PŘÍLOHA · NAIVNÍ OTÁZKY · PRAGMATICKÁ PRAVIDLA

Deset *otázek* které lidé bojí ptát

Naivní otázky jsou ty nejlepší — protože pod nimi se skrývá celá komplexita problému. Tato příloha bere deset běžných firemních otázek o LLM bezpečnosti, dává jim férovou odpověď a doplňuje je deseti praktickými do/don't pravidly.

ČÁST A	Deset otázek s odpovědí "proč ano / proč ne"
ČÁST B	Deset do/don't pravidel pro denní práci s LLM
PRO KOHO	Účastníky workshopu, kteří chtějí konkrétní vodítka, ne abstraktní principy

ČÁST A

Deset otázek, kterých se lidé bojí ptát

Sebrané z firemních workshopů. U každé je férová odpověď — proč částečně ano, proč částečně ne, a co s tím prakticky.

Q01 Pokud vložím dokument do ChatGPT, může se objevit někomu jinému?

PROČ ANO

V konzumentské verzi (Free, Plus, Pro) je toggle "Improve the model" defaultně zapnutý — vaše prompty jdou do trénovacích dat. Carlini et al. (arXiv:2311.17035, 2023) ukázali, že z trénovaných modelů lze extrahovat verbatim memorizované sekvence. Navíc od května 2025 platí preservation order soudu v NYT v. OpenAI: i smazané chaty jsou drženy a 20 milionů jich už soud nařídil vydat třetí straně.

PROČ NE

Pokud používáte Team, Enterprise, Edu nebo API tier, OpenAI defaultně netrénuje. Citace: "*By default, we do not use data from ChatGPT Enterprise, Business, Edu, or our API platform for training or improving our models.*" S API + ZDR jsou data smazána do hodin a preservation order se na ně nevztahuje.

VERDIKT

Záleží na tieru. Konzumentský = rizikové, Enterprise/API+ZDR = OK.

Q02 Když dám firemní data do Claude, je to lepší než ChatGPT?

PROČ ANO

Anthropic má historicky konzervativnější data politiku a komerční tier netrénuje na uživatelských datech. API retence byla 14. září 2025 zkrácena z 30 na 7 dní — kratší než OpenAI. Konstituce Claude obsahuje explicitní privacy commitments. Nepodléhá NYT preservation order.

PROČ NE

Anthropic 8. října 2025 změnil konzumentskou politiku — toggle "You can help improve Claude" má v UI **přednastaveno na zapnuto**. Pokud to nevypnete, retence je 5 let a data jdou do tréninku. Claude.ai Free/Pro/Max je nyní v default stavu rizikový stejně jako ChatGPT Free/Plus.

VERDIKT

Komerční Claude (Team, Enterprise, API) je marginálně lepší. Konzumentský od října 2025 vyžaduje opt-out.

Q03 Můžou mít čínské modely zadní vrátka když je používám lokálně?

PROČ ANO

Cenzurní bias je dokumentovaně zapečený přímo ve vahách (R1dacted, arXiv:2505.12625). DeepSeek-R1 lokálně suprimuje Tiananmen, Tchaj-wan, Ujgury i bez síťového připojení. Sleeper Agents (Hubinger et al., 2024) ukázali, že záměrně vložené backdoory přežijí standardní safety training a nemáme dnes škálovatelnou metodu detekce. Trigger může spustit cokoliv — tichou sabotáž kódu, manipulaci výstupu, denial-of-service.

PROČ NE

Lokální deployment nemá žádný síťový kanál ven — weights jsou numerické tenzory, nemůžou iniciovat spojení. "Posílání dat domů" je technicky nemožné u offline inference. Open-weights čínský model na vlastním GPU má tedy stejný data-egress profil jako Llama nebo Mistral. Pickle exploit je riziko bez ohledu na původ modelu.

VERDIKT

Data neutečou, ale cenzurní bias zůstává. Skryté backdoory dnes neumíme detekovat — týká se to ovšem všech modelů, jen u čínských je threat model adversariálnější.

Q04 Když je to TLS šifrované, je to bezpečné, ne?

PROČ ANO

TLS 1.2/1.3 s moderními cipher suites (AES-GCM, ChaCha20-Poly1305) účinně chrání obsah před odposlechem v síti. Žádný útočník mezi vámi a serverem prompt nepřečte. Poskytovatelé navíc přidávají AES-256 šifrování at rest pro uloženou data.

PROČ NE

TLS chrání pouze *tranzit* — provozovatel inference serveru musí prompt číst v plaintextu, jinak by model nemohl odpovědět. End-to-end šifrování (kde by ani provozovatel neviděl) je u LLM služeb **nemožné z principu**. CLOUD Act, subpoeny, breaches, preservation orders — to vše obchází TLS, protože útočník nesedí v síti.

VERDIKT

TLS je nutná podmínka, ne dostatečná. Skutečné riziko je u provozovatele, ne v síti.

Q05 Lokální Llama na našem serveru je automaticky bezpečná, že ano?

PROČ ANO

Lokální deployment eliminuje cloud provider data risk — žádný CLOUD Act, subpoeny, preservation orders. Vlastní weights, vlastní GPU, plná kontrola nad logy. Pro klasifikovaná data, regulovanou financi, healthcare PHI, advokátní privilege je to často jediná compliance cesta.

PROČ NE

FuzzingLabs našel přes **270 000 internet-exposed Ollama instancí** bez autentizace. CVE-2024-37032 v Ollamě byl RCE přes path traversal. llama.cpp má 11+ security advisories 2024–2026 v GGUF parseru. Frontendy jako Open WebUI ukládají chat history neomezeně. Pickle modely z HuggingFace mohou spustit kód při načtení (JFrog: 100 malicióźních modelů, únor 2024).

VERDIKT

Lokální = bezpečnější jen pokud správně nasazeno. Bez reverse proxy s autentizací, scanu modelů a opt-out telemetrie je to jiná attack surface, ne menší.

Q06 Můžu nahrát smlouvu s klientem do ChatGPT, abych ji shrnul?

PROČ ANO

V ChatGPT Enterprise nebo Team s podepsaným DPA, EU rezidencí (od 6. února 2025) a opt-out tréninku ano — funguje to obdobně jako kdybyste použili Microsoft 365 Copilot s Office 365 backendem. Pro standardní obchodní smlouvy bez extra senzitivních klauzulí je to legitimní use case.

PROČ NE

Obsahuje-li smlouva PII protistran, M&A; nepublikované informace, osobní údaje pacientů, advokátní privilege nebo cokoliv pod NDA, pak **ne v konzumentském tieru** (preservation order, trénink, lidský review). NYT v. OpenAI ukázal, že i smazané chaty drží OpenAI pro discovery — což může porušit NDA i bez vašeho zavinění.

VERDIKT

Záleží na obsahu smlouvy a tieru. Pro citlivé klauzule jen Enterprise+ZDR+EU. Pro běžný šablonový dokument OK i Team.

Q07 Když ChatGPT řekne, že mé data nepoužije, můžu mu věřit?

PROČ ANO

OpenAI Terms jsou právně závazný dokument — Italský Garante vydal €15M pokutu za jejich porušení (prosinec 2024). EDPB Opinion 28/2024 stanoví, že provideři musí dodržet to, co slíbí. Soudní precedenty existují. Politika je vymahatelná.

PROČ NE

"Nepoužijeme" znamená "nepoužijeme k tréninku" — neznamená to "neuložíme". Abuse monitoring drží i v Enterprise tieru typicky 30 dní obsah (u Anthropic 7 dní, u Google 55 dní paid API). Court order přebije terms — preservation order v NYT případu donutil OpenAI držet i to, co by jinak smazali. Breach se může stát komukoliv (ChatGPT Redis bug, březen 2023).

VERDIKT

Můžete věřit, že netrénují. Nemůžete věřit, že vaše data jsou v daném okamžiku trvale smazaná — soud, breach, abuse review jsou výjimky z policy.

Q08 Naše firma má vlastní GPT v Azure — to je jiný level bezpečnosti?

PROČ ANO

Azure OpenAI Service je technicky odlišný produkt než openai.com — běží v Microsoft Azure tenantech, podléhá Microsoft DPA (ne OpenAI Terms), nesdílí preservation order z NYT případu (Microsoft jako separátní entita). EU rezidence (Frankfurt, Amsterdam) je standardní. Pro EU subjekt často **nejlepší volba** mezi cloudovými GPT službami.

PROČ NE

Stále Microsoft = US provider = CLOUD Act exposure. Microsoft také indexuje obsah pro abuse monitoring (default 30 dní), pokud nezažádáte o "modified abuse monitoring" (vyžaduje schválení). Schrems II / EU-US DPF kontroverze platí. Microsoft Copilot měl EchoLeak (CVE-2025-32711, červen 2025) — zero-click prompt injection s CVSS 9.3.

VERDIKT

Lepší než openai.com pro EU enterprise, ale ne fundamentálně bezpečnější. Stále US cloud, stále agentic risks.

Q09 AI nahradí naše bezpečnostní oddělení, takže si můžeme šetřit, ne?

PROČ ANO

AI dokáže škálovat security analytics — log review, anomaly detection, SOC tier-1 triage. Microsoft Security Copilot, CrowdStrike Charlotte, Google Sec-Gemini Workbench přinášejí reálnou efektivitu. Pro obrovské objemy logů je AI rychlejší než člověk a méně chybuje v rutinních úlohách.

PROČ NE

AI samo přidává **novou attack surface**: prompt injection, sleeper agents, supply chain útoky na modely, indirect injection přes RAG. EchoLeak, Replit (smazal produkční DB v code freeze), Microsoft Copilot ASCII smuggling — všechno z roku 2024–2025. Tradiční firewall, EDR a DLP tyto vektory nepokrývají. Bezpečnost AI vyžaduje **specializované lidi**, ne méně lidí.

VERDIKT

AI rozšíří kapacitu týmu, ale tým potřebujete víc, ne méně. Threat model se rozšířil o agentic risks.

Q10 Když nepoužíváme AI, nemusíme řešit AI bezpečnost, že?

PROČ ANO

Jisté úspory v compliance overhead — nemusíte řešit EU AI Act klasifikaci jako provider, GPAI Code of Practice, ISO 42001. Pro malou firmu mimo regulovaná odvětví je "počkáme rok dva" legitimní strategie.

PROČ NE

Vaši zaměstnanci AI **používají bez ohledu na firemní politiku** — přes osobní účty na konzumentském tieru. Samsung 2023 měl zákaz, ale tři incidenty se přesto staly během 20 dní. Vaši protistrany, dodavatelé a klienti AI používají na vás — vaše data tečou skrz. EU AI Act se vás dotkne jako *deployer*, i když nejste provider. Bez politiky budete v horší pozici za rok než s pomalou ale řízenou adopcí dnes.

VERDIKT

Strategie "shadow AI" je realita každé firmy bez politiky. Lepší řízená adopce s tréninkem než zákaz, který se obchází.

ČÁST B

Deset do/don't pravidel

Praktické vodítko pro denní práci s LLM. Levý sloupec dělejte. Pravý sloupec nedělejte. Nic mezi tím není.

01. Tier matters

DO

Pro firemní data používej Team, Enterprise nebo API tier se ZDR. Plať za to — je to levnější než jeden incident.

DON'T

Nehraj se na to, že "Pro tier je už business" — Pro u všech tří hráčů je konzumentský tier s tréninkem.

02. Opt-out kontrolovat

DO

Po každé update terms zkontroluj, jestli toggle "Improve the model" zůstal vypnutý. Anthropic ho 8. 10. 2025 předzaškrtl ZAP.

DON'T

Nepředpokládej, že toggle, který jsi vypnul před rokem, je stále vypnutý. Politiky se mění, defaulty se mění.

03. Citlivá data filtrovat

DO

Před vložením do LLM odstraň jména klientů, čísla smluv, PII, IP adresy interních systémů. Použij PII redaction tool nebo regex.

DON'T

Neházej do LLM raw export z CRM, e-mailovou korespondenci s klientem, ani přílohy z due diligence. Nikdy.

04. Lokální deployment chránit

DO

Ollama dej za reverse proxy s autentizací. Nikdy OLLAMA_HOST=0.0.0.0 bez auth. Skenuj modely fickling/picklescan.

DON'T

Nepouštěj llama-server na public IP. Nestahuj pickle (.bin, .pt) modely z neznámých autorů. Preferuj safetensors a GGUF.

05. Agentům dávat málo

DO

Princip nejmenších oprávnění. Agent má read-only přístup pokud nepotřebuje psát. Human-in-the-loop pro destruktivní akce.

DON'T

Nedávej agentovi root přístup k DB, ani když "to bude jen na chvíli". Replit smazal produkční DB v code freeze za jediný den.

06. Cizí obsah validovat

DO

Pokud LLM čte e-mail, web, PDF od třetí strany, předpokládej že to obsahuje prompt injection. Sandboxuj. Allowlist linků v outputu.

DON'T

Neumožňuj agentovi auto-render markdown image z neznámých zdrojů. EchoLeak v M365 Copilot tímto exfiltroval data nulovým klikem.

07. Logy retence omezit

DO

V API nastav data retention na minimum (Anthropic 7 dní, OpenAI 30 dní default, ZDR pokud máš enterprise kontrakt).

DON'T

Nenechávej výchozí "neomezeně" v Open WebUI nebo AnythingLLM. Lokální chat history je breach risk jako každá DB.

08. EU rezidenci řešit

DO

Pro GDPR data preferuj Bedrock Frankfurt, Vertex Belgie, OpenAI EU regiony (od 6. 2. 2025). Schrems II tě jinak může v auditu bolet.

DON'T

Nepřesouvej PII přes US-only endpoints. CLOUD Act platí i když "máte adekvátnost" — DPF přežil první výzvu, ale další jsou v cestě.

09. Čínské modely rozlišovat

DO

Open-weights na vlastním GPU = OK pro hobby, kód, brainstorming. Dokumentuj cenzurní bias pro každý use case.

DON'T

Necpaj firemní data přes deepseek.com, kimi.com, qwen.aliyun.com. NIL čl. 7 + DSL čl. 36 = de facto nesplnitelný GDPR transfer.

10. Lidi školit místo zakazovat

DO

Měj jasnou politiku, jaké tiery jsou OK pro jaká data. Trénuj lidi na konkrétních scénářích. Audit shadow AI jednou za kvartál.

DON'T

Nezakazuj plošně — vede to ke shadow AI přes osobní účty (viz Samsung 2023). Politika musí být použitelná, jinak ji nikdo nepoužije.

*Pravidla nejsou náhrada za myšlení — jsou jeho startovní bod. Pokaždé, když si nejste jistí, vraťte se k matrici **tier × citlivost dat × agentic úroveň** z hlavní publikace. Bezpečnost LLM v polovině 2026 není o paranoii ani o slepé důvěře. Je o tom, vědět, co provider opravdu dělá s vašimi daty, a podle toho volit konfiguraci.*
