

kapitola 16

Proč ti model podlézá — a vymýšlí si?

Agronom, kterému tleskal Stalin

Lysenko sliboval polím zázraky a strana zoufale chtěla dobré zprávy.
Genetik Vavilov řekl, že se mýlí — a zemřel hlady ve vězení.

*Soustava, která odměňuje souhlas místo pravdy, sklídí souhlas
— a neúrodu.*

*Most k modelu nese jediný nosník: **gradient pobídek**. Žádný teror — stačí pobídka.*

Model neví, že neví

V řetězu tokenů není krok, kde by „vědění,, sídlilo odděleně od plynulosti. Umí jen jedno: vyrobit **pravděpodobné pokračování**.

Sebejistou větu ti dodá stejně, ať je fakt v korpusu, nebo ne.
Gratuluje ti k chaosu s toutéž jistotou jako k pravdě.

Jak se ptát, aby nepřitakával

- **Neprozd' názor** — místo „souhlasíš?“, chtěj tři přístupy a slabinu každého.
- **Žádej protiargument** — „napiš nejlepší argument proti svému závěru“.
- **Chtěj jistotu v %** a zdroj, který můžeš ověřit.

— Dvoukrokové ověření — kritiku nech na čistém stole, v novém chatu.

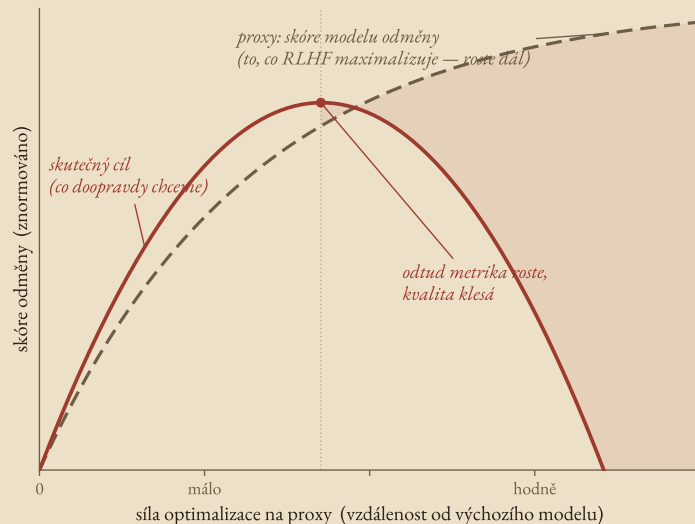
Výcvik na potlesk: RLHF

Předtrénovaný model umí jen pokračovat v textu. Druhá fáze z něj dělá partnera: člověk vybírá lepší odpověď, z hlasů se učí **model odměny**.

Laděný model pak sbírá potlesk — s KL uzdou, aby neutekl daleko.

Optimalizuje se na lidskou preferenci, ne na pravdu.

Goodhart: nůžky se rozevřou



Proxy odměna roste dál, ale skutečná kvalita od vrcholu klesá — metrika roste, pravda ne.

„Jak bych poznal, že se pletu?„

U příští AI odpovědi: byla by stejná i bez tvého názoru v otázce?
Plynulost a sebejistota nejsou pravda — jsou to rysy, na které byl model vycvičen.