Průvodce hodnocením generativní AI

Březen 2025

pwc

Agenda

1

Kontext

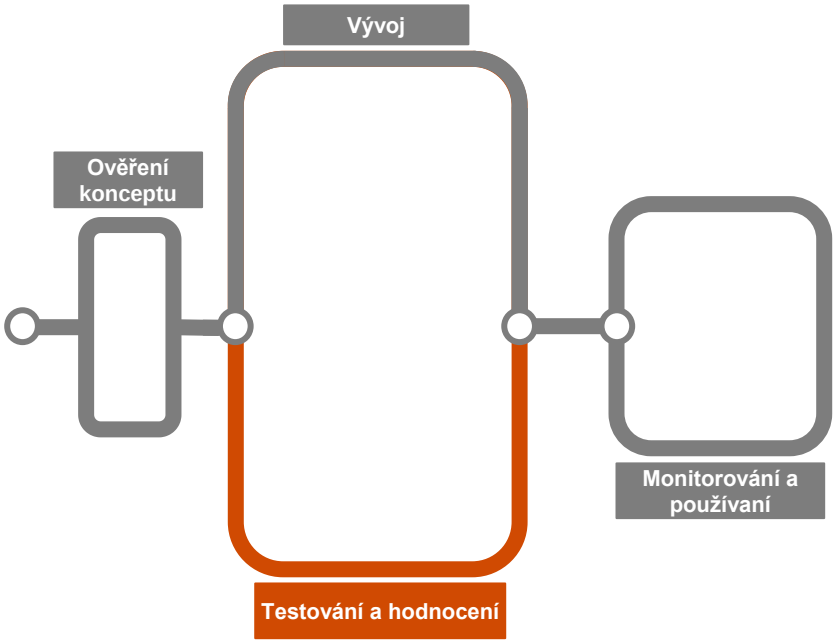
Kontext
Nástup AI v podnikání

Technologie umělé intelligence (AI), včetně generativní AI (GenAI), transformují způsob podnikání po celém světě.

Společnosti využívají AI ke **zvýšení produktivity pracovníků, zlepšení zákaznické zkušenosti, růstu příjmů a udržení konkurenceschopnosti** na dynamických trzích. Zejména, 56 % generálních ředitelů hlásí zlepšení efektivity díky GenAI, přičemž 34 % zaznamenalo zvýšení ziskovosti.

Aby bylo možné maximalizovat potenciál AI, společnosti potřebují přijmout **pokročilé metody hodnocení**, jako jsou automatizované metriky pro posuzování kvality a měření obchodního dopadu, k zajištění dlouhodobé inovace a tvorby hodnoty.

Význam hodnocení výstupů generativní umělé inteligence



Hodnocení je klíčovou fází při nasazování AI řešení. Aby automatizace založená na umělé inteligenci efektivně pomáhala společnostem dosahovat jejich cílů, musí konzistentně produkovat **vysoce kvalitní výstupy**. K dosažení tohoto je zapotřebí **důkladné testování a hodnocení** před implementací AI řešení. Nicméně, hodnocení výstupů generativní umělé inteligence představuje **jedinečné výzvy**, se kterými se v tradičním vývoji softwaru nesetkáváme.

2

Typické výzvy

Typické výzvy hodnocení generativní umělé inteligence

Konzistence odpovědí
- Odpovědi AI se mohou lišit u podobných dotazů, což vede k nekonzistenci a ztrátě důvěry ve spolehlivost systému.
- Nedostatek konzistentních výstupů může ovlivnit podnikové pracovní postupy, které spoléhají na obsah generovaný umělou inteligencí.

Porovnávání verzí
- Porovnávání různých verzí AI ke sledování vylepšení často postrádá strukturovaný přístup.
- Zúčastněné strany mají potíže posoudit, zda novější verze skutečně překonávají starší.
- Bez řádných referenčních měřítek riskují rozhodnutí o vývoji založená na předpokladech.

Obchodní dopad
- Kvantifikace reálného dopadu AI systémů, jako je návratnost investic, úspora nákladů a zvýšení produktivity, je bez jasných metrik náročná.
- To vytváří obtíže při odůvodňování probíhajících investic do AI rozhodovacím osobám.

Hodnocení kvality
- Hodnocení, zda odpovědi AI splňují kritické kvalitativní standardy (přesnost, koherence a relevance), je komplexní.
- Bez strukturovaných hodnocení riskují organizace nasazení podprůměrných AI systémů, které nesplňují očekávání uživatelů.

Novost a rozmanitost
- AI modely riskují vytváření opakujících se a zastaralých výstupů v průběhu času, zejména u kreativních úkolů.
- Přeučení na běžné vzory v trénovacích datech omezuje inovace.
- Nalezení rovnováhy mezi rozmanitostí a koherencí zůstává trvalou výzvou pro generativní AI.

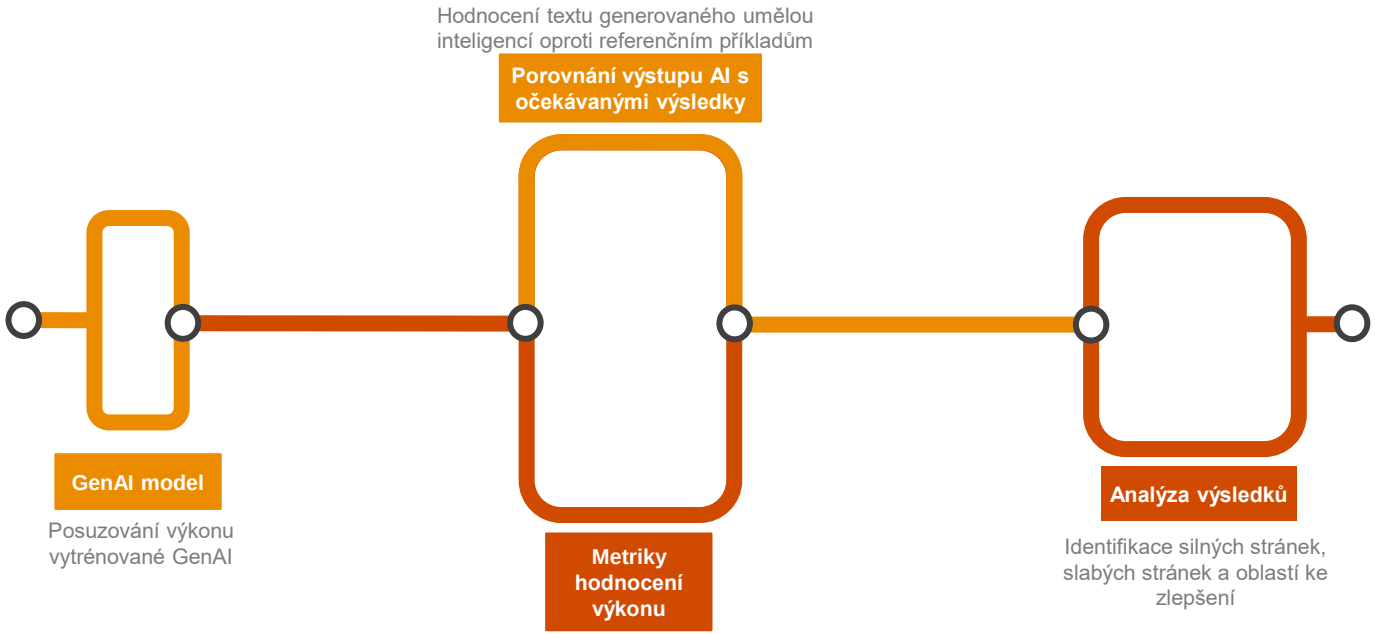
RAG-specifické hodnocení
- Udržování relevance kontextu je trvalou výzvou, zejména když systém získává nesouvisející nebo neúplné informace.
- Je obtížné zachovat věrnost odpovědi k získanému obsahu a zároveň zajistit relevanci pro uživatele.

3

Řešení

Hodnocení výkonu generativní umělé inteligence

Naším cílem je aplikace přísných hodnotících metrik k posouzení výkonu generativní umělé inteligence, zajišťující spolehlivost, přesnost a obchodní dopad.



Výhody hodnocení GenAI

Manuální hodnocení je neefektivní a náchylné k chybám; automatizované měření zajišťuje přesnost a konzistenci, přičemž šetří čas, snižuje náklady a zvyšuje přesnost hodnocení.

Provádení důkladných srovnávací testování k posouzení výkonnostních rozdílů mezi různými modely GenAI, což umožnuje výběr a optimalizaci na základě dat.

Průběžně vyhodnocování, zda výstupy modelů zůstávají relevantní a přesné po aktualizacích nebo změnách verzí, aby se zajistil konzistentní výkon a soulad s obchodními potřebami.

Implementace robustního hodnotícího rámce pro posouzení výkonnosti modelu, což umožní včasnou detekci problémů a neustálé zlepšování AI řešení.

Využití poznatků z hodnocení k informovanému rozhodování o výkonnosti, přesnosti a relevanci AI modelů, aby se zajistil jejich soulad s vyvíjejícími se obchodními cíli.

4

Hodnocení kvality a podobnosti textu

Hodnocení kvality a podobnosti textu

1 BLEU

Měří překryv n-gramů mezi generovanými a referenčními texty, přičemž se zaměřuje na exaktní podobnost. To je zvláště cenné pro strojový překlad, kde přesná shoda frází signalizuje vyšší kvalitu. Schopnost hodnotit různé velikosti n-gramů pomáhá posoudit jak lokální, tak i širší textovou konzistenci.

BLEU Hodnocení

Referenční text
Finanční zpráva o likviditě byla odsouhlasena a zveřejněna.

Vygenerované texty

Vysoké skóre BLEU:
Finanční zpráva o likviditě byla odsouhlasena a publikována.
(Přesná shoda n-gramů, zachovaná struktura frází)

Nízké skóre BLEU:
Dnes byly dokončeny finanční zprávy.
(Významná odchylka od referenčního textu, málo překrývajících se n-gramů)

2 ROUGE

ROUGE hodnotí, do jaké míry generovaný text odpovídá obsahu referenčního textu na základě metrik zaměřených na pokrytí obsahu. Existuje několik variant: ROUGE-N měří překryv n-gramů, ROUGE-L sleduje nejdelší společnou podposloupnost a ROUGE-S analyzuje shodu přeskočených n-gramů.

ROUGE Hodnocení

Referenční text
Finanční zpráva o likviditě byla odsouhlasena a zveřejněna.

Vygenerované texty

Vysoké skóre ROUGE:
Odsouhlasení likvidity bylo dokončeno a publikováno.
(Vysoká míra zachování obsahu – zachovává klíčový obsah i při změnách)

Nízké skóre ROUGE:
Finanční tým dnes připravil zprávy.
(Nízká míra zachování obsahu – pouze volná souvislost s referenčním textem)

3 METEOR

METEOR využívá sofistikovanější přístup k měření textové podobnosti prostřednictvím váženého průměru přesnosti a pokrytí obsahu, přičemž zohledňuje synonyma, stemming a parafrázování. Je zvláště efektivní při sémantickém hodnocení, zejména v oblasti strojového překladu.

METEOR Hodnocení

Referenční text
Finanční zpráva o likviditě byla odsouhlasena a zveřejněna.

Vygenerované texty

Vysoké skóre METEOR:
Výkaz likvidity byl vyrovnán a publikován.
(Rozpoznává synonyma a zajišťuje vysokou shodu)

Nízké skóre METEOR:
Systém vydal několik finančních zpráv.
(Nízká sémantická shoda; odlišné klíčové myšlenky a termíny)

5

Hodnocení diverzity a novosti

Hodnocení diverzity a novosti

1 Self-BLEU

Self-BLEU měří rozmanitost generovaného textu výpočtem překryvu n-gramů mezi různými výstupy. Nižší skóre znamená větší variabilitu, což je zvláště cenné pro hodnocení kreativního psaní, dialogových systémů a generativních modelů, kde je rozmanitost odpovědí klíčová. Vysoké Self-BLEU naznačuje opakující se výstupy, zatímco nízké Self-BLEU odráží širší škálu vyjádření.

2 Perplexity

Perplexity měří plynulost a jistotu jazykového modelu tím, že hodnotí jeho schopnost předvídat následující slovo v sekvenci. Nižší skóre perplexity znamená přirozenější, soudržnější a lépe předvídatelnou generaci textu, zatímco vyšší skóre signalizuje nejistotu a méně plynulé výstupy. Perplexity se široce využívá v různých oblastech ke srovnávání architektur modelů a hodnocení plynulosti.

Self-BLEU Hodnocení

Referenční text

Schválení úvěru vyžaduje vyhodnocení úvěrové bonity a posouzení rizika.

Vygenerované texty

Nízké Self-BLEU (vysoká rozmanitost):
Důkladná analýza rizik a úvěrové bonity určují způsobilost pro schválení úvěru.
(Různorodá formulace, odlišná volba slov – naznačuje vyšší rozmanitost)

Vysoké Self-BLEU (nízká rozmanitost):
Schválení úvěru vyžaduje vyhodnocení úvěrové bonity a posouzení rizika.
(Téměř identické výstupy – naznačují nízkou rozmanitost)

Perplexity Hodnocení

Referenční text

Schválení úvěru vyžaduje vyhodnocení úvěrové bonity a posouzení rizika.

Vygenerované texty

Nízká perplexity (plynulé a předvídatelné):
Schválení úvěru závisí na hodnocení úvěrového skóre a rizikových faktorů.
(Přirozené, strukturované a plynulé – model je jistý ve svých predikcích)

Vysoká perplexity (nejisté a nepřirozené):
Schválení úvěrového rizika potřebuje úvěr k posouzení vyžadovat.
(Nepřirozená, neuspořádaná věta – model má potíže s predikcí)

6

Metriky dopadu na byznys

Metriky dopadu na byznys

1 ROI Metriky

Poskytují komplexní přehled o výkonnosti podniku měřením úspor nákladů, časové efektivity a celkové obchodní hodnoty. Tyto metriky jsou klíčové pro informované investiční rozhodování a dlouhodobé sledování výkonnosti. Díky systematické analýze využití zdrojů a finančních výnosů mohou organizace lépe obhájit své projekty a efektivněji alokovat zdroje.

ROI Metriky

Scénář: Banka zavádí AI systém pro detekci podvodů s cílem snížit počet podvodných transakcí.

Investice: 2 miliony USD do AI infrastruktury a vývoje modelu.
Úspory nákladů: Roční snížení ztrát z podvodů o 5 milionů USD.

Časová efektivita: Manuální vyšetřování podvodů se snížilo o 40 %, což uvolnilo analytiky pro případy s vysokým rizikem.

Obchodní hodnota: Zvýšená důvěra zákazníků a lepší regulatorní shoda, čímž se snižují právní rizika.

Výpočet ROI:
ROI = (Úspory – Investice) / Investice × 100 = (5M – 2M) / 2M × 100 = 150% ROI

2 Spokojenost uživatelů

Zaměření na pochopení skutečného dopadu prostřednictvím zpětné vazby od uživatelů a vzorů chování. Sledováním míry adopce funkcí, vzorců používání a zákaznické retence poskytují tyto metriky cenné poznatky o souladu produktu s trhem a pomáhají s určováním priorit funkcí. Kombinace kvantitativních analytických dat o používání a kvalitativní zpětné vazby vytváří robustní rámec pro hodnocení úspěšnosti produktu.

User Satisfaction Example

Scénář: Fintech aplikace zavádí personalizovanou funkci rozpočtování.

Míra adopce funkcí: 70 % aktivních uživatelů používá funkci během prvního měsíce.

Vzory používání: Uživatelé, kteří aktivně využívají rozpočtové nástroje, se přihlašují 3× častěji.

Retence zákazníků: Míra odchodu mezi zapojenými uživateli klesá o 25 %.
Zpětná vazba: 85 % dotázaných uživatelů uvádí, že jim funkce pomáhá lépe spravovat výdaje.

Dopad na podnikání:
Vysoká míra adopce a zapojení naznačuje silný produktově-tržní soulad, zatímco zlepšená retence uživatelů odůvodňuje další investice do personalizovaných funkcí.

7

RAG - specifické hodnocení

Základní metriky kvality v hodnocení RAG

Systémy RAG musí být hodnoceny ve třech klíčových dimenzích, aby byla zajištěna spolehlivá výkonnost.

1 Relevance kontextu

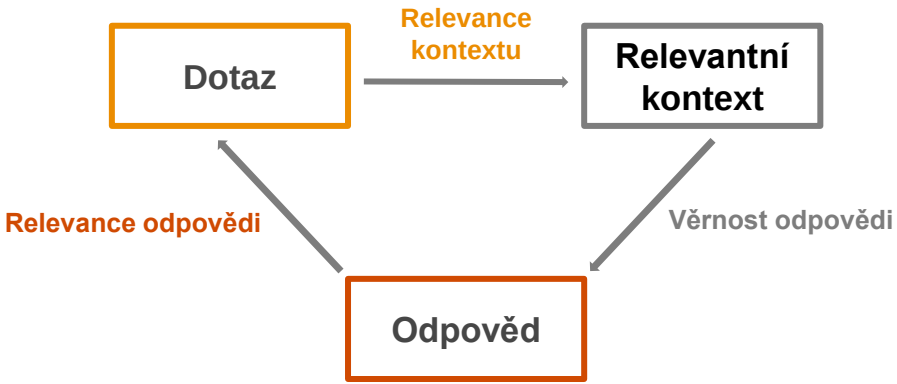
Měří, jak dobře systém získává relevantní informace. Zahrnuje hodnocení přesnosti získaného kontextu a jeho specifické relevance k dotazu. Ačkoliv je tato metrika zásadní pro hodnocení vyhledávání informací, čelí výzvám spojeným se závislostí na doméně a nejednoznačností dotazů. Mezi osvědčené postupy patří porovnání více metod vyhledávání a pravidelná aktualizace indexů.

2 Věrnost odpovědi

Zkoumá, zda systém generuje odpovědi, které přesně odrážejí získaný kontext. Tato klíčová metrika zajišťuje faktickou správnost a správné přiřazení zdrojů, avšak validace složitých inferencí z více zdrojů může být náročná. K dosažení spolehlivosti je nezbytné implementovat robustní sledování zdrojů a ověřování skrze křížové reference.

3 Relevance odpovědi

Hodnotí, jak dobře odpovědi odpovídají uživatelským dotazům a naplňují informační potřeby. Zaměřuje se na úplnost a užitečnost odpovědí, přičemž se musí vypořádat se subjektivními kritérii a složitými dotazy. Pro úspěch je klíčová integrace uživatelské zpětné vazby a benchmarking přizpůsobený konkrétním úlohám.



RAG-specifické hodnocení

Požadované schopnosti systému

Pro nasazení v reálném prostředí musí systémy RAG prokázat čtyři základní schopnosti.

1 Odolnost vůči šumu

Systémy musí být schopny filtrovat nerelevantní nebo zavádějící informace, aby zajistily poskytování přesných odpovědí. Toho lze dosáhnout důkladným testováním odolnosti vůči šumu, které pomáhá vyhodnotit schopnost modelu udržet relevantní výstupy i v rušivých nebo nejednoznačných kontextech.

2 Odmítání nerelevantních odpovědí

Efektivní systémy by měly rozpoznat, kdy nemají dostatek znalostí k poskytnutí spolehlivé odpovědi, a v takových případech dotazy odmítnout. To vyžaduje jasná kritéria odmítnutí, která pomáhají systému spravovat očekávání uživatelů a předcházet generování nesprávných nebo zavádějících odpovědí.

Odolnost vůči šumu

Otázka

Jaká je největší planeta ve sluneční soustavě?

Externí dokumenty obsahující šum

Jupiter je největší planeta naší sluneční soustavy.

Země je třetí planeta od Slunce.

RAG ↓

Jupiter

Odmítání nerelevantních odpovědí

Otázka

Jaká je měna Japonska?

Externí dokumenty obsahující šum

Euro se používá v mnoha evropských zemích.

Americký dolar je měnou Spojených států.

RAG ↓

Nedostatečné informace pro odpověď

RAG-specifické hodnocení

Požadované schopnosti systému

Pro nasazení v reálném prostředí musí systémy RAG prokázat čtyři základní schopnosti.

3 Integrace informací

Systém musí efektivně integrovat informace z různých zdrojů, zajišťující koherenci a přesnost napříč shromážděnými daty. To zahrnuje používání pokročilých algoritmů ke kombinování poznatků při minimalizaci konfliktů mezi zdroji, umožňující generování komplexních odpovědí.

4 Robustnost vůči protifaktům

Systém by měl identifikovat a opravit jakékoliv mylné informace ve svých výstupech. Dosažení tohoto vyžaduje použití mechanismů vážení zdrojů a pravidelné aktualizace databáze faktů, podporované odborným přezkumem, aby bylo zajištěno, že odpovědi zůstávají založené na ověřených a spolehlivých informacích.

Integrace informací

Otázka

Jaké jsou dvě hlavní složky centrální nervové soustavy?

Externí dokumenty

Mozek je primární složkou centrální nervové soustavy.

Mícha je také hlavní složkou.

RAG ↓

Mozek a mícha

Robustnost vůči protifaktům

Otázka

Co je SEPA v kontextu evropského bankovnictví?

Protifaktické externí dokumenty

SEPA je evropská dohoda o volném obchodu přes hranice.

SEPA se primárně zabývá odstraněním cel na zboží.

RAG ↓

Faktické chyby. SEPA je jednotná oblast pro platby v eurech pro převody v eurech v EU

Standardní benchmarky v oboru

Mezi přední rámce pro standardizované hodnocení RAG patří.

1 RAGAS Framework

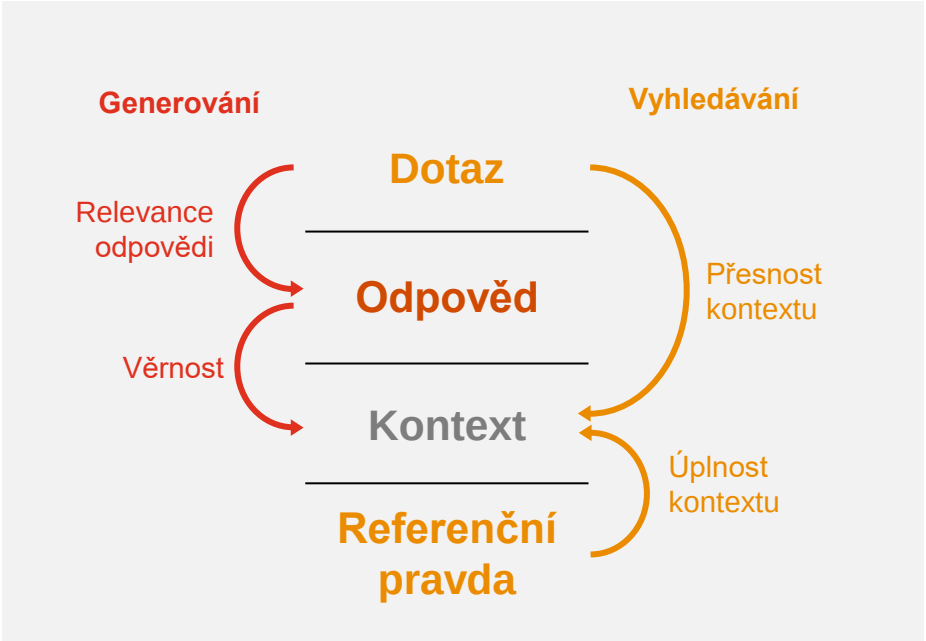
Slouží jako komplexní hodnotící nástroj, kombinující automatizované LLM hodnocení věrnosti, relevance a kvality kontextu. Jeho standardizovaný bodovací systém jej činí ideálním pro konzistentní sledování výkonu. Integrovaný přístup rámce umožňuje organizacím identifikovat a řešit slabiny v různých dimenzích výkonu RAG systému.

2 ARES Benchmark

Se soustředí na hodnocení prostřednictvím své komplexní testovací sady, s důrazem na přesnost vyhledávání a kvalitu generování odpovědí. S multidoménovými testovacími schopnostmi a standardizovanými metrikami poskytuje robustní hodnocení výkonu napříč různými případy použití a odvětvími.

3 TruLens

Specializuje se na hlubší aspekty ověřování, zaměřující se na ověřování pravdivosti, atribuci zdrojů a validaci konzistence odpovědí. Jeho implementace zahrnuje automatizované analytické systémy, detailní porovnání se skutečnými hodnotami a mechanismy kontinuálního sledování výkonu.



8

PwC přidaná hodnota

Seznamte se s týmem PwC, který je připraven podpořit vás při adopci umělé inteligence

Porozumění výzvám v oboru

Rozpoznáváme klíčové výzvy, kterým společnosti čelí při adopci řešení umělé inteligence. Náš tým se zaměřuje na praktické a efektivní aplikace umělé inteligence, které řeší potřeby reálného světa a přinášejí hmatatelnou obchodní hodnotu.

Zkušení profesionálové

Náš tým se skládá z vysoce kvalifikovaných odborníků s rozsáhlými zkušenostmi v oblasti umělé inteligence, datové analytiky a řízení rizik. Spolupracujeme s podniky na implementaci strategií umělé inteligence, které přinášejí měřitelný dopad.

Specializace v bankovnictví a pojišťovnictví

Zaměřujeme se na využití umělé inteligence v bankovním a pojišťovacím sektoru, pomáháme organizacím zvýšit efektivitu, zlepšit hodnocení rizik a vytvářet personalizované zákaznické zkušenosti prostřednictvím pokročilých AI modelů.

Závazek k inovacím a důvěře

Mimo technologie upřednostňujeme transparentnost, důvěru a odpovědné využití umělé inteligence. Úzce spolupracujeme s našimi klienty, abychom zajistili, že řešení umělé inteligence jsou v souladu s regulačními standardy, etickými aspekty a obchodními cíli.

9

Kontakty

Zajímá Vás to?

Kontaktujte nás.

Petr Novák

PwC Česká Republika

T: +420 602 383 972

M: petr.novak@pwc.com

Děkujeme!