National Czech Programme

Výstup LLM_1

Technická zpráva: Sběr požadavků a návrh architektury pro využití velkých jazykových modelů

Výstup byl vytvořen za podpory Ministerstva školství, mládeže a tělovýchovy a Operačního programu Jan Amos Komenský v rámci projektu Open Science II (reg. č. CZ.02.01.01/00/24_030/0015041)

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Projekt Open Science II
CZ.02.01.01/00/24_030/0015041
Univerzita Karlova
Ovocný trh 560/5, 116 36 Praha 1
eosc2@ruk.cuni.cz; www.eosc.cz

Manažerské shrnutí

Podaktivita Velkých jazykových modelů v rámci projektu Open Science II reaguje na překotný rozvoj technologií umělé inteligence, které umožňují lepší automatizaci práce a chytřejší interakci s uživateli, a zároveň i růst infrastruktury v souvislosti s datovými sadami, národní repozitářovou platformou a zvyšující se mezioborové spolupráci výzkumníků a výzkumnic nejen v ČR v režimu otevřené vědy. Aktuální trendy kladou mnohem větší nároky na uživatelskou přívětivost a komfort a zároveň se zvyšuje množství technických požadavků a případů, které musí řešit uživatelská podpora. Z těchto důvodů je naplánovaný vývoj AI asistenta, který bude schopen pomoci uživatelům při řešení jejich technických problémů a dotazů, a odlehčí odbornému personálu a zároveň umožní lépe škálovat kapacity uživatelské podpory.

V rámci projektu plánujeme vývoj technologie, která bude stavět na existujících open source nástrojích, jejichž průzkum proběhl v prvních měsících realizační fáze projektu. Doména AI nástrojů je aktuálně velice dynamická, a proto je nezbytné pokračovat v monitorování dostupných možností a nově vznikající nástroje a technologie případně zakomponovávat do cílového řešení AI asistenta. Důležitou komponentou výsledného řešení jsou procesy uživatelské podpory, eskalace obtížných případů na odborníky a kontinuální zlepšování znalostní databáze a dokumentačního úložiště, což by měl AI asistent komplexně zastřešovat.

Tento dokument slouží jako prvotní průzkum, analýza a high-level návrh plánovaného výsledku, který vychází ze studie proveditelnosti a návrhu projektu. Obsahem jsou definované případy užití, kvantifikační kritéria a výkonnostní parametry a nakonec i seznam požadavků vycházející z aktuálních potřeb týmu uživatelské podpory a zástupců spolupracujících týmů. Mezi požadavky jsou akcentovány potřeby na zabezpečení a compliance s platnými právními předpisy.

Obsah dokumentu

Úvod

Tento dokument představuje výstup podaktivity PKA 10.1 Velké jazykové modely v rámci projektu Open Science II (OS II). Projekt OS II přímo navazuje na předchozí aktivity Národní repozitářové platformy (NRP), kde byla vybudována víceúrovňová uživatelská. Zatímco původní projekt NRP se soustředil primárně na budování expertních kapacit pro řešení komplexních technických problémů na úrovních L2 a L3 podpory, aktivita T10.1 v rámci OS II cílí na automatizaci a asistenci pro základní úroveň uživatelské podpory (L1) pomocí využití velkých jazykových modelů.

Hlavní motivace spočívá v usnadnění řešení rutinních a často se opakujících dotazů a problémů, aby bylo možné čas pracovníků podpory efektivněji využít na řešení složitějších technických výzev. Kromě efektivnějšího využití lidských zdrojů dojde i ke zkrácení reakční doby uživatelské podpory a zajištění základní podpory v režimu 24/7.

Hlavním cílem podaktivity PKA 10.1 je vývoj komplexního systému virtuálního asistenta založeného na velkých jazykových modelech (LLM) a konceptu Retrieval-Augmented Generation (RAG). V RAG konceptu jazykový model pracuje primárně s informacemi, které získá z dedikované znalostní databáze. Tím je podpořeno generování přesnějších odpovědí jazykovým modelem. Dedikovaná znalostní báze bude obsahovat nejen dokumentaci k repozitářům a jejich infrastruktuře, ale i znalosti o řešení předchozích problémů a znalosti o metadatech v repozitářích. Rámec projektu OS II zároveň klade důraz na bezpečnost a minimalizaci rizik (např. únik informací), proto bude systém provozován on-premise v zabezpečeném prostředí e-INFRA CZ.

Během úvodních čtyř měsíců realizační fáze proběhla analýza technologického základu a sběr požadavků prostřednictvím setkání s klíčovými aktéry, včetně manažerů, datových stewardů a právních expertů. V rámci tohoto procesu bylo identifikováno několik případů užití.
Na základě těchto zjištění byly definovány funkční, výkonnostní a bezpečnostní požadavky a vytvořen první návrh architektury procesů pro strojové zpracování uživatelských vstupů. Souběžně byla zprovozněna základní verze jazykového procesoru na testovacím serveru, která slouží jako první funkční prototyp.

Provedený sběr informací a definice požadavků poskytly nezbytný podklad pro zajištění compliance s právními a etickými standardy, zejména v oblasti nakládání s citlivými daty a licencování modelů. Návrh architektury nyní umožňuje přejít k fázi intenzivního vývoje a testování na interní skupině uživatelů, jejichž průběžná zpětná vazba bude využita k hodnocení výkonnosti modelů, rozšiřování dostupné znalostní báze a úprav postupů zpracování informací v systému. Na tyto kroky navážeme plnou integrací rozhraní pro operátory

uživatelské podpory, automatizací správy znalostní báze a podporou nasazení systému v produkčním prostředí.

Kontext

Tato kapitola popisuje koncepci národního datového prostředí v České republice a zachycuje klíčové komponenty a aktivity s touto koncepcí spojené. Následně je kapitola zaměřena na uživatele a současnou architekturu uživatelské podpory.

Základem datového prostředí v ČR je Národní Datová Infrastruktura (NDI), která představuje komplexní systémový rámec pro správu a sdílení dat v rámci vědecké komunity v ČR. Součástí NDI je Národní Repozitářová Platforma (NRP), která tvoří technický základ pro ukládání a správu dat. Pro indexaci záznamů ze všech repozitářů v rámci NRP i z dalších zdrojů (např. CLARIN nebo Zenodo pro české autory) slouží Národní Metadatový Adresář (NMA). Konkrétním aktuálně dostupným datovým úložištěm v rámci NRP je Národní Repozitář (NR), který slouží jako repozitář poslední instance, tj. pokud pro data neexistuje dedikovaný doménový repozitář a není poptávka/komunita pro jeho vytvoření, pak taková data jsou vhodným kandidátem pro uchování v NR.

Na budování datového prostředí v ČR se podílelo-podílí několik projektů. Projekt IP EOSC je zaměřený na vytvoření a provoz NDI, Sekretariátu EOSC CZ, Národního metadatového adresáře a Školicího centra EOSC CZ. Projekt IP CARDS (Czech Academic and Research Discovery Services) je zaměřen na vývoj a provoz vyhledávače a navazujících služeb pro oblast vědy, výzkumu a inovací (VaVaI). Samotný projekt OS II buduje oborové repozitáře a přidává pokročilé služby nad rámec základních služeb NRP. Jednou z jeho klíčových podaktivit je právě využití velkých jazykových modelů pro automatizaci uživatelské podpory. Projekt OS III pak přináší aktivity v oblasti vzdělávání a zvyšování kompetencí v oblasti Open Science.
Z pohledu budování využití velkých jazykových modelů pro automatizaci uživatelské podpory je vhodné shrnout i typické uživatelské role v NRP. Tvůrci a správci repozitářů (datoví kurátoři) — zjišťují informace, jak mají založit repozitář, jak nastavit pravidla repozitáře. Tvůrci a majitelé dat (datoví stewardi) - tvoří data management plan, zjišťují informace o repozitářích, pravidlech, politikách (jaká metadata je nutné vyplnit, jak je nutné data zpracovat), zajišťují kvalitu a ochranu dat, řeší etické, legislativní a licenční aspekty. Vědci pak zjišťují informace o tom, kde jsou data, jak data vypadají, za jakých podmínek (licence) a jak s nimi mohou pracovat (např. u citlivých dat bude možné s daty pracovat pouze ve vyhrazeném prostředí, nebo pracovat pouze s agregovanými výsledky na základě schválených workflow). Jako uživatele automatizace můžeme rovněž zařadit pracovníky samotné uživatelské podpory, kteří řeší dotazy a problémy reportované výše zmíněnými uživateli.

Úlohy pracovníků uživatelské podpory se liší dle jejich zařazení a jejich specializace. Pro soulad se standardním víceúrovňovým modelem podpory byly definovány úrovně L0 až L3. Nicméně

k víceúrovňovému modelu je potřeba vzít v úvahu ortogonálně specializaci pracovníků pracujících na vyšších (L2 a L3). Tito pracovníci budou typicky buď technicky zaměření na řešení problémů s infrastrukturou, službami a software nebo metodologicky zaměření na řešení problémů ohledně standardů, regulací v oblasti datové analýzy apod. Např. L2 technicky zaměřený pracovník nebude schopen řešit problém týkající se citlivosti dat eskalovaný z L1.

Úroveň L0 je implementovaná pomocí FAQ a webové dokumentace (www.eosc.cz), která slouží jako primární zdroj informací, návodů, obsahuje FAQ a kontakty na další úrovně podpory. Service Desk CESNET poskytuje L1 podporu a jeho pracovníci mohou uživatele případně odkázat na L0. Jedním z úkolů Service Desk CESNET je kategorizace požadavků a jejich předání odpovídajícím osobám (správce konkrétního repozitáře, compliance, atd.). Podporu úrovně L1 zajišťují také správci jednotlivých repozitářů, kteří svým uživatelům poskytují support týkající se např. přihlašování a přístupových práv, a řeší technické problémy svých repozitářů. Úroveň L2 slouží jako pokročilejší podpora jak pro uživatele, tak pro úroveň L1. Úroveň L3 slouží jako pokročilejší technická podpora, která se zabývá nejnáročnějšími technickými požadavky a zároveň poskytuje poradenství a konzultace pro úrovně L1 a L2.

Systém Request Tracker (RT) je hlavním nástrojem pro správu a sledování uživatelských požadavků v rámci uživatelské podpory NRP. Tento systém je provozován sdružením CESNET a je integrován do standardní podpory e-infrastruktury, což usnadňuje kooperaci a eskalaci problémů.

Metodologie

Pro sběr požadavků bylo kontaktováno několik klíčových osob podílejících se na projektu OS II, se kterými následně proběhla setkání a společná diskuze. Z diskuzí byly vedeny poznámky, ze kterých byly identifikovány případy užití a následně požadavky. Kromě přímých diskuzí ohledně technického využití LLM byly prozkoumáváno i samotné prostředí projektu OS II, právní stanoviska, současný stav a práce Service Desku a jiné.

Setkání byla uspořádána s osobou odpovědnou za Service Desk v rámci OS II, manažerkou kvality, produktovým manažerem pro kritický sektor, osobou zodpovědnou za vývoj národního repozitáře, datovou stevardkou a osobami zodpovědnými za compliance a právní náležitosti.

Setkání s osobou zodpovědnou za Service Desk, manažerkou kvality a produktovým manažerem se zaměřovalo na současný stav pracoviště Service Desku na CESNETu, způsoby práce a využívané systémy. Dále bylo diskutován možný způsob využití LLM, a to jako AI asistent pro zodpovídání dotazů uživatelů, ale i jako podporu pracovníků Service Desku a zvýšení efektivity jejich práce.

Setkání s osobou zodpovědnou za vývoj národního repozitáře se zaměřovalo na možné využití LLM k podpoře nejen samotných uživatelů, ale také správců jednotlivých repozitářů. Jedná se například o návrh metadatového modelu pro jednotlivé repozitáře a jeho validaci, pomoc se zakládáním repozitářů, apod. Dále se řešila role datových kurátorů a poskytování supportu pro jednotlivé repozitáře, který má každý své specifické prostředí a kontext.

Setkání s datovou stevardkou bylo zaměřeno na představení existujících datových domén, odlišností specifické potřeby repozitářů, především pak z pohledu citlivých dat. Rovněž byla navržena spolupráce s datovými úložišti a spolupráce s tými, které již mají LLM nápovědu zprovozněnou nad svou dokumentací.

Setkání s právními a compliance experty se zaměřovalo na licenční podmínky jednotlivých LLM modelů a licencí výstupů LLM modelů. Dále se diskovaly způsoby podmínek užití, (polo)automatické zpracovávání a bezpečnost dat. V neposlední řadě se diskutovaly také potenciální limitace využití LLM v rámci OS II.

Zjištěné případy užití

1. LLM pro validaci metadatového modelu

V rámci NRP existuje tzv. Catch-All repozitář, který je repozitářem poslední volby. Pokud neexistuje specifický repozitář pro určitá data a zároveň není vhodné založit úplně novou instanci specifického repozitáře pro daná data, pak Catch-all repozitář je poslední volbou vlastníka dat, kam data uložit. Správcem Catch-all repozitáře je Národní technická knihovna (NTK).
Správce repozitáře, resp. datoví kurátoři z NTK komunikují mezi sebou a s vlastníkem dat a domlouvají odpovídající metadatový model pro daná data. Mohou vycházet např. z existujícího Czech core metadata modelu, který je odvozen od DataCite modelu.

Případ užití LLM spočívá v návrhu vhodného metadatového modelu a v podpoře ladění metadatového modelu vlastníkem dat a kurátory z NTK. Následující příklad demonstruje scénář tvorby metadatového modelu pro specifický oborový repozitář.

Vlastník dat popíše charakter dat (typ dat, obor, způsob sběru, citlivost, vztah k publikacím), definuje základní požadavky (licence, přístupová práva, etická omezení).

AI asistent analyzuje popis dat, navrhne výchozí metadatový model na základě existujících metadatových modelů (např. Czech Core Metadata Model, DataCite, případně jiných relevantních oborových schémat, označí povinná, doporučená a volitelná metadata, identifikuje chybějící nebo problematická datová pole. Zároveň bere v úvahu vyhledatelnost repozitáře v rámci NMA, soulad s politikami NRP a EOSC. Výstupem je model s dokumentací obsahující vysvětlení jednotlivých položek modelu.

Datoví kurátoři NTK zkontrolují návrh na základě výše uvedených hledisek a vytvoří připomínky. Tyto připomínky poskytnou AI asistentovi, který navrhne upravenou verzi modelu. Připomínkovaná a upravená verze je iterativně opět kontrolována, připomínkována a upravována až vznikne finální verze modelu.

2. Validace dat při vkládání do repozitáře

V rámci NRP a jednotlivých repozitářů představuje validace dat a metadat při jejich vkládání klíčový krok pro zajištění kvality, konzistence a dlouhodobé využitelnosti uložených dat. Tento proces může být obzvláště důležitý v případě specifických datových sad, které jsou ukládány do Catch-all repozitáře, kde nelze předpokládat jednotný oborový kontext.

Případ užití LLM v této oblasti spočívá v automatizované sémantické kontrole vstupních dat a metadat, a to zejména v situacích, kdy jsou data ukládána prostřednictvím strojového rozhraní bez přímé interakce s uživatelem.

AI asistent integrovaný do procesu ukládání dat do repozitáře by mohl plnit následující funkce:

- kontrolu úplnosti metadat vůči metadatovému modelu,
- identifikaci chybějících, prázdných nebo nečekaných hodnot v datech,
- interpretaci validačních chyb a jejich převod do srozumitelného vysvětlení pro uživatele, potenciálně s návrhem opravy.

Užití LLM je demonstrováno na následujícím příkladu scénáře validace dat při procesu jejich ukládání do repozitáře.

Uživatel (vlastník dat nebo systém jednající jeho jménem) nahrává data do repozitáře prostřednictvím strojového API. V průběhu nahrávání některé povinné položky metadat nejsou vyplněny nebo chybí, některé hodnoty neodpovídají očekávanému typu, významu, případně i rozsahu.

Repozitář data technicky přijme a uloží, avšak LLM analyzuje vstupní data a metadata, identifikuje konkrétní problémy, připraví srozumitelné vysvětlení zjištěných nedostatků včetně návrhů možného řešení.

Na základě výstupu validačního procesu repozitář automaticky odešle uživateli notifikaci (např. e-mail), součástí notifikace je odkaz na webové rozhraní, ve kterém má uživatel k dispozici:

- přehled zjištěných problémů,
- vysvětlení, proč jsou jednotlivé položky problematické,
- doporučení, jak by měla oprava vypadat.

Zapojení LLM do validačního procesu při vkládání dat může přinést zvýšení kvality a úplnosti dat a metadat, lepší uživatelskou zkušenost při práci se strojovým API, obohacení auditní stopy při ukládání dat.

3. LLM jako L0 a L1 support v rámci RT systému

Pracoviště Service Desku řeší i požadavky a dotazy, které se dají zodpovědět pomocí FAQ nebo veřejně dostupné dokumentace. Tyto požadavky mají proto vysokou pravděpodobnost být vyřešeny automaticky. Nabízí se tedy L0 (a případně i L1) support nahradit technologií LLM, která bude sloužit jako první vrstva interakce mezi uživateli vytvářející požadavky a lidskými pracovníky Service Desku.

V tomto případě užití je tedy LLM použito jako první vrstva řešení požadavků, kdy LLM nově příchozí požadavek přijme, zpracuje a vygeneruje odpověď. Tato odpověď může být odeslána

automaticky nebo po následné kontrole lidským pracovníkem Service Desku. Pokud uživatel s odpovědí není spokojen a požadavek vyřešen nebyl, uživatel má možnost zvolit eskalaci požadavku na další úroveň uživatelské podpory, který je zajišťován lidskými pracovníky Service Desku. V takovém případě má lidský pracovník Service Desku, který požadavek přejal, k dispozici kontext z LLM a odpověď poskytnutou nasazeným LLM.

V tomto případu užití tedy LLM:

- Přijímá nově příchozí požadavky na pracoviště Service Desku
- Požadavky zpracovává a vybírá ty, které souvisí s prostředím OS2
- Za pomoci znalostní databáze generuje odpovědi
- Odpověď může být volitelně nejprve zkontrolována lidským pracovníkem Service Desku a případně odeslána
- Odpovědi jsou odesílány jako odpovědi uživatelům
- Uživatel může reagovat navazující otázkou v rámci stejného požadavku a pokračovat tak v konverzaci
- Uživatel v případě nespokojenosti s odpovědí může požadavek eskalovat na lidského pracovníka Service Desku

Využití LLM pro zodpovídání častých dotazů a asistence pro uživatele, aby se rychle zorientovali v dokumentaci a materiálech, má potenciál snížit objem požadavků, které musí pracoviště Service Desku zpracovat, a kapacity odborníků bude možné využít na pokročilejší a náročnější požadavky.

4. LLM jako L0 a L1 support na webu

Tento případ užití je velmi podobný jako předchozí *LLM jako L0 a L1 support v rámci RT systému* s tím rozdílem, že uživatelské požadavky jsou řešeny formou rozhraní přímo na webové stránce. Vybrané webové stránky v rámci OS II tedy poskytují rozhraní, pomocí kterého se mohou uživatelé ptát na dotazy a LLM jim poskytuje odpovědi a shrnutí, a případně uživatele rovnou přesměrovává na relevantní místa na webových stránkách.

Jedná se tedy o uživatelsky přívětivější rozhraní, které má zároveň potenciál snížit počet příchozích požadavků na pracoviště Service Desku jako takové, protože základní uživatelská podpora a nasměrování je dostupná přímo na webových stránkách jednotlivých repozitářů a dalších vybraných stránkách v prostředí OS II.

V tomto případu užití tedy LLM:

- Přijímá otázky uživatelů pomocí webového rozhraní
- Požadavky zpracovává a za pomoci znalostní databáze generuje odpovědi
- Poskytuje krátké výstižné odpovědi a shrnutí
- Případně uživatele přímo přesměrovává na relevantní části webových stránek

LLM dostupné přímo na vybraných webový stránkách (např. stránky jednotlivých repozitářů) má potenciál snížit objem požadavků, které jsou odeslány na pracoviště Service Desku, a umožňuje pracovníkům Service Desku věnovat se složitějším a technicky náročnějším požadavkům.

5. LLM pro zpracovávání příchozích požadavků do systému RT

Národní repozitářová platforma poskytuje služby a Service Desk těmto službám zajišťuje podporu. Pro zajištění podpory se využívají RT fronty. Příchozí požadavky se mapují na jednotlivé RT fronty. Jedna RT fronta může sloužit pro požadavky více služeb, daná služba by ale měla být přiřazena pouze do jedné fronty.

Pracovník Service Desku musí při příchodu nového požadavku identifikovat o jakou službu se jedná a přiřadit ji do správné RT fronty. Při velkém množství služeb a front se tento úkol stává náročným a jako jeden z možných případů užití bylo identifikováno využití LLM pro identifikaci služby, které se příchozí požadavek týká, a RT fronty, kam daný požadavek patří.

Jedna z možností využití LLM v tomto případě je nasazení LLM tak, aby zpracovávalo nově příchozí požadavky na Service Desk, automaticky rozpoznalo, ke které konkrétní službě požadavek patří, a na základě toho požadavkům přiřazovalo štítek odpovídající služby a zařadilo požadavky do správných RT front. Pracovníci Service Desku díky tomu mohou začít roztříděné požadavky řešit ve správném kontextu a opatřené metadaty, které pomohou s orientací v problému. V případě nepřesné klasifikace mohou pracovníci požadavek přesunout do správného kontextu a tuto skutečnost použít k vylepšení klasifikátoru.

Využití LLM pro automatickou klasifikaci příchozích požadavků může urychlit zpracovávání požadavků tím, že požadavky budou již od začátku zpracovávány odpovídajícím pracovištěm, zvýšit efektivitu a snížit chybovost rozřazování požadavků mnoha služeb člověkem do velkého množství RT front. Zároveň se tím zkrátí čas, který musí při ručním rozřazování požadavky čekat, než se dostanou k relevantním pracovníkům.

6. LLM pro vyhledávání relevantních částí dokumentace

Při řešení požadavků využívají pracovníci Service Desku dokumentaci, návody a ostatní materiály poskytnuté správci jednotlivých repozitářů (např. dokument s FAQ). Těchto zdrojů může být více a očekává se, že jejich počet bude růst s rostoucím počtem repozitářů začleněných do NRP.

Identifikovaným případem užití je zde předzpracování dokumentace a ostatních materiálů pomocí LLM, kdy jsou tyto dokumenty poskytnuty LLM a následně mohou sloužit pro uživatelsky přívětivé procházení dokumentace pomocí série otázek v přirozeném jazyce. Vyhledávání relevantních částí dokumentace může být navíc provedeno automaticky při

příchodu nového požadavku. Pracovník Service Desku tedy bude mít dostupné relevantní části zdrojů, které může rovnou využít pro vyřešení požadavku.

Takovéto využití LLM opět usnadňuje pracovníkům Service Desku řešení příchozích požadavků a navíc jim dovoluje požadavky řešit rychleji. Tento případ užití má smysl i v momentě, kdy LLM jako takové nedokáže vyřešit příchozí požadavek samo a uživatel zvolí eskalaci na L1 support, protože pracovníkovi supportu dává dodatečné informace a pomáhá mu se rychleji zorientovat.

7. LLM pro vyhledávání podobných požadavků v historii

Uživatelé mohou mít podobné problémy a požadavky a může se tedy stát, že příchozí požadavek byl v minulosti už jednou řešen jiným pracovníkem Service Desku. To znamená, že se stejná (či velmi podobná) práce bude vykonávat znova a zbytečně ubírat čas pracovníkům Service Desku potřebný pro řešení jiných nových požadavků.

Možné využití LLM zde je vyhledávání v historii vyřešených požadavků a nalezení požadavku, v rámci kterého se řešil podobný či stejný problém. Pracovník Service Desku tedy může nejdříve prozkoumat předchozí požadavky, které LLM identifikovalo jako podobné, a případně využít již existující znalosti a možná řešení z historie.

Pracovník Service Desku může z již vyřešených požadavků s podobným cílem čerpat informace, předvídat budoucí problémy a rovnou na ně daného uživatele upozornit. Toto využití LLM opět poskytuje více informací pro pracovníky Service Desku a urychluje řešení požadavku.

Požadavky

Na základě uskutečněných setkání a analýzy aktuálního stavu a identifikovaných nedostatků byly vytvořeny (a výše popsány) případy užití, ze kterých vychází seznam požadavků níže. Požadavky jsou v následujících tabulkách kategorizované pomocí metodologie MoSCoW. V této metodologii jsou požadavky členěny do následujících čtyř kategorií:

- M — Must Have — Musí mít
- S — Should Have — Měl by mít
- C — Could Have — Mohl by mít
- W — Won't Have — Nebude mít

Kategorie M označuje základní a kritické požadavky, které navrhovaný systém mít musí. Kategorie S označuje požadavky, které jsou důležité, případně přidávají velkou hodnotu, a systém by je měl splňovat, ale ne nutně pro prvotní verzi. Do kategorie C spadají požadavky, které jsou žádané, ale mají nízkou prioritu a malý dopad na výsledný systém. Také se označují jako "nice-to-have". Poslední kategorie W obsahuje požadavky, které byly identifikovány jako nejméně důležité a které případně systém nebude naplňovat.

Dopadové požadavky

Dopadové požadavky zachycují přínosy zavedení AI asistenta do prostředí NRP v rámci projektu OS II.

ID	Název	Typ požadavku
D01	Snížení zátěže	M
D02	Zkrácení doby reakce	S
D03	Základní uživateslká podpora 24/7	M
D04	Zrychlení vyhledávání informací	C
D05	Zvýšení přehlednosti RT	C

Požadavek D01: Snížení zátěže

AI asistent poskytuje řešení a odpovídá uživatelům na základní nebo opakující se otázky a problémy. Snižuje tedy počet požadavků, které musí být obslouženy lidskými pracovníky Service Desku, kteří se tak mohou věnovat pokročilejším nebo specifičtějším požadavkům.

Požadavek D02: Zkrácení reakční doby
AI asistent analyzuje příchozí požadavky na pracoviště Service Desku. Na jednoduché otázky a problémy uživatelů odpovídá rovnou, čímž zkracuje průměrnou reakční dobu uživatelské podpory a poskytuje tak uživateli odpovědi rovnou bez výrazné časové prodlevy.

Požadavek D03: Základní uživatelská podpora 24/7

AI asistent reaguje a odpovídá na dotazy uživatelů neustále. Základní podpora je tedy zajištěna i mimo pracovní dobu uživatelské podpory.

Požadavek D04: Zrychlení vyhledávání informací

AI asistent poskytne rozhraní pro Service Desk pro sémantické vyhledávání ve svých znalostních databázích. Stává se tak jedním z prostředků, jak Service Desk může rychle získávat informace pro řešení uživatelských požadavků.

Požadavek D05: Zvýšení přehlednosti RT

Automatizace klasifikace požadavků do RT front povede k snížení počtu špatně zařazených požadavků do RT front.

Funkční požadavky

Funkční požadavky definují hlavní identifikované vlastnosti, které by systém AI asistenta měl splňovat.

ID	Název	Typ požadavku
F01	Konverzace angličtina	M
F02	Konverzace čeština	S
F03	Automatické vyhledávání v dokumentaci	M
F04	L0 a L1 support na webových stránkách	M
F05	L0 a L1 support v rámci systému RT	M
F06	Kategorizace příchozích požadavků do RT	M
F07	Správa externích zdrojů vědomostí LLM	C
F08	Validace úplnosti metadat	C
F09	Identifikace neobvyklých hodnot	C
F10	Textový popis chyb během validace dat	C
F11	Metadatové modely jako zdroj informací	S
F12	Eskalace na L1 podporu	S
F13	Indikace stavu při komunikaci na webu	C

Požadavek F01: Jazyk konverzace angličtina

Pokud uživatel komunikuje s AI asistentem v anglickém jazyce, pak AI asistent odpovídá rovněž v anglickém jazyce. Pokud ve svých odpovědích AI asistent předává uživateli odkaz na dokumentaci a tato dokumentace existuje v anglické verzi, pak předává odkaz na anglickou verzi.

Požadavek F02: Jazyk konverzace čeština

Pokud uživatel komunikuje s AI asistentem v českém jazyce, pak AI asistent odpovídá rovněž v českém jazyce. Pokud ve svých odpovědích AI asistent předává uživateli odkaz na dokumentaci a tato dokumentace existuje v české verzi, pak předává odkaz na českou verzi.

Požadavek F03: Automatické vyhledávání v dokumentaci

AI asistent vyhledává relevantní části dokumentace pro nově příchozí požadavky do systému RT. Pokud AI asistent dokáže odpovědět na daný požadavek, odkazy na vyhledáné části dokumentace jsou součástí odpovědi uživateli (viz dále F04). Pokud LLM na požadavek odpovědět nedokáže, odkazy na vyhledané relevantní sekce uživatelské dokumentace jsou poskytnuty člověku (viz dále F05), který požadavek přebírá k dalšímu řešení.

Požadavek F04: L0 a L1 support na webových stránkách

AI asistent poskytuje základní L0, a částečně L1, podporu přímo na webových stránkách uživatelské dokumentace. Dokáže uživateli odpovídat na otázky, např. typu “kde najdu v dokumentaci X?”, “jak udělám Y?” nebo na FAQ. Dokáže navíc uživatele přesměrovat na relevantní části webových stránek (např. jednotlivých repozitářů).

Požadavek F05: L0 a L1 support v rámci systému RT

Na příchozí požadavky (které se týkají prostředí OS2) do systému RT s otázkami či problémy, případně FAQ, odpovídá AI asistent sám, poskytuje tedy L0, a částečně L1, podporu. Zároveň dokáže do odpovědi přidat odkaz na relevantní části webových stránek (např. jednotlivých repozitářů). Systém tak nahrazuje L1 lidskou podporu, ale pro některé úkony podpory počítáme se zachováním L1 podpory, jejíž úloha se změní na dozor nad AI asistentem a na řešení úloh, které vyžadují vykonání akce.

Požadavek F06: Kategorizace nově příchozích požadavků do RT

AI asistent automaticky zpracovává příchozí požadavky do systému RT, které interně využívá pracoviště Service Desku. AI asistent takovým požadavkům přiřadí štítek s názvem služby, které se daný požadavek týká, a zařadí ho do odpovídající RT fronty pro danou službu.

Požadavek F07: Správa externích zdrojů vědomostí LLM

Systém umožňuje správu externích zdrojů (např. dokumentů) v tzv. vektorové databázi, které jsou využívány v rámci systému pro dodání kontextu AI asistentovi a zvyšují kvalitu odpovědí. Externí zdroje však mohou zastarávat (aktualizace webových stránek / nově vzniklé nebo naopak smazané repozitáře / FAQ apod.) a je třeba přidávat nové zdroje, a naopak neaktuální odebírat. Externí zdroj může obsahovat i způsob řešení příchozího (již vyřešeného) požadavku do RT systému.

Požadavek F08: Validace úplnosti metadat

AI asistent kontroluje validitu dat během procesu vkládání dat do repozitáře. Repozitář má definovaný metadatový model a asistent kontroluje, zda vkládaná data splňují definované náležitosti daným metadatovým modelem.

Požadavek F09: Identifikace neobvyklých hodnot

AI asistent během procesu vkládání dat do repozitáře kontroluje, zda vkládaná data neobsahují chybějící, prázdné nebo neočekávané hodnoty podle metadatového modelu.

Požadavek F10: Textový popis chyb během validace dat

AI asistent poskytuje textové srozumitelné vysvětlení problémů vzniklých během vkládání dat do repozitáře. Jedná se chyby způsobené pokusem o vložení dat, které nejsou v souladu s metadatovým modelem daného repozitáře nebo obsahují chybějící, prázdné nebo neočekávané hodnoty. Souvisí s předchozími funkčními požadavky F08 a F09.

Požadavek F11: Metadatové modely jako zdroj informací

AI asistent využívá jako zdroj dat i metadatové modely jednotlivých repozitářů. Jedná se o nutnou prerekvizitu a souvisí s požadavky F08, F09 a F10.

Poznámka: Funkční požadavky F08, F09 a F10 jsou vedeny jako typ C, protože se jedná o pokročilejší požadavky, které navíc vyžadují již zprovozněný systém AI asistenta, ale také existenci jednotlivých datových repozitářů. Ostatním požadavkům se proto dává vyšší důležitost a zmiňované F08, F09 a F10 se plánují řešit v rámci projektu OS II později. Požadavek F11 je veden jako typ S, protože se jedná o nutnou prerekvizitu těchto požadavků.

Požadavek F12: Eskalace na L1 podporu
Pokud AI asistent nedokáže odpovědět, nedokáže vyhledat relevantní informace nebo uživatel není spokojen, eskaluje požadavek dále, tzn. na člověka. V případě řešení přes email je systém nastaven tak, že před odesláním je vygenerovaná odpověď zkontrolována lidským pracovníkem Service Desku. V případě webového rozhraní je uživatel přesměrován na Service Desk přes kontaktní údaje Service Desk nebo AI asistent připraví email s kontextem a požádá uživatele o doplnění a schválení odeslání emailu přímo na Service Desk k manuálnímu zpracování (řeší se přes RT, ale takový požadavek by již neměl být podruhé odbavován AI asistentem).

Požadavek F13: Indikace stavu AI asistenta při komunikaci na webu

AI asistent v interaktivní komunikaci s uživatelem přes webové rozhraní indikuje vhodným způsobem své stavy přijetí zprávy, její zpracování, generování odpovědi.

Výkonnostní požadavky

Výkonnostní požadavky definují nároky na rychlost.

ID	Název	Typ požadavku
V01	Odpověď na webu v řádu desítek sekund	M
V02	Odpověď v RT systému řádu desítek minut	M

Požadavek V01: Odpověď na webu v řádu desítek sekund

AI asistent integrovaný na webových stránkách poskytuje uživateli odpovědi v řádu desítek sekund.

Požadavek V02: Odpověď v RT systému řádu desítek minut

AI asistent integrovaný do systému RT generuje uživatelské odpovědi v řádu desítek minut.

Poznámka: U V02 je volnější časový limit na odpověď kvůli zpracování možných příloh (obrázek, PDF, …) odeslaných uživatelem. Zároveň se nejedná o konverzaci v reálném čase (na rozdíl od V01), ale prostřednictvím emailové komunikace. Odpověď v řádu sekund tedy není nutná, nicméně uživateli již dnes přijde emailem oznámení o přijetí a zpracovávání požadavku. Systém navíc může být nastaven tak, že generované odpovědi mohou být kontrolovány lidskými pracovníky Service Desku, časový limit pro odpověď pak vychází z jiných parametrů, nad rámec tohoto dokumentu.

Integrační požadavky

Integrační požadavky definují interoperabilitu systému AI asistenta s ostatními systémy v rámci OS II.

ID	Název	Typ požadavku
I01	Integrace na webových stránkách	M
I02	Integrace se systémem RT	M
I03	Integrace na repozitáře	C

Požadavek I01: Integrace na webových stránkách

AI asistent je integrován na vybrané webové stránky v rámci OS II. Předpokládá se integrace na webové stránky jednotlivých repozitářů, případně na další. AI asistent odpovídá na otázky a přesměrovává uživatele na adekvátní části dokumentace.

Požadavek I02: Integrace s RT

AI asistent je integrován do systému RT. Asistent klasifikuje příchozí požadavky do jednotlivých RT front a přidává k nim štítek (F06), určující o kterou službu se jedná. V rámci F03 a F05 se také předpokládá obohacování příchozích požadavků o relevantní informace od AI asistenta a poloautomatické odpovídání na vybrané příchozí požadavky.

Požadavek I03: Integrace na repozitáře

AI asistent je schopen získávat aktuální veřejně dostupná data o vybraných (spravovaný seznam webových rozhraní repozitářů s připravnými dotazy pro vyhledání určitých dat) repozitářích, pokud to je relevantní zdroj informace pro tvorbu odpovědi pro uživatele.

Bezpečnostní požadavky

Bezpečnostní požadavky zachycují nutné náležitosti týkající se ochrany uživatelských dat a fyzických zdrojů, na kterých se AI asistent plánuje provozovat.

ID	Název	Typ požadavku
B01	Ochrana dat uživatelů	M
B02	Ochrana provozních zdrojů	M

Požadavek B01: Ochrana dat uživatelů

Pro navrhovaný systém budou navržena a realizována nezbytná bezpečnostní opatření pro zajištění ochrany citlivých dat při automatickém i manuálním zpracování. Součástí těchto opatření bude například izolace dat a výpočetních prostředků, nastavení a řízení přístupů a bezpečnostní audit činností systému. AI asistent nebude využívat citlivá uživatelská data ke tvorbě odpovědí na nově příchozí požadavky.

Požadavek B02: Ochrana provozních zdrojů

AI asistent odpovídá pouze na relevantní dotazy týkajících se uživatelské dokumentace nebo relevantních problémů. Pro ochranu fyzických zdrojů se nechová jako běžný univerzální chatbot dostupný pro libovolnou konverzaci, ale je zajištěno a kontrolováno, že asistent odpovídá pouze

na dotazy v rámci kontextu OS II. V případě přetížení sníží AI asistent svou dostupnost pro uživatele, který generuje více než stanovené množství dotazů za minutu.

Compliance požadavky

Compliance požadavky zachycují právní a přidružené náležitosti tak, aby systém AI asistenta fungoval podle nejnovějších zákonů a nařízení.

ID	Název	Typ požadavku
C01	Licence LLM modelu	M
C02	Podmínky užití	M
C03	Upozornění uživatele o využití LLM	M
C04	Citlivá data	M
C05	Auditní stopa	S

Požadavek C01: Licence LLM modelu

Zvolený LLM model pro asistenta je používán v rámci své definované licence a nejsou porušeny podmínky použití.

Požadavek C02: Podmínky užití

AI asistent má jasně definované podmínky užití, které jsou uživateli jasně a veřejně kdykoliv dostupné.

Požadavek C03: Upozornění uživatele o využití LLM

Uživatel je jasně informován o tom, že na jeho dotaz odpovídá AI asistent a ne skutečný člověk. Uživatel má navíc právo eskalovat své požadavky na člověka, pokud není s odpovědmi AI asistenta spokojen nebo nejsou odpovědi dostačující.

Požadavek C04: Citlivá data

S citlivými uživatelskými daty je nakládáno podle zákona a podle definovaných podmínek užití.

Požadavek C05: Auditní stopa

AI asistent loguje informace o své činnosti od statistických informací, přes uživatelské prompty a odpovědi, až po selhání a chyby.

Validace splnění požadavků

Pro sledování splnění dopadových požadavků navrhujeme sledovat následující metriky v rámci AI asistenta:

- snížení počtu tiketů řešených člověkem,
- zkrácení průměrné doby první odpovědi,
- počet eskalovaných konverzací s AI asistentem a celkový počet konverzací.

Pro budoucí možnou validaci požadavků popisujeme dva možné způsoby komunikace uživatele se systémem AI asistenta, ve kterých se odkazujeme na konkrétní požadavky definované v této kapitole a navrhujeme tedy způsoby pro pozdější ověření, že byly dané požadavky splněny. Jedná se o případ komunikace s AI asistentem skrze webové stránky a případ komunikace s asistentem skrz uživatelskou podporu. Poslední případ popisuje správu systému AI asistenta.

Případ 1 - Webový Chatbot

AI asistent je dostupný na vybraných webových stránkách, např. jednotlivých repozitářů (I01). Uživatel se může ptát AI asistenta na otázky (F01, F02) z prostředí OS2, konkrétního repozitáře, FAQ (F04), či využívat AI asistenta k validaci dat (F08-F11). Uživatel je zároveň upozorněn, že se ptá AI asistenta a ne živého člověka (C03), případně prostudovat podmínky užití (C02). AI asistent poskytuje uživateli odpovědi v reálném čase (V01). Uživatel navíc může být AI asistentem přesměrován na konkrétní stránky dokumentace či jiné relevantní stránky, pokud je to v rámci odpovědi a vyhovění požadavku žádoucí (F03). Na dotazy, které nejsou z prostředí OS2, AI asistent neodpovídá a chrání tak výpočetní zdroje proti zneužití (B02).

Případ 2 - Kontakt pomocí emailu

Uživatel využije formu emailu pro kontakt pracoviště Service Desku. Tento email je automaticky zpracováván systémem RT, do kterého je integrovaný AI asistent (I02). Ten příchozí email zpracuje, vyhodnotí o jakou službu se jedná a do jaké RT fronty má být tento požadavek zařazen (F06). Pokud se jedná o jednoduchý dotaz či část opakovaný problém a AI asistent dokáže odpovědět, asistent vygeneruje v řádu desítek minut (V02) odpověď s potenciálním řešením, opět případně s relevantní částí uživatelské dokumentace (F05). Odpověď může být před odesláním volitelně zkontrolována lidským pracovníkem Service Desku. V odpovědi je jasně indikováno, že odpověď byla vygenerována AI asistentem (C03) a obsahuje odkaz na podmínky užití (C02). Pokud dotaz či problém není vyřešen nebo je uživatel nespokojen, má možnost eskalovat požadavek na člověka, opět pomocí emailové odpovědi. S uživatelskými daty v zaslaném požadavku je nakládáno podle platných právních předpisů ČR a podmínek užití (C04).

Případ 3 - Správa systému AI asistenta

Webové rozhraní pro správce systému AI asistenta je dostupné pro jeho správu. V tomto systému může přidávat nové zdroje externích znalostí AI asistenta (např. PDF), či případně mazat již zastaralé zdroje (F07). Dále je možné provést aktualizaci informací, které vychází z webových stránek, např. po aktualizaci webových stránek s uživatelskou dokumentací je třeba aktualizovat záznamy v databázi externích znalostí, které z těchto webových stránek vychází (F07). Do databáze může být navíc přidáno řešení/odpověď na vybrané dotazy a požadavky, typicky ty, které se opakovaně objevují v RT požadavcích. S takovými daty je opět nakládáno podle platných právních předpisů ČR a podmínek užití (C04).

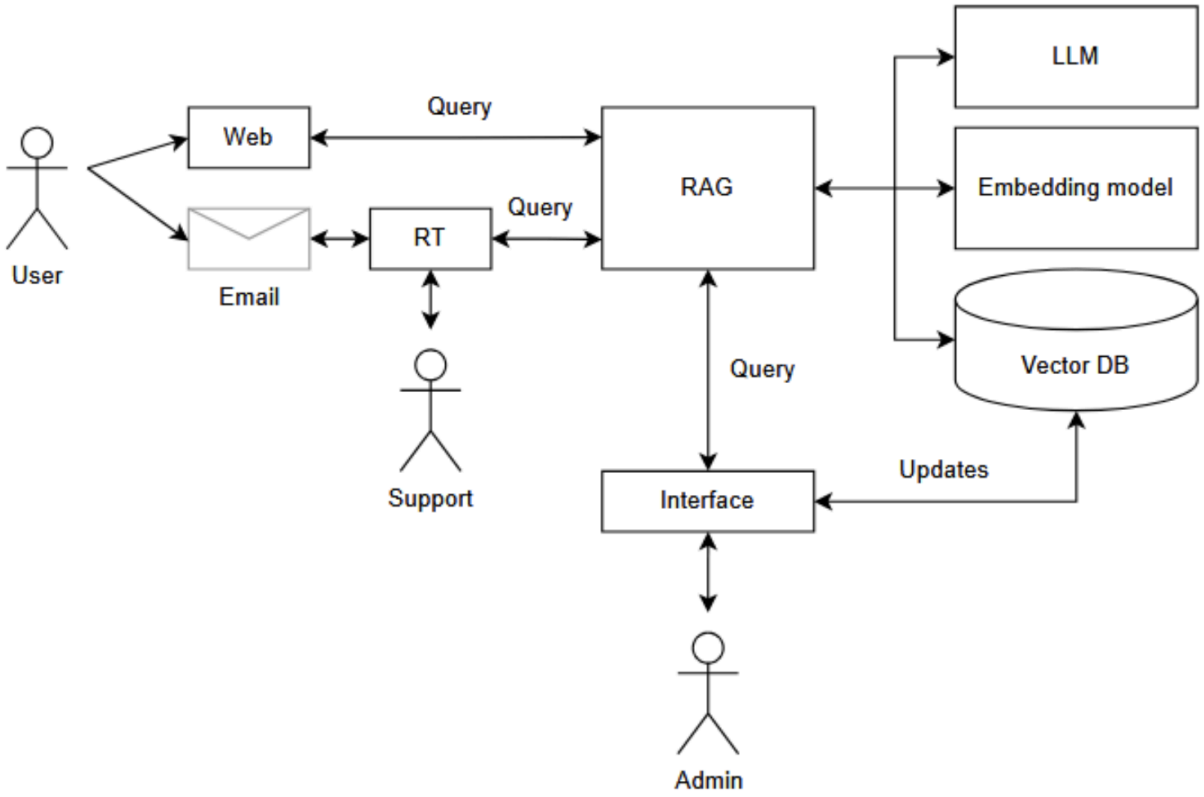
Návrh

Podle sesbíraných požadavků a případů užití se předpokládá v rámci projektu využití technologie Retrieval-Augmented Generation (RAG). Nejdříve jsou pomocí LLM vytvořeny tzv. "*embeddingy*", které jsou vloženy do vektorové databáze. Při dotazování LLM se nejdříve v této vektorové databázi vyhledají související zdroje (např. PDF, konkrétní sekce webové dokumentace) a předají se s původním uživatelským dotazem vybranému LLM modelu. Vygenerované odpovědi jsou poté informačně bohatší, přesnější a dovolují LLM pracovat s informacemi, které nebyly v době trénování dostupné.

Díky využití technologie RAG bude AI asistent umět řešit uživatelské dotazy a problémy specifické pro prostředí NRP, aniž by musel být LLM model dotrénováván. Další výhoda tohoto návrhu je také v tom, že externí zdroje mohou být do vektorové databáze v čase průběžně přidávány, vylepšovány a zastaralé zdroje naopak odebírány, opět bez drahého dotrénovávání LLM modelů.

Architektura

První verze navrhované architektury je níže na Obrázku 1. Hlavní částí systému je technologie RAG, která je dále napojená na vektorovou databázi, embedding model a samotný LLM model. Embedding model slouží ke zpracovávání externích zdrojů znalostí (relevantní webové stránky, webová uživatelská dokumentace, FAQ, apod.), ze kterých vytváří *embeddingy* (číselné vektory). Tyto vektory jsou, společně s původním zdrojem informací, uloženy do vektorové databáze. Příchozí dotaz (query) je nejdříve zpracován embedding modelem, který z něho rovněž vytvoří embedding. Ve vektorové databázi je následně vyhledáno *K* nejbližších embeddingů a jim odpovídající zdrojová data jsou přidána k původnímu dotazu. Tento obohacený dotaz je poté předán samotnému LLM k vyhodnocení a vygenerovaný výstup je vrácen jako odpověď na původní dotaz.



Obrázek 1: První verze navrhované architektury systému AI asistenta.

Administrátoři mají možnost spravovat externí zdroje znalostí (část *Interface*). Pomocí tohoto rozhraní mohou být do vektorové databáze vkládány nové dokumenty (např. PDF) nebo aktualizovat informace z relevantních webových stránek (např. pokud bude zveřejněna nová verze webové dokumentace). Nebo naopak odebírat staré, už neaktuální, zdroje znalostí. Předpokládá se také vytěžování informací z již vyřešených předchozích požadavků ze systému RT, např. admin doplní vyřešený RT požadavek s často se opakující otázkou/problémem.

Vstupy systému AI asistenta vytvářejí samotní uživatelé, kteří s AI asistentem komunikují buď skrz webovou stránku, kde je AI asistent integrovaný, nebo pošlou email na pracoviště Service Desku, kde je jejich požadavek zpracováván AI asistentem. Webová stránka, případně systém RT nebo jeho (potenciální budoucí) nadstavba, komunikuje pomocí API se systémem RAG, popsaným výše.

Navrhovaný systém se navíc plánuje implementovat tak, aby byl jednoduše nasaditelný a škálovatelný. Tedy tak, aby se daly jednotlivé instance systému AI asistenta jednoduše nasazovat, např. jako Docker kontejnery. Systém AI asistenta tak bude moci být nasazen buďto jako centrální instance nebo separátní instance pro jednotlivé (či vybrané) repozitáře. Výhodou separátních instancí může být kontext pro AI asistenta pouze v rámci daného repozitáře, a tím

pádem snížení halucinací, zvýšení kvality odpovědí asistenta a menší velikost vektorové databáze.

Shrnutí

Tento dokument shrnuje metodologii a způsob sběru požadavků pro vytvoření AI asistenta jakožto uživatelské podpory v rámci průřezové aktivity jazykové modely v projektu OS II. Na základě analýzy aktuálního stavu a sběru požadavků od relevantních pracovníků v souvislosti s problematikou projektů NRP a OS II jsme identifikovali celkem sedm případů užití pro použití navrhovaného AI asistanta, který je plánovaným výstupem pro podporu týmu odborníků z uživatelské podpory. Na základě nasbíraných informací jsme identifikovali a definovali dopadové, funkční, výkonnostní, bezpečnostní a compliance požadavky.

Zároveň jsme provedli počáteční průzkum vhodných technologií v oblasti moderních technologií strojového učení a umělé inteligence a v souvislosti s potřebami na zpracování dokumentace a vyhledávání obsažených faktů jsme pro návrh prvního prototypu AI asistenta vybrali konkrétně technologii RAG (Retrieval Augmented Generator), od které se odvíjí i první verze navrhované architektury AI asistenta.