Komplexní průvodce lokálními velkými jazykovými modely (LLM) pro osobní počítače

Část I: Základy pro provoz lokální umělé inteligence

Tato část pokládá základní znalosti nezbytné pro orientaci ve světě lokální umělé inteligence. Zabývá se otázkami "jak" a "proč" předtím, než se ponoří do "co", a zajišťuje, aby byl uživatel vybaven pro informovaná rozhodnutí o hardwaru, softwaru a formátech modelů.

Sekce 1: Ekosystém lokální AI: Klíčové koncepty pro ne-experty

Úvod do lokálních LLM

Provozování velkého jazykového modelu (LLM) "lokálně" znamená, že veškerý software a data modelu jsou uloženy a zpracovávány přímo na osobním počítači uživatele, nikoli na vzdálených serverech v cloudu. Tento přístup nabízí zásadní výhody, které jsou u cloudových služeb, jako je ChatGPT, nedosažitelné: absolutní soukromí, protože žádná data neopouštějí zařízení; možnost offline použití bez nutnosti připojení k internetu; a plnou kontrolu nad systémem umělé inteligence, včetně výběru modelu, jeho konfigurace a způsobu použití.[1]

Demystifikace základní terminologie

Pro orientaci v oblasti lokálních LLM je nezbytné porozumět několika klíčovým termínům:

- **Parametry (např. 7B, 70B):** Parametry jsou proměnné, které se model učí během tréninku. Jejich počet, obvykle udávaný v miliardách (B), zhruba odpovídá velikosti a potenciálním schopnostem modelu. Je však důležité si uvědomit, že vyšší počet parametrů automaticky neznamená lepší výkon. Moderní, efektivní modely často dosahují srovnatelných nebo lepších výsledků s menším počtem parametrů díky pokročilejší architektuře a kvalitnějším tréninkovým datům.[2]
- **Kontextové okno (např. 8k, 128k):** Kontextové okno představuje krátkodobou paměť modelu, měřenou v "tokenech". Definuje maximální množství informací (vstupní text a předchozí konverzace), které model může zohlednit při generování odpovědi. Větší kontextové okno umožňuje modelu zpracovávat delší dokumenty, udržovat koherenci v rozsáhlých konverzacích a lépe sledovat složité instrukce.[2]
- **Tokeny:** Tokeny jsou základní jednotky textu (nebo kódu), které model zpracovává. V angličtině jeden token zhruba odpovídá ¾ slova. Pochopení konceptu tokenů je klíčové pro porozumění limitům kontextového okna a metrikám výkonu, jako je rychlost generování (tokeny za sekundu).[6]

Role kvantizace: Jak vměstnat obry na váš stůl

Kvantizace je proces snížení přesnosti (bitové hloubky) vah (parametrů) modelu. Tento proces je naprosto klíčový, protože dramaticky snižuje velikost modelu a jeho nároky na paměť, což umožňuje jeho spuštění na běžném spotřebitelském hardwaru.[8]

- **GGUF (GGML Universal Format):** GGUF se stal de facto standardem pro kvantizované modely určené pro běh na CPU a GPU. Jedná se o formát souboru, který byl vyvinut v rámci projektu llama.cpp pro efektivní ukládání a načítání těchto optimalizovaných modelů. Jeho univerzálnost umožňuje snadné sdílení a používání modelů napříč různými systémy a aplikacemi.[10]
- **Úrovně kvantizace (Q2_K, Q4_K_M, Q8_0 atd.):** Názvosloví kvantizačních úrovní popisuje kompromis mezi velikostí souboru, výkonem a ztrátou kvality. Například model Llama 3 8B ve formátu Q8_0 (8bitová kvantizace) bude mít velikost přibližně 8.54 GB a nabídne téměř shodnou kvalitu jako původní model. Naproti tomu verze Q4_K_M (přibližně 4bitová kvantizace) bude mít velikost jen 4.92 GB, což ji činí vhodnou pro hardware s menší pamětí, avšak za cenu mírné ztráty přesnosti. Nižší úrovně, jako Q2_K, dále snižují velikost (3.17 GB), ale již s výraznějším dopadem na kvalitu odpovědí.[12]

Celý ekosystém lokálních LLM byl formován jediným dominantním omezením: limitovanou kapacitou VRAM (Video RAM) na spotřebitelských grafických kartách. Toto hardwarové omezení je hlavní příčinou univerzálního přijetí kvantizace a formátu GGUF. Původní velké jazykové modely vyžadovaly hardware na úrovni datových center.[15] Snaha o jejich lokální provoz z důvodů soukromí a nákladů vedla k poptávce po menších a efektivnějších modelech.[1] Hlavním úzkým hrdlem se ukázala být právě VRAM, která musí pojmout váhy modelu pro rychlé zpracování.[15] Komunita proto vyvinula kvantizační techniky, přičemž projekt

llama.cpp se stal průkopníkem a vytvořil formát GGUF jako přenositelný a efektivní standard.[10] Z toho vyplývá, že cesta uživatele nezačíná výběrem modelu jako "Llama 3", ale odpovědí na otázku: "Kolik VRAM mám k dispozici?". Tato odpověď následně určuje, jakou kvantizovanou verzi GGUF daného modelu je reálné spustit – například verzi

Q4_K_M pro 8GB GPU oproti Q8_0 pro 12GB GPU.[12] Tento pohled přerámovává celý proces výběru modelu kolem praktické hardwarové reality.

Sekce 2: Sestavení vaší lokální AI pracovní stanice: Hloubkový pohled na hardware

Centrální role GPU

Ačkoli je možné provozovat LLM pouze na procesoru (CPU), výkon je dramaticky pomalejší. Grafické karty (GPU) jsou pro přijatelný výkon nezbytné, protože jejich architektura je optimalizována pro paralelní maticové násobení, které tvoří jádro operací LLM.[15]

- **VRAM je král:** Množství VRAM je nejdůležitějším parametrem. Určuje maximální velikost modelu (nebo jeho kvantizované verze), který lze plně nahrát do paměti GPU pro nejrychlejší generování odpovědí (inferenci).[15]
- **NVIDIA vs. AMD vs. Apple Silicon:**
  - **NVIDIA:** Grafické karty NVIDIA mají nejširší podporu díky dominantní softwarové platformě CUDA, která je standardem v oblasti AI.[15]
  - **AMD:** Podpora pro GPU od AMD se neustále zlepšuje, ale může vyžadovat složitější

konfiguraci.
    - **Apple Silicon (M1/M2/M3):** Tyto čipy jsou silným konkurentem díky své architektuře sjednocené paměti. Ta umožňuje CPU a GPU sdílet velký objem systémové RAM, která efektivně funguje jako VRAM, což je ideální pro provoz velkých modelů.[18]

Další kritické komponenty

- **Systémová RAM:** Pokud je model příliš velký pro VRAM, jeho části jsou "přesunuty" do systémové RAM. Tento proces je pomalejší než čistá inference z VRAM, ale výrazně rychlejší než provoz pouze na CPU. Dostatek systémové RAM (minimum 16 GB, doporučeno 32 GB a více) je proto klíčový jako záloha a pro celkovou stabilitu systému.[15]
- **CPU:** Ačkoli je pro rychlost inference druhořadý, moderní vícejádrový procesor (např. Intel Core i5/Ryzen 5 nebo lepší) je stále důležitý pro předzpracování dat a celkovou správu systému.[15]
- **Úložiště:** Doporučuje se rychlý SSD disk (ideálně NVMe) kvůli velké velikosti souborů modelů (kvantizované verze často zabírají 4–10 GB). Rychlé úložiště výrazně zkracuje dobu načítání modelů.[15]

Zohlednění operačního systému (Windows vs. Linux vs. macOS)

Všechny tři hlavní operační systémy jsou podporovány hlavními nástroji pro lokální AI.[22]

- **Windows:** Nejsnazší pro začátečníky díky jednoduchým instalátorům pro nástroje jako LM Studio a Oobabooga. Nastavení ovladačů GPU je přímočaré.[21]
- **Linux:** Často preferován vývojáři pro své výkonné příkazové řádky a robustní podporu pro Docker a ovladače NVIDIA, což je ideální pro nástroje jako Ollama.[27]
- **macOS:** Nabízí vynikající "out-of-the-box" zkušenost, zejména na hardwaru Apple Silicon, kde nástroje jako Ollama a Jan fungují velmi dobře díky sjednocené paměti.[18]

Tabulka 1: Hardwarové úrovně pro lokální LLM

Tato tabulka poskytuje jasné a praktické hardwarové konfigurace založené na rozpočtu a požadovaných schopnostech, čímž převádí abstraktní koncepty VRAM a RAM do konkrétních příkladů.

Úroveň	Případ použití	GPU (VRAM)	Systémová RAM	CPU	Úložiště	Příklad modelů
Základní	Základní chat a generování textu (modely	NVIDIA RTX 3060 (12 GB) / AMD RX 7600 XT	16–32 GB DDR4	Moderní 6jádrový (Ryzen 5 / Core i5)	1 TB NVMe SSD	Llama 3 8B (Q5), Mistral 7B (Q6)

	do 8B)	(16 GB)				
Střední třída	Pokročilé úkoly a kódování (modely 13B–34B)	NVIDIA RTX 4070 Super (12 GB) / RTX 4080 (16 GB)	32–64 GB DDR5	Moderní 8jádrový (Ryzen 7 / Core i7)	2 TB NVMe SSD	Phi-3 Medium (14B), Qwen 32B (Q4)
Prosumer /Nadšenec	Provoz největších modelů (70B+)	2x NVIDIA RTX 3090 (24 GB každý) / RTX 4090 (24 GB)	64–128 GB DDR5	High-end 12+ jádrový (Ryzen 9 / Core i9)	4 TB+ NVMe SSD	Llama 3 70B (Q4), Mixtral 8x7B (Q4)

Sekce 3: Brána k lokálním modelům: Průvodce běhovými prostředími a rozhraními

Tato sekce je praktickým průvodcem softwarem, který spouští modely, přizpůsobeným různým úrovním dovedností uživatelů.

Pro začátečníky: Aplikace s grafickým rozhraním

- **LM Studio:** Prezentováno jako vyladěné "vše v jednom" řešení. Mezi jeho klíčové vlastnosti patří integrovaný prohlížeč modelů pro stahování z Hugging Face, jednoduché chatovací rozhraní, lokální inferenční server a podpora pro Windows, macOS a Linux. Jeho hlavní výhodou je snadné použití pro netechnické uživatele.[11] Instalace pro Windows je jednoduchá: stačí stáhnout a spustit instalační soubor, v aplikaci přejít na kartu "Discover", vyhledat a stáhnout požadovaný model (např. Llama 3) a poté na kartě "Chat" model načíst a začít konverzaci.[21]
- **Jan:** Pozicováno jako open-source alternativa k LM Studio s čistým uživatelským rozhraním a aktivním komunitním vývojem. Zaměřuje se na uživatelskou zkušenost, automatickou detekci hardwaru pro doporučení vhodných modelů a má vestavěné schopnosti pro RAG (Retrieval-Augmented Generation).[1] Pro začátečníky stačí stáhnout aplikaci z webu jan.ai, která je sama navede při výběru a stažení prvního modelu.[1]
- **GPT4All:** Stručně zmíněno jako další populární aplikace, která poskytuje kurátorované modely fungující "out of the box", což dále zjednodušuje první kroky.[10]

Pro vývojáře a pokročilé uživatele: Nástroje pro příkazový řádek a webová rozhraní

- **Ollama:** Popsáno jako výkonný nástroj pro příkazový řádek inspirovaný Dockerem. Jeho silnými stránkami jsou snadná správa modelů (např. ollama run llama3), vestavěný API server pro integraci s dalšími aplikacemi (např. Python skripty, rozšíření pro VS Code) a silná komunitní podpora. Je to preferovaný nástroj pro vývojáře a "homelab" nadšence.[10] Instalace na Linuxu a macOS je otázkou jediného příkazu. Po instalaci lze model spustit příkazem ollama run <název_modelu>. Vlastní modely lze definovat pomocí souboru Modelfile, kde lze nastavit systémový prompt a další parametry, a vytvořit je příkazem ollama create <vlastní_název> -f./Modelfile.[31]
- **Oobabooga's Text Generation WebUI:** Představeno jako nejbohatší a nejrozšiřitelnější webové rozhraní. Vyniká podporou široké škály formátů modelů a backendů, pokročilými funkcemi jako je fine-tuning, multimodální podpora a obrovským ekosystémem rozšíření.[10] Ačkoli je velmi výkonné, jeho nastavení může být pro začátečníky složitější, i když instalátory na jedno kliknutí tento proces zjednodušily.[25]

Soutěž mezi běhovými prostředími jako LM Studio, Jan a Ollama není jen o funkcích, ale o vytvoření dominantní "abstrakční vrstvy" nad základní složitostí technologie llama.cpp. Vítěz této soutěže definuje uživatelskou zkušenost pro další vlnu uživatelů lokální AI. Jádrem technologie pro spouštění kvantizovaných modelů je llama.cpp [10], ale jeho přímé použití vyžaduje kompilaci C++ kódu a složité argumenty příkazového řádku, což je pro většinu uživatelů překážkou.[28] Nástroje jako Ollama, LM Studio a Jan jsou v podstatě uživatelsky přívětivé obaly, které tuto složitost automatizují.[1] LM Studio a Jan sázejí na grafické, "aplikaci podobné" rozhraní, aby oslovily netechnické uživatele.[1] Ollama naopak cílí na vývojáře s "Dockeru podobným" přístupem, s cílem stát se standardem pro programový přístup a integraci.[10] To staví uživatele před klíčovou volbu: chtějí "aplikaci" pro chatování (LM Studio/Jan), nebo "motor" pro stavbu vlastních řešení (Ollama)? Tato volba má dlouhodobé důsledky. Uživatel, který začne s LM Studio, může narazit na potíže při integraci modelů do vlastního skriptu, což je s Ollamou triviální úkol.

Tabulka 2: Porovnání běhových prostředí pro lokální LLM

Tato tabulka poskytuje rychlé srovnání, které uživateli pomůže vybrat správný software pro jeho technickou úroveň a cíle.

Nástroj	Cílová skupina	Nastavení	Klíčové vlastnosti	API/Rozšiřitelnost
LM Studio	Začátečníci, ne-vývojáři	GUI instalátor (Win/Mac/Linux)	Integrovaný hub modelů, jednoduché chat UI, lokální server	OpenAI-kompatibilní API server
Jan	Začátečníci, open-source	GUI instalátor (Win/Mac/Linux)	Čisté UI, open-source,	Lokální API server

	nadšenci		RAG integrace	
Ollama	Vývojáři, pokročilí uživatelé	Příkazový řádek (Linux/Mac), instalátor (Win)	CLI správa modelů, robustní API, vlastní Modelfiles	Prvotřídní API, Python/JS knihovny
Oobabooga's WebUI	Výzkumníci, pokročilí hobbyisté	Instalátor na jedno kliknutí nebo manuální Python setup	Rozsáhlé možnosti fine-tuningu, rozšíření, multimodální podpora	API, rozsáhlý ekosystém rozšíření

Část II: Komplexní přehled lokálních velkých jazykových modelů

Tato část je jádrem zprávy a poskytuje podrobné profily nejdůležitějších a nejdostupnějších modelů, které si uživatel může spustit lokálně. Každá sekce se bude věnovat jiné modalitě (text, obraz, řeč, generování obrázků).

Sekce 4: Modely pro generování textu: Konverzační jádro

Tato sekce profiluje přední open-source rodiny textových modelů. Každý profil obsahuje informace o vývojáři, klíčových vlastnostech, dostupných velikostech (parametrech), licenci a shrnutí silných a slabých stránek.

- **Meta Llama 3 & 3.1/3.2:** Aktuální průmyslový standard pro otevřené modely. Jsou známé vysokým výkonem napříč širokou škálou úloh a škálovatelností.[2] Zahrnuty jsou velikosti 8B a 70B, s poznámkou o jejich hardwarových nárocích: 8B vyžaduje přibližně 8 GB VRAM, zatímco 70B vyžaduje přibližně 42 GB VRAM pro 4bitovou kvantizaci.[8] GGUF verze těchto modelů lze snadno nalézt na platformě Hugging Face.[12]
- **Google Gemma & Gemma 2:** Rodina otevřených modelů od Googlu, postavená na stejném výzkumu jako Gemini. Jsou dostupné v různých velikostech (např. 2B, 9B, 27B) a jsou známé silným výkonem na spotřebitelském hardwaru.[2] Existují různé verze s odlišnými schopnostmi.[37]
- **Modely od Mistral AI (Mistral 7B, Mixtral 8x22B):** Významný hráč z Francie. Mistral 7B je proslulý svou efektivitou, překonává větší modely a přitom je velmi nenáročný na zdroje (přibližně 6 GB VRAM pro kvantizaci Q5).[2] Modely Mixtral využívají architekturu SMoE (Sparse Mixture-of-Experts) pro vysoký výkon s nižšími inferenčními náklady.[34] K dispozici jsou různé GGUF verze.[40]
- **Alibaba Qwen (Qwen1.5, Qwen2.5, Qwen3):** Výkonná rodina modelů od společnosti Alibaba.

Jsou známé silnými vícejazyčnými schopnostmi a širokou škálou velikostí od 0.5B do 72B parametrů.[2] Tento model je klíčový pro svou podporu českého jazyka.[42] GGUF verze jsou dostupné.[44]

- **Microsoft Phi-3:** Rodina "malých jazykových modelů" (SLM), které i přes svou malou velikost (např. 3.8B "mini") prokazují vynikající schopnosti v oblasti uvažování a porozumění jazyku. Jsou navrženy pro inferenci na zařízení a v prostředích s omezenými zdroji, což je činí ideálními pro lokální použití.[2]
- **Další významné modely:**
  - **Cohere Command R+:** Optimalizovaný pro RAG a podnikové použití, s otevřenou verzí pro nekomerční účely.[2]
  - **Falcon 2:** Nabízí vícejazyčné a multimodální schopnosti.[2]
  - **DeepSeek:** Silné modely zaměřené na kódování a uvažování.[2]

Tabulka 3: Porovnání vlajkových textových modelů (malá/střední třída)

Tato tabulka poskytuje přímé srovnání nejoblíbenějších a nejdostupnějších modelů, které si uživatel pravděpodobně vyzkouší jako první. Zaměřuje se na třídu s ~7–14 miliardami parametrů, což je optimální pro spotřebitelský hardware.

Model	Vývojář	Parametry	Přibližná VRAM (Q4_K_M)	Licence	Oficiální podpora češtiny
Llama 3.1 8B Instruct	Meta	8B	~5.0 GB	Llama 3 Community	Ne (existují komunitní fine-tunes)
Mistral 7B Instruct v0.3	Mistral AI	7.3B	~4.5 GB	Apache 2.0	Ne
Gemma 2 9B Instruct	Google	9B	~5.5 GB	Gemma Terms of Use	Ne (až Gemma 3)
Qwen1.5 7B Chat	Alibaba	7B	~4.5 GB	Apache 2.0	Ano
Phi-3 Mini 3.8B Instruct	Microsoft	3.8B	~2.5 GB	MIT License	Ne

Sekce 5: Multimodální modely: Když AI získá zrak

Úvod do Vision-Language modelů (VLM)

Vision-Language modely (VLM) jsou schopny zpracovávat jak text, tak obrázky jako vstup. To jim umožňuje provádět úkoly, jako je popisování obsahu obrázku, odpovídání na otázky týkající se vizuálních prvků nebo dokonce čtení textu z dokumentů a diagramů.[49]

Doporučený model: LLaVA (Large Language and Vision Assistant)

LLaVA je přední open-source VLM. Funguje tak, že kombinuje vizuální enkodér (typicky CLIP) s velkým jazykovým modelem (jako je Vicuna nebo Mistral), čímž propojuje schopnost "vidět" se schopností "rozumět a mluvit".[53]

- **Jak spustit LLaVA lokálně:** Díky moderním nástrojům je spuštění LLaVA překvapivě jednoduché, zejména s použitím Ollama:
  1. Nainstalujte Ollama.
  2. Spusťte v příkazovém řádku ollama run llava.
  3. Po spuštění modelu zadejte textový prompt následovaný cestou k obrázku.
     Tento jednoduchý proces demonstruje, jak moderní běhová prostředí zjednodušují dříve složité instalace.53 LLaVA je také podporována v nástrojích jako Oobabooga's WebUI a přímo v
     llama.cpp.[33]
- **Hardwarové požadavky pro VLM:** VLM vyžadují o něco více VRAM než jejich čistě textové protějšky kvůli přidanému vizuálnímu enkodéru. Model jako LLaVA 7B typicky vyžaduje alespoň 8 GB RAM/VRAM pro plynulý běh.[53]

Další open-source VLM

Kromě LLaVA se objevují i další multimodální modely, jako jsou **Falcon 2 11B VLM** [34],

InternVL [50] a multimodální verze modelů

Qwen a **Gemma**.[51]

Sekce 6: Řečové a audio modely: AI, která naslouchá

Úvod do automatického rozpoznávání řeči (ASR)

Automatické rozpoznávání řeči (ASR) je technologie, která převádí mluvené slovo z audio záznamu na psaný text.

Doporučený model: OpenAI Whisper

Whisper je považován za state-of-the-art open-source ASR model. Byl trénován na masivním a rozmanitém datovém souboru obsahujícím 680 000 hodin audio záznamů, což mu propůjčuje výjimečnou odolnost vůči různým akcentům, hluku v pozadí a technickému jazyku.[57]

- **Velikosti modelů a kompromisy:** Whisper je k dispozici v několika velikostech (tiny, base, small, medium, large). Existuje přímý vztah mezi velikostí modelu, jeho nároky na VRAM, rychlostí a přesností, která se měří pomocí metriky Word Error Rate (WER) – čím nižší WER, tím vyšší přesnost.[58]
- **Hardwarové požadavky:** Největší model, large-v3, vyžaduje pro optimální výkon přibližně 10 GB VRAM.[62] Menší modely jsou mnohem méně náročné:
  small vyžaduje ~2 GB a base ~1 GB VRAM, což činí Whisper dostupným pro širokou škálu hardwaru.[58]
- **Snadno použitelná GUI pro Whisper:** Pro uživatele, kteří preferují grafické rozhraní před příkazovým řádkem, existuje několik aplikací, které poskytují jednoduché uživatelské prostředí pro Whisper. Mezi ně patří **Whisper-GUI** [64],
  Whisper Transcription UI [65] a webová rozhraní jako
  WhisperWebUI [59] nebo
  WhisperUI [66], která často využívají uživatelův API klíč, ale zjednodušují proces transkripce.

Další open-source ASR modely

Pro historický kontext je vhodné zmínit i starší, ale stále relevantní projekty jako **Mozilla DeepSpeech** a **Kaldi**.[57]

Tabulka 4: Varianty modelu OpenAI Whisper

Tato tabulka jasně ilustruje kompromisy mezi výkonem a nároky na zdroje, což uživateli umožňuje vybrat si správný model Whisper pro jeho hardware a potřeby přesnosti.

Velikost	Parametry	Požadovaná VRAM (přibl.)	Relativní rychlost (vs. large)	Přesnost / Případ použití
tiny	39M	~1 GB	~10x	Nejlepší pro zařízení s

				nízkými zdroji
base	74M	~1 GB	~7x	Dobrá rovnováha pro obecné použití
small	244M	~2 GB	~4x	Vyšší přesnost
medium	769M	~5 GB	~2x	Velmi vysoká přesnost
large-v3	1550M	~10 GB	1x	Nejlepší přesnost pro kritické úkoly

Sekce 7: Modely pro generování obrázků: AI jako umělec

Úvod do difuzních modelů

Difuzní modely, jako je Stable Diffusion, fungují na principu postupného odstraňování šumu z náhodného obrazu a jeho přetváření podle textového zadání (promptu) až do finální podoby.[68]

Doporučený model: Stable Diffusion & SDXL

Stable Diffusion je přední open-source model pro generování obrázků. Starší verze (např. v1.5) generují obrázky v rozlišení 512x512 pixelů, zatímco novější a výkonnější modely **SDXL** produkují kvalitnější obrázky v rozlišení 1024x1024 pixelů.[69]

- **Hardwarové požadavky:** Generování obrázků je hardwarově náročné. Základní verze Stable Diffusion vyžaduje GPU s alespoň 4–6 GB VRAM. Pro SDXL je oficiálně doporučeno minimum 8 GB VRAM, přičemž 16 GB je ideální pro plynulý výkon.[20] Generování pouze na CPU je možné, ale extrémně pomalé.[20]
- **Průvodce instalací pro začátečníky:** Pro ne-experta je nejjednodušší začít s uživatelsky přívětivým webovým rozhraním, jako je **Fooocus** nebo **InvokeAI**, které jsou známé svými jednoduchými instalátory a intuitivním ovládáním.[71] Proces obvykle zahrnuje:
  1. Stažení a spuštění instalátoru webového rozhraní.
  2. Stažení základního modelu (tzv. checkpoint) z webů jako Civitai nebo Hugging Face a jeho umístění do správné složky.[69]
  3. Spuštění webového rozhraní a zadání prvního promptu.
- **Umění promptování:** Kvalita výsledného obrázku silně závisí na kvalitě promptu. Dobrý prompt by

měl obsahovat popis hlavního subjektu, uměleckého stylu, osvětlení a atmosféry. Důležitou technikou je také použití "negativních promptů", které modelu říkají, co se v obrázku nemá objevit (např. "rozmazaný, text, deformace").[68]

Část III: Pokročilá témata a praktická hlediska

Tato závěrečná část se zabývá pokročilejšími otázkami uživatele ohledně přizpůsobení, právních aspektů a bezpečnosti, čímž poskytuje kompletní a ucelenou zprávu.

Sekce 8: Učení vašeho lokálního LLM: Úvod do přizpůsobení

Předtrénované modely mají obecné znalosti, ale chybí jim specifické informace o osobních nebo firemních datech. Existují dva hlavní způsoby, jak model "naučit" nové věci.

Retrieval-Augmented Generation (RAG): Dlouhodobá paměť pro váš LLM

RAG je technika, která umožňuje LLM přistupovat k informacím z externích dokumentů (PDF, textové soubory atd.) a načítat je před generováním odpovědi. Pro ne-experta je to nejdostupnější způsob, jak model "seznámit" s vlastními daty, aniž by se měnil samotný model.[72]

- **Jak to funguje (zjednodušeně):** Dokumenty jsou rozděleny na menší části (chunky), převedeny na číselné reprezentace (embeddings) a uloženy ve vektorové databázi. Když uživatel položí otázku, systém nejprve najde nejrelevantnější části dokumentů a poskytne je LLM jako kontext spolu s původní otázkou.[72]
- **Praktické RAG pro začátečníky:** Mnoho uživatelsky přívětivých nástrojů má dnes RAG vestavěné. Například **LM Studio** a **Jan** nabízejí funkci "Chat with Documents", která tento proces zjednodušuje na nahrání souboru a položení otázky.[28] Pro pokročilejší uživatele existují frameworky jako
LangChain, které lze použít s API od Ollama pro tvorbu složitějších RAG aplikací.[72]

Fine-Tuning: Přizpůsobení chování modelu

Fine-tuning je proces dalšího tréninku předtrénovaného modelu na menším, specifickém datovém souboru s cílem přizpůsobit jeho styl, znalosti nebo chování.[76]

- **Full Fine-Tuning vs. Parameter-Efficient Fine-Tuning (PEFT):** Přetrénování všech parametrů modelu je výpočetně extrémně náročné. Techniky PEFT, jako je **LoRA (Low-Rank Adaptation)**, přidávají pouze malý počet nových trénovatelných parametrů, zatímco původní váhy modelu zůstávají zmrazené. To činí fine-tuning proveditelným i na spotřebitelském hardwaru.[78]
- **Fine-Tuning pro ne-experty:** Ačkoli se stále jedná o technický proces, nástroje jako **Oobabooga's WebUI** obsahují vestavěné tréninkové záložky, které zjednodušují proces LoRA

fine-tuningu a zpřístupňují ho pokročilým hobbyistům.[79] Proces obecně zahrnuje přípravu datového souboru ve formátu instrukcí, nastavení tréninkových parametrů a spuštění tréninku.[77]

Sekce 9: Navigace v prostředí: Licence, náklady a jazyková podpora

Náklady na lokální LLM

Ačkoli jsou modely často zdarma ke stažení, jejich provoz není bez nákladů. Primární náklad představuje počáteční investice do hardwaru a následná spotřeba elektrické energie.[16] Všechny modely diskutované v této zprávě jsou zdarma ke stažení.

Porozumění "Open Source" licencím

Toto je kritická sekce, která analyzuje licence hlavních modelů a jejich dopady na osobní a komerční použití.

- **Permisivní licence (Apache 2.0, MIT):** Modely jako **Mistral**, většina verzí **Qwen** a **Phi-3** používají tyto licence. Jsou obecně velmi benevolentní pro komerční použití a vyžadují pouze uvedení autora (atribuci).[80]
- **Vlastní komunitní licence (Llama 3, Gemma):** Tyto licence mají více omezení. **Licence Llama 3** je permisivní, ale obsahuje významnou klauzuli, která vyžaduje samostatnou licenci od společnosti Meta pro služby s více než 700 miliony aktivních uživatelů měsíčně.[83]
  Podmínky použití Gemma vyžadují dodržování Zásad zakázaného použití.[87]
- **Nekomerční licence (Command R+):** Některé modely jsou uvolněny pouze pro výzkumné nebo nekomerční účely, což je zásadní rozdíl, který je třeba brát v úvahu.[34]

Jazyková podpora se zaměřením na češtinu

Tato sekce přímo odpovídá na dotaz uživatele.

- **Llama 3:** Oficiálně podporuje 8 jazyků, ale čeština mezi nimi není. Nicméně existují komunitně upravené verze, jako je **LLaMAX3**, které explicitně přidávají podporu pro více než 100 jazyků, včetně češtiny (cs).[90]
- **Gemma:** Gemma 2 je pouze anglický model. **Gemma 3** však nabízí vícejazyčnou podporu pro více než 140 jazyků, včetně češtiny (cs).[92]
- **Qwen:** Rodina modelů Qwen má silnou vícejazyčnou podporu a oficiálně uvádí češtinu jako podporovaný jazyk.[42]
- **Mistral 7B:** Je primárně trénován na angličtině a kódu a není oficiálně vícejazyčným modelem. Ačkoli může mít určité schopnosti v jiných jazycích, pro spolehlivý výkon je nutný fine-tuning.[94]

Sekce 10: Bezpečnost a zodpovědné používání v lokálním prostředí

Obecná bezpečnostní rizika LLM

I v lokálním prostředí je bezpečnost LLM důležitá. Níže jsou uvedena hlavní rizika z OWASP Top 10 pro LLM, která jsou relevantní pro lokální použití.

- **Prompt Injection:** Útočník vytvoří vstup, který přiměje model ignorovat jeho původní instrukce a provést nežádoucí akci.[96]
- **Nezabezpečené zpracování výstupu:** Pokud je výstup LLM předán dalším systémům (např. spuštění kódu, který vygeneroval), může to vést ke zranitelnostem.[96]
- **Otrava tréninkových dat (Training Data Poisoning):** Riziko při fine-tuningu na nedůvěryhodných datových sadách. Model by se mohl naučit generovat škodlivý nebo zkreslený obsah.[96]
- **Únik citlivých informací:** Model by mohl neúmyslně odhalit citlivé informace ze svých tréninkových dat nebo z dokumentů použitých v systému RAG, pokud není správně zabezpečen.[96]

Bezpečnostní otázka "čínských modelů"

Tato část se přímo zabývá obavami uživatele ohledně modelů z Číny.

- **Technická bezpečnost vs. geopolitické zkreslení:** Analýzy ukazují, že z čistě technického hlediska neexistují důkazy o inherentních zadních vrátkách (backdoors) nebo zranitelnostech v čínských open-weight modelech, jako jsou Qwen nebo DeepSeek, ve srovnání s jejich západními protějšky.[98]
- **Skutečné riziko: Původ souboru a zkreslení dat:** Primární bezpečnostní riziko nespočívá v původu modelu, ale v **původu konkrétního souboru, který stahujete**. Zlovolný aktér by mohl nahrát upravenou, škodlivou GGUF verzi jakéhokoli modelu do repozitáře, jako je Hugging Face. Klíčem je stahovat modely z oficiálních a důvěryhodných zdrojů. Dále je třeba odlišit technickou bezpečnost od **zkreslení dat (bias)**. Modely trénované na datech z určitého regionu mohou odrážet kulturní a politické narativy tohoto regionu (např. v tématech jako Tchaj-wan). Jedná se o problém zkreslení nebo cenzury, nikoli o technickou bezpečnostní zranitelnost v tradičním smyslu.[98]

Osvědčené postupy pro bezpečné lokální nasazení

- Stahujte modely pouze z oficiálních repozitářů (např. Meta, Google, Mistral AI) nebo od vysoce renomovaných členů komunity na Hugging Face (jako je "TheBloke").
- Buďte opatrní při používání modelů ke generování a spouštění kódu. Vždy kód před spuštěním zkontrolujte.
- Při použití RAG s citlivými dokumenty se ujistěte, že je lokální počítač zabezpečený. Hlavní výhodou

lokálních LLM je, že tato data nikdy neopustí váš počítač.
- Pochopte omezení a potenciální zkreslení modelu, který používáte.

Závěr

Provozování velkých jazykových modelů na osobním počítači se stalo dostupnou realitou pro širokou škálu uživatelů, od nadšenců po vývojáře. Klíčem k této revoluci je technologie kvantizace a standardizovaný formát GGUF, které umožňují provozovat i velké modely na spotřebitelském hardwaru. Výběr správného hardwaru, zejména grafické karty s dostatečnou VRAM, je prvním a nejdůležitějším krokem. Následně si uživatel může vybrat z řady softwarových nástrojů, od uživatelsky přívětivých aplikací jako LM Studio a Jan pro začátečníky, až po výkonné nástroje pro vývojáře jako Ollama.

Ekosystém open-source modelů je bohatý a rozmanitý. Modely jako Meta Llama 3, Mistral 7B a Microsoft Phi-3 nabízejí vynikající výkon v angličtině, zatímco modely jako Alibaba Qwen poskytují robustní vícejazyčnou podporu, včetně češtiny. Kromě textových modelů jsou k dispozici i multimodální modely (LLaVA), modely pro rozpoznávání řeči (Whisper) a generování obrázků (Stable Diffusion), každý se specifickými hardwarovými nároky a nástroji pro snadné použití.

Pro pokročilé uživatele existují dostupné metody přizpůsobení, jako je RAG pro přidání znalostí z vlastních dokumentů a PEFT/LoRA pro úpravu chování modelu. Při výběru modelu je však nutné pečlivě zvážit licenční podmínky, které se liší od plně permisivních (Apache 2.0) po ty s komerčními omezeními. Z hlediska bezpečnosti platí, že technická rizika jsou stejná bez ohledu na původ modelu; klíčová je obezřetnost při stahování souborů a pochopení potenciálního zkreslení dat.

Lokální AI otevírá dveře k inovacím s důrazem na soukromí a kontrolu. S informacemi uvedenými v této zprávě může každý technologicky zdatný uživatel učinit informované rozhodnutí a úspěšně se ponořit do fascinujícího světa lokálních jazykových modelů.

Citovaná díla

1. How to run AI models locally as a beginner? - Jan, použito září 4, 2025, https://jan.ai/post/run-ai-models-locally
2. The 11 best open-source LLMs for 2025 - n8n Blog, použito září 4, 2025, https://blog.n8n.io/open-source-llm/
3. Mistral vs Llama 3: Which One Should You Choose? - Novita AI Blog, použito září 4, 2025, https://blogs.novita.ai/mistral-vs-llama-3-which-one-should-you-choose/
4. Mistral 7B vs. Llama 3 70B vs. Gemma 2 9B: A Comprehensive Benchmark Showdown | by Samir Sengupta | Medium, použito září 4, 2025, https://medium.com/@samir20/mistral-7b-vs-llama-3-70b-vs-gemma-2-9b-a-comprehensive-benchmark-showdown-9c3128f24b23
5. command-r-plus - Ollama, použito září 4, 2025, https://ollama.com/library/command-r-plus
6. The same GGUF model run in LM studio or ollama is 3-4x faster than running the same GGUF in Oobabooga - Reddit, použito září 4, 2025, https://www.reddit.com/r/Oobabooga/comments/1fy0333/the_same_gguf_model_run_in_lm_studio_or_ollama_is/
7. The same Model in text generation webui slower than in lm studio. #6330 - GitHub, použito září 4, 2025, https://github.com/oobabooga/text-generation-webui/discussions/6330
8. Unleash the Power of Llama 3 70B on Your Everyday Computer | Picovoice, použito září 4, 2025, https://picovoice.ai/blog/unleash-the-power-of-llama3-70b-on-your-everyday-computer/
9. Mistral 7B System Requirements: What You Need to Run It Locally, použito září 4, 2025, https://www.oneclickitsolution.com/centerofexcellence/aiml/run-mistral-7b-locally-hardware-software-specs
10. 50+ Open-Source Options for Running LLMs Locally | by Vince Lam | The Deep Hub, použito září 4, 2025,

https://medium.com/thedeephub/50-open-source-options-for-running-llms-locally-db1ec6f5a54f
11. LM Studio vs Ollama: Choosing the Right Tool for LLMs - Openxcell, použito září 4, 2025, https://www.openxcell.com/blog/lm-studio-vs-ollama/
12. bartowski/Meta-Llama-3-8B-Instruct-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/bartowski/Meta-Llama-3-8B-Instruct-GGUF
13. SciSharp/LLamaSharp: A C#/.NET library to run LLM (LLaMA/LLaVA) on your local device efficiently. - GitHub, použito září 4, 2025, https://github.com/SciSharp/LLamaSharp
14. second-state/Llama-3-8B-Instruct-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/second-state/Llama-3-8B-Instruct-GGUF
15. Recommended Hardware for Running LLMs Locally - GeeksforGeeks, použito září 4, 2025, https://www.geeksforgeeks.org/deep-learning/recommended-hardware-for-running-llms-locally/
16. Building a Low-Cost Local LLM Server to Run 70 Billion Parameter Models - Comet, použito září 4, 2025, https://www.comet.com/site/blog/build-local-llm-server/
17. How to Run an LLM Locally with Pieces, použito září 4, 2025, https://pieces.app/blog/how-to-run-an-llm-locally-with-pieces
18. PSA: Local LLM Hardware Requirements : r/homeassistant - Reddit, použito září 4, 2025, https://www.reddit.com/r/homeassistant/comments/1hovutx/psa_local_llm_hardware_requirements/
19. How do I understand requirements to run any LLM locally? : r/LocalLLM - Reddit, použito září 4, 2025, https://www.reddit.com/r/LocalLLM/comments/1hm15ox/how_do_i_understand_requirements_to_run_any_llm/
20. Stable Diffusion PC system requirements: what do you need to run it? - Digital Trends, použito září 4, 2025, https://www.digitaltrends.com/computing/stable-diffusion-pc-system-requirements/
21. Getting started with LM Studio: A Beginner's Guide - Micro Center, použito září 4, 2025, https://www.microcenter.com/site/mc-news/article/lm-studio-getting-started.aspx
22. Install an AI LLM on Your Computer: A Step-by-Step Guide - Adventures in CRE, použito září 4, 2025, https://www.adventuresincre.com/how-to-install-llm-locally/
23. Stable Diffusion Requirements: CPU, GPU & More for Running - Aiarty Image Enhancer, použito září 4, 2025, https://www.aiarty.com/stable-diffusion-guide/stable-diffusion-requirements.htm
24. Llama 3 performance and cost benchmarks | Parseur®, použito září 4, 2025, https://parseur.com/blog/blog-llama3-performance-cost
25. How to Run Your Own Free, Offline, and Totally Private AI Chatbot | PCMag, použito září 4, 2025, https://www.pcmag.com/how-to/how-to-run-your-own-chatgpt-like-llm-for-free-and-in-private
26. Set up LM Studio on Windows | GPT for Work Documentation, použito září 4, 2025, https://gptforwork.com/help/ai-models/custom-endpoints/set-up-lm-studio-on-windows
27. Download Ollama on Linux, použito září 4, 2025, https://ollama.com/download
28. Run LLMs Locally: 7 Simple Methods - DataCamp, použito září 4, 2025, https://www.datacamp.com/tutorial/run-llms-locally-tutorial
29. Get started with LM Studio | LM Studio Docs, použito září 4, 2025, https://lmstudio.ai/docs/app/basics
30. blog.n8n.io, použito září 4, 2025, https://blog.n8n.io/local-llm/#:~:text=Ollama%20is%20ideal%20for%20quickly,AI%20backend%20for%20various%20applications.
31. How to Run Open Source LLMs on Your Own Computer Using Ollama - freeCodeCamp, použito září 4, 2025, https://www.freecodecamp.org/news/how-to-run-open-source-llms-on-your-own-computer-using-ollama/
32. Learn Ollama in 15 Minutes - Run LLM Models Locally for FREE - YouTube, použito září 4, 2025, https://www.youtube.com/watch?v=UtSSMs6ObqY
33. oobabooga/text-generation-webui: The definitive Web UI for local AI, with powerful features and easy setup. - GitHub, použito září 4, 2025, https://github.com/oobabooga/text-generation-webui
34. Top 10 open source LLMs for 2025 - NetApp Instaclustr, použito září 4, 2025, https://www.instaclustr.com/education/open-source-ai/top-10-open-source-llms-for-2025/
35. How to Run Llama 3.1 8B with Ollama - GPU Mart, použito září 4, 2025, https://www.gpu-mart.com/blog/how-to-run-llama-3-1-8b-with-ollama
36. nvidia/Llama-3.1-Nemotron-70B-Instruct-HF · Suitable hardware config for usage this model, použito září 4, 2025, https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct-HF/discussions/22
37. QuantFactory/gemma-2-9b-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/QuantFactory/gemma-2-9b-GGUF
38. showvikdbz/gemma-2-9b-instruct-gguf - Hugging Face, použito září 4, 2025, https://huggingface.co/showvikdbz/gemma-2-9b-instruct-gguf
39. Mistral-7B-v0.1: Specifications and GPU VRAM Requirements - ApX Machine Learning, použito září 4, 2025, https://apxml.com/models/mistral-7b-v0-1
40. mistralai/Mistral-7B-Instruct-v0.1 - Hugging Face, použito září 4, 2025, https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

41. mradermacher/Mistral-7B-Instruct-v0.3-abliterated-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/mradermacher/Mistral-7B-Instruct-v0.3-abliterated-GGUF
42. Alibaba Cloud Model Studio:Translation capability (Qwen-MT), použito září 4, 2025, https://www.alibabacloud.com/help/en/model-studio/machine-translation
43. Qwen LLMs - - Alibaba Cloud Documentation Center, použito září 4, 2025, https://www.alibabacloud.com/help/en/model-studio/what-is-qwen-llm
44. Qwen/Qwen1.5-7B-Chat-GGUF at main - Hugging Face, použito září 4, 2025, https://huggingface.co/Qwen/Qwen1.5-7B-Chat-GGUF/tree/main
45. Qwen/Qwen1.5-7B-Chat-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/Qwen/Qwen1.5-7B-Chat-GGUF
46. Phi-3 is a family of lightweight 3B (Mini) and 14B (Medium) state-of-the-art open models by Microsoft. - Ollama, použito září 4, 2025, https://ollama.com/library/phi3
47. microsoft/Phi-3-mini-4k-instruct-gguf - Hugging Face, použito září 4, 2025, https://huggingface.co/microsoft/Phi-3-mini-4k-instruct-gguf
48. QuantFactory/Phi-3-mini-4k-instruct-GGUF - Hugging Face, použito září 4, 2025, https://huggingface.co/QuantFactory/Phi-3-mini-4k-instruct-GGUF
49. Top 10 Multimodal LLMs to Explore in 2025 - Analytics Vidhya, použito září 4, 2025, https://www.analyticsvidhya.com/blog/2025/03/top-multimodal-llms/
50. InternVL 3.5 : Best open-sourced Multi-Modal LLM | by Mehul Gupta | Data Science in Your Pocket | Aug, 2025 | Medium, použito září 4, 2025, https://medium.com/data-science-in-your-pocket/internvl-3-5-best-open-sourced-multi-modal-llm-bc929e2b6338
51. Multimodal AI: A Guide to Open-Source Vision Language Models - BentoML, použito září 4, 2025, https://www.bentoml.com/blog/multimodal-ai-a-guide-to-open-source-vision-language-models
52. Best Open Source Multimodal Vision Models in 2025 - Koyeb, použito září 4, 2025, https://www.koyeb.com/blog/best-multimodal-vision-models-in-2025
53. LLaVA: An open-source alternative to GPT-4V(ision) | by Yann-Aël Le Borgne - Medium, použito září 4, 2025, https://medium.com/data-science/llava-an-open-source-alternative-to-gpt-4v-ision-b06f88ce8efa
54. The EASIEST way to run MULTIMODAL AI Locally! (Ollama ❤️ LlaVA) - YouTube, použito září 4, 2025, https://www.youtube.com/watch?v=smvSivZApdI
55. Ollama Multimodal: EASILY setup Llava locally & Integrate API - YouTube, použito září 4, 2025, https://www.youtube.com/watch?v=CNGwscEsl0E
56. How to use llava-v1.5-13b-Q5_K_M.gguf with python : r/LocalLLaMA - Reddit, použito září 4, 2025, https://www.reddit.com/r/LocalLLaMA/comments/18qleak/how_to_use_llavav1513bq5_k_mgguf_with_python/
57. 3 Best Open-Source ASR Models Compared: Whisper, wav2vec 2.0, Kaldi - Deepgram, použito září 4, 2025, https://deepgram.com/learn/benchmarking-top-open-source-speech-models
58. openai/whisper: Robust Speech Recognition via Large-Scale Weak Supervision - GitHub, použito září 4, 2025, https://github.com/openai/whisper
59. Whisper Transcription & Dictation Web UI: Speech-to-Text Tool, použito září 4, 2025, https://whisperwebui.com/
60. What is OpenAI Whisper? - Gladia, použito září 4, 2025, https://www.gladia.io/blog/what-is-openai-whisper
61. What is OLMoASR and How Does It Compare to OpenAI's Whisper in Speech Recognition?, použito září 4, 2025, https://www.marktechpost.com/2025/09/04/what-is-olmoasr-and-how-does-it-compare-to-openais-whisper-in-speech-recognition/
62. OpenAI Whisper performance benchmarks - 1 QuBit, použito září 4, 2025, https://www.1qubit.de/en/ai/openai-whisper-performance-benchmarks
63. How To Install And Deploy Whisper, The Best Open-Source Alternative To Google Speech-To-Text - NLP Cloud, použito září 4, 2025, https://nlpcloud.com/how-to-install-and-deploy-whisper-the-best-open-source-alternative-to-google-speech-to-text.html
64. Pikurrot/whisper-gui: A simple GUI to use Whisper. - GitHub, použito září 4, 2025, https://github.com/Pikurrot/whisper-gui
65. Ognisty321/Whisper-Transcription-UI - GitHub, použito září 4, 2025, https://github.com/Ognisty321/Whisper-Transcription-UI
66. WhisperUI: Affordable Speech to Text powered by OpenAI Whisper, použito září 4, 2025, https://whisperui.com/
67. The 5 Best Open Source Speech Recognition Engines & APIs - Rev, použito září 4, 2025, https://www.rev.com/resources/the-5-best-open-source-speech-recognition-engines-apis
68. How to Use Stable Diffusion as a Beginner - Dorik, použito září 4, 2025, https://dorik.com/blog/how-to-use-stable-diffusion

69. Stable-Diffusion: How To Pick Your MODEL (In 2 Minutes!) - YouTube, použito září 4, 2025, https://www.youtube.com/watch?v=FeRKTKdHhCI
70. education.civitai.com, použito září 4, 2025, https://education.civitai.com/sdxl-1-0/#:~:text=Hardware%20Requirements,with%2016%20GB%20of%20VRAM.
71. A Crash Course on Local Image Generation with Stable Diffusion - Rob Laughter - Medium, použito září 4, 2025, https://roblaughter.medium.com/a-crash-course-on-local-image-generation-with-stable-diffusion-f72dfd3de3df
72. Tuning Local LLMs With RAG Using Ollama and Langchain - It's FOSS, použito září 4, 2025, https://itsfoss.com/local-llm-rag-ollama-langchain/
73. Getting Started with RAG (Retrieval Augmented Generation) | All Running Locally., použito září 4, 2025, https://www.youtube.com/watch?v=bGBJhkZfDSY
74. Chat with Documents | LM Studio Docs, použito září 4, 2025, https://lmstudio.ai/docs/app/basics/rag
75. Local RAG tutorials : r/LocalLLaMA - Reddit, použito září 4, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1e544gw/local_rag_tutorials/
76. Fine-Tuning LLMs: A Guide With Examples - DataCamp, použito září 4, 2025, https://www.datacamp.com/tutorial/fine-tuning-large-language-models
77. Fine Tune Large Language Model (LLM) on a Custom Dataset with QLoRA | by Suman Das, použito září 4, 2025, https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07
78. A beginners guide to fine tuning LLM using LoRA - Zohaib, použito září 4, 2025, https://zohaib.me/a-beginners-guide-to-fine-tuning-llm-using-lora/
79. How to Use Oobabooga Web UI to Run LLMs on Hyperstack: A Quick Tutorial, použito září 4, 2025, https://www.hyperstack.cloud/technical-resources/tutorials/how-to-use-oobabooga-web-ui-to-run-llms-on-hyperstack-a-quick-tutorial
80. mistral/license - Ollama, použito září 4, 2025, https://ollama.com/library/mistral:latest/blobs/43070e2d4e53
81. Qwen - Wikipedia, použito září 4, 2025, https://en.wikipedia.org/wiki/Qwen
82. Phi Open Models - Small Language Models | Microsoft Azure, použito září 4, 2025, https://azure.microsoft.com/en-us/products/phi
83. Cómo entender la licencia de Llama 3: detalles clave que debes conocer - Novita AI Blog, použito září 4, 2025, https://blogs.novita.ai/es/how-to-understand-the-llama-3-license-key-details-you-need-to-know/
84. Meta Llama 3 License, použito září 4, 2025, https://www.llama.com/llama3/license/
85. How to Understand the Llama 3 License: Key Details You Need to Know - Novita AI Blog, použito září 4, 2025, https://blogs.novita.ai/how-to-understand-the-llama-3-license-key-details-you-need-to-know/
86. meta-llama/Meta-Llama-3-8B - Hugging Face, použito září 4, 2025, https://huggingface.co/meta-llama/Meta-Llama-3-8B
87. Gemma Terms of Use | Google AI for Developers - Gemini API, použito září 4, 2025, https://ai.google.dev/gemma/terms
88. Gemma 2 – Vertex AI - Google Cloud Console, použito září 4, 2025, https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemma2?pli=1
89. command-r-plus/license - Ollama, použito září 4, 2025, https://ollama.com/library/command-r-plus:latest/blobs/f0624a2393a5
90. llama3.3 - Ollama, použito září 4, 2025, https://ollama.com/library/llama3.3
91. LLaMAX/LLaMAX3-8B - Hugging Face, použito září 4, 2025, https://huggingface.co/LLaMAX/LLaMAX3-8B
92. Google models | Generative AI on Vertex AI, použito září 4, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/models
93. Gemma 2 model card | Google AI for Developers, použito září 4, 2025, https://ai.google.dev/gemma/docs/core/model_card_2
94. kaitchup/Llama-2-7b-mt-Czech-to-English · Can you please make an English to Romanian translation model ? - Hugging Face, použito září 4, 2025, https://huggingface.co/kaitchup/Llama-2-7b-mt-Czech-to-English/discussions/1
95. Mistral 7B is not a multilingual model. - Gathnex - Medium, použito září 4, 2025, https://gathnex.medium.com/mistral-7b-is-not-a-multilingual-model-5df3a38b3cc3
96. What Are the Main Risks to LLM Security? - Check Point Software, použito září 4, 2025, https://www.checkpoint.com/cyber-hub/what-is-llm-security/llm-security-risks/
97. LLM Risk: Avoid These Large Language Model Security Failures - Cobalt, použito září 4, 2025, https://www.cobalt.io/blog/llm-failures-large-language-model-security-risks
98. Chinese Open-Weights Models: Security Myths vs. Reality | Datasaur, použito září 4, 2025, https://datasaur.ai/blog-posts/chinese-open-weights-models-security-myths-vs-reality