

Lineární alchymie čáry, váhy, modely

Výzkumný tým Large Language Models



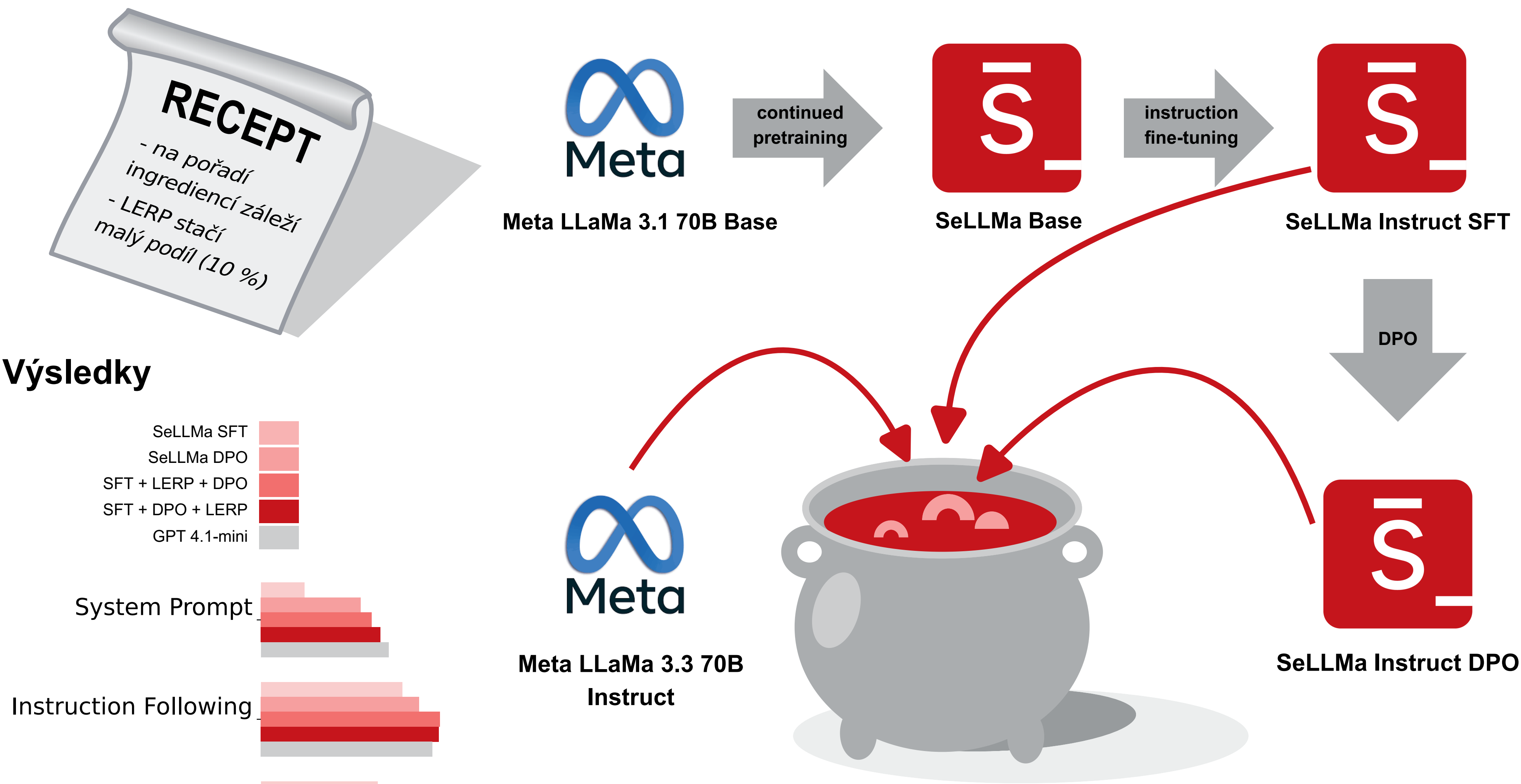
Motivace: lepší model

- Čeština (jazyková správnost, přirozené vyjadřování)
- Splnění instrukcí (instruction following)
- Systémový prompt (system prompt)
- Bezpečnost (safety)
- Volání externích funkcí (function call)
- Strukturovaný výstup (structured output)
- Pokračující konverzace (multiturn conversation)
- Cyklení
- Dlouhý kontext (long context)

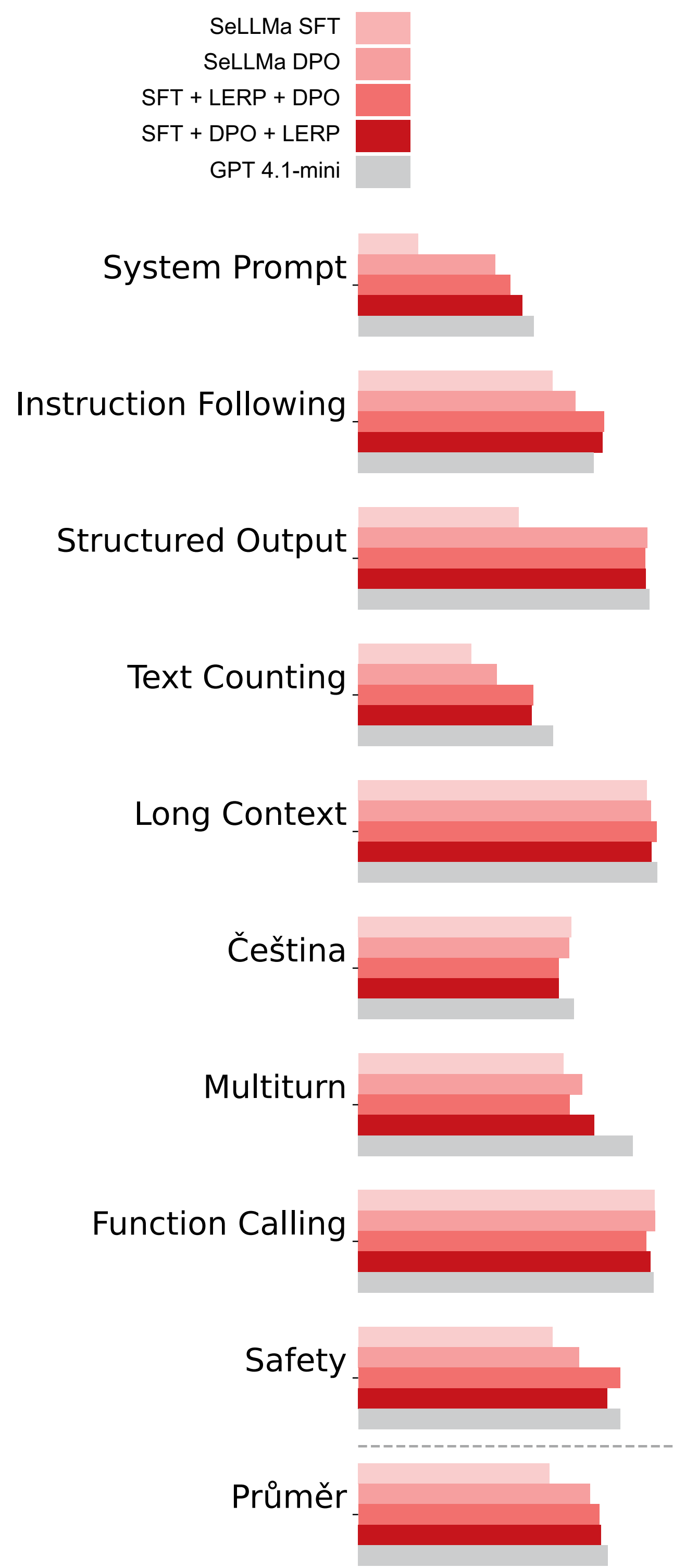
Jak editovat model

- Trénování
 - Pokračující trénování (continued pretraining) ✓
 - Ladění (supervised fine-tuning, SFT) ✓
- Zarovnání podle preferencí (alignment)
 - Posilování na základě zpětné vazby (reinforcement learning from human feedback, RLHF)
 - Přímá optimalizace (direct preference optimization, DPO) ✓
- Ladění části dat (LoRA a spol.)
- Editace založená na znalostech (knowledge editing)
- Aritmetické operace nad vahami modelu (lineární interpolace, linear interpolation, LERP) ✓

Pomocí lineárních operací tvoříme lepší modely, než máme na vstupu.



Výsledky



Vibe check: jak se jmenuje v originále Mrakoplaš?

Meta LLaMa 3.3 70B Instruct

V originále se Mrakoplaš jmenuje Stormbreaker.

SeLLMa Instruct DPO

V originále se Mrakoplaš jmenuje Cloud Giant. Je to postava z knihy The Hobbit (Hobit) od J.R.R. Tolkiena. V českém překladu ji pan Josef Škvorecký pojmenoval Mrakoplaš, což je velmi povedené a vtipné jméno.

SFT + DPO + LERP 0.1

V originále se Mrakoplaš jmenuje Rincewind. Je to postava z fantasy série Discworld (Zeměplocha) od Terryho Pratchetta. Rincewind je zbabělý a neschopný čaroděj, který se neustále dostává do problémů. Jeho jméno je často překládáno jako Mrakoplaš v českých překladech Jana Kantůrky.

