DOI: 10.35191/medsoft_2025_1_37_beranek

VELKÉ JAZYKOVÉ MODELY A AGENTNÍ SYSTÉMY V MEDICÍNĚ: PRINCIPY, SOUČASNÉ APLIKACE A PERSPEKTIVY

Pavel Beránek, Vojtěch Merunka

Abstrakt

Velké jazykové modely (Large Language Models, LLM) představují významný mezník ve vývoji umělé inteligence a otevírají nové možnosti v oblasti medicíny a zdravotnictví. Díky schopnosti porozumět přirozenému jazyku a generovat odborně relevantní texty mohou podporovat klinická rozhodnutí, usnadňovat komunikaci s pacienty a zefektivňovat administrativní procesy. Tento příspěvek nabízí srozumitelný úvod do principů fungování LLM a jejich přizpůsobení na medicínské úlohy, doplněný systematickým přehledem současných aplikací vycházejícím z rešerše odborné literatury. Zvláštní pozornost je věnována také využití LLM v roli autonomních agentů a jejich integraci do multiagentních systémů, které mohou simulovat spolupráci multidisciplinárních týmů, koordinovat komplexní úlohy a provádět křížové ověřování výsledků. Diskutovány jsou přínosy, omezení a etické aspekty těchto technologií, spolu s doporučeními pro jejich bezpečné a efektivní začlenění do lékařské praxe.

Klíčová slova

velké jazykové modely, umělá inteligence v medicíně, agentní systémy, multiagentní systémy, klinická podpora rozhodování

1 Úvod

Umělá inteligence (AI) v posledních letech zásadně proměnila řadu odvětví – od průmyslové výroby přes finance až po vzdělávání. V medicíně se zájem o využití AI objevuje již několik desetiletí, zejména v oblastech podpory diagnostiky a analýzy obrazových dat. Tradiční algoritmy strojového učení však narážely na limity: vyžadovaly rozsáhlou předpřípravu dat, byly úzce specializované a jejich výstupy nebyly snadno interpretovatelné lékařem. [1]

Nástup hlubokého učení (deep learning) po roce 2012 otevřel cestu k modelům schopným dosahovat lidsky srovnatelných výsledků v oblasti zpracování medicínských obrazů, detekce patologických struktur či predikce klinických ukazatelů. Přesto zůstávaly tyto modely většinou „černými skříňkami", které nedokázaly lékaři poskytnout vysvětlení ani širší kontext. [2, 3]

Zlom přinesl až prudký rozvoj velkých jazykových modelů (Large Language Models, LLM) po roce 2018. Tyto modely poprvé umožnily efektivně zpracovávat a generovat přirozený jazyk, a tedy komunikovat s uživatelem srozumitelným způsobem. V medicíně to znamená možnost propojení pokročilých algoritmů s každodenní klinickou praxí – od podpory rozhodování a zpracování dokumentace až po komunikaci s pacienty. [4, 5]

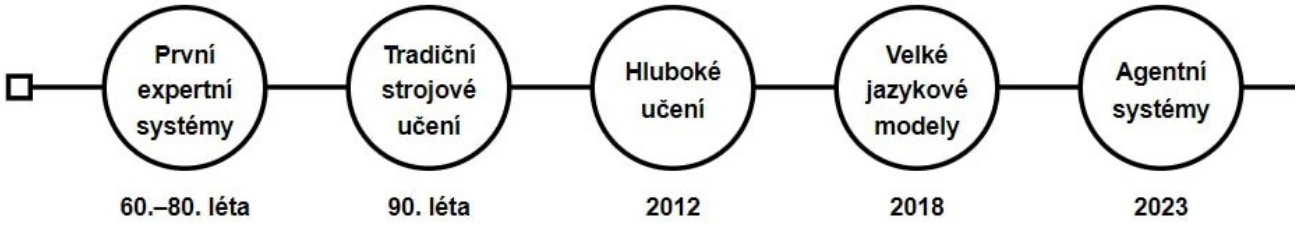
Rostoucí dostupnost výpočetních kapacit a otevřených datových sad, spolu s tlakem na efektivitu zdravotnictví a personalizovanou péči, vytvořila prostředí, ve kterém je diskuse o aplikaci AI aktuálnější než kdy dříve. Současně se však objevují otázky bezpečnosti, spolehlivosti a etické odpovědnosti. V tomto kontextu představují velké jazykové modely nejen technologickou inovaci, ale i výzvu, jak je začlenit do medicíny způsobem, který je bezpečný, důvěryhodný a skutečně přínosný pro pacienty i zdravotnický personál. [6, 7]

Velké jazykové modely představují zásadní posun oproti dřívějším přístupům k umělé inteligenci v medicíně. Na rozdíl od tradičních algoritmů, které byly úzce zaměřené na jeden typ úlohy (např. analýzu laboratorních hodnot nebo radiologických snímků), dokážou LLM pracovat univerzálně s přirozeným jazykem – tedy s hlavním prostředkem, kterým lékaři zaznamenávají, sdílejí a předávají informace. Jejich schopnost porozumět odborným textům, shrnout rozsáhlou dokumentaci či generovat srozumitelná vysvětlení znamená, že se mohou stát přirozeným prostředníkem mezi složitými daty a klinickou praxí. Díky své flexibilitě lze navíc stejný model využít pro širokou škálu úloh – od podpory klinického rozhodování přes administrativní zátěž až po komunikaci s pacientem. Právě tato univerzálnost, kombinovaná s rychlostí a stále rostoucí kvalitou výstupů, činí LLM průlomovou technologií, která má potenciál zásadně proměnit způsob, jakým je medicína vykonávána a organizována. [8–10]

Cílem tohoto přehledového článku je poskytnout lékařům a zdravotnickým pracovníkům srozumitelný úvod do problematiky velkých jazykových modelů a jejich využití v medicíně. Text je rozdělen do několika částí: nejprve představuje základní principy fungování LLM a možnosti jejich přizpůsobení na medicínské úlohy, poté nabízí přehled současných aplikací a perspektiv. Samostatná pozornost je věnována konceptům kontextového inženýrství, Retrieval-Augmented Generation a multiagentních systémů. Závěrečné kapitoly shrnují přínosy, omezení a etické aspekty těchto technologií a nabízejí doporučení, jak je bezpečně a efektivně začlenit do klinické praxe.

2 Principy velkých jazykových modelů

Velké jazykové modely (Large Language Models, LLM) jsou základem současného rozvoje umělé inteligence, protože umožňují počítačům pracovat s přirozeným jazykem podobně, jako to dělá člověk. Jejich schopnosti vycházejí z učení na obrovském množství textových dat, díky čemuž dokážou rozpoznávat významové souvislosti, vytvářet souvislé odpovědi a přizpůsobovat svůj výstup kontextu. Pro medicínu je podstatné, že tyto modely dokážou nejen reprodukovat znalosti z odborných textů, ale také je převádět do srozumitelné podoby a reagovat na konkrétní potřeby uživatele. Abychom lépe porozuměli tomu, proč jsou LLM v klinické praxi tak perspektivní, je užitečné se podívat na jejich základní principy fungování, způsoby přizpůsobení na medicínské úlohy a roli, kterou v jejich využití hraje kontextové inženýrství či propojení s externími zdroji dat.



Obrázek 1 – Časová osa vývoje AI technologií. Za posledních 60 let prošel obor od systémů založených na znalostních bázích odborníků až k agentům, kteří dokážou spolupracovat v týmu na zadaných úkolech a využívat externí data.

DOI: 10.35191/medsoft_2025_1_37_beranek

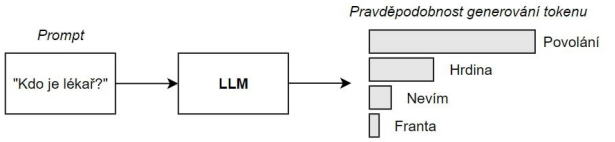
2.1 Co jsou velké jazykové modely a jak fungují

LLM jsou speciální typ neuronových sítí trénovaných na obrovském množství textových dat. Jejich základním principem je schopnost předpovídat další slovo (token) v textu na základě předchozího kontextu. Pokud tedy model dostane větu „Pacient má bolesti na …", jeho úkolem je odhadnout, že nejpravděpodobnějším pokračováním bude například „hrudi".

Formálně lze tento úkol vyjádřit jako výpočet podmíněné pravděpodobnosti:

$$P(w_t | w_1, w_2, \ldots, w_{t-1})$$

kde w_t je slovo, které má být předpovězeno, a $w_1, w_2, \ldots, w_{t-1}$ jsou předchozí slova ve větě. [11]



Obrázek 2 – Generování dalšího slova na základě existující množiny slov. Počáteční množina se nazývá výzva (prompt).

Základem LLM jsou umělé neuronové sítě – výpočetní struktury složené z vrstev propojených „neuronů". Každý neuron přijímá vstupy (slova v případě LLM), provádí jednoduchou matematickou operaci a výsledek posílá dál. Takto se z řady jednoduchých výpočtů skládá komplexní chování.

Pro velké jazykové modely se využívá především tzv. transformátorová architektura (Transformer), která se stala revoluční inovací. Její jádro tvoří mechanismus self-attention – tedy „pozornost k sobě samému". Tento mechanismus umožňuje modelu dynamicky rozhodovat, kterým částem textu má při zpracování věnovat více pozornosti.

Matematicky lze self-attention zapsat takto:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

kde Q (queries) představuje informaci, kterou slovo „hledá", K (keys) informaci, kterou slovo „nabízí", a V (values) samotný obsah slova, který se má předat dál. Tyto reprezentace mají podobu matic a zachycují různé aspekty vstupního textu, přičemž d_k je normalizační faktor. Model porovnává dotaz (Q) každého slova s „klíči" (K) všech ostatních slov a podle výsledku určuje, kolik pozornosti má danému slovu věnovat. Vážená kombinace těchto hodnot (V) vytváří novou reprezentaci slova, obohacenou o kontext celé věty. [4]

Díky této operaci dokáže model například v delší zprávě rozpoznat, že slovo „bolest" souvisí s „hrudí" a „EKG", i když se tato slova vyskytují ve větě daleko od sebe. Mechanismus pozornosti tak umožňuje automaticky propojit anamnestické údaje, laboratorní výsledky a aktuální symptomy do smysluplného celku.

Aby s textem mohla neuronová síť pracovat, musí být slova převedena do číselné podoby. To se děje pomocí tzv. embeddingů – vektorových reprezentací slov. Každé slovo nebo token je vyjádřeno jako bod v mnohorozměrném prostoru, kde blízkost vektorů odpovídá významové podobnosti.

Formálně:

$$w \rightarrow \overrightarrow{e}_w \in R^d$$

kde $\overrightarrow{e}_w$ je vektorová reprezentace (embedding) slova w v prostoru o dimenzi d. [12]

2.2 Trénink na rozsáhlých datech, parametry, škálování výkonu

Velké jazykové modely získávají své schopnosti díky tréninku na obrovském množství textových dat – zahrnujících knihy, vědecké články, webové stránky nebo databáze odborných textů. Během tréninku se model učí statistické vzorce v jazyce: tedy jaká slova se vyskytují společně, jak jsou uspořádána a jaké významové souvislosti mezi nimi existují.

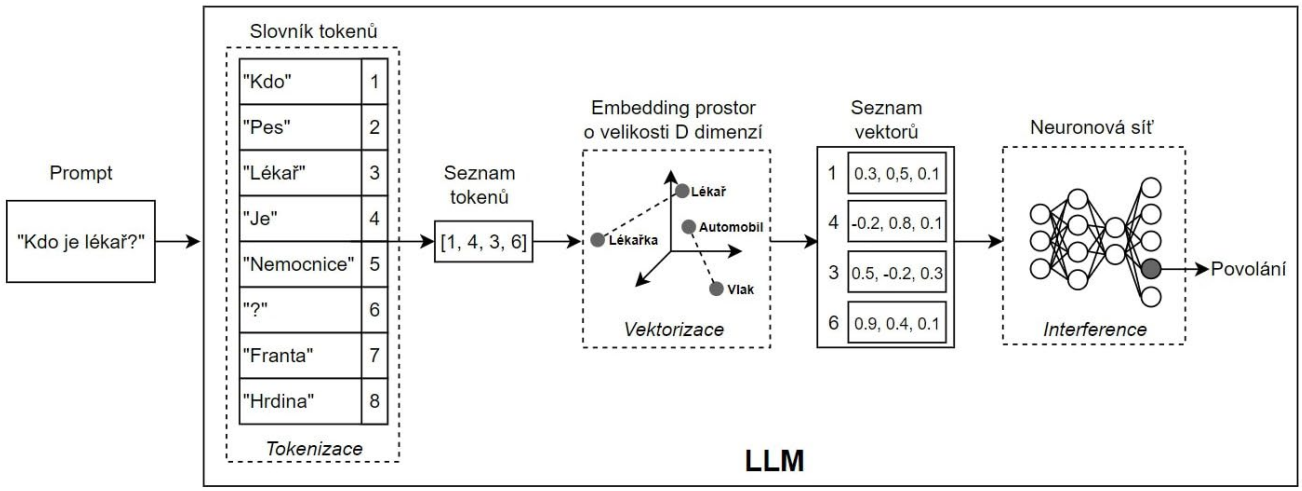
Základem je proces minimalizace chyby při předpovědi dalšího slova. Pokud má model předpovědět následující token w_t na základě předchozích slov $w_1, \ldots, w_{t-1}$ snaží se maximalizovat pravděpodobnost:

$$\prod_{T}^{t=1} P_\theta(w_1, \ldots, w_{t-1})$$

kde θ označuje parametry modelu (váhy neuronové sítě).

Parametry představují číselné hodnoty, které určují chování neuronové sítě. Moderní modely jich mají miliardy až stovky miliard – například GPT-3 měl 175 miliard parametrů. Čím více parametrů, tím jemněji dokáže model zachytit jazykové nuance a odborné souvislosti, ale tím větší jsou i nároky na výpočetní výkon a paměť. [5]

Tento vztah se označuje jako škálování výkonu: s růstem velikosti dat a počtu parametrů roste i kvalita výsledků, i když za cenu exponenciálně rostoucích nákladů. Prakticky to znamená, že zatímco první generace modelů mohla být trénována na běžných serverech, dnešní LLM vyžadují rozsáhlé výpočetní clustery s tisíci grafickými procesory (GPU). [13, 14]



Obrázek 3 – Postupné zpracování vstupního textu: nejprve převod na seznam tokenů (číselná reprezentace slov), poté na vektory zachycující význam slov. Výsledná matice čísel se vloží do neuronové sítě, která predikuje další slovo v pořadí.

DOI: 10.35191/medsoft_2025_1_37_beranek

2.3 Inženýrství výzev

Velké jazykové modely generují odpovědi na základě vstupního dotazu, tzv. promptu (výzvy). Způsob, jakým je tento prompt formulován, má zásadní vliv na kvalitu a spolehlivost výsledku. I když je model vysoce sofistikovaný, reaguje na přesně zadaný kontext – nejednoznačně nebo vágně položené otázky mohou vést k neúplným nebo zavádějícím odpovědím.

Existuje několik osvědčených technik, jak formulovat dotazy pro velké jazykové modely, tak, aby výsledky byly co nejpřesnější a nejvíce klinicky užitečné. Následující seznam obsahuje jen základní techniky psaní výzev. [15]

1. **Nulový, jednovýstřelový a vícevýstřelový prompt** – Modelu položíme dotaz bez příkladu – vhodné pro jednoduché otázky (bez výstřelu), doplníme jeden příklad – model pak snáze napodobí požadovaný styl (jednovýstřelový prompt) nebo uvedeme několik příkladů a model se naučí vytvářet odpovědi ve stejném formátu (vícevýstřelové prompty). [5]
2. **Omezování výstupu a formátování** – Pomocí instrukcí můžeme určit, v jakém formátu má být odpověď. To je užitečné zejména v administrativě a dokumentaci. Př.: „Shrň anamnézu pacienta do 5 odrážek a přidej doporučení na další vyšetření. "
3. **Rolové instrukce** – Model reaguje odlišně podle toho, jakou roli mu přisoudíme. Pokud ho požádáme, aby vystupoval jako „zkušený kardiolog" nebo „učitel medicíny pro studenty", upraví jazyk i detailnost výstupu. Př.: „Jako zkušený internista vysvětli medikovi indikace k provedení echokardiografie."
4. **Řetězení myšlenek** – Pokud model požádáme, aby vysvětloval svůj postup krok za krokem, výrazně to zvyšuje transparentnost i kvalitu výsledku. Místo jednovětné odpovědi dostane lékař přehled logických kroků vedoucích k závěru. Př.: „Vysvětli krok za krokem diferenciální diagnostiku dušnosti u pacienta se srdečním selháním." [16, 17]
5. **Iterativní upřesňování** – Někdy první odpověď modelu není dostatečná. V takovém případě je vhodné položit doplňující otázku nebo upřesnit zadání. Tím se simuluje dialog, ve kterém se odpověď postupně zpřesňuje. Př.: „Uveď stejné doporučení, ale s ohledem na pacienta s chronickým onemocněním ledvin." [18, 19]

3 Přizpůsobení velkých jazykových modelů pro medicínské úlohy

Ačkoli obecné velké jazykové modely dokážou generovat text s vysokou úrovní plynulosti a relevance, jejich využití v medicíně vyžaduje další přizpůsobení. Klinické prostředí je totiž specifické – pracuje s odbornou terminologií, citlivými daty a vyžaduje maximální přesnost a spolehlivost. Proto vznikají specializované postupy, jak LLM „naladit" na medicínské úlohy: od fine-tuningu na odborných datech, přes vývoj doménově orientovaných modelů, až po využití retrieval-augmented generation (RAG), které propojuje model s externími databázemi. Klíčovou výzvou je také jazyková bariéra, neboť většina současných modelů je trénována především na anglických textech, zatímco klinická praxe často vyžaduje práci s lokálními jazyky.

3.1 Fine-tuning

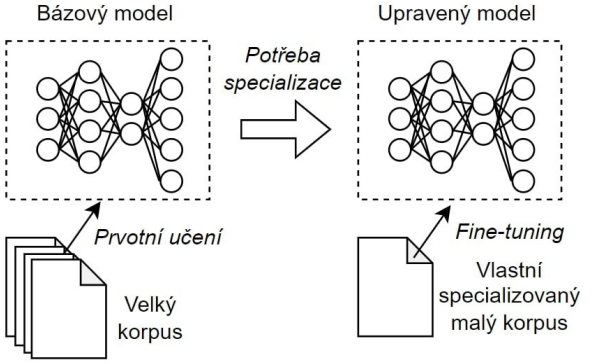
Fine-tuning je proces, při kterém se již předtrénovaný jazykový model dále učí na specializované množině dat, aby se přizpůsobil konkrétní oblasti použití. Zatímco původní model (např. GPT nebo BERT) je natrénován na obrovském množství obecného textu, fine-tuning ho „doškolí" tak, aby rozuměl specifické terminologii a kontextu medicíny.

Technicky jde o to, že se parametry neuronové sítě θ dále optimalizují nikoli na obecném jazyce, ale na doménově specifických datech. Cílem je minimalizovat chybu na specializovaném datasetu D_{med}:

$$E_{(x,y)\in D_{\text{med}}} L\big(f_\theta(x), y\big)$$

kde $f_\theta(x)$ je výstup modelu pro vstupní text x, y je správná odpověď (např. diagnóza, kategorizace či anotovaný text) a L je ztrátová funkce (loss function).

Fine-tuning na databázích jako PubMed nebo MIMIC-III umožňuje modelu lépe porozumět klinickým záznamům a odborným článkům – díky tomu dokážou doučené modely přesněji odpovídat na medicínské otázky než obecné modely. V klinické praxi může být model doučen například na anonymizovaných záznamech pacientů, aby se naučil správně sumarizovat zdravotní dokumentaci nebo podporovat rozhodování v konkrétním oboru (např. kardiologie, onkologie). [20]



Obrázek 4 – Fine-tuning je proces úpravy již natrénovaného modelu s cílem doladit jej pro specifické potřeby, například odbornou terminologii nebo konkrétní typ odpovědí.

Výsledkem doučení na doménově specifických datech jsou tzv. specializované jazykové modely, které jsou navrženy přímo pro medicínské aplikace. Tyto modely nevycházejí „z nuly", ale staví na obecně předtrénovaných architekturách, které jsou následně doladěny na odborných datech – nejčastěji z klinických databází a biomedicínské literatury.

- **BioBERT** – vychází z architektury BERT a je dále trénován na biomedicínských textech z databáze PubMed. Díky tomu lépe rozpoznává odborné pojmy a vztahy mezi nimi, což se využívá při vyhledávání informací, extrakci znalostí nebo při tvorbě znalostních grafů. [21]
- **PubMedBERT** – model natrénovaný výhradně na datech z PubMedu, tedy vědeckých článcích v oblasti medicíny a biologie. Dosahuje vysoké přesnosti v úlohách, které vyžadují hluboké porozumění biomedicínskému textu. [22]
- **ClinicalBERT** – specializovaný na klinické záznamy, zejména z databáze MIMIC-III. Umí pracovat s typicky neúplnými, zkratkovitými a různorodými záznamy pacientů. To ho činí vhodným nástrojem pro sumarizaci dokumentace a podporu klinického rozhodování. [23]
- **BioGPT** – založený na architektuře GPT, doladěný na biomedicínských textech. Je zaměřen na generování přirozeného textu a odpovídání na biomedicínské otázky. [24]
- **Med-PaLM** – model vyvinutý společností Google, navržený pro zodpovídání medicínských dotazů v souladu s klinickými standardy. V testech ukázal výkon srovnatelný s lidskými odborníky v některých typech otázek. [25]

Tyto modely jsou sice přesnější než obecné LLM, ale zůstávají omezené rozsahem dat, na nichž byly natrénovány. PubMed či MIMIC-III pokrývají jen část klinické praxe a často nereflektují lokální jazykové prostředí (například češtinu). Další výzvou je rychlé zastarávání dat – medicína se neustále vyvíjí a modely je nutné průběžně aktualizovat.

3.2 Retrieval-Augmented Generation (RAG)

Jednou z klíčových limitací velkých jazykových modelů je skutečnost, že jejich znalosti jsou omezené na data, na kterých byly natrénovány, tedy odpovědi nevycházejí z aktuálních odborných zdrojů, ale pouze z vnitřní paměti modelu. Pokud model narazí na otázku, která přesahuje jeho tréninková data, může vygenerovat odpověď, která působí věrohodně, ale není správná (tzv. halucinace). V medicíně je tento problém obzvlášť citlivý, protože nepřesné informace mohou přímo ovlivnit klinická rozhodnutí.

RAG představuje řešení tohoto nedostatku. Využívá princip propojení jazykového modelu s externí databází znalostí – například vědeckými články (PubMed), klinickými doporučeními (guidelines), nebo interními nemocničními protokoly. V praxi to funguje tak, že než model začne generovat odpověď, nejprve provede vyhledávání relevantních dokumentů, tyto texty „vloží" do svého kontextu a teprve poté vytvoří odpověď, která se opírá o nalezené zdroje.

Obrázek 5 – Schéma procesu Retrieval-Augmented Generation (RAG). Model nejprve vyhledá relevantní dokumenty v databázi, jejich obsah vloží do kontextu a poté vygeneruje odpověď podloženou citacemi odborných zdrojů.

Příkladem může být vstupní prompt lékaře: „Jaký je doporučený postup u léčby akutního infarktu myokardu u pacienta s kontraindikací k podání aspirinu?". Model vyhledá v databázi ESC (European Society of Cardiology) guidelines kapitolu týkající se léčby infarktu myokardu a antiagregační terapie a podá odpověď, např.: „Podle ESC Guidelines (2020) lze v případě kontraindikace k aspirinu zvážit podání alternativního antiagregačního léčiva, např. clopidogrelu. [zdroj: ESC Guidelines for the management of acute coronary syndromes, 2020].". Takový postup dává lékaři nejen praktické doporučení, ale i jasnou oporu v literatuře, kterou může dále zkontrolovat.

Klíčovou součástí RAG je způsob, jakým se text převádí do matematické podoby. Pro vyhledávání v databázích se používají tzv. embeddingy – vektorové reprezentace textu v mnohorozměrném prostoru. Každý dokument, odstavec nebo věta je převedena do číselného vektoru $\overrightarrow{e}_w \in R^d$. Když model obdrží dotaz, vytvoří pro něj embedding a pomocí metrik (např. kosinové podobnosti) hledá v databázi texty, jejichž embeddingy jsou nejbližší. [26]

V medicínském prostředí je generování embeddingů náročnější než v běžném jazyce. Terminologie je vysoce specifická, obsahuje mnoho synonym (např. „IM" = „infarkt myokardu" = „srdeční infarkt") a zkratek, které mohou být víceznačné. Pro kvalitní vyhledávání je proto nutné používat embeddingy, které byly trénovány přímo na biomedicínských textech. Zároveň je důležité řešit otázku multilingválních embeddingů – zatímco většina databází je v angličtině, klinická praxe často vyžaduje práci s lokálními jazyky. RAG proto musí překlenovat jazykovou bariéru, aby dokázal vyhledat relevantní informace v anglických guidelines a zároveň odpovědět česky lékaři nebo pacientovi. [27, 28]

Kromě základního principu Retrieval-Augmented Generation se v současnosti rozvíjejí i pokročilé RAG architektury, které zvyšují přesnost a spolehlivost výstupů. Patří sem například:

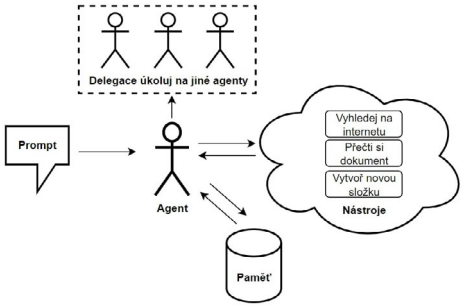
- **Multi-step RAG** – model nevyhledává zdroje pouze jednou, ale iterativně zpřesňuje dotaz a kombinuje výsledky z více kol vyhledávání. To umožňuje komplexnější odpovědi na složité klinické otázky. [29]
- **Hierarchical RAG** – pracuje s různými úrovněmi zdrojů (např. nejprve guidelines, poté podrobnější odborné články). Díky tomu dokáže vyvážit stručné doporučení a hlubší detail.
- **Domain-specific RAG** – architektura uzpůsobená konkrétní oblasti medicíny, kdy model čerpá jen z vybraných databází (např. kardiologie, onkologie). To snižuje riziko irelevantních výsledků.
- **Hybridní přístupy** – kombinace RAG s dalšími metodami, například s agentními systémy, kde různí „specializovaní agenti" vyhledávají v různých zdrojích a následně konsolidují odpovědi.

4 Autonomní agenti a multiagentní systémy

Vedle klasického použití velkých jazykových modelů pro generování textu se v posledních letech začíná prosazovat i jejich role jako autonomních agentů – tedy softwarových entit, které dokážou samostatně vykonávat úkoly, rozhodovat se a reagovat na prostředí. Na rozdíl od tradičního softwaru, který přesně následuje předem definované instrukce, mají agenti schopnost adaptace, práce s kontextem a dlouhodobější koordinace činností. Ještě výraznější možnosti nabízí propojení více agentů do multiagentních systémů, kde jednotliví agenti spolupracují podobně jako členové multidisciplinárního týmu v medicíně: předávají si informace, kontrolují navzájem své závěry a společně řeší komplexní úlohy. V kombinaci s technikami kontextového inženýrství a retrieval-augmented generation se tím otevírá prostor pro vytváření systémů, které mohou nejen podporovat klinická rozhodnutí, ale i simulovat spolupráci specialistů nebo autonomně spravovat složité procesy.

4.1 Co je agent a jak se liší od klasického softwaru

Pojem agent označuje softwarovou entitu, která dokáže samostatně vnímat své prostředí, rozhodovat se na základě získaných informací a aktivně jednat tak, aby dosáhla stanovených cílů. Zatímco tradiční software plní přesně definované instrukce podle pevně daného algoritmu, agent má určitou míru autonomie – může zvolit různé postupy, reagovat na změny okolností a koordinovat se s dalšími agenty nebo uživateli. [30, 31]



Obrázek 6 – Softwarový agent je autonomní entita s pamětí, která plní úkoly zadané ve vstupním promptu a využívá k tomu různé nástroje (algoritmy)

Jedním z klíčových konceptů při využívání LLM jako autonomních agentů jsou tzv. agentické pracovní toky (agentic workflows). Jde o procesy, ve kterých agent neřeší úkol jedním krokem, ale postupně jej rozkládá na menší části, které se mohou opakovaně vykonávat, vyhodnocovat a opravovat. Každý krok se přitom řídí výstupem předchozího a celý proces má charakter cíleně řízené smyčky.

Z technického pohledu lze agentický pracovní tok rozdělit na:

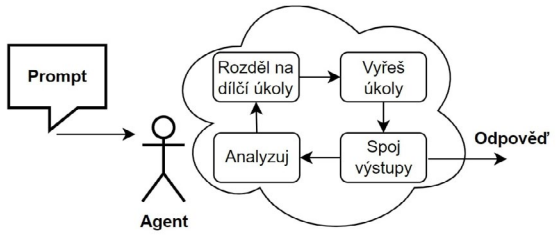
- Inicializaci cíle: agent obdrží úkol (např. „navrhni plán diagnostiky pacienta s dušností").
- Plánování: rozdělení cíle na kroky (anamnéza, laboratorní vyšetření, zobrazovací metody).
- Exekuci: volání nástrojů nebo databází pro každý krok.
- Reflexi: vyhodnocení, zda výsledek odpovídá cíli, případně návrat zpět k doplnění nebo opravě.
- Uzavření úkolu: poskytnutí finálního, koherentního výstupu.

DOI: 10.35191/medsoft_2025_1_37_beranek

Pro medicínu to znamená, že agent může například zpracovat klinický dotaz, vyhledat guidelines, vytvořit diferenciální diagnostiku a následně zkontrolovat konzistenci výsledku – nikoli v jednorázové odpovědi, ale jako sekvenci kroků připomínající práci skutečného lékaře.

Mezi typické druhy agentických pracovních toků řadíme [32]:

1. **Lineární (sekvenční) workflow:** Úkol je rozdělen na předem definovanou posloupnost kroků. Agent jde krok za krokem od začátku do konce bez zpětných návratů. *Příklad v medicíně:* sumarizace pacientovy anamnézy → strukturování dat → vytvoření zprávy pro dokumentaci.
2. **Iterativní workflow (looping):** Agent opakuje určité kroky, dokud není dosaženo cíle nebo přijatelné kvality výsledku. Využívá reflexi a sebehodnocení („zkontroluj, zda je výsledek konzistentní"). *Příklad v medicíně:* generování diferenciální diagnostiky s následnou kontrolou proti guideline, případně doplnění chybějících vyšetření.
3. **Rozhodovací (branching) workflow:** Tok se větví podle výsledků předchozích kroků. Agent volí různé strategie podle toho, jaké informace získá. *Příklad v medicíně:* pokud je EKG normální → pokračuj laboratorními testy; pokud je patologické → doporuč urgentní CT.
4. **Hierarchický workflow:** Agent rozdělí komplexní úkol na menší podúkoly, které řeší samostatně. Většinou funguje ve více úrovních (hlavní cíl → dílčí cíle). *Příklad v medicíně:* vypracování léčebného plánu, kde se zvlášť řeší farmakoterapie, nefarmakologické intervence a následná péče.
5. **Kooperativní workflow (multiagentní):** Na úkolu spolupracuje více agentů, kteří mají odlišné role nebo specializace. Agent koordinátor rozděluje práci, ostatní plní dílčí úkoly a sdílejí výsledky. *Příklad v medicíně:* agent–radiolog interpretuje snímek, agent–farmakolog navrhne léčbu, agent–koordinátor integruje doporučení do jedné zprávy.



Obrázek 7 – Agentické pracovní toky mohou mít různé podoby. Na obrázku je znázorněn agent, který iterativně řeší úkol po částech, dokud není spokojen s výsledkem. Takovéto postupy tvoří jádro návrhu agentních systémů.

4.2 Multiagentní systémy

Zatímco jeden autonomní agent dokáže samostatně vykonávat dílčí úkoly, skutečný potenciál se naplno projeví při jejich propojení do multiagentních systémů (MAS). Tyto systémy tvoří síť spolupracujících agentů, kteří si mezi sebou vyměňují informace, koordinují činnosti a společně dosahují cíle, který by pro jednoho agenta byl příliš složitý. Inspirací je zde lidská praxe – například multidisciplinární týmy v medicíně, kde internista, radiolog a farmakolog přispívají každý svými znalostmi, aby společně stanovili optimální postup péče o pacienta. [33]

Technicky lze multiagentní systémy charakterizovat třemi klíčovými vlastnostmi:

- **Decentralizace:** žádný agent nemá úplné informace o celém problému, každý pracuje se svým dílčím kontextem.
- **Koordinace a komunikace:** agenti si musí navzájem předávat výsledky, hodnotit jejich relevanci a někdy i řešit konflikty.
- **Společný cíl:** ačkoli jednotliví agenti mohou mít specializované role (vyhledávání, sumarizace, validace), jejich činnost směřuje k jednotnému výsledku – například vytvoření komplexního klinického doporučení.

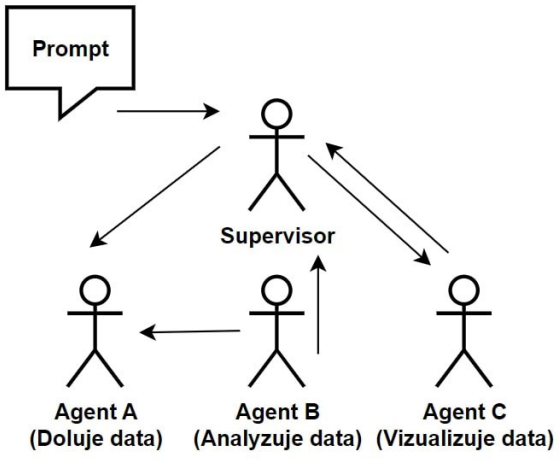
Existuje několik hlavních architektonických přístupů, které určují, jak jsou agenti v systému organizováni a jak mezi sebou spolupracují [34]:

- **Centralizovaná architektura:** Jeden „řídicí" agent nebo orchestrátor rozděluje úkoly a koordinuje ostatní agenty. *Příklad v medicíně:* agent–koordinátor sbírá výstupy od agentů specializovaných na laboratorní hodnoty, zobrazovací metody a guidelines a integruje je do jedné zprávy.
- **Decentralizovaná (distribuovaná) architektura:** Neexistuje centrální řídicí jednotka, každý agent rozhoduje autonomně a komunikuje s ostatními podle potřeby. *Příklad v medicíně:* několik agentů paralelně vyhodnocuje data pacienta (např. jeden se soustředí na farmakologii, druhý na zobrazovací výsledky) a sdílí závěry mezi sebou.
- **Hybridní architektura:** Kombinuje centrálního koordinátora pro hlavní rozhodování s částečnou autonomií agentů. Vhodná pro komplexní prostředí, kde je potřeba rovnováha mezi kontrolou a flexibilitou. *Příklad v medicíně:* tumor board agentů, kde centrální koordinátor dohlíží na proces, ale specializovaní agenti mají značnou samostatnost.

V medicíně umožňují multiagentní systémy simulovat a podporovat spolupráci odborníků, automatizovat složité procesy (např. kombinování guideline, laboratorních výsledků a anamnézy) a zvyšovat spolehlivost tím, že se jednotliví agenti navzájem kontrolují. [33]

Agenti si musí efektivně předávat informace, což může probíhat několika způsoby [34]:

- **Sdílený kontext (blackboard model):** všichni agenti zapisují a čtou informace ze společného prostoru („tabule"). Přehledné, ale může být zahlcené.
- **Přímé zprávy (message passing):** agent A pošle konkrétní dotaz agentovi B. Efektivní, ale může být složité ve větších systémech.
- **Konverzační protokoly:** agenti vedou strukturovaný dialog (např. otázka–odpověď, nabídka–proti-nabídka). Hodí se pro vyjednávání a validaci.
- **Kontextové řetězení přes LLM:** každý agent generuje textový výstup, který se stává vstupem pro dalšího agenta. Tohle je běžné u systémů postavených nad LLM.



Obrázek 8 – Multiagentní systémy mohou spolupracovat v různých architekturách, nejčastěji v hierarchii. Každý agent přitom disponuje jinými nástroji a řeší odlišné dílčí úkoly.

Model Context Protocol (MCP) je nově vznikající standard, jehož cílem je sjednotit způsob, jakým agenti (nebo aplikace založené na LLM) přistupují k nástrojům, datům a vzájemné komuni-

kaci. V podstatě jde o „jazyk" pro interoperabilitu agentů. Umožňuje, aby různé agenty odlišných vývojářů mohly spolupracovat bez složité integrace. V multiagentních systémech v medicíně by MCP mohl hrát podobnou roli, jako mají standardy HL7 nebo FHIR pro výměnu zdravotnických dat – zajistí, že agenti různého zaměření (diagnostika, farmakoterapie, administrace) budou schopni komunikovat v jednotném protokolu. [35]

5 Současné aplikace v medicíně a zdravotnictví

Velké jazykové modely již dnes nacházejí uplatnění v medicíně i zdravotnictví, kde přispívají k vyšší efektivitě práce, lepší komunikaci a podpoře rozhodování. Na rozdíl od dřívějších technologií zaměřených na úzké specializace (např. analýzu obrazových dat) nabízejí LLM univerzálnost – dokážou pracovat s přirozeným jazykem, tedy s hlavním médiem, kterým lékaři i pacienti sdílejí informace. V praxi se využívají od administrativních činností přes komunikaci s pacienty až po podporu diagnostiky, léčby i vzdělávání. [20, 36]

5.1 Administrativní podpora

Jednou z prvních oblastí využití velkých jazykových modelů v medicíně je administrativní zátěž, která zabírá podstatnou část práce zdravotnického personálu. Lékaři i sestry tráví mnoho času dokumentací, psaním zpráv či kódováním výkonů, což omezuje kontakt s pacientem. LLM tyto činnosti usnadňují – umí shrnout dlouhé zprávy, převést volný text do kódů (např. ICD-10) nebo vytvořit návrh záznamu k rychlé kontrole. Tím šetří čas a zvyšují konzistenci dokumentace.

Praktické příklady využití:

- **Automatické shrnutí hospitalizačních zpráv:** po ukončení hospitalizace model dokáže z několika stránek záznamů vygenerovat krátký a srozumitelný souhrn pro praktického lékaře nebo pacienta.
- **Generování propouštěcí zprávy:** na základě anamnézy, průběhu hospitalizace a podaných léků vytvoří návrh závěrečné zprávy, kterou lékař pouze ověří a doplní.
- **Kódování výkonů:** z volného textu v dokumentaci model rozpozná odpovídající kódy (např. ICD-10, SNOMED CT), což urychluje administraci a snižuje chybovost.
- **Strukturování anamnézy:** pacientův popis potíží převede do standardizovaných položek (rodinná anamnéza, alergie, medikace).
- **Předvyplnění elektronických formulářů:** model může automaticky doplnit části žádanky nebo laboratorního požadavku na základě již existujících údajů.

V praxi se ukazuje, že taková podpora může lékařům ušetřit desítky minut denně, což se při plném provozu nemocnic a ambulancí násobí do významného časového zisku. Současně se tím zlepšuje kvalita dat ve zdravotnických informačních systémech, protože výstupy modelů mají tendenci být konzistentnější než ruční zápisy od různých autorů. [37]

5.2 Komunikace s pacienty

Velké jazykové modely otevírají nové možnosti v oblasti komunikace mezi zdravotnickým personálem a pacienty. Tradičně byla tato komunikace časově náročná, často omezená na krátký prostor ordinace a zatížená nerovností v odborných znalostech. LLM mohou sloužit jako prostředník, který dokáže srozumitelně vysvětlit diagnózu, léčebný postup či rizika, a zároveň být pacientům k dispozici formou interaktivního asistenta.

Praktické příklady využití:

- Chatboti pro pacienty: inteligentní asistenti schopní odpovídat na časté otázky (např. jak se připravit na kolonoskopii, kdy je třeba přijít na kontrolu) a tím ulevit personálu od rutinních dotazů.
- Vysvětlování diagnóz v srozumitelném jazyce: model dokáže převést odborný lékařský zápis („ischemická choroba srdeční, NYHA II") do podoby, které pacient rozumí („Vaše srdce má omezenou schopnost pumpovat krev, což způsobuje dušnost při větší námaze, ale zvládáte běžné denní aktivity").
- Podpora adherence k léčbě: připomínání užívání léků, vysvětlování jejich účinků a varování před možnými interakcemi, a to formou přívětivé konverzace.
- Edukace pacientů: generování personalizovaných materiálů o prevenci, dietních opatřeních nebo životním stylu (např. u diabetu či hypertenze).
- Vícejazyčná komunikace: LLM mohou sloužit jako překladatel mezi pacientem a zdravotnickým personálem, což je důležité v multikulturním prostředí nebo u pacientů s jazykovou bariérou.

Využití těchto nástrojů má potenciál zvýšit srozumitelnost péče, podpořit důvěru pacientů v léčebný proces a ulevit zdravotníkům od opakujících se vysvětlení. Zároveň však vyvolává otázky bezpečnosti – je nezbytné, aby generované informace byly ověřitelné a aby pacienti vždy měli možnost konzultovat doporučení přímo s lékařem. [38]

5.3 Klinická podpora rozhodování

Jednou z nejambicióznějších oblastí využití velkých jazykových modelů je klinická podpora rozhodování, tedy asistence lékařům při diagnostice, volbě léčebných postupů a plánování péče. Na rozdíl od administrativních úloh nebo komunikace s pacientem jde o aplikaci s vysokými nároky na přesnost, bezpečnost a interpretovatelnost. LLM mohou lékaře podpořit tím, že rychle vyhodnotí rozsáhlou dokumentaci, vyhledají relevantní doporučení v guidelines a nabídnou diferenciální diagnostiku či návrhy léčebného postupu.

Praktické příklady využití:

- **Analýza klinické dokumentace:** model dokáže z dlouhé anamnézy, laboratorních výsledků a zobrazovacích zpráv vyextrahovat klíčové informace a upozornit na možné souvislosti.
- **Diferenciální diagnostika:** na základě popisu symptomů navrhne několik pravděpodobných diagnóz s odkazy na relevantní zdroje.
- **Doporučení léčebných postupů:** propojením s RAG může LLM vyhledat aktuální guidelines a nabídnout doporučení šité na míru konkrétnímu pacientovi (např. léčba srdečního selhání s ohledem na přidružené onemocnění ledvin).
- **Prediktivní podpora:** ve spojení s elektronickou zdravotnickou dokumentací může model upozornit na rizikové faktory (např. možnost sepse u pacienta s laboratorními odchylkami a horečkou).
- **Kontrola konzistence rozhodnutí:** agentní systémy mohou navzájem porovnávat své závěry a upozornit na nesoulad mezi doporučením a zavedenými postupy.

Významným přínosem je rychlost a schopnost integrovat různé typy dat, které by jinak lékař musel složitě procházet. Zároveň je však nezbytné, aby taková doporučení byla chápána jako doplněk k rozhodnutí lékaře, nikoli jako jeho náhrada. Odpovědnost za léčbu musí vždy zůstat na odbornícovi, zatímco LLM slouží jako nástroj pro zvýšení kvality a bezpečnosti péče. [39]

DOI: 10.35191/medsoft_2025_1_37_beranek

5.4 Vzdělávání a výzkum

Velké jazykové modely se začínají uplatňovat i v oblasti lékařského vzdělávání a výzkumu, kde mohou fungovat jako nástroje pro rychlé získávání znalostí, simulaci klinických situací a podporu vědecké práce. Jejich schopnost pracovat s rozsáhlými textovými korpusy a převádět složité odborné informace do srozumitelné podoby z nich činí užitečné asistenty jak pro studenty medicíny, tak pro výzkumníky.

Praktické příklady využití:

- Literární rešerše: LLM dokážou prohledat velké množství vědeckých článků, shrnout aktuální poznatky a poskytnout odkazy na relevantní zdroje.
- Simulace klinických případů: studenti medicíny mohou komunikovat s modelem, který se „chová" jako pacient se specifickými symptomy. Tím si trénují dovednosti v diagnostice a komunikaci.
- Podpora při psaní odborných textů: asistence při strukturování článků, návrhu výzkumných hypotéz nebo vytváření přehledových studií.
- Edukace formou personalizovaných vysvětlení: model může studentovi nabídnout různé úrovně obtížnosti výkladu, od jednoduchého vysvětlení po detailní odbornou analýzu.
- Výzkumná asistence: propojení LLM s databázemi (např. PubMed) umožňuje rychle identifikovat mezery v poznání a formulovat nové výzkumné otázky.

Tímto způsobem mohou LLM urychlit vzdělávací proces, poskytnout studentům možnost trénovat v simulovaném prostředí a zároveň podpořit vědecké pracovníky při orientaci v rychle rostoucím objemu odborné literatury. [40]

6 Omezení velkých jazykových modelů

Velké jazykové modely přinášejí do medicíny nové možnosti, které mohou zásadně ovlivnit klinickou praxi i organizaci zdravotní péče. Jejich schopnost pracovat s přirozeným jazykem, sumarizovat informace a generovat doporučení umožňuje zvýšit efektivitu, snížit administrativní zátěž a podpořit kvalitu rozhodování. Současně je však nezbytné reflektovat jejich limity – od technických problémů, jako jsou halucinace či zkreslení dat, přes nároky na výpočetní kapacity až po praktické otázky jejich začlenění do každodenního provozu. Tato kapitola proto nabízí vyvážený pohled na přínosy i omezení LLM a zdůrazňuje, že skutečný užitek mohou přinést jen při zodpovědném a uváženém nasazení.

6.1 Technická omezení

Navzdory svému potenciálu mají velké jazykové modely významná technická omezení, která limitují jejich spolehlivost v klinické praxi. Jedním z nejzávažnějších problémů jsou tzv. halucinace – situace, kdy model vygeneruje text, který působí důvěryhodně, ale není pravdivý. V medicíně to může vést k chybným doporučením nebo nesprávným interpretacím klinických dat. Proto je nezbytné, aby výstupy modelu vždy podléhaly odborné kontrole. [41]

Dalším problémem je bias (zkreslení), které vzniká z tréninkových dat. Pokud data obsahují systematické odchylky (např. nedostatečné zastoupení určité populace nebo genderové nerovnosti), mohou se tyto zkreslení přenést i do doporučení modelu. V klinické praxi to může znamenat, že některé skupiny pacientů dostanou méně přesná nebo méně vhodná doporučení. [42]

K tomu se přidává i omezená doménová znalost. Většina velkých jazykových modelů byla původně trénována na obecném textu, nikoli na specializovaných biomedicínských datech. I když existují modely přizpůsobené medicíně (např. PubMedBERT), jejich nasazení je zatím omezené a vyžaduje dodatečný vývoj. Modely tak mohou selhávat při práci s vysoce odbornými termíny, zkratkami nebo komplexními klinickými scénáři. [20–22]

Technická omezení proto představují zásadní výzvu: aby LLM mohly být bezpečně využívány v medicíně, je nutné kombinovat je s metodami zajišťujícími přesnost (např. RAG), zajistit pravidelnou aktualizaci dat a zavést mechanismy pro detekci a korekci biasu.

6.2 Praktická omezení (náklady, integrace do workflow, uživatelská přívětivost)

Kromě technických limitů je nasazení velkých jazykových modelů v medicíně spojeno i s řadou praktických omezení, která rozhodují o jejich reálné využitelnosti.

Prvním z nich jsou náklady. Trénink a provoz velkých modelů vyžadují značné výpočetní kapacity a infrastrukturu, což může být finančně náročné jak pro nemocnice, tak pro poskytovatele cloudových služeb. Menší zdravotnická zařízení proto často nemají přístup k plnohodnotným řešením.

Další výzvou je integrace do stávajícího workflow. LLM mohou být přínosné jen tehdy, pokud se stanou přirozenou součástí práce lékařů a sester. Pokud je jejich použití komplikované nebo vyžaduje paralelní práci s více systémy, může vést spíše k frustraci než k úsporám času.

Neméně důležitá je také uživatelská přívětivost. Lékaři nejsou IT specialisté a očekávají, že technologie bude intuitivní a snadno ovladatelná. To zahrnuje nejen jednoduché rozhraní, ale i jasně formulované výstupy, které lze rychle zkontrolovat a začlenit do klinické dokumentace.

Praktická omezení tak ukazují, že zavedení LLM není jen technickou otázkou, ale vyžaduje komplexní přístup zahrnující ekonomiku, organizaci práce a design uživatelského rozhraní. [36]

7 Etické a právní aspekty

Nasazení velkých jazykových modelů v medicíně přináší nejen technické a praktické výzvy, ale také zásadní otázky etiky a práva. Práce s citlivými zdravotnickými daty, možnost chybného doporučení či riziko zneužití generovaného obsahu vyžadují, aby tyto technologie byly používány v jasně vymezených mantinelech. Je třeba řešit ochranu osobních údajů, odpovědnost za rozhodnutí učiněná na základě výstupů modelu, stejně jako transparentnost a vysvětlitelnost jejich činnosti. V neposlední řadě vyvstává otázka, jak zajistit, aby byly tyto systémy spravedlivé a nediskriminovaly určité skupiny pacientů. Tato kapitola se proto zaměřuje na hlavní etické dilemata a právní rámce, které musí být při využívání LLM v medicíně respektovány.

7.1 Ochrana osobních údajů a důvěrnosti dat pacientů

V medicíně patří práce s daty mezi nejcitlivější oblasti. Velké jazykové modely často vyžadují přístup k rozsáhlým souborům pacientských záznamů, aby mohly poskytovat relevantní výstupy. To přináší riziko porušení ochrany osobních údajů a konflikt s právními předpisy, jako je GDPR. Zásadní otázkou je, kde se data zpracovávají – zda na lokální infrastruktuře nemocnice, nebo v cloudu – a kdo k nim má přístup. Etické využití LLM proto vyžaduje přísné anonymizace, auditní záznamy o přístupu a transparentní pravidla pro uchovávání i sdílení dat. [43, 44]

7.2 Odpovědnost za chyby a bezpečnost

Pokud model poskytne chybnou nebo zavádějící odpověď, vyvstává otázka odpovědnosti – lékaře, vývojáře nebo poskytovatele služby. V praxi musí mít odpovědnost vždy lékař, avšak je nutné jasně definovat roli technologie jako podpůrného nástroje. Stejně důležité je zajistit bezpečnost prostřednictvím pravidelného monitoringu, ochranných mechanismů proti halucinacím a testování v reálných podmínkách.

Modely trénované na nevyvážených datech mohou diskriminovat určité skupiny pacientů, např. ženy, starší osoby či menšiny. Aby se bias minimalizoval, je nezbytné používat diverzifikovaná data, zavádět mechanismy pro jeho detekci a transparentně informovat o omezeních. [45]

7.3 Transparentnost a vysvětlitelnost výsledků

LLM jsou často vnímány jako „černé skříňky", jejichž výstupy je obtížné interpretovat. V medicíně je však vysvětlitelnost zásadní – lékař musí rozumět, proč model doporučil konkrétní postup, aby mohl jeho doporučení důvěřovat. Proto je nutné, aby systémy poskytovaly odkazy na zdroje (např. guidelines), ukazovaly, z jakých informací vycházely, a umožňovaly zpětnou kontrolu. Transparentnost zvyšuje důvěru uživatelů a usnadňuje přijetí těchto technologií v klinické praxi. [46, 47]

7.4 Regulace a standardy (např. EU AI Act, zdravotnické směrnice)

Vývoj a nasazení velkých jazykových modelů v medicíně se odehrává v prostředí, které je stále více ovlivňováno regulačními rámci a standardy. V Evropské unii je klíčovým dokumentem připravovaný EU AI Act, který zavádí klasifikaci systémů umělé inteligence podle míry rizika. Aplikace v medicíně spadají do kategorie vysokého rizika, což znamená přísné požadavky na bezpečnost, transparentnost, sledovatelnost a dohled nad provozem. [48]

Vedle obecné legislativy je nutné respektovat i specifické zdravotnické směrnice a standardy. Patří sem zejména normy pro práci s elektronickou zdravotnickou dokumentací (např. HL7, FHIR), standardy ochrany dat (GDPR, HIPAA v USA) či pravidla pro uvádění zdravotnických prostředků na trh (MDR – Medical Device Regulation). Pokud je LLM součástí systému, který má přímý vliv na diagnostiku nebo léčbu, může být klasifikován jako zdravotnický prostředek a podléhá certifikaci podobně jako software pro zobrazovací metody či laboratorní analýzu. [49, 50, 51, 52, 53]

Standardizace a regulace jsou proto zásadní nejen z hlediska bezpečnosti pacientů, ale také pro důvěru zdravotníků. Poskytují jasný rámec, v němž se technologie může vyvíjet a nasazovat, a zároveň zajišťují, že systémy budou vzájemně kompatibilní a interoperabilní. [54]

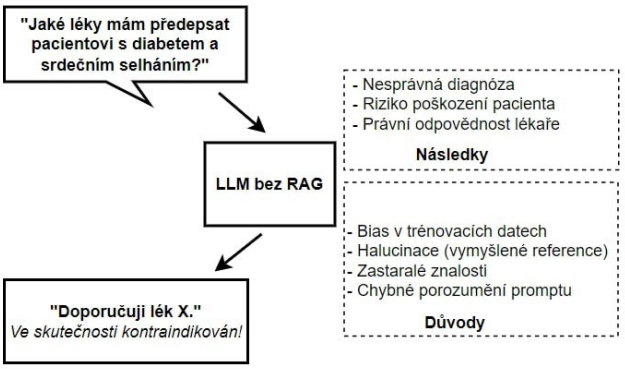
8 Doporučení pro praxi a implementaci

Aby bylo možné velké jazykové modely využívat v medicíně bezpečně a efektivně, je nutné postupovat postupně a s jasně definovanými pravidly. Zdravotnické instituce by měly začínat s omezenými scénáři, které mají nižší riziko dopadu na pacienta, například administrativní podporou či literárními rešeršemi. Klíčové je zajistit odborný dohled nad všemi výstupy a vytvořit prostředí, kde model funguje jako podpůrný, nikoli autonomní nástroj. [20, 36]

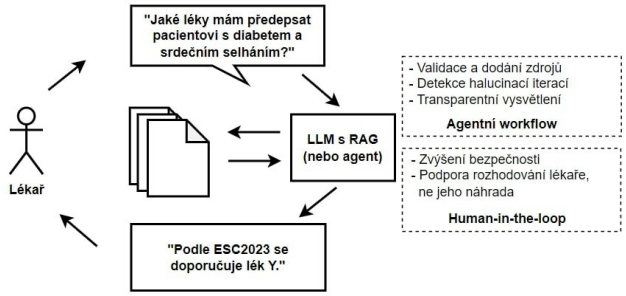
Za vhodný první krok se považuje Retrieval-Augmented Generation (RAG), které propojuje LLM s externími databázemi a umožňuje opírat odpovědi o citované zdroje. To významně snižuje riziko halucinací a zvyšuje důvěryhodnost výstupů, zejména v klinických doporučeních.

Velký potenciál má i kontextové inženýrství, tedy schopnost formulovat správné dotazy a přizpůsobovat vstupy modelu. Lékař se tak stává nejen uživatelem, ale částečně i „technologem", který aktivně řídí kvalitu výstupu prostřednictvím promyšlených instrukcí.

Před zavedením do ostrého provozu je vhodné testovat nové přístupy v rámci multiagentních simulací, kde více modelů spolupracuje, kontroluje si navzájem výsledky a napodobuje práci multidisciplinárního týmu. Tak lze identifikovat slabá místa a vyhodnotit spolehlivost systému bez rizika pro pacienty.



Obrázek 9 – Používání LLM bez promyšleného workflow, bez RAG a bez zapojení člověka je rizikové.



Obrázek 10 – Do procesu rozhodování by měl být vždy zapojen expert. Ten iterativně pracuje s promyšleným agentním workflow, využívá důvěryhodné zdroje nebo fine-tunované modely a spolupracuje na řešení úkolu.

Nezbytnou podmínkou jsou také minimální bezpečnostní a kvalitativní standardy: pravidelný audit výkonu modelů, validace na lokálních datech, ochrana osobních údajů, jasná pravidla odpovědnosti a transparentní komunikace limitů technologie vůči zdravotnickému personálu. Pouze tak lze dosáhnout, aby byly LLM skutečně přínosem, a ne potenciálním rizikem. [55]
Proces zavádění LLM do zdravotnické praxe:

Analýza potřeb a rizik

- Identifikovat, ve kterých oblastech může LLM přinést největší přínos (např. administrativní podpora vs. klinická rozhodnutí).
- Vyhodnotit rizikovost scénářů – začínat tam, kde chybné odpovědi nemají přímý dopad na zdraví pacienta.

Pilotní projekty

- Spustit omezené piloty s jasně definovaným cílem (např. shrnutí propouštěcích zpráv).
- Zajistit, aby výstupy vždy kontroloval zdravotník.
- Sbírat zpětnou vazbu uživatelů a měřit úsporu času, kvalitu výstupů a spokojenost personálu.

Integrace do workflow

- Model musí být napojen na stávající zdravotnické informační systémy (EHR, LIS, PACS), aby nebyl „dalším oknem navíc".
- Je nutné zapojit IT oddělení, aby byla zajištěna interoperabilita (FHIR, HL7).

Školení uživatelů

- Lékaře a sestry je třeba naučit základům práce s LLM – zejména principům kontextového inženýrství (jak klást správné dotazy).
- Podpora role „lékař jako prompt engineer", aby uživatelé uměli z modelu získat relevantní výstupy.

Bezpečnost a compliance

- Pravidelný audit výstupů modelu.
- Ověření souladu s legislativou (GDPR, MDR, případně EU AI Act).
- Mechanismy pro anonymizaci dat a kontrolu přístupu.

Škálování a rozšiřování

- Po úspěšném pilotu lze postupně rozšiřovat využití (od administrativy → edukace pacientů → podpora rozhodování).
- Zohlednit náklady na výpočetní infrastrukturu a dlouhodobou udržitelnost.

9 Perspektivy do budoucna

Vývoj velkých jazykových modelů se dynamicky posouvá kupředu a otevírá stále širší spektrum možností v medicíně. Jedním z nejvýraznějších trendů je vznik multimodálních modelů, které dokážou vedle textu zpracovávat také obrazová a zvuková data. V praxi to znamená, že budoucí systém bude schopen nejen shrnout anamnézu pacienta, ale současně vyhodnotit radiologické snímky, analyzovat výsledky laboratorních testů a třeba i rozpoznat změny v hlase související s neurologickým onemocněním. Tato schopnost integrovat různé modality přiblíží AI realitě klinického rozhodování, kde se lékař vždy opírá o více zdrojů informací najednou.

Další slibnou oblastí je personalizovaná medicína. Propojením jazykových modelů s genetickými daty, elektronickými zdravotními záznamy a informacemi o životním stylu pacienta se otevírá cesta k léčbě „na míru". Modely budou schopny generovat individualizované léčebné plány, predikovat reakce na farmakoterapii nebo doporučit preventivní opatření. V dlouhodobém horizontu lze očekávat, že AI pomůže řídit chronická onemocnění formou neustálého monitoringu a adaptivních doporučení, která se přizpůsobují vývoji zdravotního stavu. [56, 57]

Rychlý pokrok je vidět i v oblasti multiagentních simulací zdravotnických procesů. Ty umožňují vytvářet virtuální prostředí, kde specializovaní agenti spolupracují podobně jako lékaři v multidisciplinárním týmu – například radiolog, kardiolog a farmakolog. Díky tomu lze testovat složité klinické scénáře, hledat optimální rozdělení zdrojů nebo simulovat dopady organizačních změn v nemocnici ještě před jejich zavedením. Tento přístup má potenciál stát se novým standardem pro plánování a optimalizaci zdravotnických procesů. [58]

Tyto trendy směřují k vizi tzv. „AI-enabled hospitals", tedy nemocnic, kde je umělá inteligence pevně integrována do každodenního provozu. V takovém prostředí by AI podporovala administrativní agendu, zajišťovala rychlé vyhledávání v klinických doporučeních, pomáhala s triáží pacientů a zároveň fungovala jako partner při klinickém rozhodování. Součástí této vize je i hlubší interoperabilita – tedy schopnost systémů bezproblémově spolupracovat napříč odděleními a zdravotnickými zařízeními, ať už v rámci jedné nemocnice, nebo celých regionálních sítí péče. [59]

Budoucí vývoj tak slibuje posun směrem k inteligentnějším a efektivnějším zdravotnickým systémům. Aby se tato vize naplnila, je ale nezbytné současně řešit otázky etiky, regulace a vzdělávání zdravotníků. Jen tehdy mohou být přínosy nových technologií skutečně využity bezpečně, spravedlivě a ku prospěchu pacientů.

10 Závěr

Velké jazykové modely představují významný milník v oblasti umělé inteligence a otevírají široké možnosti pro využití v medicíně. Jak ukázaly předchozí kapitoly, jejich přínos je patrný v celé řadě oblastí – od administrativní podpory a komunikace s pacienty až po klinickou podporu rozhodování, vzdělávání a výzkum. Stejně důležité je však uvědomit si jejich omezení, ať už technická (halucinace, bias, omezená doménová znalost), nebo praktická (náklady, integrace do workflow, uživatelská přívětivost). Bez jejich pochopení a zohlednění by nasazení LLM mohlo představovat více rizik než přínosů.

Pro lékaře i zdravotnické instituce z toho vyplývají tři klíčová doporučení. Za prvé, s implementací je vhodné začínat v méně rizikových oblastech, zejména administrativní agendě a práci s dokumentací, kde lze rychle prokázat efektivitu a snížit zátěž personálu. Za druhé, vždy je nutné využívat LLM jako podpůrný nástroj, jehož výstupy podléhají odborné kontrole; odpovědnost za klinické rozhodování musí zůstat na lékaři. Za třetí, při zavádění do praxe je nezbytné respektovat principy bezpečnosti, ochrany dat a transparentnosti, a zároveň poskytovat zdravotníkům školení v práci s těmito nástroji.

Pokud budou tato doporučení dodržena, mohou velké jazykové modely významně přispět k efektivnější, kvalitnější a dostupnější péči. Jejich skutečná hodnota nespočívá v nahrazení lékaře, ale v posílení jeho role – ve vytvoření prostředí, kde technologie pomáhají zvládat rostoucí komplexitu medicíny a umožňují zdravotníkům věnovat více času tomu nejdůležitějšímu: pacientovi.

LARGE LANGUAGE MODELS AND AGENT-BASED SYSTEMS IN MEDICINE: PRINCIPLES, CURRENT APPLICATIONS, AND FUTURE PERSPECTIVES

Abstract

Large Language Models (LLMs) constitute a major milestone in the advancement of artificial intelligence and open up new opportunities within medicine and healthcare. Owing to their capacity to comprehend natural language and generate domain-specific texts, LLMs can enhance clinical decision-making, facilitate patient communication, and optimize administrative workflows. This paper provides a clear introduction to the underlying principles of LLMs and their adaptation to medical tasks, supplemented by a systematic review of current applications based on scholarly literature. Particular attention is devoted to the deployment of LLMs as autonomous agents and their integration into multi-agent systems, which can emulate the collaboration of multidisciplinary teams, coordinate complex processes, and conduct cross-validation of results. The discussion addresses the benefits, limitations, and ethical considerations of these technologies, and concludes with recommendations for their safe and effective integration into medical practice.

Keywords

large language models, artificial intelligence in medicine, agent-based systems, multi-agent systems, clinical decision support

Literatura

[1.] Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief History of Artificial Intelligence. Neuroimaging Clin N Am. 2020 Nov;30(4):393–399. doi: 10.1016/j.nic.2020.07.004. Epub 2020 Sep 18. PMID: 33038991.

[2.] Esteva, A., Kuprel, B., Novoa, R. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017). https://doi.org/10.1038/nature21056

[3.] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60–88. doi: 10.1016/j.media.2017.07.005. Epub 2017 Jul 26. PMID: 28778026.

[4.] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[5.] Brown, T., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).

[6.] Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024 Feb 15;10(4):e26297. doi: 10.1016/j.heliyon.2024.e26297. PMID: 38384518; PMCID: PMC10879008.
[7.] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023 Mar 30;388(13):1233–1239. doi: 10.1056/NEJMsr2214184. PMID: 36988602.
[8.] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7
[9.] Maity, S., & Saikia, M. J. (2025). Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering, 12(6), 631. https://doi.org/10.3390/bioengineering12060631
[10.] Nori, H., et al. (2023). Capabilities of GPT-4 on Medical Challenge Problems. 10.48550/arXiv.2303.13375.
[11.] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155.
[12.] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.
[13.] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.
[14.] Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556.
[15.] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 55(9), 1–35.
[16.] Wei, J., Wang, X., Schuurmans, D., Bosma, M., et al. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS, 35, 24824–24837.
[17.] Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.
[18.] Zhou, Y., Schick, T., Zhou, C., & Schütze, H. (2022). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. arXiv:2205.10625.
[19.] Dong, Y., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651.
[20.] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., & Tan, T. E. (2023). Large language models in medicine. Nature Medicine, 29, 1930–1940.
[21.] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
[22.] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., … & Gao, J. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23.
[23.] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv:1904.03323.
[24.] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409.
[25.] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., et al. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.
[26.] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459–9474.
[27.] Amugongo, M., et al. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health.
[28.] Ke, Y. H., et al. (2025). Clinical evaluation of retrieval-augmented generation models for surgical decision-making. npj Digital Medicine.
[29.] Zhan, X., et al. (2025). MMRAG: Multi-mode retrieval-augmented generation for medical in-context learning. arXiv:2502.15954.
[30.] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. Knowledge Engineering Review, 10(2), 115–152.
[31.] Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
[32.] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). ReAct: Synergizing reasoning and acting in language models. 10.48550/arXiv.2210.03629.
[33.] IBM. (2023). Prompt chaining in LangChain [Tutorial]. Retrieved August 25, 2025, from https://www.ibm.com/think/tutorials/prompt-chaining-langchain
[34.] Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. Autonomous Agents and Multi-Agent Systems, 1(1), 7–38.
[35.] Model Context Protocol. (2023). Retrieved from https://modelcontextprotocol.io
[36.] Jiang, Z., & Yu, K. H. (2023). Large language models in medicine: current applications and future directions. The Lancet Digital Health, 5(9), e543–e554.
[37.] Bednarczyk, L., et al. (2025). Scientific Evidence for Clinical Text Summarization Using Large Language Models. Journal of Medical Internet Research, 27, e68998.
[38.] Yang, X., et al. (2025). Enhancing doctor-patient communication using large language models: translating pathology reports into accessible language. BMC Medical Informatics and Decision Making, 25, 58.
[39.] Gaber, F., et al. (2025). Evaluating large language model workflows in clinical decision support. NPJ Digital Medicine.
[40.] Abd-alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P. M., Latifi, S., Aziz, S., Damseh, R., Alrazak, S. A., & Sheikh, J. (2023). Large language models in medical education: Opportunities, challenges, and future directions. JMIR Medical Education, 9(1), e48291.
[41.] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
[42.] Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182.
[43.] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.
[44.] Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.
[45.] Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning in health care. Nature Medicine, 26(9), 1327–1334.
[46.] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
[47.] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. Machine Learning for Healthcare Conference.
[48.] European Commission. (2023). Artificial Intelligence Act. Retrieved August 25, 2025, from https://artificialintelligenceact.eu
[49.] Health Level Seven International (HL7). (2025). HL7 Standards. Retrieved August 25, 2025, from https://hl7.org
[50.] HL7 FHIR Foundation. (2025). Fast Healthcare Interoperability Resources (FHIR). Retrieved August 25, 2025, from https://fhir.org
[51.] European Union. (2016). General Data Protection Regulation (GDPR). Retrieved August 25, 2025, from https://gdpr-info.eu
[52.] U.S. Department of Health & Human Services. (2025). Health Insurance Portability and Accountability Act (HIPAA). Retrieved August 25, 2025, from https://www.hhs.gov/hipaa
[53.] European Commission. (2017). Medical Device Regulation (EU 2017/745). Retrieved August 25, 2025, from https://medical-device-regulation.eu
[54.] Mesko, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine, 6, 105.
[55.] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 25, 30–36.
[56.] Li, H., Chen, X., & Wu, Y. (2024). Health-LLM: Retrieval-augmented generation for personalized disease prediction. arXiv preprint arXiv:2402.00746.
[57.] Rana, A., Gupta, R., & Malik, K. (2024). CoLaCare: Collaborative large language model agents for personalized healthcare predictions. arXiv preprint arXiv:2410.02551.
[58.] Carramiñana, F., López, J., García, M., & Torres, A. (2025). Enhancing healthcare infrastructure resilience through agent-based simulation methods. Journal of Medical Systems, 49, 32.
[59.] Mesko, B. (2023). The rise of the AI-powered hospital. The Lancet Digital Health, 5(7), e364–e366.

DOI: 10.35191/medsoft_2025_1_37_beranek

Kontakt

Ing. Mgr. Pavel Beránek, MBA
(dvojí afiliace)
Univerzita Jana Evangelisty Purkyně
v Ústí nad Labem
Přírodovědecká fakulta
Pasteurova 3544/1
400 96 Ústí nad Labem

Česká zemědělská univerzita v Praze
Provozně ekonomická fakulta
Kamýcká 129
165 00 Praha – Suchdol
+420 475 286 724
pavel.beranek@ujep.cz
https://ki.ujep.cz/cs/personalni-slozeni/pavel-beranek/

doc. Ing. Vojtěch Merunka, Ph.D.
Česká zemědělská univerzita v Praze
Provozně ekonomická fakulta
Kamýcká 129
165 00 Praha – Suchdol
+420 224 382 272
merunka@pef.czu.cz
https://vojtech.merunka.eu/