Constitutional AI (CAI)

Constitutional AI je přístup k trénování modelů, který nahrazuje rozsáhlé lidské hodnocení (RLHF) sadou pevných pravidel nebo principů – tzv. **Ústavou**. AI se pak učí sama sebe korigovat a hodnotit podle těchto pravidel.

Proč vznikla?

Běžné modely se ladí pomocí metody **RLHF** (Reinforcement Learning from Human Feedback), kde lidé hodnotí odpovědi. To má ale nevýhody:

- Je to velmi drahé a pomalé.
- Lidé mohou do AI nevědomky vnést své vlastní předsudky.
- Je těžké definovat, co je „správné“ pro miliony různých témat.

Jak CAI funguje? (Dvě fáze tréninku)

Trénink probíhá ve dvou hlavních krocích, kde AI v podstatě „vychovává sama sebe“:

1. Fáze: Sebekritika a revize (Supervised Learning)

Model dostane za úkol vygenerovat odpověď na potenciálně škodlivý dotaz. Následně:

1. Model si přečte svou „Ústavu“ (např. „Buď užitečný, ale nepodporuj násilí“).
2. Model sám zkritizuje svou původní odpověď.
3. Model vytvoří novou, revidovanou verzi odpovědi, která už pravidla splňuje.

2. Fáze: Posilované učení z AI zpětné vazby (RLAIF)

V této fázi se model učí vybírat lepší odpovědi z dvojic možností. Místo člověka ale o tom, která odpověď je lepší, rozhoduje jiný AI model na základě ústavních principů. Tím vzniká **RLAIF** (Reinforcement Learning from AI Feedback).

Srovnání metod trénování

Vlastnost	**RLHF (Lidská zpětná vazba)**	**CAI (Konstituční UI)**
Zdroj pravidel	Subjektivní pocity lidí	Jasně definovaná „Ústava“
Škálovatelnost	Nízká (omezeno počtem lidí)	Vysoká (běží automaticky)
Transparentnost	Nízká (nevíme, proč člověk dal bod)	Vysoká (víme, který princip byl použit)
Hlavní představitel	ChatGPT (OpenAI)	Claude (Anthropic)

[Image comparing RLHF and RLAIF architectures]

Příklady ústavních principů

Ústava modelu není jeden dokument, ale soubor instrukcí inspirovaných např.:

- Všeobecnou deklarací lidských práv OSN.
- Pravidly pro bezpečnost v digitálním prostoru.
- Etickými kodexy (např. „Nebuď arogantní“, „Nepoučuj uživatele zbytečně“).

Výhody a význam

- **Bezpečnost:** Model je mnohem odolnější vůči pokusům o „jailbreak“ (obejití pravidel).
- **Předvídatelnost:** Chování AI je určeno textem, který si vývojáři mohou přečíst a upravit.
- **Rychlost vývoje:** Nové verze modelů lze trénovat mnohem rychleji bez čekání na armádu lidských testerů.

—

> **Zajímavost:** Díky CAI je model Claude známý tím, že bývá méně „přednášející“ a více věcný než jeho konkurenti, protože jeho ústava mu přímo ukládá, jakým tónem má s lidmi mluvit.

Související: LLM, Reinforcement Learning, Původní vědecká práce Anthropic