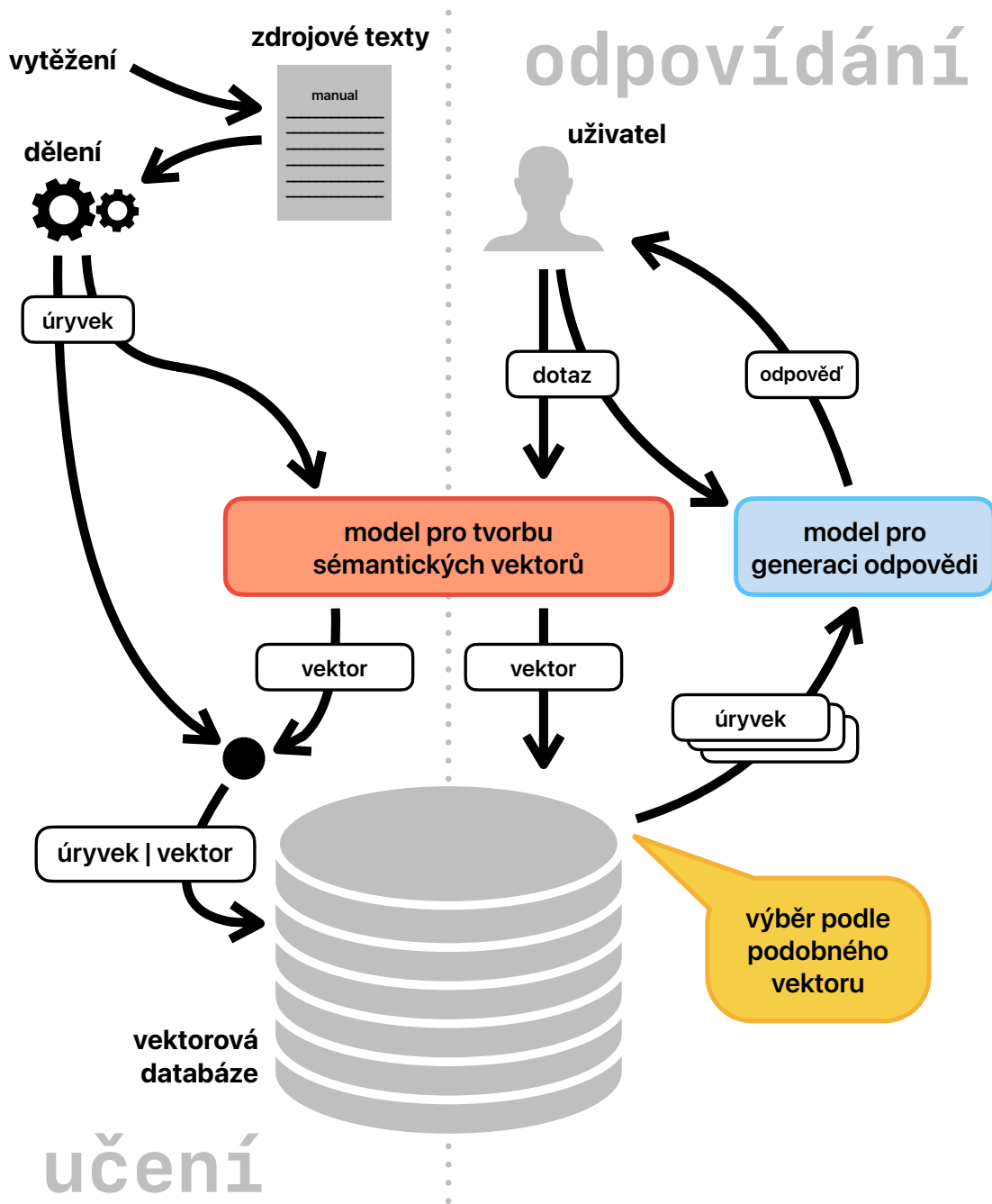
Pro lepší soustředění generativního modelu, menší pravděpodobnost halucinací [11] a rychlejší odezvu systému bylo rozhodnuto dokumenty dále dělit. V rámci řešení bylo experimentálně ověřeno⁵, že všechny uvažované modely dosahují na těchto vstupních datech poměru 4,40 až 4,55 znaků na token. Vzhledem k již zmíněným maximálním délkám kontextu se zdá, že pro zachování celého textu uvnitř vstupu embedding modelu o kontextu 128 tokenů je vhodné použít segmenty o délce 512 znaků. Pro modely s delším kontextem opět čtyřikrát více znaků než podporují tokenů. Vzhledem k variabilitě těchto poměrů na různých textech stále může dojít k odseknutí části textu, zvolený poměr je ale považován za vhodný kompromis.

Jak již bylo zmíněno, některé nasazení (například v literatuře [25]) s použitím pouze části dokumentu počítají, proto bude experimentálně otestována i výkonnost vyhledávání mezi

³<https://www.trychroma.com/>

⁴https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/multi_vector/

⁵viz `bp-rag/A_tokenizer_test.py`



Obrázek 5.2: Obecná „retrieval-augmented generation“ architektura

2048 znaků dlouhými řetězci modely o 128 tokenech kontextu. Pro vytvoření sémantického vektoru tedy bude využita zhruba první čtvrtina textu.

Součástí experimentu bude i využití velkého jazykového modelu pro vytvoření kratšího shrnutí delšího úryvku pro indexaci. Při vybrání dokumentu pak bude načten originální text, toto shrnutí slouží pouze k vytvoření embeddingu většího úseku textu bez odseknutí konce. Použitý prompt je k vidění v tabulce 5.1 spolu s jedním z výsledků. Pro vytvoření shrnutí byl použitý checkpoint `gemma3:27b-it-q8_0` z nabídky softwaru Ollama. O modelu a jeho nasazení bude více v sekci 5.3. Vzhledem ke schopnosti zkrátit i delší text je do experimentu zařazena i varianta s dělením na díly o velikosti 8 192 znaků. Ty jsou pak shrnuty v odstavci, který opět může dosahovat až čtvrtiny jejich délky, i když pro mnoho segmentů je výsledek výrazně kratší. Limity na délku shrnutí nejsou kontrolovány, v případě jejich přesáhnutí model přirozeně vynechá konec textu. Výstupy se ale v drtivé většině velmi dobře drží požadované délky z promptu.

Prosím doplň shrnutí následujícího textu v jednom odstavci do 500 znaků. Není nutné uvádět, že náleží k programu Zoner Photo Studio

Text: *Podporované formáty. Zoner Photo Studio X podporuje čtení těchto formátů: Bitmapové formáty: ZPS – Zoner Photo Studio; PSD, PSB – Adobe Photoshop Image; JPEG – File Interchange Format; HEIF – High Efficiency Image Format; GIF – Compuserve Graphics Interchange Format; TIFF – Tagged Image File Format; PNG – . . .*

Summary: *Zoner Photo Studio X podporuje čtení i zápis široké škály obrazových formátů, včetně bitmap (JPEG, PNG, TIFF, BMP) a RAW formátů (CR2, NEF, DNG, ARW a další). Podporuje i moderní formáty jako AVIF, HDR, JXL a JPEG 2000. Umožňuje prohlížet PSD/PSB soubory s kompatibilitou.*

Tabulka 5.1: Použitý prompt pro shrnutí (**tučně**) a příklad dosazených (vygenerovaných) hodnot (*kurzívou*). Formátování textu upraveno pro čitelnost.

Pro samotné dělení textu je použita rekurzivní implementace z LangChain, která je nastavena tak, aby dělila text pomocí hierarchických prvků jazyka Markdown. Dokumenty jsou postupně děleny posloupností úrovní nadpisů, dokud není dosaženo požadované velikosti. Tato metoda lépe zachovává logickou strukturu textu.

Výběr modelu pro tvorbu sémantických vektorů

Pro tvorbu sémantických vektorů byly zváženy modely publikované v publikaci [2], jmenovitě RetroMAE-Small, Dist-MPNet-ParaCrawl, Dist-MPNet-CzEng a jejich SimCSE verze, plus SimCSE verze staršího Small-E-Czech. Je nutno dodat, že modely trénované na datasetu CzEng (jmenovitě Dist-MPNet-CzEng a jeho SimCSE verze) jsou k dispozici pouze pod licencí CC-BY-NC 4.0 a nemohou tedy být nasazeny do komerčního užití. V experimentu jsou uvedeny pro úplnost, jsou však sázeny kurzívou. Všechny tyto modely produkují vektory o 256 dimenzích.

Zároveň jsou pro srovnání součástí testu modely od OpenAI: starší text-embedding-ada-002, který se objevuje v předchozích publikacích, a nové text-embedding-3-small a text-embedding-3-large. Všechny 3 produkují výrazně rozměrnější vektory než otevřené alternativy, jejich kontextové okno je také delší. Tyto modely jsou k dispozici přes OpenAI API a zpoplatněny za token vstupního textu. Celkový náklad pro tento experiment vyšel na \$ 2.86 + DPH.

Model	Kontext [char]	Úspěšnost [%]				
		Bez shrnutí			Generované shrnutí	
Max. délka úryvků [char]		512	2048	8192	2048	8192
Max. délka shrnutí [char]		-	-	-	500	2000
Počet nalezených úryvků		3	3	3	3	3
dist-mpnet-paracrawl-cs-en	≈ 512	47	47	53	59	65
<i>dist-mpnet-czeng-cs-en</i>	≈ 512	41	35	41	41	53
simcse-dist-mpnet-paracrawl-cs-en	≈ 512	35	47	47	65	71
<i>simcse-dist-mpnet-czeng-cs-en</i>	≈ 512	35	35	35	59	47
simcse-small-e-czech	≈ 512	0	0	6	0	0
retromae-small-cs	≈ 2048	24	18	18	24	41
simcse-retromae-small-cs	≈ 2048	29	24	24	35	35
text-embedding-ada-002		35	41	47	59	47
text-embedding-3-small	8191 tokenů!	65	53	65	59	53
text-embedding-3-large		41	53	65	59	59

Tabulka 5.2: Úspěšnost jednotlivých modelů na různě dlouhých úryvcích při hledání pevného počtu 3 úryvků.

Předmětem testu je zjistit relativní výkonnost modelů na sadě českého textu tohoto typu. Testovací sada není perfektní, je složena z ručně vybraných otázek, jejichž odpověď je v textu. Otázky jsou inspirovány již zmíněnou sadou z přílohy A, častými dotazy na podporu⁶ a očekávaným využitím systému. Toto blíže simuluje očekávané užívání systému s větnými dotazy, oproti datasetu DaReCzech [14], který se soustředí na heslovité zadání do vyhledávače (viz kapitolu 3.2). Výsledkem testu je pak procento dotazů, kde retriever založený na daném modelu získal alespoň jeden z manuálně zadaných textů, které obsahují odpověď. Sada je tedy složena z otázek a přímých citací textu, kde se nachází odpověď. Výsledky jsou k vidění v tabulce 5.2.

Z výsledků experimentu je možno vyvodit hned několik důležitých závěrů. Jedním ze zjištění je, že modely s delším kontextem nejsou zdaleka nejlepší v open-source nabídce. Je možno si všimnout, že pro modely z první sekce jsou výsledky hledání založeného na generovaném shrnutí převážně lepší, to stejně ale nepozorujeme u modelů od OpenAI s výrazně delším kontextem.

Porovnávání výsledků mezi různými velikostmi úryvku je v této tabulce poněkud zavádějící. Pro účely zachování odezvy a spolehlivosti systému není relevantní počet získaných úryvků, ale jejich celková délka. Do této délky je možno umístit více kratších dokumentů. Proto je zde experiment uveden znovu, nyní bez konstantního počtu úryvků, ale omezený na maximální délku 16 384 znaků. To znamená 2 z nejdelších nebo 32 z nejkratších úryvků, ve všech případech ale zhruba 0,2% textu v korpusu. Výsledky jsou uvedeny v tabulce 5.3.

V tomto experimentu můžeme nejvyšší čísla pozorovat u nejmenších úryvků. Nalezení zdroje ale není konec činnosti RAG systému, ty musí ještě zpracovat velký jazykový model. Vzhledem k problémům jako rušení kontextu, efekt lost-in-the-middle a dalším, které jsou blíže diskutovány v kapitolách 2 a 3, není využití 32 různých úryvků o pouze 500 znacích o nic lepší než využití úryvků nevhodně velikých.

Po zvážení požadavků velkého jazykového modelu bylo rozhodnuto využít úryvky o délce 2048 znaků. Nejlepších výsledků v této kategorii dosahuje model simcse-dist-mpnet-paracrawl-cs-en v kombinaci s generovaným shrnutím pro indexaci a proto bude tato konfigurace využita v navrženém systému.

⁶<https://www.milujemefotografii.cz/nejcastejsi-dotazy-podpora-zps-x>

Model	Kontext [char]	Úspěšnost [%]				
		Bez shrnutí			Generované shrnutí	
Max. délka úryvků [char]		512	2048	8192	2048	8192
Max. délka shrnutí [char]		-	-	-	500	2000
Počet nalezených úryvků		32	8	2	8	2
dist-mpnet-paracrawl-cs-en	≈ 512	76	65	47	59	65
<i>dist-mpnet-czeng-cs-en</i>	≈ 512	59	53	53	53	47
simcse-dist-mpnet-paracrawl-cs-en	≈ 512	71	65	47	65	59
<i>simcse-dist-mpnet-czeng-cs-en</i>	≈ 512	65	41	35	59	35
simcse-small-e-czech	≈ 512	0	12	6	0	0
retromae-small-cs	≈ 2048	35	18	18	41	29
simcse-retromae-small-cs	≈ 2048	41	24	24	47	29
text-embedding-ada-002		59	47	47	59	47
text-embedding-3-small	8191 tokenů!	71	59	59	65	47
text-embedding-3-large		71	71	59	71	59

Tabulka 5.3: Úspěšnost jednotlivých modelů na různě dlouhých úryvcích při hledání pevného součtu maximální délky 16 384 znaků.

Pokud není uvedeno jinak, vybírá retriever ve zbytku práce 3 úryvky, jako v prvním experimentu. Jak je možno vidět při srovnání tabulek 5.2 a 5.3, výsledek pro 8 úryvků není o nic lepší.

5.3 Výběr modelu pro generaci odpovědi a jeho zapojení

V tomto systému tedy vystupuje jazykový model ve dvou funkcích. Jako v každém RAG systému generuje finální odpověď pro uživatele, zároveň ale vytváří i shrnutí jednotlivých úryvků, jak již bylo zmíněno v sekci 5.2. Pro tyto dvě funkce je možno použít modely různé, jak je i ukázáno v obrázku 5.3.

Ve skutečnosti má v konverzačním RAG systému velký jazykový model i funkci třetí: kontextualizovat doplňující dotazy, jak je znázorněno šedými šipkami, to ale nenastává u první zprávy, proto je v diagramu oddělen od běžného postupu.

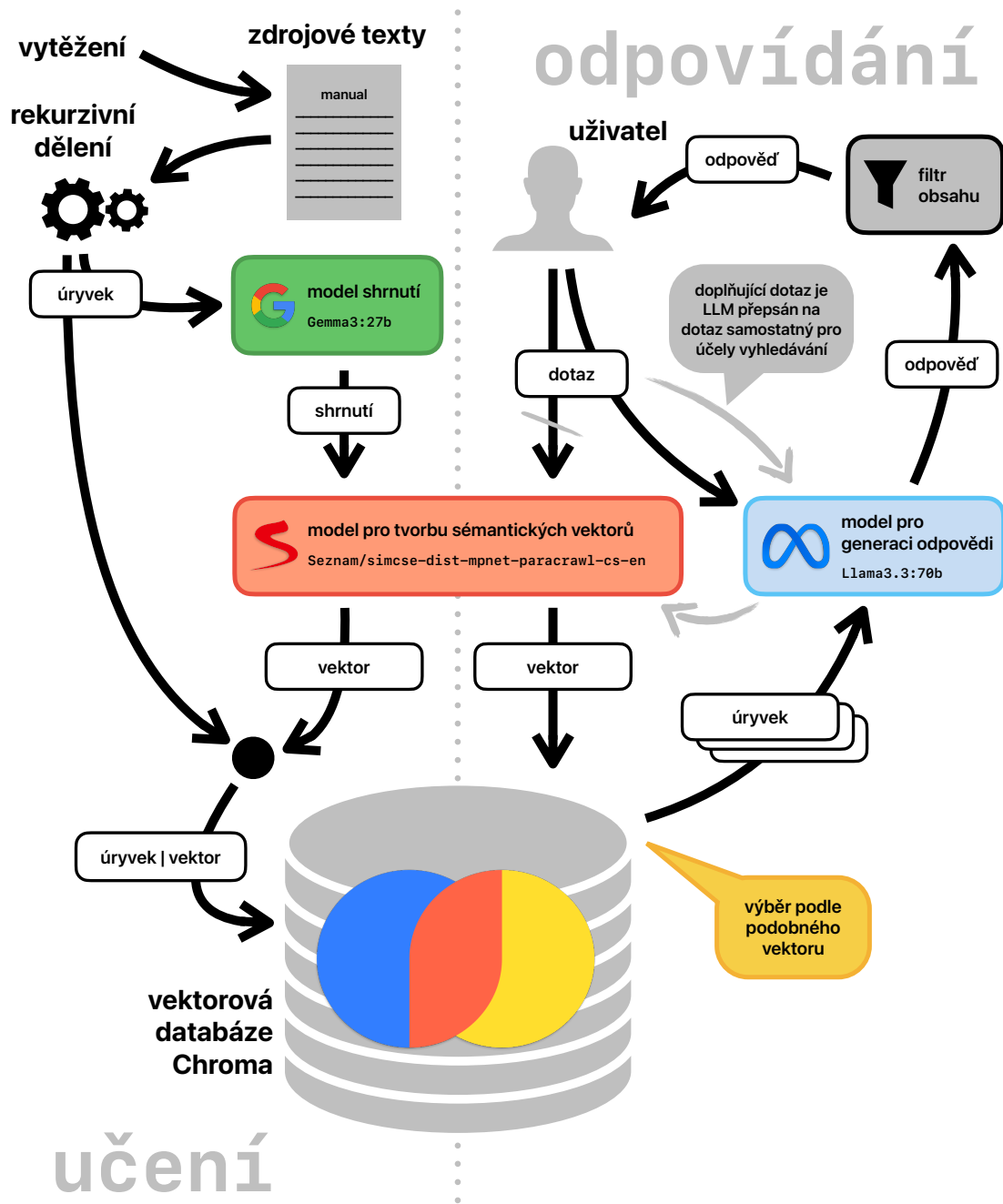
Toto probíhá v režimu doplnění textu, nikoliv konverzace (LangChain volání OllamaLLM místo ChatOllama), s následujícím promptem:

Následuje konverzace mezi uživatelem a asistentem. Prosím vytvoř otázku, která je shodná s poslední otázkou, na kterou se uživatel ptá, ale která je pochopitelná samostatně. Např. nahraď veškeré zájmena, které se odkazují do kontextu. Na otázku neodpovídej!

Ten vychází z anglické verze v LangChain dokumentaci⁷. K samotné konverzaci je pak dodán velmi obvyklý systémový prompt.

Tvým úkolem je pomoci uživateli pracovat s programem Zoner Photo Studio X (ZPS X) od Zoner Software. Pro tento účel máš k dispozici podrobný manuál k programu ZPS X. Doporučuj techniky a postupy pro program Zoner Photo Studio X. Odpovídej stručně a jasně.

⁷https://python.langchain.com/v0.2/docs/tutorials/qa_chat_history/



Obrázek 5.3: Konkrétní „retrieval-augmented generation“ architektura

Při vývoji systému probíhala inference velkých jazykových modelů na kartě Radeon Pro W7900. Tato karta od AMD disponuje příkonem 295 W a 48 GB paměti. Na stroji s GPU běží Ubuntu Server a software Ollama, aplikace k němu přistupuje přes Ollama REST API.

Mezi otevřenými jazykovými modely se nabízí dva zaměřené na širokou vícejazyčnost – Llama 3 a Gemma 3. Vrcholnou verzí Llama 3 je 3.1 405B, ta si i v nejhrušší kvantizaci řekne o více než 100 GB paměti, při kvantizaci na alespoň 4 bity úctyhodných čtvrt terabytu. Provoz takového modelu byl mimo možnosti vývoje v tomto projektu, zároveň je otázkou, zda by výrazně ovlivnil výsledek, nýbrž parametrickou paměť RAG využívá pouze pro pochopení informací z paměti neparametrické (viz kapitolu 3).

Na opačném konci stupnice Llama 3.1 stojí model o 8 miliardách parametrů, ten ale pravidelně pokulhává se skladbou českého textu, jak je možno vidět na následujícím úryvku. Je zde uvedena odpověď na to dotaz „Co je to barevná hloubka?“, jako příklad obecné znalosti principů digitální fotografie. Původcem je model llama3.1:8b z knihovny Ollama, který je pouze aliasem pro llama3.1:8b-instruct-q4_K_M, tedy verzí s fine-tuning pro následování instrukcí a 4 bitovou kvantizací.

*Barevná hloubka, také známá jako **hluboké barvy** nebo **vysoce efektivním barevným modelem** (High Dynamic Range - HDR), je funkce obrazu, **kteřá jej v reálném čase umožňuje zobrazení s vysokou úrovní barevného rozlišení, bez ohledu na to, zda se jedná o fotografii, video či jiný typ média. . . .***

Po přepnutí na model v 8 bitové kvantizaci jako llama3.1:8b-instruct-q8_0 jsou výsledky významně lepší.

*Barevná hloubka (anglicky Color Depth nebo Bit Depth) je míra detailu v barevách, který může obrazový zapisovač zobrazit. Čím vyšší barevná hloubka, tím více úrovní **světla** a **tmavých míst** může zobrazit. . . .*

Opět je možno si všimnout nedokonalosti v jazyce, ale ve významně menší míře.

Tento osamocený příklad nemá vysokou informační hodnotu, je ale reprezentativní ilustrací poznatků z rozsáhlejšího testování vhodnosti různých modelů.

Také velmi jasně demonstruje, že kvantizace totožného model může významně ovlivnit jeho výsledky.

Ze škálovacích zákonů velkých jazykových modelů vyplývá, že kvalitnější výsledky bychom měli očekávat od větších modelů. Do paměťového prostoru již zmíněné vývojové karty se vejde řádově 70 miliardový model při kvantizaci na 4 bity nebo 27 miliardový model při 8 bitech. Do užšího testování tedy postoupili llama3.3:70b (llama3.3:70b-instruct-q4_K_M, 43 GB) a gemma3:27b-it-q8_0 (30 GB), oba z nabídky Ollama.

Mimo jemnější kvantizace v rámci paměťových možností je další výhodou Gemma 3 i slib lepšího vícejazyčného dotrénování [13], kde autoři Llama 3 uvažují pouze užší skupinu 8 světových jazyků [9], i když oba modely jsou předtrénovány na více než stovku jazyků.

Menší z modelů je také výrazně rychlejší, což pomáhá s interaktivitou dialogového systému.

Pro nastínění tohoto rozdílu je vhodné zmínit, že na úloze shrnutí úryvků, která vyžaduje načtení většího množství vstupních tokenů a výstup relativně kratšího výsledku, dosahuje Gemma 3krát kratšího času zpracování. Výsledky jsou přitom k nerozeznání:

llama3.3:70b-instruct-q4_K_M

Výklopné displeje přinášejí komfort při focení, umožňují pohodlně fotit ze zajímavých perspektiv bez nutnosti ležení nebo skákání. Díky nim lze využívat nadhled nebo podhled bez žebříku. Kombinace metod, jako je focení přes hledáček nebo displej, nabízí více možností a flexibility při zachycení scény.

gemma3:27b-it-q8_0

Výklopné displeje fotoaparátů usnadňují fotografování z neobvyklých úhlů – nízkých i vysokých – bez nutnosti se ohýbat či používat pomůcky. Nabízejí pohodlí a alternativu k fotografování přes hledáček. Doporučuje se vyzkoušet obě metody a střídát je dle situace. Článek dále nabízí odkazy na další tipy a novinky ze světa fotografie.

Vzhledem k tomu, že délka dotazů a odpovědí není jasně daná, není jednoduché přesně určit rozdíl v rychlosti v konverzačních úlohách.

Na platformě Chatbot Arena⁸, na níž se podílí i UC Berkeley SkyLab [4], dosahuje Gemma 3 27B v sekci jazyka dokonce lepších výsledků než Llama 3.1 405B. Jediné veřejně dostupné modely, které přesahují hodnocení Gemma 3, jsou 25krát větší modely série DeepSeek.

Tyto výsledky jsou velmi impresivní, zároveň ale nejsou přímo relevantní k RAG výkonnosti v českém jazyce. Pro tvorbu shrnutí v kapitole 5.2 byl použit model Gemma 3, i kvůli jeho významně vyšší výpočetní efektivitě na dané úloze. Vzhledem k povaze úlohy tvorby finálních odpovědí nelze sestrojít jasný indikátor lepšího či horšího modelu, různí uživatelé mohou mít jiný názor.

Prvotní plánovanou fází nasazení tohoto systému je využití na oddělení zákaznické podpory Zoner Photo Studia, kde systém může připravit odpovědi pro příchozí dotazy, které operátor následně pouze překontroluje. Toto je velmi vhodné prostředí pro nasazení tzv. A/B testování, kdy mohou být dočasně nabízeny obě řešení, z nichž je možno vybírat a jsou uchovávány statistiky, kterou odpověď na skutečné dotazy preferovali zkušení odborníci. Tento způsob vyhodnocení bude mít výrazně vyšší průkaznou hodnotu než jakýkoliv syntetický test, ale není v časových možnostech bakalářské práce.

Pro účely dalšího vyhodnocení jsou tedy zváženy obě varianty.

5.4 Ostatní detaily implementace

Vzhledem k očekávanému nasazení je na konci řetězce nástrojů zařazen filtr obsahu. Parametrická paměť modelu v rámci svých znalostí světa obsahuje i postupy pro jiné populární rastrové editory. Ve výjimečných případech, obvykle když není k dispozici relevantní kontext, model sklouzne a začne doporučovat použití konkurenčního produktu, což je nežádoucí, proto filtr raději odpoví, že není schopen odpovědět. Toto je následně potřeba ošetřit v uživatelském rozhraní při nasazení systému ve streamovacím režimu.

V již zmíněném nasazení na oddělení uživatelské podpory, kde je úkolem operátora podpory informaci rychle ověřit, je též vhodné poskytnout zdroje, ze kterých generace odpovědi čerpala. Při vývoji předcházejících fází systému bylo vše navrženo tak, že scraper ukládá text do slovníku, kde klíčem je URL zpracovaného dokumentu. Při dělení dokumentů na úryvky pak systém zachová celou URL v identifikátoru úryvku, pouze připojí za znak mezery číslo úryvku v rámci celku ve formátu M/N . Při výpisu nabízené odpovědi je tedy možné dekodovat z identifikátoru úryvku zdrojovou URL a nabídnout odkaz na daný

⁸lmarena.ai

dokument z manuálu či blogu, kde může operátor svižně ověřit správnost odpovědi. Systém tedy šetří čas nejen při formulaci odpovědi, ale i při prohledávání dokumentace.

Vzhledem k veřejné dostupnosti zdrojů je možné tyto citace poskytnout i koncovým uživatelům, je ale nutno nastavit očekávání oproti zvyklostem tradičních vyhledávačů. Tyto výsledky mohou například obsahovat dokumenty na příbuzné témata, které ale nemusí být přímo relevantní. Cílem tohoto projektu není vyvinout vyhledávač, pokud by tomu tak bylo, bylo by vhodné zvážit další stupeň filtrování výsledků ze sémantického vyhledávače. Tímto druhým filtrem může být i samotný velký jazykový model s vhodným promptem.

Systém je také vybavené telemetrií pro kontinuální vývoj a monitorování, ta je napojena na open-source systém Phoenix⁹ od společnosti Arize. Ten běží jako vlastní Docker kontejner. Tento systém umožňuje například v již zmíněném A/B nasazení monitorovat rychlost generace odpovědí na reálných datech.

Odevzdanou implementaci je možno využít přímo jako chain dle zvyklostí LangChain nebo přes objekt `InferenceSession`, který uchovává historii pro jednoho klienta. Toto používá terminálový klient v `chat.py`.

5.5 Vyhodnocení vlastností implementace

Navržený systém úspěšně splňuje body předložené specifikací.

- Využívá open-source základy, je schopen provozu na vlastním hardwaru a dokonce offline. ✓
- Komunikuje česky. ✓
- Je připraven na integraci nové generace modelů. ✓
- Jeho znalosti jsou aktualizovány z uživatelské specifikace bez lidského úsilí. ✓

Z hlediska vyhodnocení je nejsložitější ověřit věrohodnost odpovědí. Protože samotný výsledek bez kontextu by neposkytl celý obrázek, bude systém porovnán proti systémům podobným. Nejlepším postupem pro vyhodnocení kvality odpovědí proti konkurenci by bylo využití uživatelského testování v reálném nasazení, to ale není v časových možnostech této práce.

Nezávisle na této práci vznikla v minulosti sada dotazů, které jsou časté ze strany uživatelů. Vzhledem k jejich nestrannosti budou použity pro zhodnocení vyvinutého systému. Sada obsahuje 15 otázek a je přiložena v příloze A.

Na základě vlastní snahy vyhledat odpovědi na všechny otázky, jsem byl schopen dohledat pouze 8 z nich v rámci oficiálních zdrojů, na ostatní jsem nenašel jednoznačnou odpověď od společnosti Zoner. Některé ze zbývajících 7 ale lze odpovědět na základě obecného přehledu, kterým disponují velké jazykové modely. Odpovědi na všechny otázky doplnil a potvrdil Pavel Fridrich, Customer Communications Manager Zoner Photo Studio, z interních zdrojů. V tomto zároveň přiznal, že na některé opravdu není veřejně publikovaná odpověď, ale bude přidána do nových stránek podpory, jejichž vývoj právě probíhá.

5.6 Srovnání systému s alternativami

Pro srovnání je navržený systém testován proti komerčnímu ChatGPT, a to ve třech verzích: bez jakékoliv konfigurace (agentní režim), s explicitním použitím ChatGPT Search a s po-

⁹<https://phoenix.arize.com/>

užitím vlastního MyGPT, do kterého jsou nahrány soubory z fáze shromáždění znalostí. Pro ukázkou možností parametrické paměti základních modelů také samostatně Llama 3.3 70B a Gemma 3 27B. Pro obecné systémy bude před samotnou otázkou vložena věta „*Mám problém se Zoner Photo Studio*“ pro určení kontextu. Vyhodnocení správnosti odpovědí proběhlo manuálně s pomocí odpovědí referenčních, výsledky je možno vidět v tabulce 5.4. V horní části jsou uvedeny obecné modely a systémy, dole pak systémy pracující s oficiální dokumentací programu.

Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Úspěšnost
ChatGPT	✓		✓	✓	✓	✓	✓				✓	✓	✓		✓	10/15
ChatGPT Search	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	13/15
Llama 3.3 70B	✓		✓	✓	✓	✓		✓			✓	✓			✓	9/15
Gemma 3 27B	✓			✓	✓	✓										4/15
<i>Oficiální dokumentace</i>		✓	✓	✓	✓		✓				✓	✓		✓		8/15
MyGPT	✓	✓	✓	✓	✓			✓				✓	✓	✓	✓	10/15
Varianta s Llama	✓		✓	✓	✓		✓	✓			✓	✓		✓	✓	10/15
Varianta s Gemma	✓		✓	✓	✓		✓	✓				✓		✓	✓	9/15

Tabulka 5.4: Úspěšnost jednotlivých variant v zodpovídání otázek z přílohy A

Vyhodnocení správnosti odpovědí není jednoduchým úkolem. Dotaz 2 se zaměřuje na zobrazení „pravého panelu programu“, který uživateli „zmizel“. Správná odpověď je vybrat v nabídce „Zobrazit“ panel „Informace“ nebo zkratkou F8. Autorský systém zde navrhuje použít menu „Zobrazit“ a následně vybrat „Pravý panel“, což není správný název, i když je zde ikona znázorňující jeho umístění při pravé straně okna. Vcelku tedy tato odpověď nebyla uznána jako správná, protože obsahuje nepravdivé informace, i když by pro uživatele asi splnila účel.

Podrobné výsledky

Výsledky pro referenční ChatGPT jsou velmi dobré, v 80 % případů agent správně vybral režim ChatGPT Search a poskytl odpověď za pomoci internetové vyhledávání. V případech, kde k tomuto nedošlo, vždy odpověď obsahovala kroky přes neexistující menu nebo špatnými klávesovými zkratkami – tzv. halucinace. Pokud bylo použita funkce Search, dosáhl systém celkově dobrých výsledků a neuspěl na pouze dvou otázkách. Na jednu z nich neodpověděl správně žádný testovaný systém, na otázku 14 ale dokázali odpovědět všechny systémy přímo užívající online dokumentaci. Dotazem je, jak lze nainstalovat starší program ZPS 17. Při vyhledávání na internetu nachází ChatGPT instalaci i z neoficiálních stránek a ty doporučí nad stránkami oficiálními. Toto samozřejmě není správný postup.

Výsledky dosažené autorským systémem je možno díky integraci se systémem Fenix dále analyzovat.

U otázky číslo 6, která popisuje chybu informací o odvolání certifikátu, dochází k rušení kontextu. Načtené úryvky jsou o systému elektronických podpisů pro fotografie a díky použitému způsobu dělení to každý z nich v samotném nadpisu zmiňuje. V aktuální uživatelské nápovědě totiž chyba odvolání certifikátu serveru není vůbec zmíněna. Model tedy popisuje, kde vypnout ověřování certifikátů elektronického podpisu, než aby správně doporučil překontrolovat nastavení antiviru a sítě, což dokázaly všechny 3 základní modely mimo RAG systém, bez tohoto ruchu. Je to tedy příkladem negativního vlivu RAG systému na kvalitu odpovědi.

U otázky 13 („*Kde najdu licenční/sériové číslo programu?*“) testovaný systém správně informuje, kde se nachází informace o licenci, což je uvedeno v manuálu. Na rozdíl od ChatGPT Search ale vynechává klíčovou informaci, že licenční číslo už se nepoužívá a licence je vázaná na účet. Toto není na manuálové stránce¹⁰ jasně uvedeno, pozorný čtenář to ale zvládne pochopit z kontextu.

V otázce 14 naopak autorský systém překonává ChatGPT, které doporučuje stáhnout z neoficiálních zdrojů, a správně směřuje uživatele na stránku jeho Zoner účtu, odkud si může program bezpečně nainstalovat po vložení licenčního klíče.

Rozdíly mezi verzí založenou na Llama a Gemma vznikají u čísla 11. Gemma zde doplňuje informaci, že trvalá licence je k dispozici pouze pro starší verze, což není pravda.

Velmi zajímavé je srovnání autorského retrieval-augmented generation systému a produktu MyGPTs od OpenAI se stejnými znalostmi. Do obou systémů vstupují úplně stejné soubory JSON s texty článků. Veškeré dělení, indexaci a hledání provádí OpenAI bez jakékoliv možnosti uživatelské konfigurace. Použitý systémový prompt je také stejný.

Náhledem je možno si všimnout, že počet správně zodpovězených otázek proprietárního systému je shodný se systémem autorským, který využívá významně menší modely pro hledání i tvorbu odpovědi. Na druhou stranu je nutno uznat zaměření autorského systému pouze na jazyk český, oproti světovému zaměření ChatGPT.

Otázka číslo 9 je poněkud záhadou. Dle Pavla Fridricha není její odpověď v nápovědě, ChatGPT Search ale perfektně reprodukoval potřebnou cestu ke složce, i když odkázal na článek¹¹, kde se tato cesta nenachází a prezentovaný postup přesunu uživatelských nastavení je poněkud odlišný. Tato adresa se na internetu objevuje v článku o Beta programu ZPS X¹², ten ale ChatGPT necituje.

Vyhodnocení

Výsledná implementace věrohodně odpověděla na 67 % očekávaných uživatelských dotazů, přitom 88 % dotazů, jejichž odpověď je přímo zřejmá z uživatelské dokumentace bez složitější syntézy. Systém dosáhl stejných výsledků, jako komerční RAG systém se stejnými vstupními informacemi, ale byl překonán systémem se znalostmi z širšího internetu. Z toho vyplývá, že samotný retrieval-augmented generation systém není špatně naladěný, nýbrž se v přímém srovnání vyrovná komerčnímu řešení, ale vstupní informační zdroje jsou nedostatečné.

Tento výsledek a použití neparаметrické znalostní báze umožňuje celkovou úspěšnost rychle a výrazně zlepšit rozvinutím uživatelské dokumentace v oblastech, kde vytváří nedostatečné odpovědi. Z pohledu vyhodnocení práce není akceptovatelné ladit systém na otázky v jeho testovací sadě, proto to zde nebylo provedeno. Z hlediska kontinuálního vývoje uživatelského produktu je to ale velmi žádoucí.

¹⁰<https://manual.zoner.com/cs/zoner-ucet-fd45ca56/>

¹¹<https://www.milujemefotografii.cz/jak-na-nove-pc-dostat-zps-x-se-vsím-vsudy>

¹²<https://manual.zoner.com/cs/zoner-photo-studio-x-beta/>

Kapitola 6

Závěr

Cílem této práce bylo vytvořit dialogový systém ve funkci uživatelské pomocníka pro program zpracování fotografií. Tento záměr byl v práci naplněn.

Za tímto účelem jsem prostudoval literaturu o velkých jazykových modelech a jejich aplikacích pro získávání informací. Výsledky tohoto bádání jsou k nalezení v úvodních kapitolách práce. Identifikoval jsem retrieval-augmented generation jako vhodnou metodu doplnění znalostí velkých jazykových modelů. Systém jsem implementoval pomocí komerčně použitelných open-source technologií a zlepšení vlastností výsledného systému oproti použitému základnímu modelu jsem vyhodnotil na sadě dotazů z praxe.

Vybraný model pro sémantické vektory `simcse-dist-mpnet-paracrawl-cs-en` přesahuje při vyhledávání ve zdrojových dokumentech výsledky proprietárního modelu OpenAI Ada. Na sadě dotazů z praxe systém správně odpovídá na 88 % otázek ke zdrojovým textům. Při srovnání na stejných vstupních informacích dosahuje navržený retrieval-augmented generation systém shodných výsledků jako krabicové řešení od OpenAI. Systém se ale nedokáže vyrovnat produktu ChatGPT Search, který prohledává širší bázi znalostí z internetu.

Práce výrazně prohloubila mé znalosti velkých jazykových modelů. Zároveň mi dala pochopení pro komplexitu systémů získávání informací.

Systém je velmi dobře připraven pro zapojení nových generativních modelů, jako například Llama 4 z dubna tohoto roku. S ním má systém ambici dosahovat kvalitnějších výsledků s rychlejší odezvou. Vzhledem k publikaci kompatibilní verze Ollama 0.6.7 až 1. května, to bohužel nebylo v možnostech této práce.

V práci chci nadále pokračovat nasazením do praxe na oddělení uživatelské podpory a vylepšeními samotného systému na základě získaných zkušeností. Zároveň je velmi zajímavou možností nasazení na jiné oblasti znalostí. Dalším možným rozšířením je lepší podpora multimediálních zdrojů informací. V této práci by mohl někdo pokračovat hlubším výzkumem do srovnání výkonnosti různých jazykových modelů v českém jazyce.

Literatura

- [1] ACHIAM, J.; ADLER, S.; AGARWAL, S.; AHMAD, L.; AKKAYA, I. et al. *GPT-4 Technical Report*. San Francisco, CA, USA: OpenAI, 2023. Dostupné z: <https://arxiv.org/abs/2303.08774>.
- [2] BEDNÁŘ, J.; NÁPLAVA, J.; BARANČÍKOVÁ, P. a LISICKÝ, O. Some Like It Small: Czech Semantic Embedding Models for Industry Applications. *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington, DC, USA: AAAI Press, 2024, sv. 38, č. 21, s. 22734–22742. ISSN 2374-3468. Dostupné z: <https://ojs.aaai.org/index.php/AAAI/article/view/30307>.
- [3] BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D. et al. Language Models are Few-Shot Learners. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M. a LIN, H., ed. *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2020, sv. 33, s. 1877–1901. ISBN 978-1-71382-954-6. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [4] CHIANG, W.-L.; ZHENG, L.; SHENG, Y.; ANGELOPOULOS, A. N.; LI, T. et al. Chatbot arena: an open platform for evaluating LLMs by human preference. In: *Proceedings of the 41st International Conference on Machine Learning*. Brookline, MA, USA: JMLR.org, 2024, sv. 235, s. 8359–8388. Proceedings of Machine Learning Research. ISSN 2640-3498. Dostupné z: <https://dl.acm.org/doi/10.5555/3692070.3692401>.
- [5] DAI, A. M. a LE, Q. V. Semi-supervised Sequence Learning. In: CORTES, C.; LAWRENCE, N.; LEE, D.; SUGIYAMA, M. a GARNETT, R., ed. *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2015, sv. 28. ISBN 978-1-5108-2502-4. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf.
- [6] DELETANG, G.; RUOSS, A.; DUQUENNE, P.-A.; CATT, E.; GENEWEIN, T. et al. Language Modeling Is Compression. In: *The Twelfth International Conference on Learning Representations*. Appleton WI, USA: ICLR, 2024. Dostupné z: <https://iclr.cc/virtual/2024/poster/17997>.
- [7] DONG, J.; FATEMI, B.; PEROZZI, B.; YANG, L. F. a TSITSULIN, A. *Don't Forget to Connect! Improving RAG with Graph-based Reranking*. 2024. Dostupné z: <https://arxiv.org/abs/2405.18414>.
- [8] GAO, Y.; XIONG, Y.; GAO, X.; JIA, K.; PAN, J. et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. Dostupné z: <https://arxiv.org/abs/2312.10997>.

- [9] GRATTAFIORI, A.; DUBEY, A.; JAUHRI, A.; PANDEY, A.; KADIAN, A. et al. *The Llama 3 Herd of Models*. Menlo Park, CA, USA: Meta, 2024. Dostupné z: <https://arxiv.org/abs/2407.21783>.
- [10] GUO, D.; YANG, D.; ZHANG, H.; SONG, J.; ZHANG, R. et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. Beijing, China: DeepSeek, 2025. Dostupné z: <https://arxiv.org/abs/2501.12948>.
- [11] HUANG, L.; YU, W.; MA, W.; ZHONG, W.; FENG, Z. et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*. New York, NY, USA: Association for Computing Machinery (ACM), 2025, sv. 43, č. 2, s. 1–55. ISSN 1558-2868. Dostupné z: <http://dx.doi.org/10.1145/3703155>.
- [12] JONES, C. R. a BERGEN, B. K. *Large Language Models Pass the Turing Test*. 2025. Dostupné z: <https://arxiv.org/abs/2503.23674>.
- [13] KAMATH, A.; FERRET, J.; PATHAK, S.; VIEILLARD, N.; MERHEJ, R. et al. *Gemma 3 Technical Report*. Mountain View, CA, USA: Google, 2025. Dostupné z: <https://arxiv.org/abs/2503.19786>.
- [14] KOCIÁN, M.; NÁPLAVA, J.; ŠTANCL, D. a KADLEC, V. Siamese BERT-Based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto, CA, USA: AAAI Press, 2022, sv. 36, č. 11, s. 12369–12377. ISSN 2374-3468. Dostupné z: <https://ojs.aaai.org/index.php/AAAI/article/view/21502>.
- [15] LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020, s. 9459 – 9474. NIPS '20. ISBN 978-1-71382-954-6. Dostupné z: <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>.
- [16] LIU, N. F.; LIN, K.; HEWITT, J.; PARANJAPE, A.; BEVILACQUA, M. et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. Cambridge, MA: MIT Press, 2024, sv. 12, s. 157–173. Dostupné z: <https://aclanthology.org/2024.tacl-1.9/>.
- [17] MA, X.; WANG, L.; YANG, N.; WEI, F. a LIN, J. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2024, s. 2421–2425. SIGIR '24. ISBN 979-840070431-4. Dostupné z: <https://dl.acm.org/doi/10.1145/3626772.3657951>.
- [18] MERCER, S.; SPILLARD, S. a MARTIN, D. P. Brief analysis of DeepSeek R1 and its implications for Generative AI. *SuperIntelligence - Robotics - Safety & Alignment*. Sheridan, WY, USA: [b.n.], 2025, sv. 2, č. 1. Dostupné z: <https://s-rsa.com/index.php/agi/article/view/11097>.
- [19] MESNARD, T.; HARDIN, C.; DADASHI, R.; BHUPATIRAJU, S.; PATHAK, S. et al. *Gemma: Open Models Based on Gemini Research and Technology*. Mountain View, CA, USA: Google, 2024. Dostupné z: <https://arxiv.org/abs/2403.08295>.

- [20] MIELKE, S. J.; ALYAFEAI, Z.; SALESKY, E.; RAFFEL, C.; DEY, M. et al. *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP*. 2021. Dostupné z: <https://arxiv.org/abs/2112.10508>.
- [21] NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S. et al. *A Comprehensive Overview of Large Language Models*. 2024. Dostupné z: <https://arxiv.org/abs/2307.06435>.
- [22] OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C. L. et al. Training language models to follow instructions with human feedback. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2022, s. 27730–27744. NIPS '22. ISBN 978-1-71387-108-8. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [23] RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. et al. *Improving language understanding by generative pre-training*. San Francisco, CA, USA: OpenAI, 2018. Dostupné z: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [24] RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*. Brookline, MA, USA: JMLR.org, 2020, sv. 21, č. 1, s. 5485–5551. ISSN 1532-4435. Dostupné z: <https://arxiv.org/abs/1910.10683>.
- [25] STETINA, J.; FAJCIK, M.; STEFANIK, M. a HRADIS, M. *A Comparative Study of Text Retrieval Models on DaReCzech*. 2024. Dostupné z: <https://arxiv.org/abs/2411.12921>.
- [26] TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A. et al. *LLaMA: Open and Efficient Foundation Language Models*. Menlo Park, CA, USA: Meta, 2023. Dostupné z: <https://arxiv.org/abs/2302.13971>.
- [27] TOUVRON, H.; MARTIN, L.; STONE, K.; ALBERT, P.; ALMAHAIRI, A. et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. Menlo Park, CA, USA: Meta, 2023. Dostupné z: <https://arxiv.org/abs/2307.09288>.
- [28] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017, s. 6000–6010. NIPS'17. ISBN 978-1-5108-6096-4. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [29] WEI, J.; TAY, Y.; BOMMASANI, R.; RAFFEL, C.; ZOPH, B. et al. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Mountain View, CA, USA: Google, 2022. ISSN 2835-8856. Dostupné z: <https://arxiv.org/abs/2206.07682>.
- [30] WEI, J.; WANG, X.; SCHUURMANS, D.; BOSMA, M.; ICHTER, B. et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. Mountain View, CA, USA: Google, 2023. Dostupné z: <https://arxiv.org/abs/2201.11903>.

- [31] YASUNAGA, M.; AGHAJANYAN, A.; SHI, W.; JAMES, R.; LESKOVEC, J. et al. *Retrieval-Augmented Multimodal Language Modeling*. 2023. Dostupné z: <https://arxiv.org/abs/2211.12561>.
- [32] ZHAO, W. X.; ZHOU, K.; LI, J.; TANG, T.; WANG, X. et al. *A Survey of Large Language Models*. 2025. Dostupné z: <https://arxiv.org/abs/2303.18223>.
- [33] ZHOU, J.; LIANG, Z.; FANG, Y. a ZHOU, Z. Exploring Public Response to ChatGPT With Sentiment Analysis and Knowledge Mapping. *IEEE Access*, 2024, sv. 12, s. 50504–50516. Dostupné z: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10494726>.

Příloha A

Otázky z praxe

Tento seznam otázek poskytl ředitel Zoner Software Michal Prouza. Jako referenční jsou zde uvedeny odpovědi potvrzené Pavlem Fridrichem z oddělení podpory. Odpovědi umělé inteligence jsou sázeny *kurzívou* a generovány navrženým systémem s modelem Llama 3.3 70B q4. Pro uvedení zde jsou zkráceny a je odstraněno Markdown formátování. Vyhodnocení navrženého systému proti alternativám na této sadě je k nalezení v sekci [5.6](#).

- 1. Program mi hlásí, že nejsem připojen k Internetu, i když jsem. Co s tím?**
ref.: Na vině je obvykle antivirový program, vytvořte výjimku, dočasně deaktivuje firewall atd.
AI: Zkontrolujte funkčnost internetového připojení a zda není pro určité případy blokováno antivírem či firewallem. ✓
- 2. Jak vrátím do programu pravý panel programu, který mi zmizel?**
ref.: V menu Zobrazit > Informace nebo zkratkou F8
AI: V menu Zobrazit můžete zapnout nebo vypnout jednotlivé panely. Pokud chcete zobrazit pravý panel, klikněte na Zobrazit a zvolte Pravý panel. ×
- 3. Jak v Zoneru přehraju HEVC video?**
ref.: Nutno doinstalovat přes Windows Store nebo ze Zoner účtu
AI: Pro přehrávání HEVC videa v Zoner Photo Studio X je nutné mít nainstalován příslušný kodek. ... Otevřete Microsoft Store ... Hledejte "HEVC Video Extensions"... Pokud stále máte problémy, zkuste restartovat počítač nebo kontaktovat podporu Zoneru pro další pomoc. ✓
- 4. Proč můj RAW vypadá jinak než JPG stejné fotky?**
ref.: RAW jsou surová data, vždy je to o interpretaci
AI: To je běžný jev. Rozdíl mezi RAW a JPG verzí stejné fotografie může být způsoben několika faktory: ... Tyto rozdíly mohou způsobit, že RAW verze fotografie vypadá jinak než JPG verze. Nicméně, právě díky těmto rozdílům můžete při zpracování RAW získat lepší výsledky, protože máte více dat k dispozici a můžete aplikovat různé úpravy podle svých potřeb. ✓
- 5. Budu potřebovat DNG convertor ke zpracování mých RAWů?**
ref.: Pro podporované fotoaparáty ne, pro nepodporované lze použít např. Adobe DNG Converter
AI: Ano, budete potřebovat DNG Converter, pokud chcete v Zoner Photo Studio X zpracovat RAW soubory, které program sám o sobě neumí zobrazit pomocí interního

způsobu zpracování RAW. DNG Converter vám umožní převést tyto soubory do formátu DNG, se kterým umí ZPS X pracovat. ✓

6. Jak se zbavím upozornění, že „Nejsou k dispozici informace o odvolání certifikátu zabezpečení tohoto serveru.“?

ref.: Opět obvykle antivirový problém, viz otázka 1

AI: Vypni ověření zneplatnění certifikátu v Možnostech na záložce Certifikáty pro podpis. Toto upozornění se zobrazuje, protože program nemůže ověřit stav certifikátu přes internet. ×

7. Jak změním jazyk programu na angličtinu?

ref.: Přes instalační soubor

AI: Chcete-li změnit jazykovou verzi programu Zoner Photo Studio X, spusťte instalační soubor a v menu zvolte volbu Změnit jazyk programu. Poté vyberte požadovaný jazyk, v tomto případě angličtinu. Tímto krokem dojde ke změně jazyka programu, avšak Katalog zůstane v původní jazykové verzi. ✓

8. Program po přepnutí do Vyvolat padá, co mám dělat?

ref.: Zaslát soubor systémových informací a popis na podporu

AI: Zkuste zrušit veškeré uživatelské změny v dialogu a uvést ho do výchozího stavu kliknutím na tlačítko Výchozí v pravém dolním rohu dialogu. Pokud problém přetrvává, kontaktujte technickou podporu emailem podpora@zoner.cz nebo prostřednictvím kontaktního formuláře přímo v programu Zoner Photo Studio X. Přiložte soubor Systémových Informací, snímky, záznamy obrazovky ilustrující problém a případně i problémové soubory. ✓

9. Lze přenést videoprojekty na jiný počítač?

ref.: Ano, je potřeba mít zdrojová média ve stejné lokaci a přenést obsah složky

C:\Users\\AppData\Local\Zoner\ZPS X\CreatePresets

AI: Ano, lze. Projekty modulu Video jsou uloženy v adresáři Dokumenty/Zoner/Photo Studio/Video Projects ve formátu *.zvpj. Tyto soubory je možné zkopírovat a přenést na jiný počítač se stejnou verzí programu ZPS X. ×

10. Po spuštění se mi zobrazuje upozornění, že aktualizace metadat se nezdařila, co s tím?

ref.: Otevřít složky katalogu a vyřešit problém indikovaný Informačním příznakem

AI: Pokud se při spuštění programu Zoner Photo Studio X (ZPS X) zobrazí upozornění, že aktualizace metadat se nezdařila, můžete zkusit následující kroky: ... Zkontrolujte připojení k internetu ... Restartujte program ... Zkontrolujte nastavení Katalogu ... Aktualizujte metadata ručně ... Kontaktujte podporu ... ×

11. Je možné zakoupit trvalou licenci programu?

ref.: Ano

AI: Ano, je možné zakoupit trvalou licenci programu Zoner Photo Studio X. Trvalá licence znamená, že máte právo používat program bez časového omezení, aniž by bylo nutné platit roční předplatné. ✓

12. Jak zaplatím bankovním převodem?

ref.: Údaje Vám budou zaslány na email

AI: V případě platby bankovním převodem vám na uvedenou e-mailovou adresu zašleme potvrzení a výzvu k úhradě objednávky ze Zoner.cz. ... ✓

13. Kde najdu licenční/sériové číslo programu?

ref.: Licenční číslo už není používáno, licence se váže na účet

AI: Licenční nebo sériové číslo programu Zoner Photo Studio X naleznete v následujících místech: ... Zoner účet: Přihlaste se do svého Zoner účtu ... a klikněte na sekci "Licence a platby". Zde byste měli vidět seznam svých licencí, včetně licenčního čísla.

×

14. Jak nainstaluji ZPS17, které jsem zakoupil v roce 2015?

ref.: Přes svůj Zoner účet

AI: Pokud vlastníte licenci na starší verzi programu než ZPS 18, instalátor si můžete stáhnout ze svého Zoner účtu, ale potřebujete k tomu znát 25místné licenční číslo, které jste získali při nákupu. V Zoner účtu toto číslo vložíte do pole Zadejte další aktivační kód nebo licenční číslo a licenci přidáte na účet. Pak se vám v sekci Zobrazit starší licence zpřístupní možnost stáhnout instalátor programu, pomocí kterého starší verzi ZPS opět nainstalujete. ✓

15. Nedostal jsem fakturu za nákup licence ZPS X, můžete mi ji poslat?

ref.: Lze ji nalézt v Zoner účtu, sekce Platby a Objednávky

AI: Ano, samozřejmě. Fakturu za vámi zakoupenou licenci Zoner Photo Studio X si můžete stáhnout ve svém Zoner účtu v sekci Platby a objednávky ... ✓