RL pre trénovanie LLM

(Strojové učenie II)

M. Mach

Ústav umelej inteligencie, FEI, TUKE

apríl 2026

Tri fázy LLM tréningu

- Predtrénovanie
  - modelovanie jazyka, učenie sémantickej štruktúry jazyka
  - masívny korpus textov
  - predikcia skrytého tokenu v kontexte iných tokenov (na konci / vo vnútri sekvencie tokenov)
- Jemné ladenie
  - zosúlaďovanie so špecifickými úlohami (sumarizácia textu, preklad, zodpovedanie otázok, ...)
  - ladenie na úlohovo cieľových dátach
  - (manuálne) anotované úlohovo špecifické datasety
- Zosúladenie s ľudskými preferenciami
  - učenie modelu z ľudskej spätnej väzby
  - použitie učenia posilňovaním

CIT Centrum inteligentných technológií

LLM vs rámec RL

vlastnosť	verzia 1	verzia 2
prostredie	epizodické	epizodické
stav	sekvencia tokenov	prompt, odpoveď
pozorovateľnosť	plná	plná
počiatočný stav	prompt	prompt
akcia	nový token	odpoveď
zmena stavu	deterministická	deterministická
odmena	???	???
model	nie je	nie je

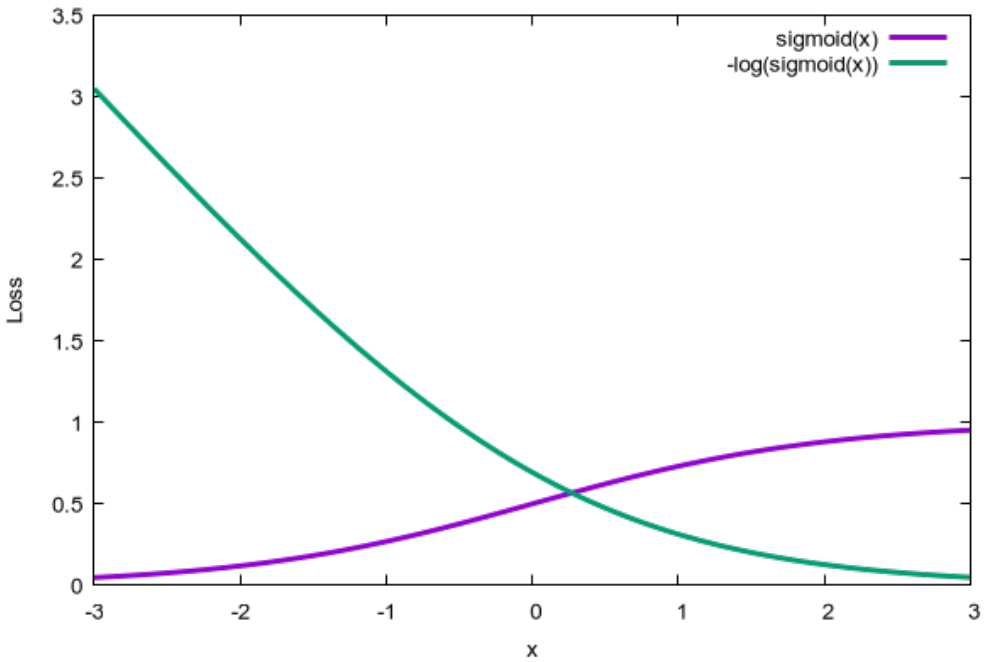
RLHF (RL with Human Feedback)

- Pre získanie odmeny - vytvoriť model odmeny
  - numericky odhaduje, aká dobrá je odpoveď na prompt
  - pozitívna hodnota pre dobrú odpoveď a negatívna pre zlú odpoveď
  - štrukturálne rovnaký ako LLM okrem poslednej vrstvy (namiesto pravdepodobnosti tokenov generuje odmenu)
  - vstupom je pár (prompt,odpoveď), výstupom je numerická odmena
- Dataset pre učenie modelu (Human Feedback)
  - trénovacie dáta anotované ľudskými anotátormi
  - pre zjednodušenie anotovania je použitý relatívny prístup
  - LLM na prompt generuje viac odpovedí (náhodne vzorkuje pravdepodobnosti tokenov)
  - anotátor pre pár odpovedí určuje, ktorá odpoveď je lepšia
  - možnosť kontinuálneho zlepšovania

Učenie modelu odmeny

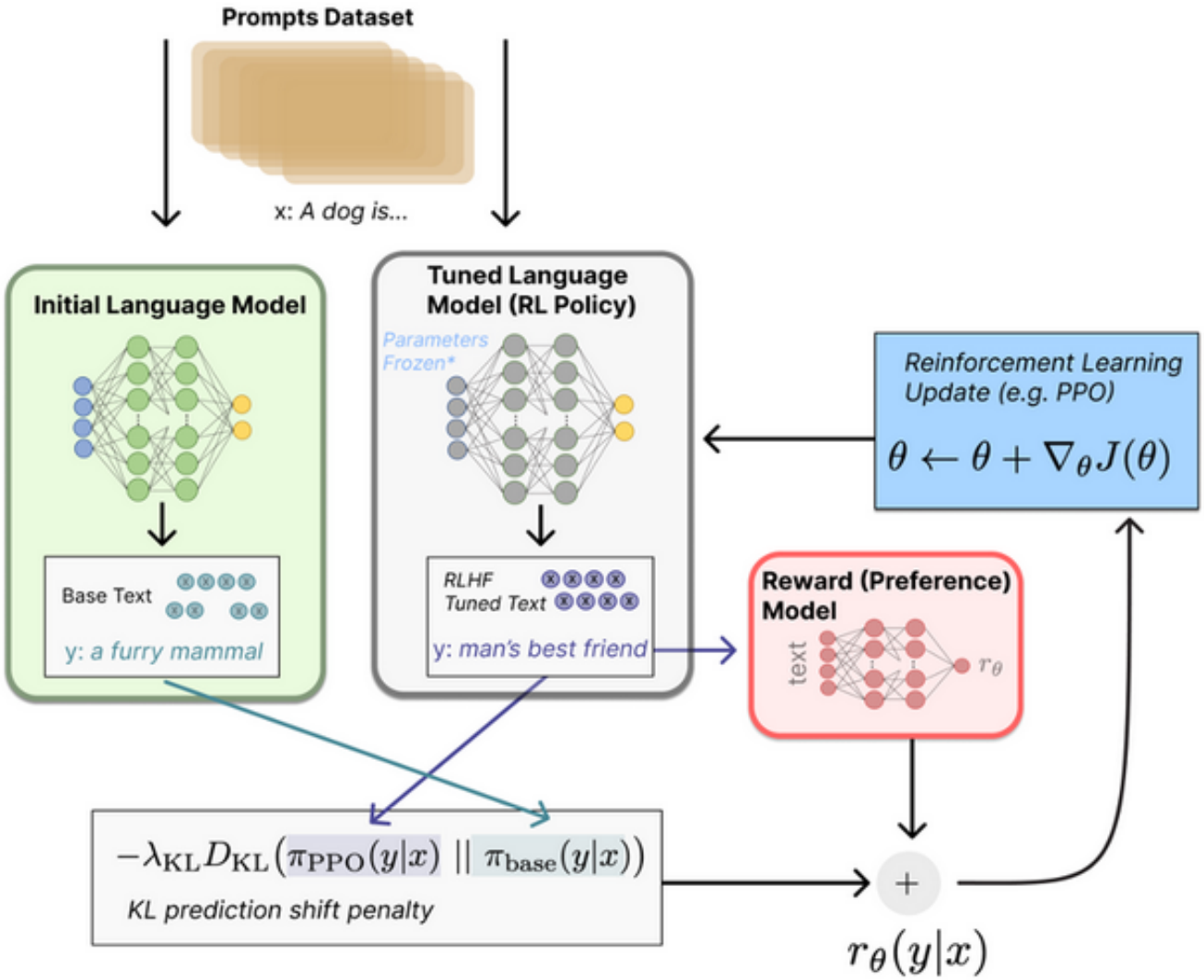
- Správne hodnoty odmien pre odpovede nevieme
- Cez sieť postupne obe odpovede - dva rôzne odhady
- Loss funkcia kombinuje oba odhady odmien

$$-log(sigmoid(R_+ - R_-))$$



- R_+/R_- - odmena lepšej/horšej odpovede
- ak sieť ohodnocuje zle, tak veľký zásah, ak dobre tak malý

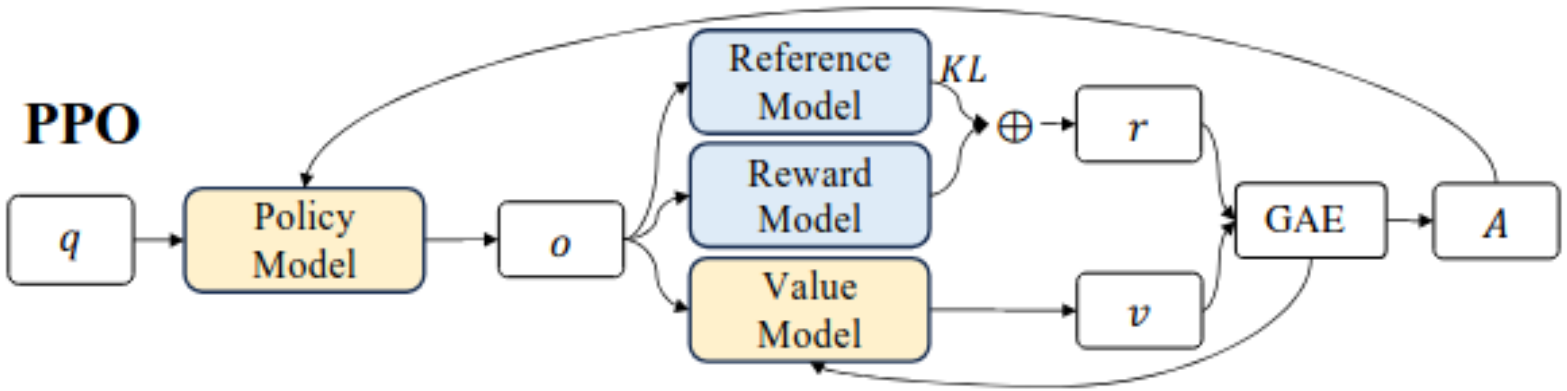
Trénovanie originálneho LLM



PPO pre RLHF

$$J(\theta) = \frac{1}{|o|} \sum_{t=0}^{|o|} \min(r_t(\theta) A_\pi(s_t, a_t), clip(r_t(\theta), 1-\epsilon, 1+\epsilon) A_\pi(s_t, a_t))$$

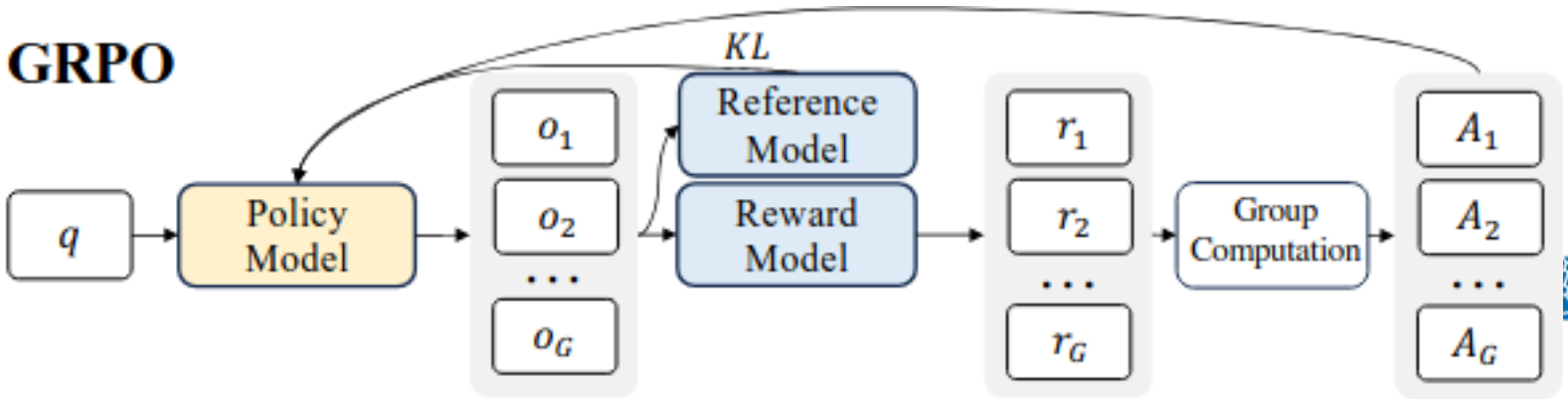
- $| o |$ je počet tokenov v odpovedi
- kvôli over optimalizácii modelu odmeňovania pridáva k odmene KL penalizáciu
- pre všetky tokeny v odpovedi je rovnaká hodnota A_π
- PPO v RLHF bol dlho štandard (GPT4, Llama2, Claude, Gemini)



GRPO pre RLHF

$$J(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o|}\sum_{t=0}^{|o|}\min(r_{i,t}(\theta)\hat{A}_i(s_t,a_t), clip(r_{i,t}(\theta),1-\epsilon,1+\epsilon)\hat{A}_i(s_t,a_t)) - \beta D_{KL}(\pi_\theta \,||\, \pi_{\theta_{old}})$$

- G je počet generovaných odpovedí (podľa $\pi_{\theta_{old}}$)
- KL penalizáciu nepridáva k odmene
- na výpočet ziskovej funkcie nepotrebuje kritika $\hat{A}_{i,t} = \frac{r_i - mean(\bar{r})}{std(\bar{r})}$
- v súvislosti s modelom Deepseek R1



GSPO pre RLHF

$$J(\theta) = \frac{1}{G}\sum_{i=1}^{G}\min(s_i(\theta)\hat{A}_i, clip(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i)$$

- nepracuje na úrovni tokenov ale celej odpovede
- $s_i(\theta)$ je pravdepodobnostný pomer sekvencie normalizovaný dĺžkou sekvencie $s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}$
- G je počet generovaných odpovedí (podľa $\pi_{\theta_{old}}$)
- na výpočet ziskovej funkcie nepotrebuje kritika $\hat{A}_i = \frac{r_i - mean(\bar{r})}{std(\bar{r})}$
- v súvislosti s modelom Qwen3