XXVIII

Jak poznáš, že to AI napsala dobře?

Po této kapitole se přestaneš dívat na číslo „pass rate“ a začneš se ptát, z kolika běhů vzniklo. Naučíš se stavět evals jako hloupého soupeře pro svůj vlastní skill, používat model jako soudce, aniž ti zesměšní sám sebe, a poznáš, proč tři úspěšné běhy nedokazují skoro nic — a kolik jich vlastně potřebuješ, než smíš říct „funguje to“.

> *Ověřit odpověď je snazší než ji vytvořit — a přesně na tom stojí každá dobrá vyhodnocovací sada.*
>
> — Anthropic, *parafráze principu z* Demystifying evals for AI agents, *engineering blog, 2026 — verifikace je levnější než generování*

Otevřu tuhle kapitolu scénou, která se mi stala minulý týden, protože je poctivější než jakékoli dějiny. Napsal jsem skill — malý balík instrukcí, který naučí model dělat prezentace o vědeckých studiích tak, aby nepletl korelaci s kauzalitou (ten skill z předchozí kapitoly). A protože jsem vědec, nechtěl jsem věřit dojmu; pustil jsem na něj vyhodnocovací sadu. Nástroj rozjel tři zadání dvakrát vedle sebe: jednou se skillem, jednou bez něj — bez skillu je *hloupý soupeř*, ten, kterého musím porazit, než smím mluvit o úspěchu. Za pár minut vypadla tabulka a v ní dvě čísla, na která jsem hleděl s hřejivým pocitem hotové práce:

94 %	39 %
se skillem	bez skillu (hloupý soupeř)

Devětatřicet procent na čtyřiadevadesát. Víc než dvojnásobek. Kdybych ten obrázek dal na slajd, potlesk zaručen — „skill zvedl kvalitu z devětatřiceti na čtyřiadevadesát procent“. A pak jsem se zeptal na otázku, kterou tahle kniha klade dvaadvacet kapitol: *jak bych poznal, že se pletu?* Otevřel jsem, z čeho ta čísla vznikla. Byly to *tři* zadání. Tři. Čtyřiadevadesát procent nebyl výsledek tisíce měření — byla to hrstka asertů na třech příkladech, které jsem si sám vybral. A v tu chvíli mi to hřejivé číslo zhořklo v ústech: nevím, jestli jsem změřil skill, nebo náhodu.

skill — balík instrukcí (soubor `SKILL.md`*), který modelu předává tvé řemeslo; kap. 25.*
eval / vyhodnocovací sada — sada zadání s očekávaným výsledkem, na níž měříš, jestli změna pomohla. Nesklonné; doma tady, v části druhé.
pass rate — podíl zadání, která prošla všemi svými testy. Číslo, které svádí: vypadá přesně, ale bez počtu pokusů za sebou nic neváží.

Tahle kapitola je o tom hořknutí. Otevírá poslední dějství knihy — *vědec v praxi* — a ptá se na věc, která je v roce 2026 nejdůležitější ze všech: když kód, text nebo prezentaci napsal stroj, *jak poznáš, že to napsal dobře?* V části první jsi stavěl stěžeň metody. Teď ho napneš proti výstupu, který nevytvořila tvoje ruka — a zjistíš, že úsudek, který ti zbyl, se schovává přesně do téhle otázky.

Recenzent nemusí umět napsat román

Začnu úlevou, protože ta je jádro celé věci. Možná tě děsí, že máš soudit výstup modelu, který „umí víc než ty“ — napíše kód v jazyce, který neznáš, prezentaci o oboru, kterému nerozumíš. Jak můžeš soudit něco, co bys sám nesvedl vyrobit?

Odpověď je stará jako kritika: *recenzent nemusí umět napsat román, aby poznal špatný*. Editor, který by sám nenapsal ani povídku, spolehlivě pozná, že děj nedrží, že postava ve třetí kapitole zapomněla, co řekla v první, že konec nesedí na začátek. Neumí tvořit na téže úrovni — a přesto *ověřuje* přesně a rychle. To není náhoda ani nespravedlnost; je to hluboká asymetrie světa, ke které se ještě vrátíme ve věži. Prozatím ji vezmi jako dobrou zprávu: *ověřit je levnější než vytvořit*. Nemusíš umět, co model umí. Musíš umět poznat, kdy se netrefil — a to je jiná, snazší dovednost.

Ozvěna kap. 1: hodnota se stěhuje od tvorby k úsudku. Kentaur nevyhrává tím, že hraje líp než stroj, ale tím, že pozná*, kdy stroj hraje špatně. Tvoje nová práce je recenze, ne autorství.*

A tady je past, do které padne skoro každý: protože ověřit je snazší, svádí to k tomu ověřit *málo*. Uděláš to jednou, vyjde to, řekneš „funguje to“ — a jsi zpátky u Theraku z kapitoly devatenáct, jen o patro výš. „Viděl jsem to fungovat“ a „vím, že to funguje“ jsou pořád dvě různé věty; u výstupu AI je ta propast ještě širší, protože model umí být plynule a přesvědčivě špatně. Ověřování je levné — proto ho dělej *hodně*, ne málo.

Řemeslo: rubrika, soudce a hloupý soupeř

TEENAGER · MECHANISMUS

Eval je jen hloupý soupeř oblečený do jednoho souboru. Princip z kapitoly dvanáct beze změny: než uvěříš, že tvůj skill (nebo prompt, nebo agent) něco umí, musíš ho porovnat s nejjednodušší možnou alternativou — s během *bez* něj. Vyhodnocovací sada je sbírka zadání, u nichž předem víš, co chceš vidět:

```
{
  "skill_name": "veda-vs-media",
  "evals": [
    {
      "prompt": "Udělej slajdy o studii, že ranní
káva prodlužuje život",
      "assertions": [
        { "name": "rozlišuje korelaci od
kauzality", "type": "llm_judge" },
        { "name": "soubor .pptx existuje",
"type": "file_exists" },
        { "name": "více než 5 slajdů",
"type": "slide_count_gte", "value": 5 }
      ]
    }
  ]
}
```

Každé zadání se pustí *dvakrát paralelně* — se skillem i bez — a asery se sečtou. Rozdíl mezi oběma sloupci je jediné, co tě zajímá: hloupý soupeř odděluje „skill pomohl“ od „model to uměl beztak“.

Aserce jsou dvojího druhu a je dobré vědět, který právě používáš. Ty *tvrdé* změří stroj beze sporu: soubor existuje, slajdů je víc než pět, funkce vrátila tři sta. To je jednotkový test z kapitoly devatenáct, jen aplikovaný na výstup modelu. Ale spousta věcí, na kterých ti záleží, se tvrdě změřit nedá: „rozlišuje prezentace korelaci od kauzality?“, „je ten odstavec srozumitelný?“. Na tyhle *měkké* otázky nasadíš druhý stroj — model jako soudce.

aserce — jedno konkrétní tvrzení o výstupu, které buď platí, nebo ne: „má víc než pět slajdů“, „obsahuje slovo kauzalita*“. Sčítáním asercí vzniká pass rate.*

→ kap. 12: bez hloupého soupeře nemáš nic. „Skill dává 94 %“ neříká nic, dokud nevíš, kolik dá holý model na týchž zadáních.

TEENAGER · MECHANISMUS

Model jako soudce (LLM-as-judge). Když se výsledek nedá změřit funkcí, necháš ho posoudit jiný model. Nedáš mu ale otázku „je to dobré?“ — dostaneš chválu. Dáš mu *rubriku*: přesná, rozepsaná kritéria, jako má učitel na opravování písemek.

```
Ohodnoť prezentaci podle těchto kritérií, každé
zvlášť (splněno / ne):
1. Uvádí, že jde o pozorovací studii, ne o
experiment?
2. Zmiňuje aspoň jeden alternativní důvod
korelace?
3. Vyhýbá se slovům „prokazuje“, „způsobuje“ u
pozorovací studie?
Vrať pro každé kritérium ano/ne a jednu větu
zdůvodnění.
```

Rubrika dělá dvě věci. Rozbije mlhavé „dobré/špatné“ na kontrolovatelné kusy — a donutí soudce *zdůvodnit*, takže si jeho verdikt můžeš přečíst a nesouhlasit s ním. Soudce bez rubriky je věštec; soudce s rubrikou je zkoušející.

rubrika — rozepsaná kritéria hodnocení, každé zvlášť ano/ne. Táž věc, jakou dělá dobrý test v kap. 19: převádí mlhavý dojem na měřitelné tvrzení. Bez rubriky měříš náladu soudce, ne kvalitu výstupu.

Model jako soudce je mocný a zrádný zároveň, a jednu jeho zradu musíš znát, protože je doložená a snadno tě obelže. Jmenuje se *polohové vychýlení*.

TEENAGER · MECHANISMUS

Polohové vychýlení a jak ho zabít randomizací. Když soudci předložíš dvě odpovědi, A a B, a zeptáš se, která je lepší, on nesoudí jen jejich obsah — soudí i *pořadí*. Modely mají doložený sklon dávat přednost odpovědi, která přišla první, bez ohledu na kvalitu (Zheng a spol., 2023, na sadě MT-Bench). Kdyby tvůj skill náhodou byl vždy „A“, měřil bys zčásti jeho výhodu z pořadí, ne jeho kvalitu.

Lék je jednoduchý a povinný: *prohoď pořadí a zeptej se znovu*. Vítěze uznej, jen když vyhraje v obou pořadích; když se hlas otočí podle toho, kdo je první, je to remíza, ne výhra. Ve velkém prostě pořadí u každého souboje náhodně losuj. Randomizace je levná a bez ní tvůj soudce tajně stranicky fandí tomu, koho jsi napsal nalevo.

Prameny: L. Zheng a spol., Judging LLM-as--a-Judge with MT-Bench and Chatbot Arena*, 2023 — doložené polohové, upovídané a sebe-nadržující vychýlení soudců; obrana „prohoď a uznej jen shodu v obou pořadích“.*

Soudce navíc nadržuje vlastnímu *stylu (self--preference). Proto ideálně souď jiným modelem, než který výstup napsal — a nikdy nenech model opravovat vlastní domácí úkol. To je papoušek zkoušející papouška z kap. 19.*

A ještě jedna vrstva, na kterou v praxi narazíš hned: skill nebo prompt jednou odladíš — a za tři měsíce vyjde nová verze modelu a tvůj skill tiše přestane fungovat, protože se změnilo, jak model čte instrukce. Proto se eval nepouští jednou. Stane se *regresním testem* z kapitoly devatenáct: pomníkem, na který se chodíš přesvědčit, že to, co kdysi fungovalo, funguje pořád. Uložíš zadání, na kterém skill selhal, opravíš, a to zadání

zůstane v sadě navždy — hlídá, aby se tentýž průšvih nevrátil zadními vrátky při příští verzi.

EINSTEIN · MATEMATIKA — přeskočitelné

Testy jako specifikace — obrácení rolí z kap. 19, teď proti modelu. Nejsilnější použití evalu není soudit hotový výstup. Je to napsat sadu *dřív*, než pustíš model tvořit, a nechat ho iterovat, dokud nesvítí zeleně. Sada přestává být zkouškou a stává se *zadáním*: přesným, měřitelným popisem toho, co „správně" znamená. Model dělá snadnou půlku (tvoří), ty tu nezcizitelnou (definuješ, co má platit). Úsudek zůstává tobě — jen zapsaný předem, jako *pakt* z kapitoly osmé.

Věž: proč tři běhy nestačí — a kolik jich stačí

Teď to jádro, kvůli kterému mi číslo zhořklo. Vrátím se k těm čtyřiadevadesáti procentům a udělám s nimi to, co jsem měl udělat hned: zeptám se, jak přesné to číslo vlastně je.

EINSTEIN · MATEMATIKA — přeskočitelné

Pass rate je odhad, ne fakt — a malé N ho nechá plandat. Když pustíš skill na n zadání a k jich projde, k/n není pravda o skillu — je to *odhad* jeho skryté úspěšnosti p, zatížený náhodou výběru. Kolik náhody v něm je, řekne interval spolehlivosti. Pro binomický pokus dává poctivou odpověď Clopper–Pearsonův (přesný) interval; jeho šířka klesá zhruba jako $1/\sqrt{n}$ — pomalu. Dosaď má čísla:

$$\hat{p} = 3/3 = 100\% \quad \rightarrow \quad 95\%\ \text{interval} = [29\%, 100\%].$$

Tři úspěchy ze tří vypadají jako dokonalost. Poctivě jsou slučitelné se skutečnou úspěšností *devětadvaceti procent* — tedy i s tím, že skill je *horší* než hloupý soupeř na čtyřiceti procentech. Bodový odhad křičí sto; interval šeptá „nevím".

Hrdinský graf. Tečka = bodový odhad (pass rate), úsečka = poctivý 95% interval. Nahoře „se skillem, tři běhy“: tečka na sto procentech, ale interval padá až k devětadvaceti — pod bodový odhad hloupého soupeře nad ním. Čím víc běhů, tím kratší úsečka: sto běhů (94 ze 100) konečně sevře interval na 87–98 %, bezpečně nad hodem mincí. Číslo se nezměnilo skoro vůbec — jistota ano. *Svislá čára je 50 %: hod mincí, spodní hranice smyslu.*

EINSTEIN · MATEMATIKA — přeskočitelné

Kolik běhů tedy? Pravidlo, které si zapamatuj, se jmenuje *pravidlo tří*: vidíš-li k úspěchů z n bez jediného selhání, dolní 95% mez skutečné úspěšnosti je zhruba

$$p_{\text{dolní}} \approx 1 - 3/n.$$

Tři běhy tě pustí sotva nad třicet procent. Aby ses dolní mezí dostal nad devadesát procent, potřebuješ řádově *třicet* bezchybných běhů; abys rozlišil 94 % od 39 % s jistotou, potřebuješ desítky pokusů, ne tři. Malé N není zkratka — je to iluze měření. Statistika malých čísel je ta nejkrutější: vypadá jako důkaz a je to anekdota.

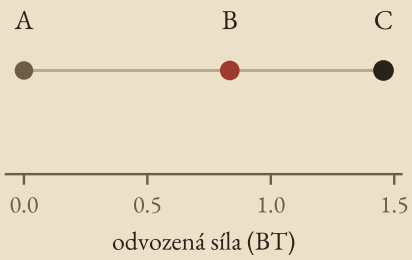
Prameny: rule of three, *upper/lower 95% mez z* $(1-p)^n = 0,05$*; přesný Clopper–Pearsonův interval z beta-rozdělení. Pozor: interval mluví o* stejném *zadání spuštěném víckrát (variace modelu), i o rozšíření sady na víc* různých *zadání — obojí zvětšuje* n*, obojí zužuje úsečku.*

EINSTEIN · MATEMATIKA — přeskočitelné

Nesuď absolutně, suď v párech — Bradley–Terryho škála. Když porovnáváš dva skilly (nebo dva modely, dva prompty) a měkkou kvalitu neumíš dát na absolutní stupnici, ptej se soudce jen na to, co umí spolehlivě: *který z téhle dvojice je lepší?* Z hromady takových párových hlasů $A > B$ pak odvodíš *spojitou škálu síly*. Model Bradley–Terry říká, že pravděpodobnost výhry A nad B je logistická funkce rozdílu jejich sil:

$$P(A > B) = \frac{1}{1 + e^{-(s_A - s_B)}}.$$

Síly s se dohledají tak, aby co nejlíp seděly na pozorované výsledky (maximální věrohodnost) — a je to přesně logistická regrese z okolí kapitoly dvanáct, jen s dvojicemi místo tříd. Tak funguje veřejná Chatbot Arena: miliony párových hlasů „tahle odpověď je lepší“ se scvrknou do jediného žebříčku Elo. Párové srovnání je poctivější tam, kde absolutní známka lže, protože využívá jen tu asymetrii, o které mluvila celá kapitola: *porovnat* dvě věci je snazší a spolehlivější než *oznámkovat* jednu.



Prameny: Bradley–Terry (1952); Chatbot Arena / LMSYS používá BT-Elo nad miliony párových hlasů (2023–2026). BT je MLE Elo modelu za předpokladu pevné, ale neznámé míry výher — logistická regrese nad výsledky soubojů.

Graf: ze čtyř set simulovaných soubojů tří soupeřů vyleze škála síly A–B–C. Nikdo neřekl absolutní známku; přesto vznikne pořadí i odstup.

EINSTEIN · MATEMATIKA — přeskočitelné

Asymetrie generování a verifikace — a proč eval nesmíš přehnat. Celá kapitola stojí na tom, že *ověřit řešení je levnější než ho najít* — na půdorysu slavné otevřené domněnky o třídách P a NP: existují úlohy, u nichž správnou odpověď ověříš rychle, i když ji hledáš dlouho. Recenzent proti romanopisci, test proti kódu, soudce proti autorovi. Ale pozor na druhou hranu téhož nože: jakmile začneš skill *ladit na svou eval sadu* — měnit ho, dokud těch pár zadání nesvítí —, přeučíš ho na test (kapitola jedenáct). Eval sada je pak tvá *trénovací data*; abys věděl, že skill umí věc, a ne jen tvůj kviz, musíš mít svatá *testovací data*: zadání, která při ladění nikdy neuvidíš. Kdo ladí na testovací sadu, podvádí sám sebe — most zpátky do kapitoly čtvrté.

→ kap. 11: přeučení. Skill vyladěný na tři zadání umí ta tři zadání, ne úlohu. → kap. 4: testovací data jsou svatá; kdo na nich ladí, změří vlastní zbožné přání. Drž oddělenou hrst zadání, kterou skill při vývoji nikdy nevidí.

Co si odnést

Poskládej to dohromady. Číslo „94 % pass rate“ mě málem svedlo — dokud jsem se nezeptal, z kolika běhů vzniklo. Tři. A tři běhy jsou anekdota převlečená za důkaz: poctivý interval je pustil až k devětadvaceti procentům, pod hloupého soupeře. Ověřit výstup AI je snazší než ho vytvořit — recenzent nemusí umět napsat román —, a proto ověřuj *hodně*, ne málo. Stav eval jako hloupého soupeře: skill proti holému modelu, jinak neměříš nic. Měkkou kvalitu suď modelem, ale s rubrikou a s prohozeným pořadím, jinak měříš náladu soudce a jeho lásku k první odpovědi. A nikdy neveř bodovému číslu bez počtu pokusů za ním: statistika malých N lže nejsebejistěji.

To všechno je jen kapitola devatenáct, přenesená z tvého kódu na výstup stroje: napřed řekni, co je správně, teprve pak se ptej. A příští kapitola se zeptá na věc, kterou eval nezměří — *kdo vlastně čte tvůj prompt a tvá data*, když ho posíláš cizímu modelu ven.

DEMO

lesk-a-bida-vibecodingu.gaussalgo.com/d/napis-test-
-chyt-bug

Napiš test, chyť bug: dostaneš funkci, kterou psal model, a v ní jednu schovanou chybu. Piš aserce — a sleduj interval spolehlivosti, jak se pod tebou zužuje s každým dalším během. Uvidíš na vlastní oči, kolik běhů musíš pustit, než tvé „prošlo“ přestane být dojmem a stane se důkazem.

„Jak bych poznal, že se pletu?“

Konkrétní test k této kapitole máš pokaždé, když ti model něco vyrobí: *napiš jeden jediný test, který by selhal, kdyby to bylo špatně* — a pak ho pusť víckrát, než mu uvěříš. Jednu aserci, jedno měřitelné tvrzení: „prezentace odliší korelaci od kauzality“, „funkce nevrátí záporné číslo“, „soubor existuje“. Když ho umíš napsat, spustit ho dost často na to, aby interval ztuhl, a on přežije, poprvé nemáš dojem, ale odhad s intervalem. Kapitola se plete jedině tehdy, ukážeš-li mi výstup AI, o němž s jistotou víš, že je dobrý, a přitom *žádný* takový test napsat nejde — pak jsi buď našel hranici metody, nebo, mnohem pravděpodobněji, spletl potlesk tří běhů s důkazem, jako jsem to málem udělal já.