# OD PRONÁJMU OPEN LLM MODELŮ K TOKEN AS A SERVICE

## Zkušenosti z praxe s používáním open source modelů

Robert Mitka, Head of AI Product Development, České Radiokomunikace

# CRA Cloud Services

Certified Security
Inside EU
Czech Support
Low Latency
High SLA

| CPU | RAM | STORAGE | SERVERY |
|---|---|---|---|
| 12 000 CPU JADER | 210 TB RAM | >9 PB úložiště | 8 000 virtuálních serverů |

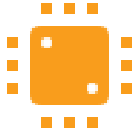
| | | |
|---|---|---|
| Podpora 24/7 v ČJ i EN | SLA 99,999% (max. 5 minut výpadek za rok) | 7 Datacenter v ČR |

## CRA CLOUD SERVICES SINCE 2010

# CRA
# AI Cloud Services

Certified Security
Inside EU
Czech Support
Low Latency
High SLA

**GPU**



500+
GPU karet

**AI MODELY**



50+
OpenSource

Fixní cena

Kritická infrastruktura

Data v ČR

## CRA AI CLOUD SERVICES SINCE 2025

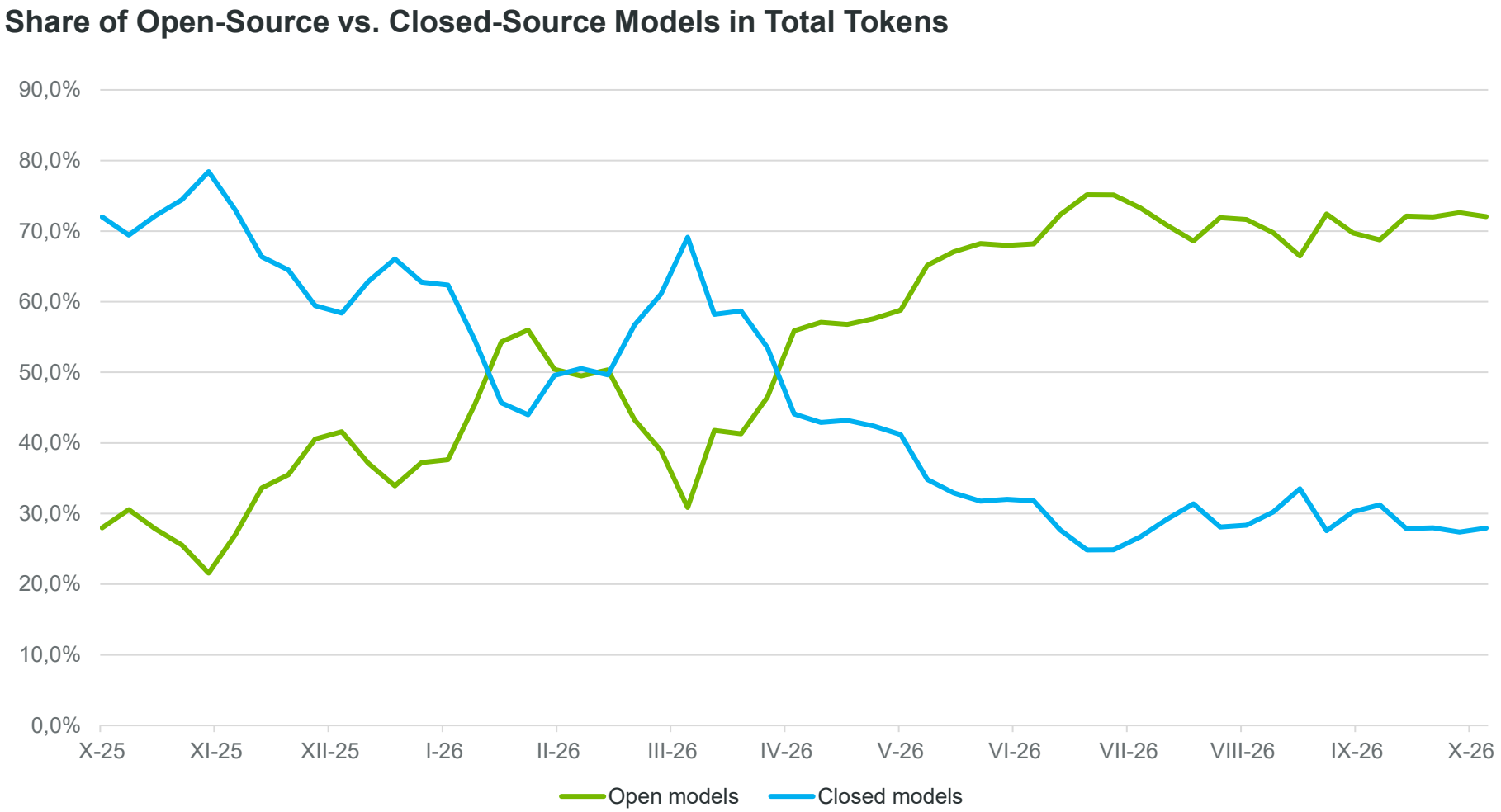
# Open Models vs Foundation Models




Source: Nvidia CES 2026

**OPEN MODELS HAVE CAUGHT UP WITH THE FRONTIER**

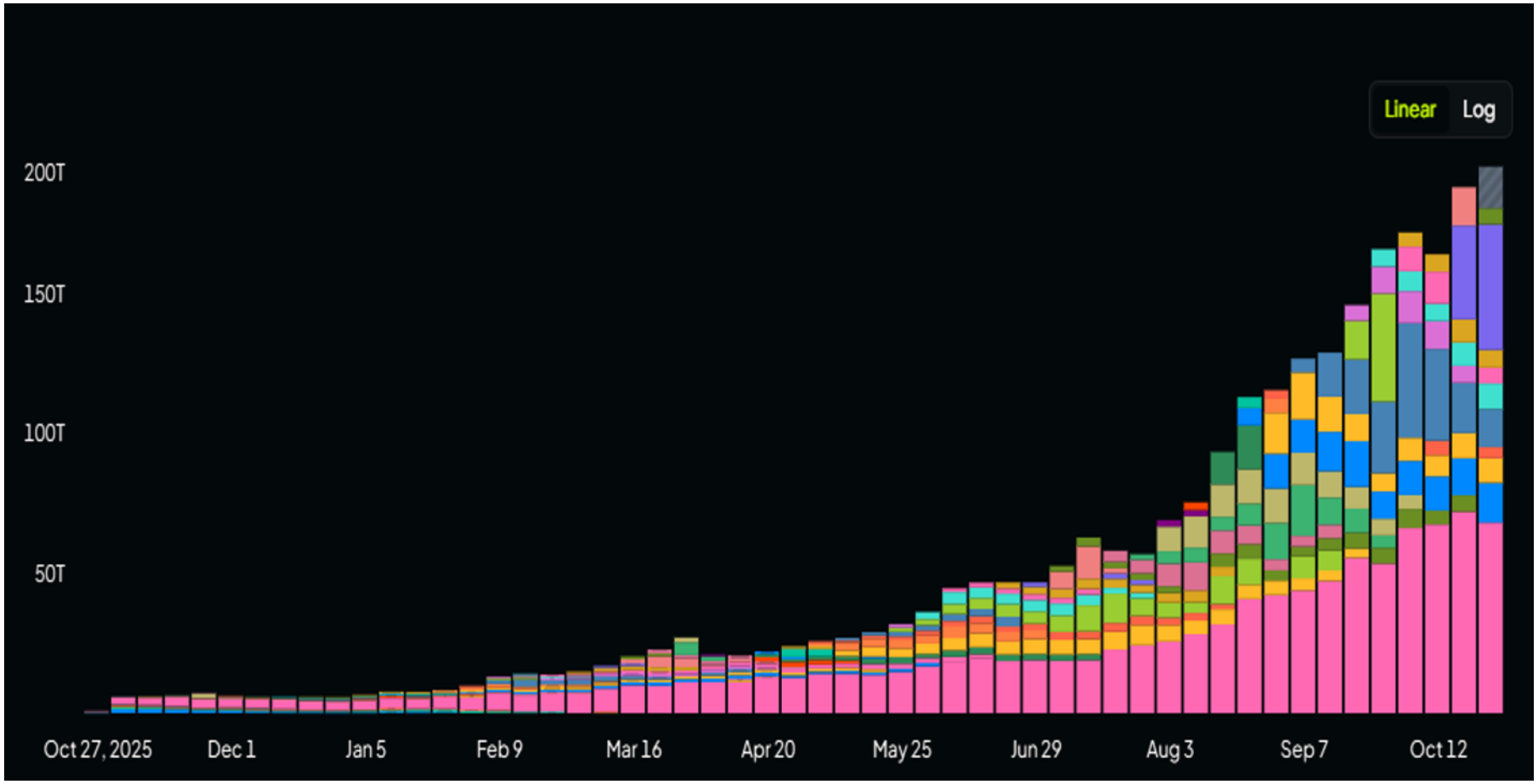
# Total Tokens Share




Source: OpenRouter.ai

## OPEN MODELS GREW FROM 28 % TO 72 % OF TOKENS IN ONE YEAR

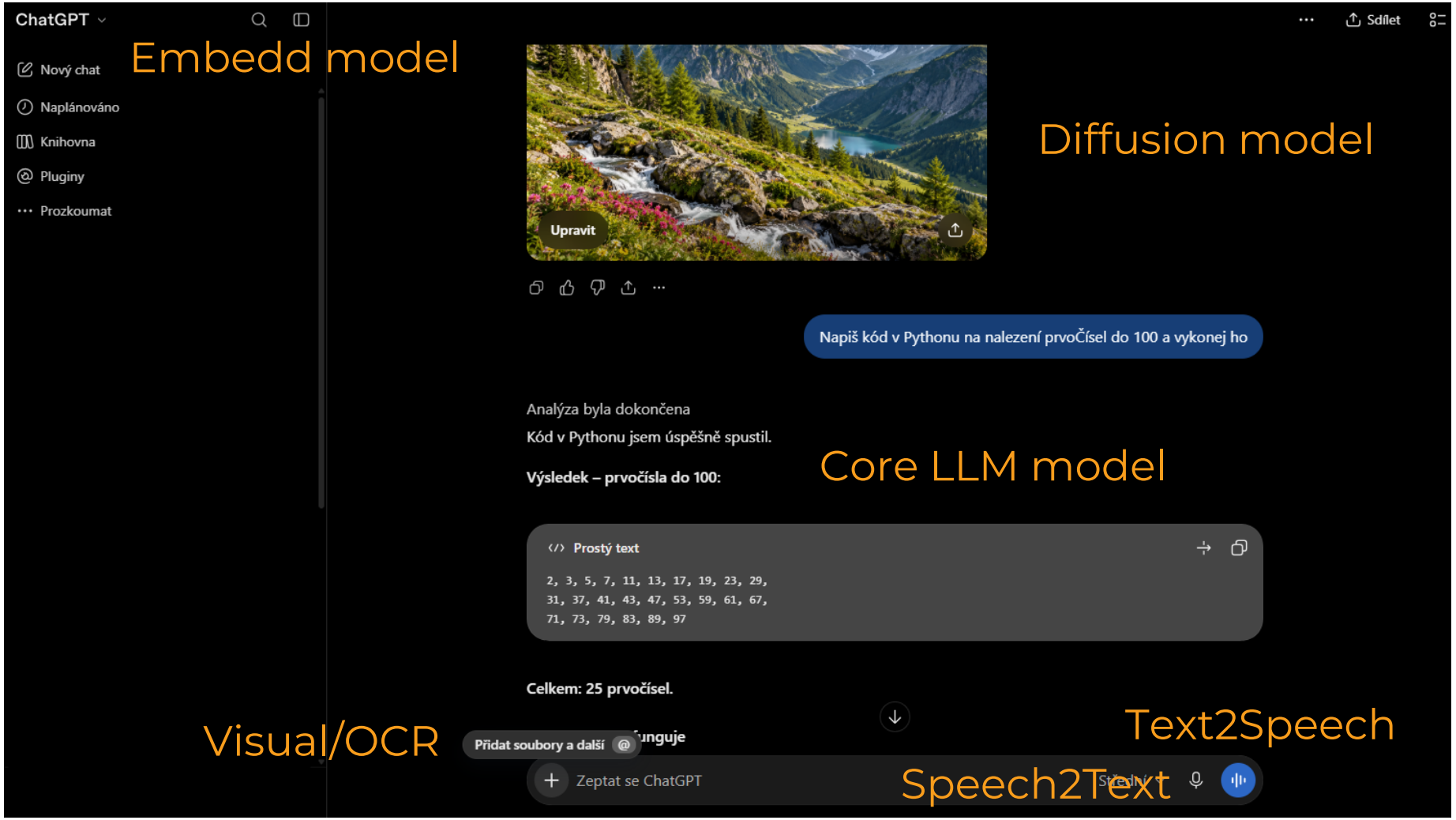
# Exponential Token Spend




Source: OpenRouter.ai

## AI TOKEN CONSUMPTION DOUBLES EVERY ~10 WEEKS

# Models Behind Chat




**ONE MODEL ON ONE GPU IS NOT ENOUGH TO RUN A REAL CHAT SERVICE**

# Core Czech Model

**77.0 %**
**best leaderboard score**
DeepSeek-V4-Flash, no reasoning

**87.5 %**
**nvidia/Qwen3.8-27B-NVFP4**
same task with reasoning

## Czech Bench leaderboard – top 5 by Cermat Czech MultiChoice (CCL-MC acc)

| # | Model | Submitted | CCL-MC acc |
|---|---|---|---|
| 1 | **deepseek-ai/DeepSeek-V4-Flash-0731** | 2026-08-07 | **77.0 %** |
| 2 | **deepseek-ai/DeepSeek-V3-0324** | 2025-07-03 | **75.3 %** |
| 3 | **google/gemma-4-31B-it** | 2026-05-19 | **74.6 %** |
| 4 | **deepseek-ai/DeepSeek-R1-0528** | 2025-07-01 | **74.4 %** |
| 5 | **meta-models/Muse-Glimmer-30B** | 2026-10-26 | **71.0 %** |

## nvidia/Qwen3.8-27B-NVFP4: with vs. without reasoning

| Task | with reasoning | without reasoning |
|---|---|---|
| **cermat_czech_mc** | **87.5 %** | 69.95 % |
| **cermat_math_mc** | **97.0 %** | 37.30 % |
| **klokan** | **92.8 %** | 36.68 % |

Detail: Without reasoning: accuracy (BenCzechMark, 3-shot, max 2048), With reasoning**:** pass@1 (CRA, 0-shot, 5 repeats, bootstrap mean from 10k resamples), pass@1 here is estimated as the probability that a single randomly generated output is correct, so it corresponds to accuracy

## THE BEST CZECH BENCHMARK LACKS REASONING-MODEL SUPPORT

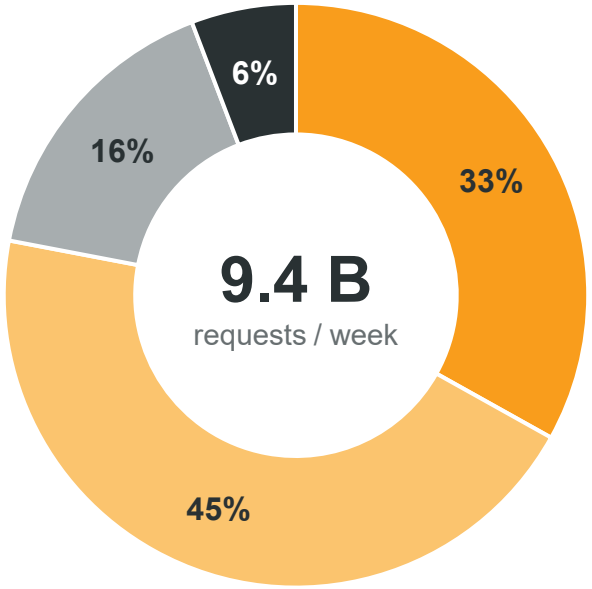
# Context length

**78 %**

**of requests use < 10K tokens**

94 % fit within 100K

## OpenRouter requests by context length – week of 19 Oct 2026




| Context length | Requests | Share | Cumulative |
|---|---|---|---|
| < 1K tokens | 3,102,269,331 | 33.1 % | 33.1 % |
| 1K – 10K tokens | 4,211,194,522 | 44.9 % | **78.0 %** |
| 10K – 100K tokens | 1,524,785,813 | 16.2 % | **94.2 %** |
| 100K – 1M tokens | 546,424,354 | 5.8 % | ≈ 100 % |
| 1M – 10M tokens | 146,462 | < 0.01 % | 100 % |
| **Total** | **9,384,822,482** | **100 %** | |

Source: OpenRouter, 9,384,822,482 requests in the week beginning 19 Oct 2026.

**SIZE THE CONTEXT TO REAL DEMAND, NOT TO THE MAXIMUM**

# Nvidia NVFP4 Quantisation on B300

**2.0×**

**concurrency**
89.5 → 182.9 requests @32k

**+25 %**

**output throughput**
2,419 → 3,035 tok/s

## Model size, KV cache capacity and concurrency

| Model | Weights | KV cache capacity (tokens) | Max concurrency @32k | KV cache usage (mean) |
|---|---|---|---|---|
| **Qwen/Qwen3.8-27B**<br>baseline · BF16 weights · BF16 KV cache | 50.22 GiB | 2,932,053 | 89.5× | 35.4 % |
| **nvidia/Qwen3.8-27B-NVFP4**<br>NVFP4 weights · FP8 KV cache | 19.05 GiB | 5,994,116 | 182.9× | 20.5 % |
| **NVFP4 vs. baseline** | **−62 %** | **+104 % (2.0×)** | **+104 % (2.0×)** | **−14.9 %** |

## Throughput and response time

| Model | Output throughput (tok/s) | Per item (tok/s) | Mean answer length (tok) | Mean answer time (s) |
|---|---|---|---|---|
| **Qwen/Qwen3.8-27B**<br>baseline · BF16 weights · BF16 KV cache | 2,419.1 | 16.8 | 2,847.5 | 169.5 |
| **nvidia/Qwen3.8-27B-NVFP4**<br>NVFP4 weights · FP8 KV cache | 3,035.2 | 21.1 | 3,225.9 | 153.0 |
| **NVFP4 vs. baseline** | **+25.5 %** | **+25.6 %** | **+13.3 %** | **−9.7 %** |

Setup: 1× NVIDIA B300 · max-model-len 32,768 · max seqs 144 · max batched tokens 4,096 · gpu-memory-utilization 0.97

**NVFP4 ON B300: 2× THE CAPACITY, +25 % THROUGHPUT, FASTER ANSWERS**

# Nvidia NVFP4 Quantisation on H100

**up to −14 %**
**output throughput**
319.9 → 276.4 tok/s @32 seqs

**up to +34 %**
**time to first token**
3.7 → 4.8 s @32 seqs

**up to +16 %**
**time per output token**
96.1 → 111.4 ms @32 seqs

### Output throughput vs. concurrency

| Concurrent seqs | Qwen/Qwen3.8-27B Base (tok/s) | nvidia/Qwen3.8-27B-NVFP4 NVFP4 (tok/s) | Δ output |
|---|---|---|---|
| 16 | 270.1 | 253.2 | −6.3 % |
| 24 | 303.3 | 265.8 | −12.4 % |
| 32 | 319.9 | 276.4 | −13.6 % |

### Latency: TTFT and TPOT (mean)

| Seqs | TTFT base | TTFT NVFP4 | Δ TTFT | TPOT base | TPOT NVFP4 | Δ TPOT |
|---|---|---|---|---|---|---|
| 16 | 2,538 ms | 3,411 ms | +34 % | 56.6 ms | 59.9 ms | +5.8 % |
| 24 | 2,983 ms | 3,985 ms | +34 % | 75.6 ms | 86.0 ms | +14 % |
| 32 | 3,717 ms | 4,803 ms | +29 % | 96.1 ms | 111.4 ms | +16 % |

Setup: 1× NVIDIA H100 · Qwen/Qwen3.8-27B (base) vs. nvidia/Qwen3.8-27B-NVFP4 · concurrency 16 / 24 / 32 seqs

## NVFP4 ON H100: SLOWER AND HIGHER LATENCY, NO NATIVE FP4 SUPPORT

# Open Models in Production

Six things to get right before going live

### One model is not enough

A real chat service needs core, embedding, STT, OCR, diffusion and TTS models

### Benchmarks mislead

The best Czech benchmark skips reasoning – same model, same task

### Context costs capacity

Concurrent full-length requests per B300 at 32K vs. 1M context

### Quantisation issues

NVFP4 throughput on B300 vs. H100

### Czech narrows the choice

Quality for Czech models varies widely between models

### Demand grows fast

Token consumption doubles every ~10 weeks – capacity must scale

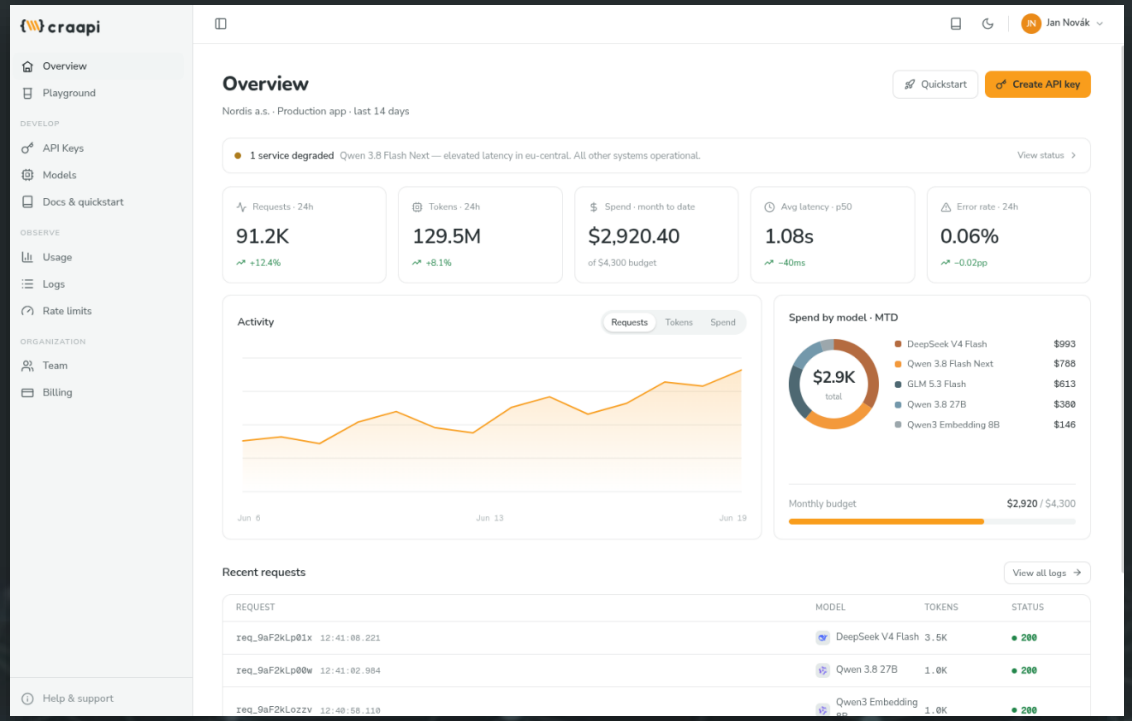
## DEPLOYING OPEN MODELS IS MORE THAN DOWNLOADING WEIGHTS

NEW SERVICE

# craapi

## Token as a Service for open models

- Open models via one API
- Multimodal, Speech2Text, diffusion, embedding
- Focused on Czech, data in the Czech Republic




## CRAAPI.AI – LAUNCHING THIS YEAR

# Token vs. Model as a Service

**Two ways to run open models at CRA**

| | Token as a Service NEW | Model as a Service |
|---|---|---|
| **Models** | Curated model catalog | Almost any open model |
| **Pricing** | Pay per 1M tokens | Fixed price |
| **Cost** | Scales with usage | Higher entry cost |
| **Best for** | Low volume | High volume |
| **Environment** | Shared | Dedicated |
| **Customers** | PoC, startups, developers | Enterprise, production |

**Both services**

- ✓ Certified Security
- ✓ Inside EU
- ✓ Czech Support
- ✓ Low Latency

## START WITH TOKENS, SCALE TO DEDICATED MODELS

# craapi.ai
# Launch



**GET NOTIFIED WHEN CRAAPI LAUNCHES**

r.mitka@cra.cz