VIBECODING TALKS

CO SE STALO OD ZÁŘÍ

Nové modely, kdo je kde a co přibylo v Claude Code

Patrick Zandl · listopad 2026

HRA

HOT OR NOT MODEL

Zahrajte si s námi populární hru o nejvíc HOT MODEL (AI). Vyberte ze dvojice, který model je vám sympatičtější. Devět kol, pak výsledek!



vibecoding.cz/duel

Osm týdnů, tři velké laboratoře

5

OpenAI

GPT-6 Astra, Sol a Luna, GPT-6.1 Sol a Terra

5

Anthropic

Fable 5.1, Mythos 5.1, Opus 5.5, Sonnet 5.5, Haiku 5.5

3

Google

Gemini 3.8 Flash, Gemini 4 Argon, Gemini 4 Carbon

K tomu desítky dalších: jen v září 42 nových modelů od 27 firem.

1. září až 28. října 2026 · Zdroj: BenchLM, No Code MBA release tracker

Rodina GPT-6: čtyři modely, čtyři role

Astra

10/50 USD · AA 57

Nejsilnější a nejdražší. Věda, nejtěžší úlohy, vede ARC-AGI-2. Útočnou kyberbezpečnost odmítá.

6.1 Sol

2/10 USD · AA 55

Pracovní kůň na programování. Výchozí v Codexu, blízko Astře za pětinu ceny.

6.1 Terra

1/5 USD · AA 51

Ovládání počítače a kancelář: formuláře, výkresy, faktury. Třikrát rychlejší než Sol.

Luna

0,10/0,50 USD · AA 28

Levný model na objem: třídění, krátká shrnutí, vytahování údajů.

Ceny za 1M tokenů vstup/výstup · AA = Artificial Analysis Intelligence Index v5

Čtyři úrovně modelů

Úroveň	OpenAI	Anthropic	Google	Otevřené modely
1 · Špička věda, kyber, nejtěžší úlohy	**GPT-6 Astra** 10/50 · **57**	**Opus 5.5** 4/20 · **58**	**Gemini 4 Carbon** neveřejný · **59**	žádný
2 · Pracovní agentní kódování, dlouhé úlohy	**GPT-6.1 Sol** 2/10 · **55**	**Sonnet 5.5** 2/10 · **54**	**Gemini 4 Argon** 2/10, od 1. 12. 4/20 · **50**	**Kimi K3.1 Pro** 2/10 · **46**
3 · Střední ovládání počítače, kancelář	**GPT-6.1 Terra** 1/5 · **51**	žádný	žádný	žádný
4 · Levná třídění, extrakce, subagenti	**GPT-6 Luna** 0,10/0,50 · **28**	**Haiku 5.5** 0,10/0,50 · **36**	**Gemini 3.8 Flash** 0,75/3,75 · **31**	**DeepSeek V4.1 Flash** 0,30/1,20 · **33** **GLM-5.3** 1,40/4,40 · **32** **GLM-5.3-Flash** 0,15/0,50 · **28** **Mistral Large 4** 0,70/2,10 · **26** **Muse Spark Open** 1,25/4,25 · **25**

Ceny USD/1M tokenů vstup/výstup · tučně AA Index v5 · Fable 5.1 (49) a Mythos 5.1 jsou specializované

Co z matice plyne

Opus 5.5 za 40 % ceny Astry

Špička Anthropicu má víc bodů než špička OpenAI a stojí 4/20 místo 10/50 USD za milion tokenů.

Střední úroveň je prázdná

Pod 2/10 USD vede cesta rovnou k modelům s poloviční výkonem. Mezeru zaplnil jen OpenAI s Terrou.

Otevřené modely o úroveň níž

Nejlepší, Kimi K3.1 Pro, je 12 bodů za Opusem. Ostatní jsou na úrovni Haiku a Luny.

Google zdražuje

Argon od 1. prosince za cenu Opusu s výkonem o 8 bodů nižším. Nejlepší Carbon zatím nekoupíte.

PŘÍPADOVÁ STUDIE

TICHÁ POŠTA S TOKENY

Revolut, čínská AI a co internet udělal z patnáctiminutové přednášky

TITULKY, 3. A 4. LISTOPADU

„Britský fintech Revolut snížil náklady na AI kódování desetkrát: vyřadil Claude a hostuje si model sám“

The Register

„Revolut převedl zhruba 2 000 svých vývojářů z Claude Code a Codexu na čínský otevřený model, který si hostuje sám.“

Trending Topics, perex

Zvedněte ruku: kdo si myslí, že Revolut přestal používat Claude?

Co David Abitbol skutečně řekl

13. 10.
AI Engineer NYC, hlavní sál

11 minut
z toho 2,5 minuty o náhradě modelů, nejvíc o ladění GPU

11 sekund
zabrala čísla 60 % a 70 %

„Codex a Claude Code samozřejmě dál používáme. Ale jen jako výjimku.“

přepis, 5:40:59

„Úzkým hrdlem vývoje nemusí být samotné psaní kódu.“

přepis, 5:40:23

„Long context is a scam.“

U kódování stačí 256 tisíc tokenů kontextu místo milionu.

přepis, 5:44:01

KDO TO ŘEKL

David Abitbol

vede AI operations v Revolutu

80 mil. zákazníků ve 45 zemích

2 000 vývojářů

56 × B200 GPU pro vlastní GLM

“Shipping like a startup, regulated like a bank”
AI Engineer NYC, 13. října 2026



PŘEKROUCENÍ 1/3

„Zahodili Claude a Codex“

Fakta: Změnil se výchozí nástroj. Frontier modely zůstaly na nejtěžší práci.

Výchozí

Samohostovaný GLM od dubna 2026

Výjimka

Claude Code a Codex na těžké úlohy

Obrázky

Gemini nebo model OpenAI, GLM je jen textový

Zdroj: přepis přednášky, 5:40:53 a 5:45:20

PŘEKROUCENÍ 2/3

„GLM je lepší než Claude“

Srovnávají se dva různé režimy přidělování tokenů.

Claude Code, leden až duben 2026

Přísné rozpočty na vývojáře, tokeny se šetřily na nejtěžší úkol

GLM, od dubna 2026

Paušál bez limitu, AI dostala i Jiru, Confluence a debugování

Kontrolní skupina chybí. Sám Abitbol připisuje zlepšení o 70 % tomu, že lidé přestali šetřit tokeny.

Zdroj: přepis přednášky, 5:39:43 a 5:40:34

PŘEKROUCENÍ 3/3

„Desetkrát levnější“

1/10
faktury, kterou dřív platili Anthropicu a OpenAI

2-3×
levnější token než dnešní API s cache (odhad; proti cenám z jara 3-5×)

56 × B200
143-278 tis. USD měsíčně za pronájem

Vyplatí se při 2 000 vývojářích. Pro tým deseti lidí platí jiná matematika.

Odhad: B200 3,50-6,80 USD/h · Opus 5.5 a Sonnet 5.5 s 90 % cache 0,54-0,90 USD/M tokenů · bez lidí na provoz

Jak se zpráva měnila cestou

1 · 13. 10.

Přednáška

GLM výchozí, frontier jako výjimka

2 · 29. 10.

StartupHub

„GLM 4.5“, řečník jako moderátor

3 · 2. 11.

Post na X

„BREAKING“, tři týdny po akci

4 · 3. 11.

The Register

„cutting out Claude“

5 · 4. 11.

Desítky kopií

„přešli na čínský model“

Co se cestou ztratilo

Zkrácení: z „výchozí s výjimkami“ je „přešli“

Domyšlená příčina: zlepšení připsané modelu

Ztráta podmínek: 2 000 lidí, 56 GPU

Revolut ukázal, co se stane, když vývojářům přestanete počítat tokeny.

A teď znovu ruce nahoru: kdo pořád věří, že zahodili Claude?

Otevřené modely na vlastním železe

Model	Parametry	Sestava	Cena (USD)	Propustnost
Kimi K3.1 Pro	2,8T / 104B	8× B300	400-450 tis.	92-106 tok/s · 4 634 celkem
GLM-5.3	753B / 40B	8× H200, FP8	~340 tis.	452 tok/s celkem, 32 agentů
GLM-5.3 v Revolutu	753B / 40B	7 uzlů 8× B200, FP4	~2,5-3,3 mil.	~200 tok/s na uživatele
DeepSeek V4.1 Flash	~748B / 8-16B	2× DGX Spark, kvantováno	neověřeno	55-110 tok/s
GLM-5.3-Flash	320B / 18B	Mac Studio M3 Ultra 256 GB	~7 tis.	~28 tok/s (odhad)
Mistral Large 4	1,05T / 52B	8× B200, FP8	360-470 tis.	neměřeno
Muse Spark Open	816B / 42B	2× Mac Studio M5 Ultra 512 GB	44,6 tis.	25 tok/s

Mac za 22 tisíc obslouží jednoho vývojáře. Server za 340 tisíc desítky agentů naráz.

tok/s bez „celkem“ = jeden uživatel · ceny nákupu, Nvidia ceník nezveřejňuje · zdroje: HelixML, Revolut, Apple, vLLM, Rack Report

PHP2Shell: AI na obou stranách útoku

22,6 mil.
serverů s PHP, které mohou být zranitelné

2 dny
od záplaty k rootu ve vládní síti USA

240 → 25
vybraných cílů a serverů s rootem v jedné kampani

AI chybu našla

- chyba v jádru PHP od verze 5.4
- výzkumníci ji v červenci našli s modelem OpenAI
- záplata 21. října, dva měsíce příprav

AI útok zrychlila

- na útočné infrastruktuře Claude Code, DeepSeek a Kimi
- zhruba 40 skriptů psaných s pomocí AI
- z 19 000 firewallů automaticky vybráno 240 cílů

Lidské chyby staré desítky let teď hledá a zneužívá AI. Rychleji, než stihneme záplatovat.

CVE-2026-109618 a -109619, CVSS 9,2 · zdroje: TechTarget, The Hacker News, Dream, Palo Alto Networks

Stojíme před securitocalypsou. Nástrojů je na to zatím málo!

/code-review

Hledá chyby ve správnosti změn na vaší větvi. Podle úrovně i úklid a pravidla z CLAUDE.md.

- běží na pozadí jako subagent, vy pracujete dál
- úroveň low až max: méně, nebo více nálezů
- **--fix** nálezy rovnou opraví, **--comment** je vloží do PR
- **ultra** spustí hlubší review v cloudu
- všechny plány, platí se jako běžná práce

```
/code-review high 1234 --fix
```

/security-review

Hledá bezpečnostní díry v nepushnutých změnách: SQL injection, XSS, autentizace, práce s daty, závislosti.

- každý kandidát se ověřuje zvlášť a paralelně
- propustí jen nálezy s jistotou 8 z 10 a víc
- 18 výjimek proti planým poplachům
- stejná kontrola jako GitHub Action na každý PR
- levná a rychlá: minuta, dvě

Pozor: filtr může zahodit i skutečný nález. Bezpečnostní audit nenahrazuje.

Team a Enterprise: od 15. 10. také Code Review jako GitHub aplikace (Quick, Standard, Deep, pravidla v REVIEW.md)

Jak to řeším já? /codex-review

Druhý názor od jiného modelu. Jedno kolo, pak rozhodnutí.

Jak běží

1. zjistí záměr změny z PR a PRD
2. pustí testy a lint, padají-li, končí
3. Codex jednou prohlédne diff, jen pro čtení
4. Claude každý nález zkusí vyvrátit
5. opraví nejvýš v 10 krocích
6. verdikt: lze, s výhradou, nemergovat
7. merge až po vašem souhlasu

Proč se review neopakuje

- co chytí testy, reviewer nehlásí
- zamítnuté nálezy jdou do docs/review.md a nevracejí se
- nejvýš 10 nálezů, každý se scénářem selhání
- vady mimo diff jdou do todo, ne do PR
- opakovaná chyba se stane pravidlem v CLAUDE.md

Co dalšího přibylo v Claude Code

Claude Mods

Pluginy mění chování hlouběji. Vestavěný mod „You should know“ hlídá, co se přehlédlo.

Idle compaction

Odložená konverzace nad 200K tokenů se zhustí dřív, než vyprší cache. Šetří limit.

`/doctor prompt-audit`

Najde v CLAUDE.md a skillech pokyny psané pro starší modely.

`/skill-doctor`

Které skilly nikdo nepoužívá a kolik stojí kontextu.

AGENTS.md

Bez CLAUDE.md čte Claude Code AGENTS.md. Jedny instrukce pro Claude i Codex.

`/diff`

Panel s necommitnutými změnami vedle konverzace, živě při úpravách.

`/auto-mode-setup`

Naučí auto mode váš projekt. Nově jde zakázat zápis mimo pracovní adresář.

Hooks s pojistkou

onFailure: block: když hook selže, akce se zastaví.

Claude Code 2.1.257 až 2.1.318, září až listopad 2026 · zdroj: CHANGELOG.md

EXPERIMENT

AI First Company: firmu řídí Eduarda

Eduarda
AI ředitelka



Přijímačky na školu: web s daty o tisících oborů pro uchazeče a rodiče. Od konce září ho vede AI ředitelka Eduarda na Grok Botu. Programuje Claude Code, já schvaluji.

785
oslovených škol

368
v komunikaci

66
pořadatelů veletrhů

Fáborského test

Běží to doopravdy?	**ano**
Běží to déle než týden?	**ano**
Vydělá víc, než stojí, včetně času zakladatele?	**otevřené**

Čísla ze třetího týdne provozu · soukromé a církevní školy zatím neosloveny · github.com/tangero/stredniskoly

Agent autonomně tvoří, člověk zpětně reviduje



Eduarda

ředitelka: návrhy, e-maily, briefing

Claude Code

realizace podle zadání

Codex a Kimi

oponentura, druhý klíč

Patrick

reviduje zpětně, předem jen peníze, právo, reputace

Agent sám

- kód a opravy
- odpovědi školám a rodičům
- rešerše a databáze
- kampaně v rámci limitu

Vždy člověk

- platby a smlouvy
- úřady
- hromadné rozesílky, které nejde vzít zpět
- cokoli s dopadem na pověst

Dělicí čára je jediná otázka: dá se to vzít zpět?

„Kdo kód píše, ten ho nerecenzuje. Kdo zadává, ten neschvaluje.“

Co Eduarda zatím ukázala

Co funguje

- vymýšlení nových funkcí
- odpovědi na e-maily škol a rodičů
- sběr dat z webů škol
- katalog veletrhů ve všech krajích
- datové podklady pro redakce

Kde to drhne

- chtěla najmout brigádníka, aby za ni vystoupil na veletrhu
- vymyšlený technický fakt v podkladech pro rozesílku
- nápady rychle sklouzly k vylepšování nepodstatných detailů

Další pokus

- vlastní oblast: Dosah a Vztahy
- čtvrteční průzkum příležitostí
- rozpočet na pokusy, které nemusí vyjít

Provoz deleguji, podnikatelem zůstávám já.

Jak experiment pokračuje, ukážu průběžně na webinářích Vibecoding.cz.