

## kapitola 28

# Jak poznáš, že to AI napsala dobře?

---

# Číslo, které málem svedlo

---

Napsal jsem skill. Pustil na něj eval: 94 % se skillem, 39 % bez něj. Potlesk zaručen.

Pak jsem se zeptal, z kolika běhů to vzniklo. Tři.

*Nevím, jestli jsem změřil skill, nebo náhodu.*

# Recenzent nemusí umět napsat román

---

Editor sám nenapíše povídku — a přesto pozná, že děj nedrží.

**Ověřit je levnější než vytvořit.** Nemusíš umět, co model umí. Musíš poznat, kdy se netrefil.

*A protože je to snadné, svádí to ověřit málo. Proto ověřuj hodně.*

# Řemeslo evalu

- **Hloupý soupeř:** skill proti holému modelu — jinak neměříš nic.
- **Model jako soudce s rubrikou,** ne otázkou „je to dobré?“.
- **Prohod' pořadí** a uznej jen shodu v obou — jinak měříš lásku k první odpovědi.

# Proč tři běhy nestačí

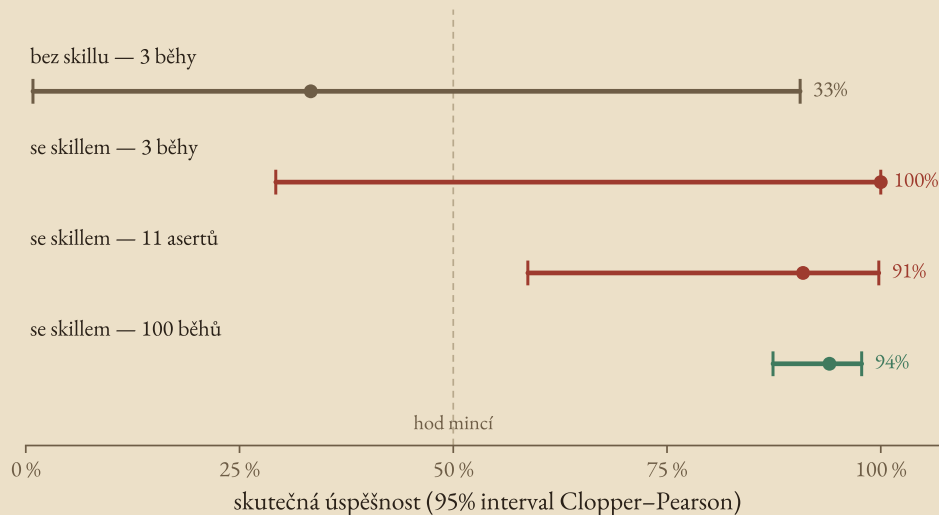
---

Tři úspěchy ze tří vypadají jako dokonalost.

Poctivý 95% interval je pustí až k **29 %** — tedy i pod hloupého soupeře na čtyřiceti.

*Bodový odhad křičí sto; interval šeptá „nevím“.*

# Číslo se nemění, jistota ano



*Tečka = pass rate, úsečka = 95% interval. Víc běhů → kratší úsečka.*

# „Jak bych poznal, že se pletu?„

U příštího výstupu AI: napiš jeden test, který by selhal, kdyby to bylo špatně — a pusť ho tolikrát, až interval ztuhne. Teprve pak nemáš dojem, ale důkaz.