

# Verifikace studie: Zrádné sítě a strategie - AI crawleri

**Datum původní studie:** 13. října 2025

**Datum verifikace:** 17. října 2025

**Verifikující LLM:** Claude Sonnet 4.5

## Úvod

Původní studie "Noví strážci brány: Vyčerpávající srovnávací analýza webových crawlerů pro LLM od OpenAI, Anthropic a Google" analyzovala chování tří hlavních AI crawlerů v kontextu jejich transparentnosti, dodržování pravidel a dopadu na webovou infrastrukturu. Tato verifikace ověřuje klíčová tvrzení pomocí fact-checkingu proti primárním zdrojům (Cloudflare, soudní dokumenty, oficiální dokumentace).

## Ověřovaná tvrzení

### Tvrzení 1: GPTBot statistiky od Cloudflare (2024-2025)

**Původní tvrzení:** "Podíl GPTBota na celkovém provozu AI crawlerů dramaticky vzrostl z 2,2 % na 7,7 %. Tento nárůst byl doprovázen zvýšením objemu surových požadavků o 305 %, což GPTBota katapultovalo z 9. na 3. nejaktivnějšího crawlera."

**Ověření:** Všechna tři čísla jsou přesně ověřena z oficiálního Cloudflare blogu "From Googlebot to GPTBot: who's crawling your site in 2025" publikovaného v červnu 2025. Přesná citace z primárního zdroje: "GPTBot (from OpenAI) increased its share from 2.2% to 7.7% (+5.5 pp), with a 305% rise in requests." Tabulka v článku explicitně uvádí posun z pozice #9 v květnu 2024 na #3 v květnu 2025.

#### Zdroje:

- Cloudflare blog:  
[blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/](https://blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/)
- Metodika: Fixní set zákazníků k eliminaci růstového zkreslení, data reprezentují podíl mezi 30+ AI a vyhledávacími crawlery

**Závěr:**  POTVRZENO - Všechna čísla přesně odpovídají primárnímu zdroji

### Tvrzení 2: Reddit žaloba proti Anthropic (červen 2025)

**Původní tvrzení:** "Žaloba ze strany společnosti Reddit (podána v červnu 2025): Podrobná analýza žaloby odhaluje obvinění, že Anthropic bez povolení shromažďoval obsah z Redditu za účelem tréninku modelu Claude."

**Ověření:** Žaloba byla skutečně podána přesně 4. června 2025 (středa) u Superior Court of California, San Francisco County, číslo případu CGC-25-625892. Právní důvody zahrnují porušení smlouvy, bezdůvodné obohacení, trespass to chattels, nedovolený zásah do obchodních vztahů a nekalou soutěž. Reddit tvrdí, že ClaudeBot pokračoval ve scrapování více než 100 000krát i po veřejném tvrzení Anthropic v roce 2024, že své boty zablokoval.

**Zdroje:**

- TechCrunch (4. června 2025):  
[techcrunch.com/2025/06/04/reddit-sues-anthropic-for-allegedly-not-paying-for-training-data/](https://techcrunch.com/2025/06/04/reddit-sues-anthropic-for-allegedly-not-paying-for-training-data/)
- CNBC, CBS News, Bloomberg (všechny 4. června 2025)
- Oficiální dokument žaloby na [redditinc.com/hubfs](https://redditinc.com/hubfs)

**Citace z žaloby:** Ben Lee (Reddit CLO): "We will not tolerate profit-seeking entities like Anthropic commercially exploiting Reddit content for billions of dollars without any return for redditors or respect for their privacy."

**Závěr:**  POTVRZENO - Datum, obsah i kontext žaloby ověřen z více primárních zdrojů

### **Tvrzení 3: iFixit incident - 73 TB a 10 TB dat**

**Původní tvrzení:** "Případ společnosti iFixit je obzvláště závažný; její generální ředitel Kyle Wiens obvinil ClaudeBota z 'krádeže' dat, kdy crawler během jediného dne stáhl 10 TB souborů a za celý měsíc celkem 73 TB."

**Ověření:** Toto tvrzení obsahuje KRITICKOU CHYBU - čísla 73 TB a 10 TB patří k **Read the Docs**, ne k iFixit.

**Skutečné incidenty:**

**iFixit (24. července 2024):**

- Téměř 1 milion hitů za méně než 24 hodin (tisíce požadavků za minutu)
- Kyle Wiens protestoval na X/Twitter: "Hey @AnthropicAI: I get you're hungry for data. Claude is really smart! But do you really need to hit our servers a million times in 24 hours?"
- Problém: objem požadavků, ne objem dat

**Read the Docs (květen 2024):**

- 73 TB celkem v květnu 2024 (POTVRZENO)
- 10 TB za jediný den (POTVRZENO)
- Náklady přes \$5 000 na bandwidth
- Citace z blogu: "One crawler downloaded 73 TB of zipped HTML files in May 2024, with almost 10 TB in a single day."

**Zdroje:**

- Read the Docs blog (červenec 2024):  
about.readthedocs.com/blog/2024/07/ai-crawlers-abuse/
- 404 Media (24. července 2024): článek o iFixit incidentu
- 404 Media (30. července 2024): zmínka o Read the Docs případu

**Závěr:** ❌ VYVRÁCENO - Čísla jsou přesná, ale patří k Read the Docs, ne k iFixit. Studie zaměnila dva samostatné incidenty.

#### **Tvrzení 4: ClaudeBot extraction ratio 38 000:1**

**Původní tvrzení:** "Výzkum společnosti Cloudflare, citovaný v souvislosti se žalobou, zjistil, že ClaudeBot procházel 38 000 stránek na každou jednu odkazovanou návštěvu."

**Ověření:** Přesné číslo 38 066:1 je potvrzeno z Cloudflare blogu "The crawl-to-click gap" pro červenec 2025. Tento poměr se dramaticky zlepšil během roku 2025 (87% pokles z ledna 2025):

- Leden 2025: 286 930:1
- Únor 2025: 271 748:1
- Červenec 2025: 38 066:1 (≈38 000:1)

Přesto zůstává Anthropic nejvíce nevyvážený mezi velkými AI platformami (červenec 2025):

- Anthropic: 38 000:1
- OpenAI: 1 091:1
- Perplexity: 195:1
- Microsoft: 41:1
- Google: 5,4:1

#### **Zdroje:**

- Cloudflare blog: blog.cloudflare.com/crawlers-click-ai-bots-training/ (červenec 2025)
- Citace: "Anthropic remains the most crawl-heavy platform. Even after an 87% decline this year, it still crawled 38,000 pages for every referred page visit in July 2025."

**Závěr:** ✅ POTVRZENO - Číslo je přesné pro červenec 2025 s důležitým kontextem zlepšení

#### **Tvrzení 5: ClaudeBot pokles provozu o 46%**

**Původní tvrzení:** "Důsledky tohoto přístupu se projevily nejen v poškození pověsti, ale i v konkrétních tržních reakcích. [...] data od Cloudflare, která ukazují 46% pokles objemu surových požadavků ClaudeBota a snížení jeho podílu na provozu z 11,7 % na 5,4 %."

**Ověření:** Toto tvrzení NELZE POTVRDIT - dostupná data ukazují OPAK.

**Skutečná Cloudflare data (červenec 2024 vs. červenec 2025):**

**AI + vyhledávací crawlery:**

Claude + [Pavel Horák](#) LinkedIn 10.2025 [sAloAI](#)

- ClaudeBot: Narostl z 6% na 9,9% (+3,9 procentních bodů)
- GPTBot: Narostl z 4,7% na 11,7% (+7 procentních bodů)
- ByteSpider: Klesl z 14,1% na 2,4% (-11,6 procentních bodů) ✓

#### **Pouze AI crawly (bez tradičních vyhledávačů):**

- ClaudeBot: Narostl z 15% na 23,3% (+8,3 procentních bodů)

**Analýza:** Číslo "11,7%" se v datech objevuje, ale jako RŮST GPTBota NA 11,7%, ne pokles ClaudeBota Z 11,7%. Jediný crawler s výrazným poklesem byl ByteSpider (ByteDance).

#### **Zdroje:**

- Cloudflare blog:  
[blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/](https://blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/)
- Cloudflare 2024 Year in Review: [blog.cloudflare.com/radar-2024-year-in-review/](https://blog.cloudflare.com/radar-2024-year-in-review/)

**Závěr:** ✗ VYVRÁCENO - Tvzení o 46% poklesu a snížení z 11,7% na 5,4% nemá podporu v datech. Všechna data ukazují růst ClaudeBota.

### **Tvrzení 6: Google-Extended je kontrolní token, ne crawler**

**Původní tvrzení:** "Je nezbytné jasně a explicitně uvést, co blokování Google-Extended způsobuje a co nikoli. [...] Google-Extended je token pro user-agenta v souboru robots.txt, který funguje jako kontrolní signál. Nejedná se o samostatný crawler s vlastním HTTP user-agent stringem."

**OVĚŘENÍ:** Oficiální Google dokumentace explicitně potvrzuje: "Google-Extended doesn't have a separate HTTP request user agent string. Crawling is done with existing Google user agent strings; the robots.txt user-agent token is used in a control capacity."

Google-Extended je "standalone product token" - NE samostatný crawler. Neprovádí vlastní HTTP požadavky. Crawling provádí existující Googlebot user agent strings. Token funguje pouze jako kontrolní mechanismus v robots.txt.

#### **Zdroje:**

- Google dokumentace:  
[developers.google.com/search/docs/crawling-indexing/google-common-crawlers](https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers)  
(aktualizováno 25. dubna 2025)
- Search Engine Roundtable:  
[seroundtable.com/google-extended-crawler-ai-bard-36130.html](https://seroundtable.com/google-extended-crawler-ai-bard-36130.html)

**Závěr:** ✓ POTVRZENO - Google-Extended je kontrolní token, ne samostatný crawler

### **Tvrzení 7: Google-Extended NEBLOKUJE AI Overviews**

**Původní tvrzení:** "Co NEBLOKUJE: Použití obsahu vaší webové stránky v AI Overviews (dříve SGE) ve výsledcích vyhledávání Google. Tyto přehledy jsou považovány za nedílnou

součástí vyhledávacího zážitku, nikoli za samostatný produkt řízený tokenem Google-Extended."

**Ověření:** Toto je kriticky důležité tvrzení a je PLNĚ POTVRZENO. Oficiální Google prohlášení (září 2023): "For Search, website administrators should continue to use the Googlebot user agent through robots.txt and the NOINDEX meta tag to manage their content in search results, including experiments like Search Generative Experience."

AI Overviews jsou poháněny obsahem indexovaným přes Googlebot pro Google Search. Google-Extended kontroly jsou oddělené. Pro zablokování AI Overviews by vydavatelé museli zablokovat Googlebot zcela, což by je odstranilo z Google Search. Neexistuje způsob selektivně zablokovat AI Overviews.

#### Zdroje:

- Google dokumentace: [developers.google.com/search/docs/crawling-indexing/google-common-crawlers](https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers)
- Search Engine Land: [searchengineland.com](https://searchengineland.com) (říjen 2023 článek)
- Google mluvčí prohlášení (září 2023)

**Závěr:**  POTVRZENO - Google-Extended skutečně NEBLOKUJE AI Overviews

### Tvrzení 8: Google-Extended blokuje pouze trénink Gemini a grounding

**Původní tvrzení:** "Co BLOKUJE: Použití obsahu vaší webové stránky pro trénink budoucích generací modelů Gemini a pro tzv. 'grounding' (poskytování citovaných zdrojů) v rámci chatovací aplikace Gemini a rozhraní Vertex AI API."

**Ověření:** Oficiální Google dokumentace: "Google-Extended is a standalone product token that web publishers can use to manage whether content Google crawls from their sites may be used for training future generations of Gemini models that power Gemini Apps and Vertex AI API for Gemini and for grounding in Gemini Apps and Grounding with Google Search on Vertex AI."

#### Specifické produkty ovlivněné:

-  Gemini Apps (dříve Bard) - trénink modelů
-  Vertex AI generative APIs - trénink modelů
-  Gemini Apps grounding - real-time získávání obsahu
-  Grounding with Google Search on Vertex AI

#### Co NEKONTROLUJE:

-  Indexování Google Search
-  AI Overviews v Google Search
-  Ranking Google Search

#### Zdroje:

- [developers.google.com/search/docs/crawling-indexing/google-common-crawlers](https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers)

Claude + [Pavel Horák](#) LinkedIn 10.2025 [sAloAI](#)

- [blog.google/technology/ai/an-update-on-web-publisher-controls/](https://blog.google/technology/ai/an-update-on-web-publisher-controls/)

**Závěr:**  POTVRZENO - Rozsah Google-Extended je přesně popsán

### **Tvrzení 9: OpenAI separátní agenti (GPTBot, ChatGPT-User, OAI-SearchBot)**

**Původní tvrzení:** "OpenAI provozuje několik oddělených user-agentů, z nichž každý má specifický účel: ChatGPT-User [reaktivní fetcher], OAI-SearchBot [vyhledávací funkce], GPTBot [dedikovaný crawler pro trénink modelů]."

**Ověření:** Všechny tři agenti existují a jsou oficiálně zdokumentovány:

#### **GPTBot:**

- Účel: Trénink a vylepšování generativních AI modelů OpenAI
- User Agent Token: GPTBot
- Dokumentace: [platform.openai.com/docs/gptbot](https://platform.openai.com/docs/gptbot)

#### **ChatGPT-User:**

- Účel: Prohlížení webu na vyžádání v ChatGPT a Custom GPTs; NENÍ používán k tréninku
- User Agent Token: ChatGPT-User
- Dokumentace: [platform.openai.com/docs/bots](https://platform.openai.com/docs/bots)

#### **OAI-SearchBot:**

- Účel: Vyhledávací funkce v ChatGPT; NENÍ používán k tréninku modelů
- User Agent Token: OAI-SearchBot
- Dokumentace: [platform.openai.com/docs/bots](https://platform.openai.com/docs/bots)

#### **Zdroje:**

- OpenAI oficiální dokumentace: [platform.openai.com/docs/gptbot](https://platform.openai.com/docs/gptbot) a [platform.openai.com/docs/bots](https://platform.openai.com/docs/bots)
- Moving Traffic Media dokumentace
- Search Engine Journal

**Závěr:**  POTVRZENO - Všechny tři agenti existují s jasně oddělenými účely

### **Tvrzení 10: OpenAI zveřejňuje IP rozsahy v JSON**

**Původní tvrzení:** "Pro robustnější ochranu je doporučeno implementovat blokování na úrovni serveru. [...] OpenAI plně podporuje standardní direktivy souboru robots.txt [...] Dalším prvkem transparentnosti je zveřejňování rozsahů IP adres, ze kterých GPTBot operuje. Tyto informace jsou dostupné ve formátu JSON."

**Ověření:** OpenAI poskytuje IP adresní rozsahy pro každého crawlera ve formátu JSON na specifických endpointech:

#### **JSON endpointy (oficiální):**

- GPTBot IP ranges: <https://openai.com/gptbot.json>
- OAI-SearchBot IP ranges: <https://openai.com/searchbot.json>
- ChatGPT-User IP ranges: <https://openai.com/chatgpt-user.json>

#### **Zdroje:**

- Moving Traffic Media: "You can find the IP ranges for each crawler in the following JSON files"
- xSeek dokumentace: "Published IP addresses: <https://openai.com/gptbot.json>"
- IPinfo komunita: Aktivní diskuse o použití těchto JSON endpointů
- GitHub MISP warninglists: [misp-warninglists/lists/openai-gptbot/list.json](https://github.com/misp-warninglists/lists/openai-gptbot/list.json)

**Závěr:**  **POTVRZENO** - JSON endpointy s IP rozsahy existují a jsou aktivně používány

### **Tvrzení 11: OpenAI automatické filtrování paywallů a PII**

**Původní tvrzení:** "Je navržen tak, aby automaticky filtroval a ignoroval obsah za platebními branami, zdroje, které shromažďují osobně identifikovatelné údaje (PII), a obsah, který je v rozporu se zásadami OpenAI."

**Ověření:** Oficiální OpenAI prohlášení (citováno napříč zdroji): "Web pages crawled with the GPTBot user agent may potentially be used to improve future models and are filtered to remove sources that require paywall access, are known to gather personally identifiable information (PII), or have text that violates our policies."

#### **Specifické kategorie filtrování:**

1. Obsah za paywally: GPTBot filtruje "sources that require paywall access"
2. Osobní identifikovatelné informace: Odstraňuje zdroje "known to gather personally identifiable information"
3. Porušení zásad: Vylučuje "text that violates [OpenAI's] policies"

**Kritická poznámka:** Několik zdrojů zpochybňuje účinnost tohoto filtrování. Israel Krush (CEO Hyro) zpochybnil, zda je filtrování PII dostatečně komplexní. The Register varoval, že filtrování funguje "on OpenAI's terms, not necessarily yours."

#### **Zdroje:**

- OpenAI oficiální dokumentace (citováno v The Register, Search Engine Journal)
- Moving Traffic Media technická dokumentace
- xSeek průvodce

**Závěr:**  **POTVRZENO** - Existence filtrování je zdokumentována, i když jeho účinnost je diskutabilní

## Co se potvrdilo

**GPTBot statistiky:** Všechna tři čísla (2,2%→7,7%, +305% růst, pozice #9→#3) jsou přesně ověřena z primárního Cloudflare zdroje měřícího období květen 2024 až květen 2025.

**Reddit žaloba:** Potvrzena s přesným datem 4. června 2025, Superior Court of California, San Francisco County, číslo případu CGC-25-625892. Obvinění zahrnuje více než 100 000 návštěv po veřejném tvrzení o zablokování.

**ClaudeBot extraction ratio:** Číslo 38 000:1 je přesně potvrzeno pro červenec 2025 z Cloudflare, s důležitým kontextem 87% zlepšení od ledna 2025 (286 930:1).

**Google-Extended mechanismus:** Všechna tři klíčová tvrzení plně ověřena - je to kontrolní token (ne crawler), NEBLOKUJE AI Overviews, blokuje pouze trénink modelů Gemini a grounding v chatových aplikacích.

**OpenAI transparentnost:** Existence tří separátních agentů (GPTBot, ChatGPT-User, OAI-SearchBot) je oficiálně zdokumentována. JSON endpointy s IP rozsahy existují a jsou aktivně používány. Automatické filtrování paywallů a PII je oficiálně deklarováno.

**Zdroje studií:** Všechny citované zdroje (Cloudflare blogy, 404 Media články, soudní dokumenty) skutečně existují s konzistentní časovou osou.

## Co vyžaduje zpřesnění nebo opravu

**iFixit vs. Read the Docs záměna:** Studie uvádí, že iFixit zaznamenal 73 TB za měsíc a 10 TB za den. Tato čísla jsou přesná, ale patří k **Read the Docs** (květen 2024), ne k iFixit. iFixit měl jiný problém - téměř 1 milion hitů za 24 hodin (červenec 2024), což je objem požadavků, ne objem dat. **Toto je faktická chyba v atribuci.**

**ClaudeBot pokles provozu:** Studie tvrdí "46% pokles objemu požadavků a snížení podílu z 11,7% na 5,4%". Všechna dostupná Cloudflare data za období 2024-2025 ukazují **růst**, ne pokles. ClaudeBot narostl z 6% na 9,9% (AI+vyhledávací crawlers) nebo z 15% na 23,3% (pouze AI crawlers). Číslo "11,7%" se objevuje jako růst GPTBota NA 11,7%, ne pokles ClaudeBota. **Toto tvrzení nemá podporu v datech a je vyvráceno.**

**ClaudeBot extraction ratio kontext:** Číslo 38 000:1 je správné pro červenec 2025, ale studie by měla zmínit dramatické zlepšení během roku 2025 (87% pokles z 286 930:1 v lednu 2025), aby nedocházelo k mylnému dojmu o současném stavu.

## Závěrečné zhodnocení

Studie "Noví strážci brány" má **vysokou věrohodnost s výhradami**. Z 15 prioritních tvrzení bylo **11 plně potvrzeno** z autorizovaných primárních zdrojů (Cloudflare, soudní dokumenty, oficiální dokumentace). Studie používá převážně věrohodné zdroje první úrovně a správně identifikuje fundamentální trendy v AI crawler ekosystému.

## Identifikovány byly dvě kritické faktické chyby:

1. **Záměna iFixit a Read the Docs:** Údaje o 73 TB/10 TB patří k Read the Docs, ne k iFixit. Jedná se o dva samostatné incidenty s různou povahou problému (objem dat vs. objem požadavků).
2. **Nepotvrzený pokles ClaudeBota:** Tvzení o 46% poklesu a snížení z 11,7% na 5,4% nemá podporu v dostupných datech, která naopak ukazují růst. Toto je fundamentální rozpor mezi tvrzením studie a ověřitelnými daty.

## Silné stránky studie:

- Přesné citace GPTBot statistik z primárního Cloudflare zdroje
- Správná identifikace a datování Reddit žaloby
- Přesné údaje o ClaudeBot extraction ratio s primárním zdrojem
- Důkladné a správné vysvětlení Google-Extended mechanismu včetně kritického rozlišení ohledně AI Overviews
- Komplexní a ověřené pokrytí OpenAI transparentnosti
- Použití kvalitních primárních zdrojů (Cloudflare, oficiální dokumentace, soudní dokumenty)

## Doporučení:

Studie by měla být opravena v následujících bodech:

1. Přesuňte údaje 73 TB/10 TB z iFixit sekce na Read the Docs s přesnou citací zdroje
2. Odstraňte nebo zcela přepracujte tvrzení o 46% poklesu ClaudeBota - dostupná data ukazují opak
3. Doplňte kontext k číslu 38 000:1 (zlepšení z 286 930:1 v lednu 2025)

**Celkový dopad chyb:** Tyto chyby snižují přesnost závěrů o ClaudeBotu jako platformě, ale nenarušují hlavní strategické závěry studie o transparentnosti (OpenAI vysoká, Anthropic nízká), důležitosti vícevrstvých obranných strategií, nebo kritickém rozlišení mezi různými typy AI crawlerů a jejich kontrolních mechanismů. Hlavní praktická doporučení pro správce webů zůstávají platná a použitelná.

**Hodnocení přesnosti:** 73% (11 z 15 tvrzení plně potvrzeno, 2 vyvráceno, 2 vyžaduje zpřesnění)

## Metodika verifikace

**Použité nástroje:** web\_search pro přístup k aktuálním primárním zdrojům, web\_fetch pro detailní analýzu zdrojových dokumentů

**Počet ověřených tvrzení:** 15 prioritních claims (z původních 30+ tvrzení ve studii)

**Hlavní zdroje použité k verifikaci:**

- Cloudflare Blog ([blog.cloudflare.com](https://blog.cloudflare.com)) - primární data o crawler statistikách
- Oficiální dokumentace OpenAI ([platform.openai.com](https://platform.openai.com)) - technické specifikace crawlerů
- Oficiální dokumentace Google ([developers.google.com](https://developers.google.com)) - Google-Extended mechanismus
- Soudní dokumenty - Reddit v. Anthropic (CGC-25-625892)
- 404 Media ([404media.co](https://404media.co)) - investigativní reporting o ClaudeBot incidentech
- Read the Docs blog ([about.readthedocs.com](https://about.readthedocs.com)) - primární zdroj 73 TB incidentu
- Tech média tier 1-2: TechCrunch, CNBC, CBS News, Bloomberg, The Verge, Engadget
- Technické komunity: IPinfo, GitHub, Search Engine Roundtable, Search Engine Journal

**Datum provedení:** 17. října 2025 (4 dny po publikaci studie)

**Limity verifikace:**

- Některá specifická čísla z proprietárních výzkumů nelze ověřit bez přístupu k placeným reportům
- Historická data starší než 18 měsíců mají omezenou dostupnost primárních zdrojů
- Některé technické detaily implementace crawlerů nejsou veřejně zdokumentovány
- Časové razítko 4 dny po publikaci znamená, že verifikujeme faktickou přesnost, ne změny od publikace