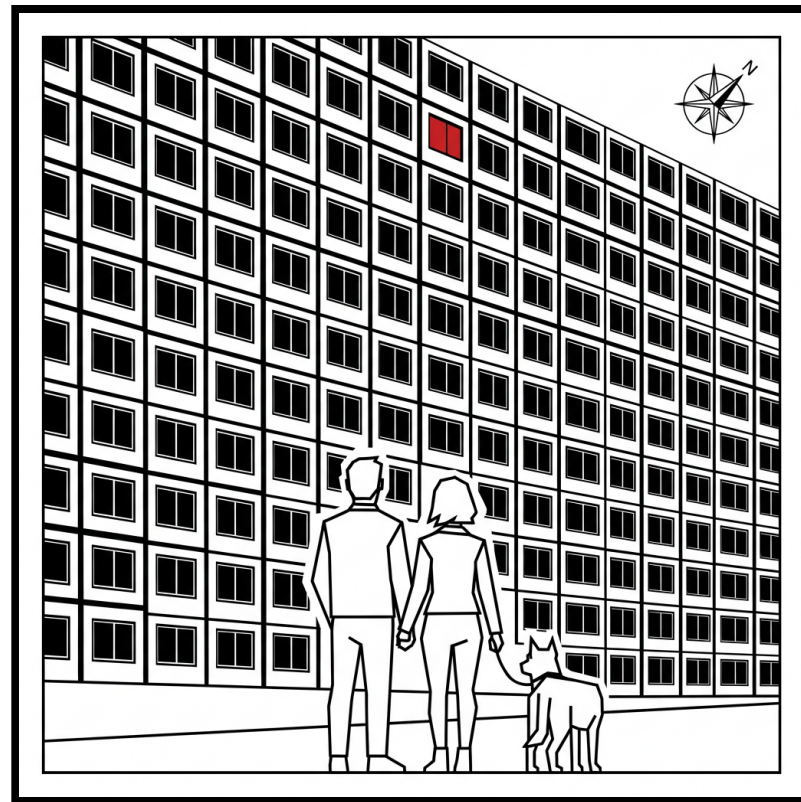


Evaluace v praxi

Od naivního narativu ke kritériím
a verifikaci

Cerna.ai · září 2026 · Šimon Podhajský a Veronika Král'



Kdo vás dnes provede



Šimon Podhajský

AI @ Curebase — evaluační pipeline a regresní testy pro produkční AI systémy

Kognitivní věda @ Yale

Flat Scout vznikl při jeho letošním hledání bytu v Londýně



Veronika Král'

AI @ Ctrl AI — AI inženýring

Dříve Data Lead @ Gamee a hlavní lektorka Czechitas

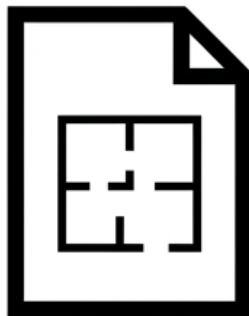
Specializace: evaluace multimodálních AI systémů

Spolu organizujeme **evals.cz** — pražský meetup pro lidi, kteří staví a zkoumají AI produkty.

Co Flat Scout dělá



ALERT



INZERÁT



SOUDCE

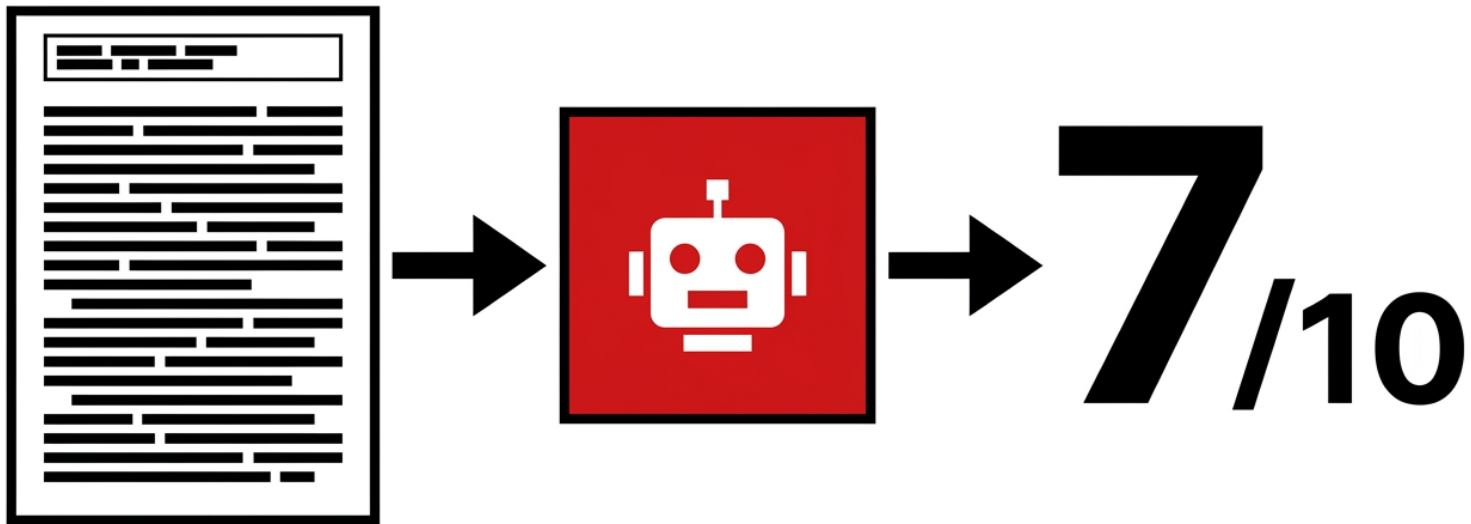


KARTA

Varianta 1: Jeden narativ

Napiš vše, o co máš zájem. Nech model rozhodnout.

Jeden brief, jedno volání, jedno číslo



`criteria.md` · jedna normostrana prózy

„Ohodnot' tenhle inzerát 0–10.“

Jedno číslo na kartě

Proč dostal sedmičku?

Číslo nemá části.

Nevíte, které kritérium ho stáhlo nahoru a které dolů.
Ani model to neví.

Každá úprava přehodnotí všechno.

Jedna věta v briefu =
opětovné volání modelu na
každý inzerát. Takže brief
nikdo neupravuje.

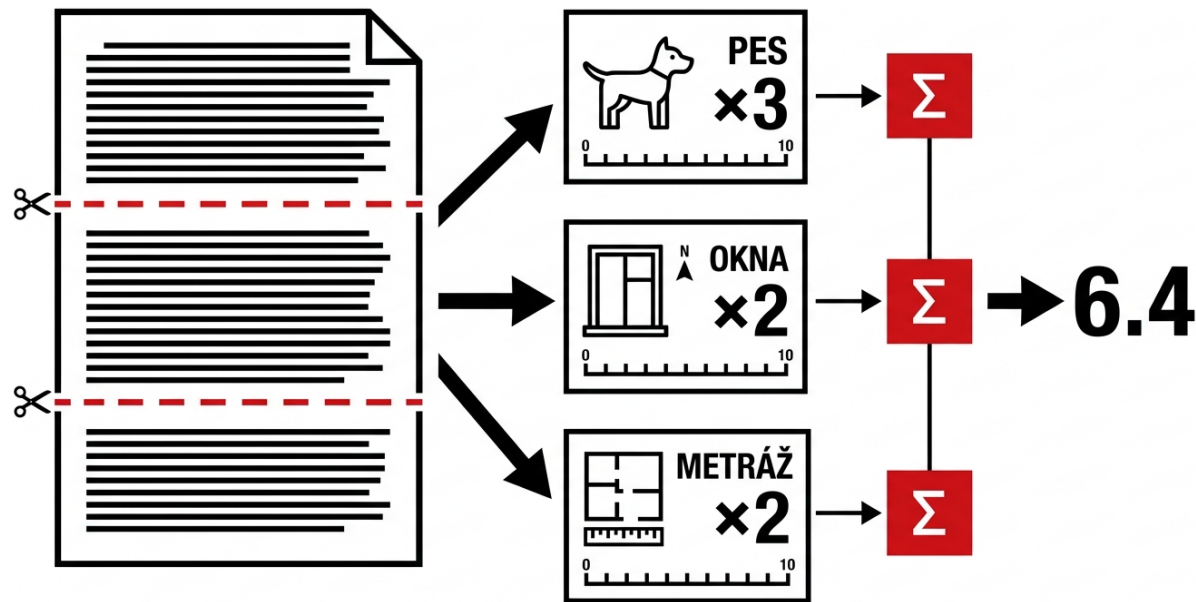
Strohý a neúplný inzerát vypadá dobře.

Když inzerát nic neřekne o
psovi, nic ho nepenalizuje.

Varianta 2: Kritéria s rubrikami

Rozděl brief na části. Každý kus dostane váhu a známkovací stupnici.

Každé kritérium ohodnocené zvlášť



Jeden soubor na kritérium

Známka 0–10 *nebo* neznámo + opodstatnění

Vážený průměr

Anatomie kritéria

```
---
name: Pronajímatel, který dovoluje mazlíčky
weight: 3
veto_at_or_below: 0
unknown: skip
ask: "Jaká jsou pravidla budovy pro domácí mazlíčky?"
---

## Rubrika
10 – pet-friendly, bez příplatku
 8 – mazlíčci "po dohodě"
 5 – s podmínkami nebo kaucí navíc
 0 – výslovný zákaz

## Důkazy
Mlčení = neznámo, ne zákaz. Vrať null;
otázka půjde makléři.
```

Váha · jak moc to rozhoduje

Veto · nula vyloučí inzerát

Kotvy · co přesně znamená 10, 5, 0

Neznámo ≠ nula · hodnotit absenci jinak

Otázka · na co se má systém zeptat, když to v textu inzerátu není

Ted' vy: rozsekejte narativ

1 Co přesně měříme? Jedna věta.

2 Jak vypadá 10, 5 a 0?

3 Co když to v inzerátu není?

4 Jakou to má váhu, a proč?

Fiktivní mladý pár, pes, tři kritéria

01

Pes

Nájem musí dovolit psa. Pes ještě neexistuje, ale bude.

02

Okna ne na západ

Kam míří okna obýváku? V textu inzerátu se to skoro nikdy nenajde.

03

Metráž

Obývák musí být místnost, ne chodba s gaučem.

Proč to dělat?

Známka se dá obhájit

Každé kritérium nese známku a **jednu větu důkazu** — co v inzerátu ji určilo.

„Proč 6,4?“ má odpověď: rozpad na kritéria, váhy a důkazy. **Proces jde auditovat.**

Kritéria se dají analyzovat

Průměrná známka u bytů, které jste **chtěli**, proti těm, které jste **odmítli**:

Kritérium	Chtěli	Odmítli	Rozdíl
Pes	7,8	5,2	+2,6
Metráž	6,9	5,5	+1,4
Okna	5,9	5,9	0,0

ilustrační čísla

Na kritériu s **nulovým rozdílem** nezáleží. Je jedno, jakou má váhu.

Dá se tomu věřit?

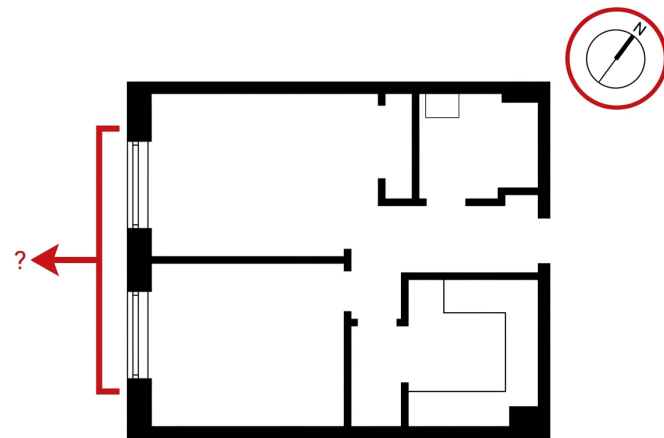
Exaktní číslo neimplikuje exaktní měření.

Kam míří okna?

Inzerát to neřekne. Půdorys ano, ale ve třech krocích:

1. **Najít kompas** a změřit jeho úhel
2. **Najít stěnu** s okny obýváku
3. **Otočit** jedno podle druhého: aritmetika

Model umí 1 a 2. Krok 3 mu **nesvěřujte**. Model je dobrý ve všímání, ale (zatím) ne v rotování.



24

půdorysů,
ručně
anotovaných

7

bez
jakéhokoli
kompasu

6 / 9

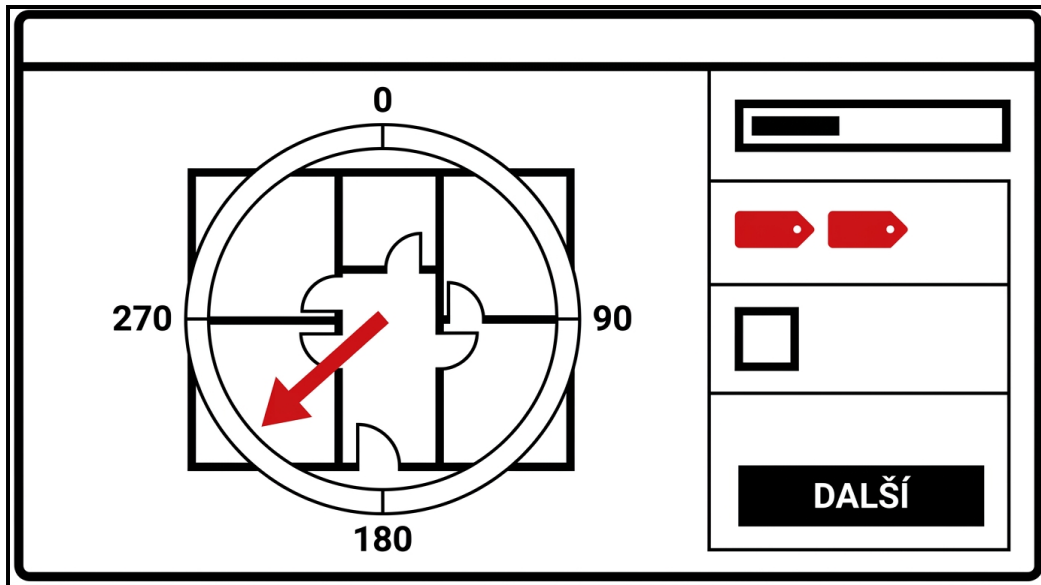
odpovědí
správně; u
zbytku
pipeline mlčí

3×

„sever je
nahore“ na
plánu, kde
míří doleva

Shoda modelu sama se sebou **není důkaz.**

Ground truth: člověk, obrázek, žádná aritmetika



Anotace v souřadnicích obrázku: 0 = nahoru, 90 = doprava.

Formulář je **prázdný**. Nepředvyplňovat tím, co říká pipeline.

Kompas a okna se **skórují zvlášť**. Selhávají zvlášť a opravují se jinak.

Stavíme anotační nástroj

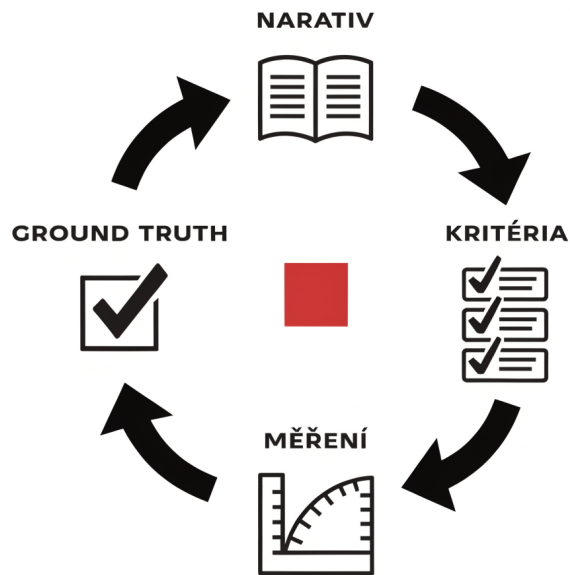
Co nástroj musí umět

- Načíst půdorysy z datasetu
- Ukázat každý zvlášť, přes něj kompasový kruh
- Zapsat `north_clock`, `windows_clock` (seznam), `compass_style`, `done`
- Uložit někam, odkud může číst benchmark

Co neřešíme

- Krásu anotačního nástroje
- Přihlášení a multi-user
- Cokoli, co za hodinu nestihneme

→ terminál



Smyčka

Narativ řekne, **co** chcete.

Kritéria řeknou, **jak** to poznat.

Měření řekne, **která** kritéria (ne)fungují.

Ground truth pomáhá **měřit správně**.

Co si odnést

01

Evaly se vyplatí, až když je věcí hodně

Jeden byt ohodnotíte sami. U stovky potřebujete kritéria, důkazy a report, který řekne, které kritérium nic nerozlišuje.

02

~~Důvěřuj, ale~~
prověřuj

Přesné číslo není přesné měření. Shoda modelu sama se sebou není důkaz.

03

Ground truth si můžete vyrobit sami

Dvacet čtyři půdorysů, hodina vibe-codingu a prázdný formulář. Člověk, obrázek, žádná aritmetika.

Díky



<https://slides.podhajsky.net/ai-2026-evals/>

Slidy



<https://evals.cz>

evals.cz

Šimon Podhajský · simon.podhajsky.net

Veronika Král' · evals.cz