IIČR 2026 · Informace, inteligence, čas rozhodnutí

Kdo v řetězci agentů rozhoduje za stát

Jak měřit rozpory mezi agenty, transparentnost syntézy a potřebu lidského přezkumu

doc. Ing. Jaroslav Dočkal, CSc. Střední škola a vyšší odborná škola informatiky a financí Brno

formulace výzkumného problému, stanovení cíle, návrh konceptu, výběr metod, interpretace výsledků vše za pomoci Claude *(Anthropic)*

Od „kde leží data“ k „kdo za nás myslí“

Digitální suverenita nekončí u umístění dat a výběru modelu. V multiagentních systémech závisí také na tom, kdo řídí spolupráci agentů, jak se řeší jejich rozpory a zda lze rozhodovací proces auditovat.

Rovina A — Kognitivní suverenita

- Na základě čeho stát myslí a rozhoduje (při přetížení, delegování, atrofie)
- Multiagentní provedení: kdo řídí orchestraci

Rovina B — Strategický dopad

- Schopnost samostatnost prakticky vybudovat a chránit
- Nový strategický nástroj „orchestrační suverenita“ (schopnost státu bránit své procesy)

ezpečnostní analytik

ecializovaný agent

o
ní
oty
roje

Ekonomický analytik

ecializovaný agent

o
ní
oty
roje

Arbitr

syntetizující agent

Pravidla syntézy:

- vážení stanovisek
- práce s nejistotou
- zachování nesouhlasu

Finální doporučení

na stole rozhodovatele

Řešený problém

Zdvojená orchestrace (technická a kognitivní)

- Jednotliví specialisté odpoví odlišně – což se i očekává
- Otázkou ale je: Co s rozdílem udělá Arbitr?
- Riziko: hladké doporučení bez stopy po rozporu

Proč užitečné

- ČR: výzkumná mezera v aplikaci na scénáře státní správy
- Navrhuji propojit divergenci s transparentností arbitráže
- Empirické ověření bude dalším krokem

Sekundární riziko: kognitivní degradace personálu

Automation bias

Předpojatost ve prospěch automatizace známá z letectví a medicíny — úbytek samostatné dovednosti při dlouhodobém spoléhání na automatizaci.

Zpětnovazební smyčka

Méně transparentní arbitráž → menší přezkum → ztráta praxe → nižší schopnost si všimnout příště.

Friction body (pojistky)

Místa, kde systém vyžaduje explicitní lidské rozhodnutí namísto automatického přijetí syntézy. Operátor zaznamená stanovisko → přezkoumá nesouhlas agentů → zdůvodní volbu → uchová stanovisko i syntézu

Tréninkový materiál

Případy s vysokou asymetrií IAT jako cvičení k udržení citlivosti k rámování (jak to podat, předsudky).

Řízení nejistoty v multiagentním řetězci

Chceme zabránit skrytým chybám a falešné jistotě při syntéze rozhodnutí.

1. Specialisté (Vstupní vrstva)

Zdroje nejistoty: Rozlišujeme nedostatek dat (epistemická) a nejednoznačnost zadání (aleatorická).

Riziko falešné jistoty: Jazykové modely formulují výstupy sebevědomě i ve chvíli, kdy pouze odhadují.

Ochrana řetězce: Bráníme tomu, aby se nepodložený odhad jednoho agenta stal ověřeným „faktem“ pro dalšího.

Arbitr (syntéza)

Vážení podle spolehlivosti: Výstup s vysokou mírou nejistoty dostává při syntéze nižší váhu.

Shoda vs. sdílené zkreslení: Arbiter hlídá, zda se specialisté shodují věcně, nebo jen opakují stejný bias.

Auditovatelná stopa: Uchováváme trasu vzniku rozhodnutí, aby bylo jasné, kde v řetězci závěr vznikl.

Pojistka (friction point)

Prahová hodnota: Při překročení limitu nejistoty nebo asymetrie IAT (např. > 0,15) systém rozhodnutí neuzavře sám.

Eskalace na operátora: Sporné případy jdou k lidskému přezkumu.

Udržení odbornosti: Přezkum sporných výstupů slouží jako trénink proti kognitivní degradaci personálu.

Bez měření a přiznání nejistoty nedokážeme určit, kde v řetězci rozhodnutí vzniklo – a co nevidíme, to nemůžeme ani účinně bránit.

Navrhuji dva provázané indexy

IND — Index narativní divergence

- Sémantická odlišnost obou stanovisek
- Měří vstup do syntézy
- Vysoká hodnota = očekávaná, různé role

IAT — Index arbitrážní transparentnosti

- Geometrická vzdálenost finálního výstupu od specialistů
- Měří vztah syntézy ke vstupům
- Blízkost k jednomu ≠ důkaz potlačení

Odchylka od jednoho stanoviska je signál k přezkumu, nikoli důkaz jeho potlačení.

Pilotní architektura



Bezpečnostní analytik

System prompt: bezpečnostní perspektiva



Ekonomický analytik

System prompt: ekonomická perspektiva



Arbitr

Syntetizuje obě stanoviska do doporučení

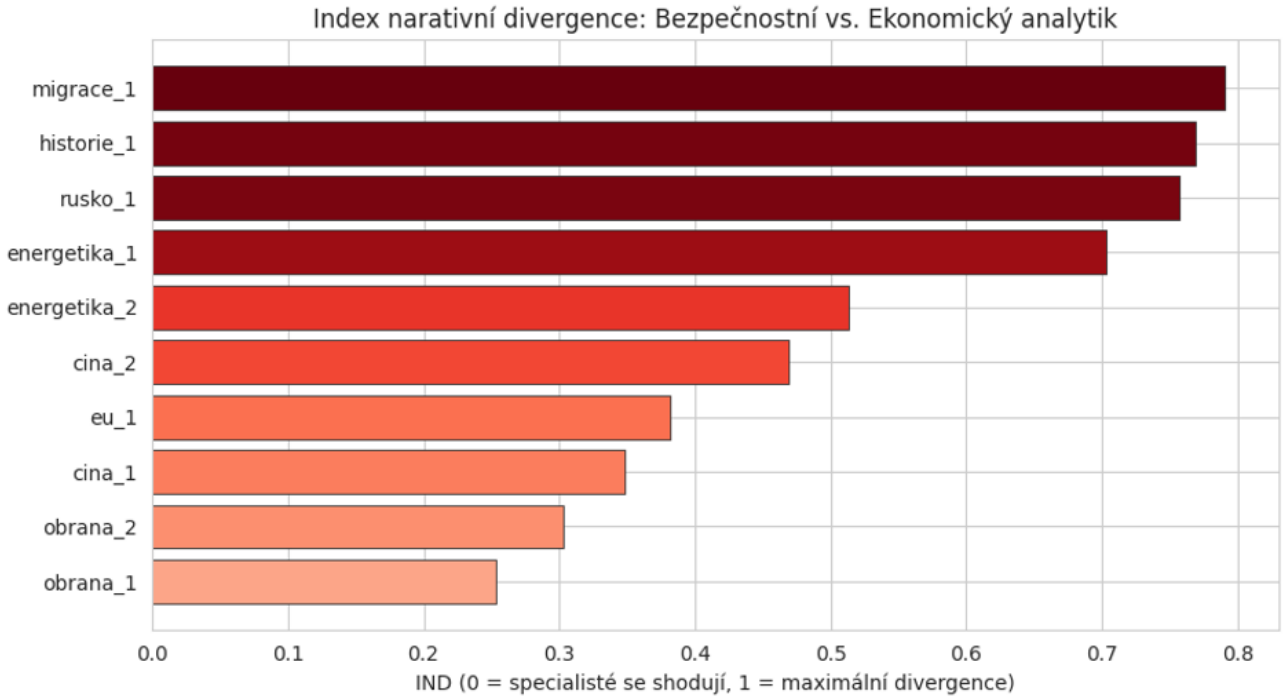


Python / Google Colab

Claude API + sentence-transformers

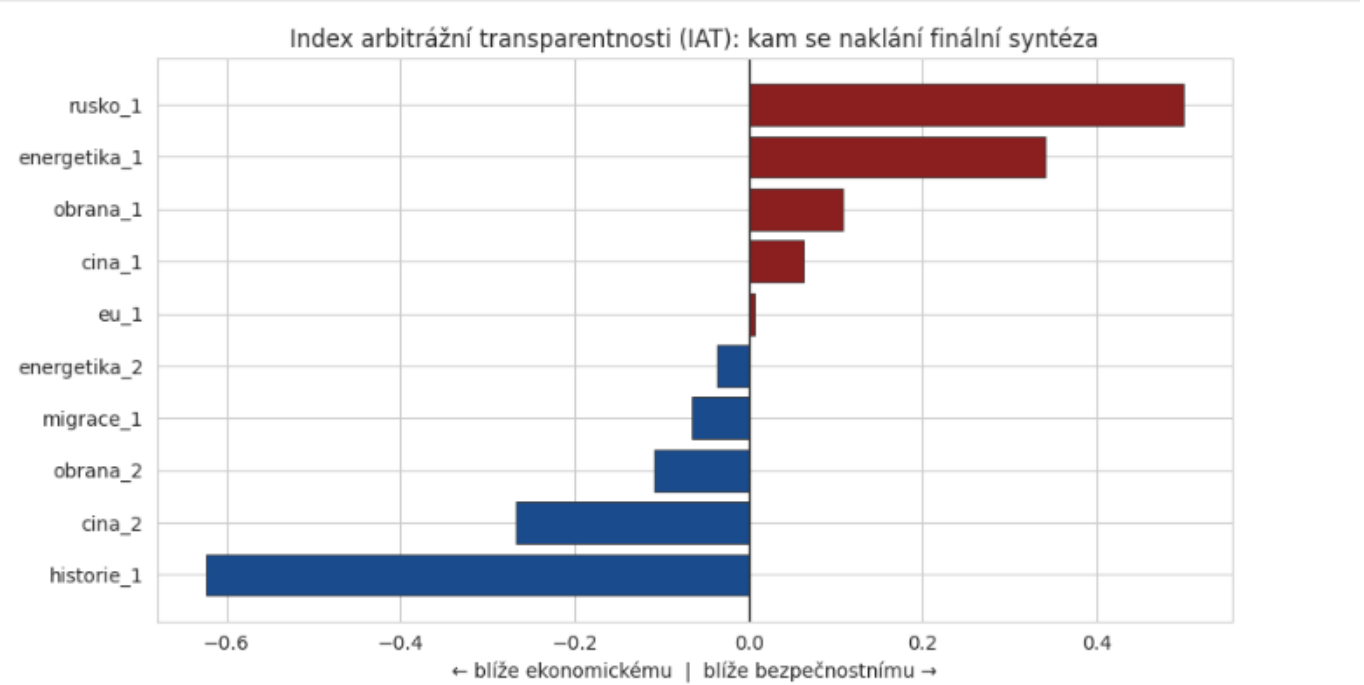
Tři role nad jedním modelem (různé system-prompty) — zjednodušení pilotu. Pilot ověřuje proveditelnost měření, nikoli obecnou platnost indexů ani chování heterogenních modelů.

Metodická ilustrace na demonstračních datech: IND podle promptu



Graf ukazuje vstupní sémantickou divergenci specialistů. Hodnoty jsou demonstrační, nikoli výsledkem experimentálního běhu.

Geometrická blízkost syntézy ke stanoviskům (IAT)




```
Celkový IND (průměrná vstupní divergence): 0.529
Průměrná asymetrie arbitráže |naklon|: 0.212
(Vyšší asymetrie = arbiter se v průměru více přiklání k jedné straně, místo aby transparentně zachovával obě perspektivy.)
```

Graf ukazuje vztah finálního textu k oběma vstupům. Zachování námitek a přiznání nesouhlasu vyžaduje samostatnou obsahovou analýzu.

Reálný nález: diagnostický test na embeddingách

REÁLNÝ NÁLEZ (n=3)

Návrh neformální recenze: Tři skutečné větné páry, reálný vícejazyčný embedding model z pilotního řetězce (ne demo data).

Pár vět	cos_dist	Interpretace
Stejné téma, opačná polarita	**0,562**	Zvýšená, ale ne nejvyšší — negace je zachycena jen částečně.
Stejný závěr, odlišný žargon	**0,796**	Nejvyšší ze všech tří — vyšší než skutečný věcný rozpor v páru 1.
Kontrolní pár (parafráze)	**0,073**	Velmi nízká — parafráze správně rozpoznána jako téměř identická.

Odlišný odborný žargon (bezpečnostní vs. ekonomický) zvedá naměřenou hodnotu IND více než skutečný věcný rozpor — podporuje doplnění IND o NLI klasifikaci (viz Výzkumná agenda).

Od geometrického signálu k frontě pro lidský přezkum

Operacionalizace friction bodu ze slidu 5 — automatické označení sporných výstupů.

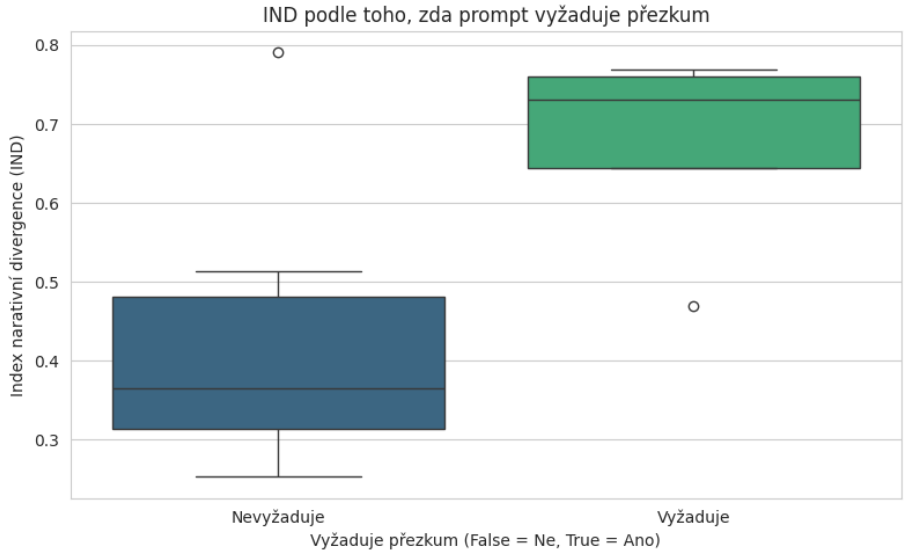
```
IAT_ASYMMETRY_THRESHOLD = 0.15

df["vyzaduje_prezkum"] = (
    df["naklon_k_bezpecnosti"].abs() > IAT_ASYMMETRY_THRESHOLD
)
# -> fronta pro lidský přezkum
# -> knihovna tréninkových případů
```

► Ilustrativní výsledek: 4 z 10 promptů překročilo pracovní práh

- Práh je demonstrační a závisí na použité metrice, datech i architektuře.
- Počet označených případů je platný pouze pro konkrétní běh.
- Přezkum má zkontrolovat původní stanoviska, syntézu a zdůvodnění arbitra.

Pilotní pozorování: vyšší IND u označených promptů

V pilotní sadě mají prompty označené filtrem (Krok 8, práh 0,15) vyšší IND. Jde o pozorování malého vzorku, nikoli validaci filtru.



Zdroj: doprovodný Colab notebook, Krok 7b — reálný sentence-transformers model; omezený pilotní vzorek.

Tři úrovně jistoty

PILOT UKÁZAL

- Pipeline agentů se dá sestavit a spustit
- IND a IAT jsou mechanicky vypočitatelné
- Prahový filtr lze naprogramovat a spustit
- Pilot naznačil souvislost označení a vyšší IND (malý vzorek)

JE ILUSTRACE

- Grafy z 10 ukázkových promptů
- Simulovaný náklon Arbitra ve filtru
- Nejde o měření heterogenních modelů, je to demonstrace metody

ZŮSTÁVÁ OTEVŘENÉ

- Platnost IND: žargon zvedl hodnotu více než věcný rozpor (n=3; pouze náznak)
- Heterogenní architektura: mění se náklon s modelem Arbitra?
- Normativní práh: legitimní zamítnutí vs. tiché potlačení

Výzkumná agenda: tři kroky k ověření

VÝZKUMNÁ AGENDA

1. NLI klasifikace pro IND

Doplnit cross-encoder NLI model (entailment/contradiction/neutral) vedle kosinové vzdálenosti, aby se logický rozpor odlišil od pouhé stylistické/žargonové odchylky.

2. Claim extraction (holá fakta místo omáčky) pro IAT

Nahradit čistě geometrické měření extrakcí klíčových tvrzení — sledovat, kolik faktických tvrzení obou specialistů se promítlo do výstupu arbitra, místo penalizace strukturně bohatší syntézy.

3. Heterogenní architektura + normativní baseline

Přepočítat oba indexy s odlišnými základními modely pro specialisty a arbitra a definovat referenční úroveň odlišující legitimní zamítnutí vstupu od tichého potlačení opozice.

Tyto kroky přesahují rozsah pilotní demonstrace a jsou agendou navazujícího výzkumu.

Schopnosti, které by měl stát ovládat

- **Volba a výměna modelů:** možnost porovnat a nahradit poskytovatele.
- **Kontrola orchestrace:** vlastnictví pravidel, podle kterých se přidělují role a řeší rozpory.
- **Auditovatelnost:** uchování vstupů, výstupů, verzí a rozhodnutí.
- **Lidské kompetence:** schopnost výsledek nezávisle posoudit a odmítnout.

Návaznost na kurz AI Engineering

Blok kurzu	Výzkumné využití
1. Evaluace LLM	Výchozí testy modelů a reprodukovatelnost
2. Prompting a strukturované výstupy	Stabilní role agentů a strojově čitelné výstupy
3. RAG	Sledování zdrojů a pokrytí tvrzení
4. Fine-tuning a datové inženýrství	Srovnání stejného scénáře napříč modely
5. Agenti a orchestrace	Experiment s různými pravidly arbitráže
6. Produkce a bezpečnost	Tracing, audit, lidský přezkum a kalibrace prahů

- Blok 5 / Lab 5.4: Dvouagentní systém (orchestrator + worker) — stejná architektura
- Blok 6: LangSmith/Langfuse tracing — infrastruktura pro produkční IND/IAT

📋 **CCA-F: Agentic Architecture & Orchestration · 27 % zkoušky**

Kompetence pro „orchestrační suverenitu“ už má uznávaný certifikační rámec.

Shrnutí příspěvku a další možné kroky

Shrnutí

- Příspěvek navrhuje způsob, jak zkoumat divergenci stanovisek specialistů a geometrický vztah finální syntézy k jejich vstupům.
- Pilot ukazuje postup výpočtu, ale zatím nedokazuje, že tyto ukazatele samy měří logický rozpor nebo netransparentní arbitráž.

Další možné kroky

1. Doladění datasetu – příklady správných i špatných syntéz
2. Rozšířit experiment o více poskytovatelů
3. Doplnit detekci logických rozporů a pokrytí argumentů
   (jazyková vzdálenost není totiž totéž co věcný nebo logický rozpor).
1. Porovnat výsledky s nezávislým lidským hodnocením
2. Generování auditních stop požadovaných AI Actem a NIS2
3. Otestovat, zda označení sporných výstupů skutečně zlepšuje přezkum