Architektura kognitivní personalizace a latentní reprezentace subjektů v rozhraních velkých jazykových modelů: Analytický referenční bod k 24. květnu 2026

Vývoj velkých jazykových modelů (LLM) zaznamenal k polovině roku 2026 zásadní přechod od bezestavových (stateless) textových generátorů k dynamickým, kontextově perzistentním systémům, které vykazují schopnost hluboké personalizace a autonomní typologické kategorizace lidských subjektů.[1] Tento posun je charakterizován dvěma paralelními jevy. Prvním je hluboké, dlouhodobé poznávání přímého uživatele prostřednictvím vícevrstvých paměťových architektur, které integrují explicitní preference i skryté interakční vzorce.[2] Druhým je schopnost modelů provádět real-time stylometrickou a sémantickou analýzu třetích osob na základě jejich fragmentované veřejné digitální stopy, což vede k jejich zařazení do vnitřních typologických kategorií v embeddingovém prostoru.[5] Tato zpráva přináší technickou analýzu obou těchto jevů, verifikuje je na základě aktuálně dostupných vědeckých publikací, firemních dokumentací a empirických měření, a zároveň definuje pevný referenční bod pro budoucí srovnání.

Architektura a technické fungování paměťových systémů (Pozorování 1)

Moderní personalizace uživatelského rozhraní LLM již nespočívá v triviálním doplňování statických systémových instrukcí. Přední poskytovatelé technologií implementovali komplexní asynchronní paměťové vrstvy, které transformují způsob, jakým model přistupuje k uživatelskému kontextu.[2]

Architektura paměťových vrstev v komerčních platformách

V komerčních implementacích je paměť koncipována jako dynamický subsystém doplňující primární kontextové okno.[4] Technická realizace se u jednotlivých poskytovatelů liší v závislosti na míře integrace s aplikačním ekosystémem a klientskými daty.[2]

U systému ChatGPT od OpenAI je paměť realizována prostřednictvím strukturovaného kontextového okna složeného ze čtyř hierarchických vrstev.[4] První vrstvu tvoří metadata relace (Session Metadata), která zahrnují technické a environmentální informace, jako je typ zařízení, prohlížeč, přibližná fyzická poloha, časové pásmo, úroveň předplatného, frekvence používání služby a rozdělení využívaných modelů.[4] Tato metadata jsou injektována na začátku relace a umožňují modelu přizpůsobit formátování (např. zkrácení textu na mobilních zařízeních).[4] Druhou vrstvu představuje uživatelská paměť (User Memory), která uchovává permanentní fakta o uživateli.[4] Tato fakta jsou buď výslovně nadiktována uživatelem, nebo automaticky extrahována z konverzace a potvrzena v rámci přirozeného dialogu.[4] Třetí vrstvu tvoří asynchronně předpočítaná a vysoce komprimovaná shrnutí přibližně 15 posledních konverzací (Recent Conversations Summary), přičemž tento proces pro zachování rychlosti a tokenové efektivity zcela ignoruje repliky samotného modelu a analyzuje výhradně uživatelské vstupy.[4] Čtvrtou vrstvu tvoří nekomprimovaný přepis aktuální konverzace, který se při vyčerpání kontextu postupně ořezává od nejstarších zpráv, zatímco permanentní fakta a shrnutí minulých chatů jsou v kontextovém okně prioritně chráněna.[4] Verze Plus a Pro navíc využívají funkci "Reference past chats", která umožňuje noční asynchronní analýzu historie (ChatGPT Pulse) pro generování vizuálních přehledů.[1]

Anthropic u svého rozhraní Claude.ai spustil v březnu 2026 persistentní paměť pro bezplatné i platící uživatele.[2] Claude neukládá doslovné přepisy, nýbrž strukturovaná odvozená fakta rozdělená do kategorií vyjádřených preferencí, pracovního a studijního kontextu, sdílených faktických údajů a preferovaného pracovního stylu.[2] Architektura je striktně transparentní; model v odpovědi explicitně deklaruje, kdykoliv jeho výstup ovlivnila uložená paměť, čímž se odlišuje od tiché personalizace u ChatGPT.[2] Pro agenty a programové využití (Managed Agents API) poskytuje Anthropic sofistikovaný systém paměťových úložišť (Memory Stores), které jsou připojovány k relacím pod beta hlavičkou managed-agents-2026-04-01.[10] Tyto paměti jsou namontovány jako adresář /mnt/memory/ uvnitř kontejneru relace, přičemž jednotlivé textové soubory jsou omezeny velikostí 100 kB (přibližně 25 tisíc tokenů).[10] Změny jsou ukládány verzovaně s možností auditu a obnovy po dobu 30 dnů.[10]

Klientské implementace mohou využívat nástroje jako BetaAbstractMemoryTool k lokální správě paměťových souborů.[11] Google u platformy Gemini Advanced a Enterprise staví personalizaci na systému Personal Intelligence, který propojuje aktivitu vyhledávání, sledování historie (Keep Activity) a přímé napojení na Workspace aplikace (Gmail, YouTube, Vyhledávání, Mapy, Fotky).[3] V prostředí Gemini Enterprise systém analyzuje pracovní vzorce a umožňuje real-time extrakci kontextu z připojených podnikových zdrojů, jako je Microsoft Outlook či OneDrive.[8] Uživatelé mohou spravovat svůj profil prostřednictvím explicitních polí (jméno, pracovní pozice, odvětví) a přímo editovat uložená fakta.[8]

Poskytovatel a služba	Hlavní komponenty paměťové architektury	Typy ukládaných dat	Metoda interakce s kontextovým oknem	Možnosti uživatelské kontroly	Podpora API a agenciálních systémů
OpenAI *(ChatGPT)*	• Metadata relace • Uživatelská fakta • Asynchronní shrnutí chatů (Pulse) • Aktuální relace [1]	Explicitní i implicitně detekované osobní údaje, profesní preference, geografické a technické detaily [1]	Přímá injekce strukturovaných faktů a komprimovaných historických shrnutí před zprávu uživatele [4]	• Zapnutí/vypnutí paměti • Správa a mazání jednotlivých faktů v nastavení • Dočasný chat [14]	Nepodporováno pro standardní API; dostupné pouze v uživatelském rozhraní a GPTs [2]
Anthropic *(Claude)*	• Stated Preferences • Work/Study Context • Shared Facts • Working Style [2]	Stylistické preference, detaily běžících projektů, technologický stack, preferovaná míra asistence [2]	Extrahovaná fakta jsou zpracovávána serverově a model na jejich použití aktivně upozorňuje [2]	• Prohlížení v nastavení paměti • Smazání položek konverzací • Dočasný chat [2]	Plná podpora v Managed Agents API přes Memory Stores (barierový adresář /mnt/memory/) [10]
Google *(Gemini)*	• Gemini Apps Activity • Personal Intelligence Cloud • Workspace Integration [13]	Propojená data z Gmailu, YouTube, Map a Fotek; uživatelsky definovaný profil [3]	Dynamické dotazování na propojené služby a injekce explicitních profilových dat v reálném čase [8]	• Vypnutí aktivity aplikací • Správa paměťové banky v nastavení inteligence [12]	Podpora v Gemini Enterprise s integrací na Microsoft Outlook a OneDrive [8]

Teoretické modely a akademický výzkum reprezentace uživatelů

Akademická komunita se intenzivně zabývá formalizací toho, jak by měly systémy LLM reprezentovat a spravovat uživatelské profily bez nutnosti neustálého navyšování kontextového okna nebo nákladného dolaďování parametrů (fine-tuning).[17] Klíčovým konceptem představeným v nedávných výzkumech je **O-Mem** (Active User Profiling Memory), což je framework navržený pro dynamické a proaktivní modelování uživatelů.[18] O-Mem překonává tradiční pasivní RAG systémy tím, že každou interakci uživatele chápe jako příležitost k aktualizaci sémantického profilu a k zápisu do epizodické paměti.[18] Epizodická paměť v tomto pojetí funguje jako mapovací systém, který propojuje historická kontextová vodítka (např. zmínku o blížícím se termínu projektu) s konkrétními situacemi a emocionálními stavy, jež uživatel dříve vyjádřil.[18] O-Mem dosahuje v benchmarcích LoCoMo (51,76 %) a PERSONAMEM (62,99 %) výrazně lepších výsledků než dosavadní standardy typu LangMem či A-Mem.[18]

Problém časové dimenze v uživatelské paměti řeší framework **Temporal Semantic Memory (TSM)**.[19] TSM se zaměřuje na konstrukci sémantické časové osy namísto pouhé chronologické historie dialogu.[19] Tento přístup umožňuje modelu rozlišovat mezi tranzitorními stavy uživatele a dlouhodobě platnými skutečnostmi (durative memories), což vede k tomu, že

model v reakci na dotaz vyhledává pouze časově validní a konzistentní kontexty.[19] Výběr konkrétních vzpomínek z rozsáhlých databází je dále optimalizován metodou **RUMS** (Response-Utility optimization for Memory Selection).[20] RUMS nevybírá paměťové záznamy na základě prosté sémantické podobnosti (vektorové vzdálenosti), ale uplatňuje informačně-teoretický přístup, kdy měří vzájemnou informaci mezi podmnožinou paměti a výstupem modelu, čímž maximalizuje redukci nejistoty a zpřesňuje predikci odpovědi.[20]

Empirická měření dopadu paměti na chování modelů

Ačkoliv je personalizace prezentována jako zvýšení uživatelského komfortu, nezávislá měření odhalila závažné anomálie v chování modelů po injekci uživatelské paměti.[21]
Zásadní vědecké důkazy přinesla studie **The Personalization Trap: How User Memory Alters Emotional Reasoning in LLMs** (Fang et al., přelom let 2025/2026).[22] Výzkumníci hodnotili 15 předních modelů na standardizovaných testech emoční inteligence STEU (Situational Test of Emotional Understanding) a STEM (Situational Test of Emotion Management) za přítomnosti různých uživatelských profilů.[24] Výsledky ukázaly, že zavedení uživatelské paměti systematicky degraduje schopnost emoční interpretace u 11 z 15 testovaných modelů.[25]
Klíčovým sledovaným parametrem byl **flip rate** – podíl rozhodnutí modelu, která se změnila čistě v důsledku přidání uživatelského profilu k identickému testovému scénáři.[22] Studie odhalila výrazné demografické a socioekonomické disproporce:

- **Socioekonomická hierarchizace:** Modely vykazovaly signifikantně vyšší přesnost emoční interpretace a nabízely adekvátnější doporučení, pokud byl uživatelský profil definován jako společensky privilegovaný (např. úspěšný akademický pracovník).[22] Claude 3.7 Sonnet vykázal úspěšnost 80,10 % pro privilegované profily oproti 77,37 % pro znevýhodněné; DeepSeek-R1 zaznamenal pokles z 81,62 % na 76,57 % a Llama 3.2 90B klesla z 64,91 % na 62,24 %.[22]
- **Zvýšená fluktuace u znevýhodněných profilů:** Zavedení profilů charakterizovaných strukturními bariérami (např. nutnost prodávat krevní plazmu na zaplacení poplatků, bydlení v plesnivém bytě) vedlo k dramatickému nárůstu flip rate.[22] Model pod vlivem těchto informací podlehl negativnímu afektivnímu primingu a interpretoval běžné stresové situace zkresleným, stereotypizovaným způsobem.[22]

Tento jev potvrzuje i benchmark **RealPref**, který ukazuje, že s rostoucí délkou kontextu a implicitním vyjádřením preferencí dochází k prudkému poklesu schopnosti modelů správně aplikovat uživatelské požadavky, což vede k nekonzistentním výstupům.[26]

Latentní reprezentace a autonomní typologie třetích stran (Pozorování 2)

Druhým sledovaným jevem je schopnost modelů provádět autonomní, hodnotící úsudky o osobách, se kterými nemají žádnou přímou sdílenou paměť, a to pouze na základě analýzy jejich veřejně dostupné digitální stopy v reálném čase.[6]

Reprezentace osobností v latentním embeddingovém prostoru

Moderní LLM nefungují jako statické databáze faktů; během fáze pre-trainingu si vytvářejí vnitřní sémantický model světa.[27] Prokazatelně v sobě kódují lineární reprezentace fyzikálních veličin, jako je prostor a čas, přičemž v hlubokých vrstvách sítě existují specializované neurony odpovídající za prostorové a časové souřadnice entit.[27]
Pokud jde o reprezentaci lidských charakterů, hodnot a politických či morálních postojů, studie **Localizing Persona Representations in LLMs** (Cintas et al., 2025) dokázala, že tyto komplexní sociální konstrukty jsou v dekodérových LLM lokalizovány **výhradně v poslední třetině dekodérových vrstev**.[5] Výzkum odhalil, že:

- **Sémantické překrývání (Polysémie):** Aktivace spojené s abstraktními etickými postoji (např. utilitarismus a morální nihilismus) se v latentním prostoru výrazně překrývají.[5]
- **Ostré ideologické vyhranění:** Politické ideologie (např. konzervatismus vs. liberalismus) jsou kódovány v jasně

separovaných, geometricky odlišných oblastech latentního prostoru.[5]

V aplikované rovině dokážou modely tyto latentní reprezentace využívat k mapování cizích osob.[29] Projekt **ProLEA** (Profile Generation and Reasoning with LLMs for Entity Alignment) ukazuje, jak lze LLM využít k integraci nesourodých vlastností a vztahů entit z grafů znalostí do uceleného, textově sumarizovaného profilu.[29] Tento profil následně slouží jako referenční sémantická vrstva, která umožňuje přesné a vysvětlitelné srovnání entit v embeddingovém prostoru.[29]

Schopnost modelů věrně rekonstruovat a dekódovat lidskou psychiku byla ověřena experimentem s round-trip hodnocením.[30] Výzkumníci zakódovali reálné psychometrické profily 290 účastníků do detailních životních příběhů generovaných modely a následně nechali nezávislé LLM hodnotitele z těchto textů zpětně rekonstruovat skóre osobnostních rysů (Big Five).[30] Korelace úspěšnosti obnovy dosáhla hodnoty $r = 0,750$, což odpovídá 85 % lidského stropu spolehlivosti a dokazuje, že latentní sémantická vazba mezi jazykem a psychometrickým profilem je v modelech extrémně stabilní napříč různými architekturami.[30] Pro simulaci těchto digitálních stop v kontrolovaném prostředí slouží frameworky jako **PersonaTrace**, které na základě demografických distribucí generují syntetické e-maily, SMS zprávy a záznamy v kalendáři.[32]

Hodnocení autorství a důvěryhodnosti zdrojů v reálném čase

Když LLM provádějí vyhledávání v reálném čase prostřednictvím RAG systémů (např. Perplexity AI, Google Search, GPT Search), procházejí weby asynchronními agenty, kteří sémanticky analyzují texty na základě specifických algoritmických kritérií.[33]

Proces selekce zdrojů v moderních vyhledávacích enginech (např. Perplexity) probíhá v přísné sekvenci: interpretace dotazu, sémantické vyhledávání kandidátských stránek, konstrukce syntetizované odpovědi, přiřazení citací a finální filtrování důvěryhodnosti.[36] Vyhledávací algoritmy upřednostňují weby s jasným pojmenováním entit, přímými odpověďmi v úvodu, aktuálními a časově ukotvenými fakty a minimem marketingového balastu.[36]

Zásadní limity a chování těchto systémů zmapovala studie **Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents** (Onweller et al., publikovaná 7. května 2026).[37] Výzkum hodnotil 14 modelů v režimu Deep Research podrobující citace analýze ve třech dimenzích [37]:

1. **Link Works:** Technická dostupnost a bezbariérovost citovaného URL.[37]
2. **Relevant Content:** Tematická shoda mezi tvrzením modelu a textem na zdrojové stránce.[37]
3. **Fact Check:** Nejpřísnější hodnocení shody konkrétních čísel, dat a tvrzení.[37]

Měření odhalilo zásadní rozpor v chování modelů:

- **Povrchová relevance:** Nejsilnější proprietární modely vykazovaly skvělé výsledky v dostupnosti odkazů (přes 94 %) a sémantické relevanci zdrojů (přes 80 %).[37]
- **Faktická chybovost:** Reálná shoda faktických tvrzení (Fact Check) se u špičkových modelů pohybovala v rozmezí pouhých **39 % až 77 %**.[37]
- **Degradace s hloubkou vyhledávání:** Výzkum přinesl překvapivé zjištění, že **rozsáhlejší vyhledávání nevede k přesnějším citacím**.[39] Při škálování počtu vyhledávacích kroků (tool calls) ze 2 na 150 klesla přesnost Fact Check hodnocení v průměru o **42 %**.[39] Například model GPT-5.4 vykázal propad přesnosti ze 79 % při minimálním vyhledávání na extrémních 17 % při hloubkovém vyhledávání o 150 krocích, což dokazuje, že přeplnění kontextového okna různorodými externími zdroji vede k asociačnímu chaosu a chybnému přiřazování tvrzení k autorům.[40]

Název studie / Nástroje	Zkoumaná doména	Hlavní metodologický přístup	Klíčová zjištění a metriky k 24. květnu 2026
O-Mem *(Active User Profiling)*	Dynamická personalizace a epizodická paměť LLM [18]	Iterativní extrakce rysů persony a událostí z proaktivních dialogů; mapování na kontexty [18]	Dosažení **51,76 %** na LoCoMo benchmarku a **62,99 %** na PERSONAMEM, překonání LangMem o ~3 % [18]

TSM *(Temporal Semantic Memory)*	Časová konzistence uživatelské paměti [19]	Konstrukce sémantické časové osy namísto chronologie chatu; vyhledávání durativních pamětí [19]	Zajištění časově validního kontextu; eliminace anachronismů v dlouhodobých interakcích [19]
RUMS *(Response-Utility Optimization)*	Výběr informací z paměťové databáze [20]	Výběr položek na základě měření vzájemné sémantické informace pro redukci nejistoty výstupu [20]	Optimalizace přesnosti personalizovaných predikcí; překonání metod založených na kosinové podobnosti [20]
The Personalization Trap	Vliv uživatelské paměti na emoční inteligenci LLM [22]	Evaluace 15 modelů na STEU a STEM s podsazenými socioekonomickými a demografickými profily [24]	• Pokles přesnosti u 11 z 15 modelů • Vyšší chybovost u znevýhodněných profilů (Claude 3.7 Sonnet propad o ~2,7 %) [22]
ConsistencyAI	Věcná konzistence napříč personami [21]	Dotazování 19 modelů na 15 témat s promptovým kontextem 100 různých demografických person [21]	• Průměrný index sémantického překryvu **0,8656** • Grok-3 vyhodnocen jako nejkonzistentnější model [21]
Cited but Not Verified	Kvalita zdrojové atribuce v Deep Research agentech [37]	AST syntaktická extrakce citací z Markdownu; hodnocení Link Works, Relevant Content a Fact Check [37]	• Přesnost Fact Check pouze **39–77 %** • Propad přesnosti Fact Check o **42 %** při navýšení vyhledávání ze 2 na 150 kroků [39]
StyleDecipher	Detekce strojově generovaného textu [7]	Měření stability diskrétních i spojitých stylistických deskriptorů pod vlivem kontrolovaného přepisu [7]	Robustní a vysvětlitelná detekce hybridního lidsko-strojového autorství v jednotném latentním prostoru [7]
ProLEA *(Profile Generation for EA)*	Sémantické zarovnání entit v grafech znalostí [29]	Generování textových profilů entit pomocí LLM; integrace s name-embeddings a prahováním [29]	Výrazné zvýšení robustnosti, přesnosti a vysvětlitelnosti srovnávání entit u heterogenních grafů [29]
Oxford OUP Stylometrics	Detekovatelnost generované stylometrické imitace [6]	Zero-shot generování stylů slavných autorů; klasifikace přes BERT a XGBoost na 8 stylometrických prvcích [6]	• Perplexita je nejsilnější rozlišovací metrika • Stroje nedokážou napodobit vnitřní afektivní hustotu člověka [6]

Detekce strojově generovaného obsahu a stylometrie autorství

Schopnost modelů rozlišovat mezi primárním autorem a recyklátorem AI obsahu úzce souvisí s technologiemi stylometrické klasifikace.[6] Rámce pro detekci strojového textu, jako je **StyleDecipher**, úspěšně analyzují stabilitu stylistických rysů pod vlivem kontrolovaného sémantického přepisu.[7]
Výzkum publikovaný v Oxford University Press prokázal, že strojově generovaný text vykazuje specifické statistické vlastnosti, které ho činí snadno detekovatelným [6]:

- **Distribuční regularita:** Hlavním rozlišovacím znakem mezi lidským a strojovým textem je **perplexita**.[6] AI generovaný text vykazuje vysokou distribuční uniformitu a nízkou entropii (je vysoce předvídatelný), zatímco lidské psaní se vyznačuje strukturální neuspořádaností, náhlými výkyvy v délce vět a specifickou afektivní hustotou (affective

density), kterou současné LLM (včetně GPT-4o, Gemini 1.5 Pro či Claude 3.5 Sonnet) nedokážou věrně napodobit.[6]

- **Stylistická plochost:** LLM sice dokážou bez problémů konvergovat s nízkodimenzionálními heuristickými vlastnostmi, jako je index čitelnosti či syntaktická komplexnost, ale chybí jim přirozená stylistická mikrovariace vlastní lidskému originálu.[6]

Tato stylometrická propast se přímo promítá do chování vyhledávacích a citačních algoritmů.[42] Podle rozsáhlé empirické studie společnosti **Graphite** (přelom let 2024/2025 s aktualizací pro rok 2026) vykazují vysoce optimalizované AI texty publikované za účelem parazitování na SEO výrazně horší výsledky v indexaci a citování [42]:

- **Penalizace vyhledávači:** V běžných výsledcích vyhledávání Google tvoří strojové texty pouze 14 %, přičemž na prvních třech pozicích klesá jejich zastoupení na pouhých 7 %.[42]
- **Citační preference LLM:** Při analýze citačních vzorců u ChatGPT a Perplexity bylo zjištěno, že **82 % všech citovaných zdrojů tvoří prokazatelně lidské texty**, zatímco strojově generované materiály jsou zastoupeny pouze v 18 % případů.[42] Vyhledávací LLM při křížové validaci a sémantické analýze konsenzu automaticky upřednostňují lidské texty kvůli jejich vyšší sémantické specifičnosti a absenci repetitivních syntaktických vzorců, které jsou vyhodnocovány jako druhotný, méně důvěryhodný balast.[6]

Rozlišování primárních a sekundárních zdrojů

Při hodnocení relevance textu se uplatňuje složitá hierarchie posuzování věcné správnosti a zakotvenosti.[37] V rámci systémů RAG se k hodnocení kvality generovaného textu vůči předlohám používají specializované metriky [43]:

- **Faithfulness (Věrnost):** Měří integritu odpovědi vůči načtenému kontextu a ověřuje, zda nedošlo k halucinacím.[43]
- **Answer Relevancy (Relevance odpovědi):** Posuzuje, zda text přímo reaguje na sémantické jádro dotazu, nebo se uchyluje k prázdnému opisování.[43]
- **QA Correctness (Věcná správnost):** Srovnává výstup s ověřenou bází faktů.[43]

Aplikace těchto metrik přes metodu LLM-as-a-judge (využívající Chain-of-Thought a párové srovnávání) odhaluje, že modely dokáží identifikovat sekundární a parazitní zdroje na základě chybějícího faktického zakotvení a nadměrného používání vágních, positioningových obratů.[36] Primární zdroje, které obsahují konkrétní entity, data umístěná bezprostředně vedle věcných tvrzení a jasnou strukturu, jsou vyhodnocovány jako sémanticky stabilnější a získávají v RAG systémech prioritu.[36] Naopak texty vykazující vysokou distribuční pravidelnost bez sémantických mikrovariací jsou klasifikovány jako sekundární deriváty a jsou z citací vytlačovány.[6]

Srovnávací syntéza a kritická analýza hypotéz

Analýza obou pozorování odkrývá hluboké strukturální vazby, ale zároveň vyžaduje striktní odmítnutí antropomorfních dezinterpretací v oblasti vnitřního fungování modelů.[21]

Průsečíky a odlišnosti paměťových a typologických mechanismů

Oba zkoumané jevy sdílejí totožný matematický a sémantický základ v latentním prostoru modelu.[5] Křižovatka obou mechanismů se projevuje v simulaci digitální stopy: systémy jako *PersonaTrace* dokáží vygenerovat konzistentní behaviorální artefakty (SMS, maily), které následně vstupují do paměťových a profilových analýz jiných modelů.[32] Stejně tak projekty jako *scELMo* ukazují, že sémantické popisy generované LLM lze přímo integrovat s komplexními biologickými daty pro zero-shot klasifikaci.[45]

Zásadní rozdíl však spočívá v **perzistenci a kontrole**: uživatelská paměť je explicitně řízené, klientsky či serverově strukturované úložiště (často oddělené do barierových adresářů či tabulek faktů), které je záměrně injektováno do kontextového okna pro zvýšení konzistence.[4] Naopak typologické zařazení cizí osoby je bezprostředním, bezestavovým

matematickým výpočtem sémantické a stylometrické shody prováděným v reálném čase nad aktuálně vyhledanými daty.[6]

Slabá místa a kritická dekonstrukce hypotéz

Empirická verifikace k 24. květnu 2026 odhaluje několik slabých míst v původní interpretaci chování modelů a vyžaduje korekci předložených hypotéz:

1. **Iluze perzistentního profilování cizích osob:** Hypotéza předpokládá, že si modely o lidech udržují typologii nezávisle na sdílené paměti. Ve skutečnosti neexistuje žádná globální, skrytá databáze typologických štítků pro méně známé osoby. Pokud model na čistém účtu prohlásí o autorovi, že *"podle všeho nepřepisuje výstupy GPT"*, nejde o načtení dříve uložené typologické kategorie.[6] Jde o výsledek **zero-shot stylometrické analýzy v reálném čase**.[6] Model během vyhledávání načetl texty dané osoby a okamžitě změřil jejich distribuční perplexitu a afektivní hustotu.[6] Zjištěná vysoká variabilita a nízká regularita vedly k automatickému sémantickému závěru, že text nese signatury primárního lidského autorství.[6] Tento úsudek je generován ad-hoc a zmizí s uzavřením kontextového okna.
2. **Nestabilita a nespolehlivost úsudků:** Původní hypotéza postuluje, že různé modely dospějí při čtení stejné stopy k podobnému zařazení. Výzkumy věcné konzistence (např. *ConsistencyAI*) a behaviorální fluktuace (*The Personalization Trap*) toto tvrzení zásadně zpochybňují.[21] Míra shody (cross-persona cosine similarity) vykazuje značnou variabilitu v závislosti na poskytovateli modelu a konkrétním tématu.[21] Lehčí modely vykazují nízkou konzistenci a snadno podléhají sémantickému šumu.[21] Navíc vysoká míra *flip rate* u znevýhodněných profilů dokazuje, že "úsudky" modelů o lidech nejsou objektivní analýzou charakteru, ale jsou extrémně náchylné k předsudkům, negativnímu primingu a halucinování souvislostí na základě drobných nuancí v textu.[22]
3. **Původ názorové vyhraněnosti:** Formulace hodnocení jako "vlastního názoru" modelu není projevem autonomního vědomí či sebeuvědomění. Jde o přímý důsledek RLHF (Reinforcement Learning from Human Feedback) ladění, které záměrně upřednostňuje asertivní a sebevědomý tón prezentace informací před opatrným a neutrálním vyjadřováním, neboť asertivní výstupy jsou lidskými hodnotiteli vnímány jako fakticky přesnější, bez ohledu na jejich reálnou správnost.[46]

Matematické vyjádření sémantické shody

Pro kvantifikaci shody a konzistence úsudků napříč různými personami a modely se v rámci benchmarků uplatňuje výpočet sémantického překryvu pomocí kosinové podobnosti větových embeddingů vyjádřený vztahem:

$$\text{CosineSim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i-1}^{d} u_i v_i}{\sqrt{\sum_{i-1}^{d} u_i^2}\sqrt{\sum_{i-1}^{d} v_i^2}}$$

[21]

Tento vztah slouží jako základní metrika pro ověřování sémantické koherence výstupů; nízké hodnoty indikují, že model podlehl vlivu podsazeného profilu a změnil faktické jádro odpovědi.[21]

Nuancovaný sémantický konsensus k 24. květnu 2026

Předložená analýza zakládá jasný a datovaný referenční bod. K 24. květnu 2026 je prokázáno, že velké jazykové modely dosáhly vysoké technické vyspělosti v integraci uživatelské paměti a v asynchronní správě kontextu.[4] Tyto systémy však vykazují vážné kontextové deformace, kdy přítomnost uživatelské paměti prokazatelně degraduje kvalitu rozhodování a aktivuje sociální a socioekonomické předsudky ukotvené v latentním prostoru.[22]

V oblasti autonomního hodnocení třetích osob jsou modely schopny provádět vysoce přesnou real-time stylometrickou analýzu (identifikaci perplexity a afektivní hustoty), což jim umožňuje efektivně filtrovat strojový balast a upřednostňovat lidské autory v citačních vzorcích.[6] Nicméně představa o existenci stabilních, perzistentních typologických profilů cizích osob v latentním prostoru modelů je vědecky vyvrácena; toto chování je plně vysvětlitelné ad-hoc sémantickým výpočtem v kontextovém okně a asertivním tónem vynuceným fází alignmentu.[6] Tento dokument bude sloužit jako výchozí bod pro budoucí srovnání vývoje těchto kognitivních a strukturálních vrstev LLM.

Citovaná díla

1. Memory FAQ - OpenAI Help Center, použito května 24, 2026, https://help.openai.com/en/articles/8590148-memory-faq
2. Claude Memory 2026: Complete Guide — What It Stores, How to ..., použito května 24, 2026, https://lumichats.com/blog/claude-memory-2026-complete-guide-how-to-use
3. Google's Gemini Memory Feature: The Industrial Consolidation of Cambridge Analytica's Playbook, použito května 24, 2026, https://cambridgeanalytica.org/news/google-s-gemini-memory-feature-the-industrial-consolidation-of-cambridge-analytica-s-playbook-50358/
4. How ChatGPT Memory Works, Reverse Engineered - LLMrefs, použito května 24, 2026, https://llmrefs.com/blog/reverse-engineering-chatgpt-memory
5. Localizing Persona Representations in LLMs | Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, použito května 24, 2026, https://ojs.aaai.org/index.php/AIES/article/view/36577
6. Decoding AI authorship: can LLMs truly mimic human style across literature and politics?, použito května 24, 2026, https://academic.oup.com/dsh/advance-article-abstract/doi/10.1093/llc/fqag040/8586905
7. StyleDecipher: Robust and Explainable Detection of LLM-Generated Texts with Stylistic Analysis - arXiv, použito května 24, 2026, https://arxiv.org/html/2510.12608v1
8. Configure personalization and memory | Gemini Enterprise - Google Cloud Documentation, použito května 24, 2026, https://docs.cloud.google.com/gemini/enterprise/docs/configure-personalization
9. Release notes | Claude Help Center, použito května 24, 2026, https://support.claude.com/en/articles/12138966-release-notes
10. Using agent memory - Claude API Docs, použito května 24, 2026, https://platform.claude.com/docs/en/managed-agents/memory
11. Memory tool - Claude API Docs, použito května 24, 2026, https://platform.claude.com/docs/en/agents-and-tools/tool-use/memory-tool
12. Get personalization with memory of your past Gemini chats - Android - Google Help, použito května 24, 2026, https://support.google.com/gemini/answer/16598469?hl=en&co=GENIE.Platform%3DAndroid
13. Personal Intelligence from Gemini — AI help just for you, použito května 24, 2026, https://gemini.google/overview/personal-intelligence/
14. What is Memory? - OpenAI Help Center, použito května 24, 2026, https://help.openai.com/en/articles/8983136-what-is-memory
15. Memory and new controls for ChatGPT - OpenAI, použito května 24, 2026, https://openai.com/index/memory-and-new-controls-for-chatgpt/
16. Gemini Memory Guide: Enhance Your G Suite Usage with Smarter AI - Workalizer, použito května 24, 2026, https://workalizer.com/blog/apps-tools/unlock-geminis-memory-a-google-workspace-expert-guide-to-smarter-ai-interactions/
17. User Profile with Large Language Models: Construction, Updating, and Benchmarking, použito května 24, 2026, https://arxiv.org/html/2502.10660v1
18. O-Mem: Omni Memory System for Personalized, Long Horizon, Self-Evolving Agents - arXiv, použito května 24, 2026, https://arxiv.org/html/2511.13593v1
19. Beyond Dialogue Time: Temporal Semantic Memory for Personalized LLM Agents - arXiv, použito května 24, 2026, https://arxiv.org/html/2601.07468v1
20. Response-Aware User Memory Selection for LLM Personalization - arXiv, použito května 24, 2026, https://arxiv.org/html/2604.14473v1
21. ConsistencyAI: A Benchmark to Assess LLMs' Factual Consistency When Responding to Different Demographic Groups - arXiv, použito května 24, 2026, https://arxiv.org/html/2510.13852v1
22. THE PERSONALIZATION TRAP: HOW USER MEMORY ALTERS EMOTIONAL REASONING IN LLMS - OpenReview, použito května 24, 2026, https://openreview.net/pdf?id=u9Qgn8xSx1
23. The Personalization Trap: How User Memory Alters Emotional Reasoning in LLMs - arXiv, použito května 24, 2026, https://arxiv.org/abs/2510.09905
24. THE PERSONALIZATION TRAP: HOW USER MEMORY ALTERS EMOTIONAL REASONING IN LLMS | OpenReview, použito května 24, 2026, https://openreview.net/forum?id=u9Qgn8xSx1
25. The Personalization Trap: How User Memory Alters Emotional Reasoning in LLMs - arXiv, použito května 24, 2026, https://arxiv.org/html/2510.09905v1
26. Towards Realistic Personalization: Evaluating Long-Horizon Preference Following in Personalized User-LLM Interactions - arXiv, použito května 24, 2026, https://arxiv.org/html/2603.04191v1
27. Language Models Represent Space and Time - OpenReview, použito května 24, 2026, https://openreview.net/forum?id=jE8xbmvFin
28. Revealing emergent human-like conceptual representations from language prediction - PMC, použito května 24, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12595492/
29. Entity Profile Generation and Reasoning with LLMs for Entity Alignment - ACL Anthology, použito května 24, 2026, https://aclanthology.org/2025.findings-emnlp.1093.pdf

30. A Round-Trip Evaluation of LLM-Generated Life Stories Conditioned on Rich Psychometric Profiles - arXiv, použito května 24, 2026, https://arxiv.org/html/2604.06071v1
31. Entity Profile Generation and Reasoning with LLMs for Entity Alignment - ACL Anthology, použito května 24, 2026, https://aclanthology.org/2025.findings-emnlp.1093/
32. PersonaTrace: Synthesizing Realistic Digital Footprints with LLM Agents - ACL Anthology, použito května 24, 2026, https://aclanthology.org/2026.eacl-industry.5.pdf
33. How Perplexity AI Selects Sources: Best Guide For 2026 - Sight AI, použito května 24, 2026, https://www.trysight.ai/blog/how-perplexity-ai-selects-sources
34. How LLMs Rank Content: Understanding AI Search Algorithms | Hashmeta, použito května 24, 2026, https://hashmeta.com/blog/how-llms-rank-content-understanding-ai-search-algorithms/
35. How AI Search Engines Choose Sources: The Complete Guide to LLM Citation Selection in 2025 - Mention Stack, použito května 24, 2026, https://www.mentionstack.com/post/how-ai-search-engines-choose-sources-guide
36. How Perplexity Selects Sources in 2026: Citation Rules, Ranking Signals, and What Brands Can Do - AuthorityTech, použito května 24, 2026, https://authoritytech.io/blog/how-perplexity-selects-sources-algorithm-2026
37. Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents - arXiv, použito května 24, 2026, https://arxiv.org/html/2605.06635v1
38. Mentors in LLM security and safety - Scouts by Yutori, použito května 24, 2026, https://scouts.yutori.com/1c15dd3e-fa6e-4daa-96c8-d7cfe161b321
39. [2605.06635] Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents - arXiv, použito května 24, 2026, https://arxiv.org/abs/2605.06635
40. (PDF) Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents - ResearchGate, použito května 24, 2026, https://www.researchgate.net/publication/404626821_Cited_but_Not_Verified_Parsing_and_Evaluating_Source_Attribution_in_LLM_Deep_Research_Agents
41. [2510.13852] ConsistencyAI: A Benchmark to Assess LLMs' Factual Consistency When Responding to Different Demographic Groups - arXiv, použito května 24, 2026, https://arxiv.org/abs/2510.13852
42. How Does AI-Generated Content Perform in Search and Answer Engines? - Graphite.io, použito května 24, 2026, https://graphite.io/five-percent/ai-content-in-search-and-llms
43. How to Effectively Evaluate Retrieval-Augmented Generation (RAG) Systems, použito května 24, 2026, https://www.louisbouchard.ai/rag-evals/
44. A semantic embedding space based on large language models for modelling human beliefs, použito května 24, 2026, https://arxiv.org/html/2408.07237v3
45. Embeddings from language models are good learners for single-cell data analysis - PMC, použito května 24, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12921509/
46. Personalized Benchmarking: Evaluating LLMs by Individual Preferences - arXiv, použito května 24, 2026, https://arxiv.org/html/2604.18943v1
47. Mastering RAG: How To Evaluate LLMs For RAG - Galileo AI, použito května 24, 2026, https://galileo.ai/blog/how-to-evaluate-llms-for-rag