Technické dovednosti v AI: Principy, dotazování a vlastní modely

Michal Vašinek,

Katedra informatiky, FEI,

VŠB – Technická Univerzita Osrava

Umělá inteligence – predikce dalšího slova

Úloha: Pro text na vstupu neuronové sítě vygeneruj další slovo

Umělá inteligence je obor …

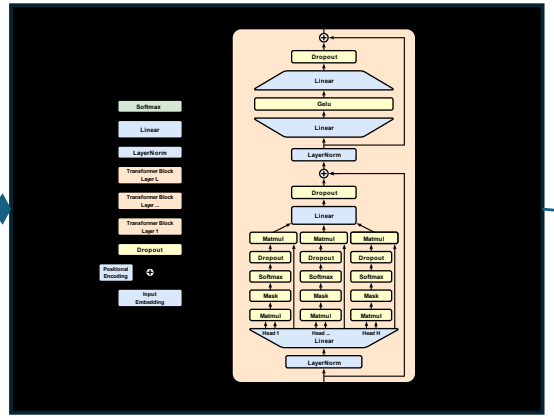
jaké bude další slovo?

Umělá

inteligence

je

obor



Neuronová síť složená z Transformer bloků

Slovo	P(slovo|vstup)
informatiky	0,45
matematiky	0,12
lingvistiky	0,000007
veterinářství	0,0000005
počítačových	0,02
koťátko	0,000000001
…	…

Token

- Je elementární jednotkou zpracování textu
- Může být slovem, kombinací slov či n-gramem
- Důležitou účetní jednotkou

Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between models. Newer models like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and will produce different tokens for the same input text.

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

Enter some text

GPT-4o

GPT-4o is our most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Learn about GPT-4o ↗

Model	Input	Output
gpt-4o	0,01 US$ / 1K tokens	0,02 US$ / 1K tokens
gpt-4o-2024-05-13	0,01 US$ / 1K tokens	0,02 US$ / 1K tokens

Vision pricing calculator

Set width: 150 px by Set height: 150 px = 0,001275 US$

☐ Low resolution

OpenAI - tokenizer

- platform.openai.com/tokenizer

- Vyzkoušejte si tokenizer, pozorujte, kolik tokenů generuje, jaký text.
- Porovnejte „Hello world.“ a „Ahoj Světe.“
- Vyzkoušejte bez diakritiky
- Napište vše malým písmem
- Odstraňte tečku

Kontext – klíč k relevantní odpovědi

Můj oblíbený … je … fiction.

Jakými slovy byste nahradili tečky?

Kontext – klíč k relevantní odpovědi

Můj oblíbený ... je ... fiction.

Jakými slovy byste nahradili tečky?

Doplň mi v následující větě místo tří teček nějaká slova, tak ať věta dává smysl:

Můj oblíbený .. je ... fiction

Můj oblíbený žánr je science fiction.

Kontext – klíč k relevantní odpovědi

Můj oblíbený **film** je **Pulp** fiction.

Jakými slovy byste nahradili tečky?

Doplň mi v následující větě místo tří teček nějaká slova, tak ať věta dává smysl:

Můj oblíbený .. je ... fiction

Můj oblíbený žánr je science fiction.

Kontext – experiment s rolema

- **ChatGPT předpovídá další token.**

- **Vyzkoušejte:**
- „Doplň mi další slovo v následující větě: Umělá“
- „Předpokladej, že jsi ortoped, doplň mi další slovo v následující větě: Umělá“
- „Předpokladej, že jsi kadeřnice, doplň mi další slovo v následující větě: Umělá“

Kontext – role a pravděpodobnost tokenu

Bez role

Token	P
int	0.998
intel	0.000892
intelligence	0.000109

Ortoped

Token	P
k	0.645
kl	0.342
ky	0.0116

Kadeřnice

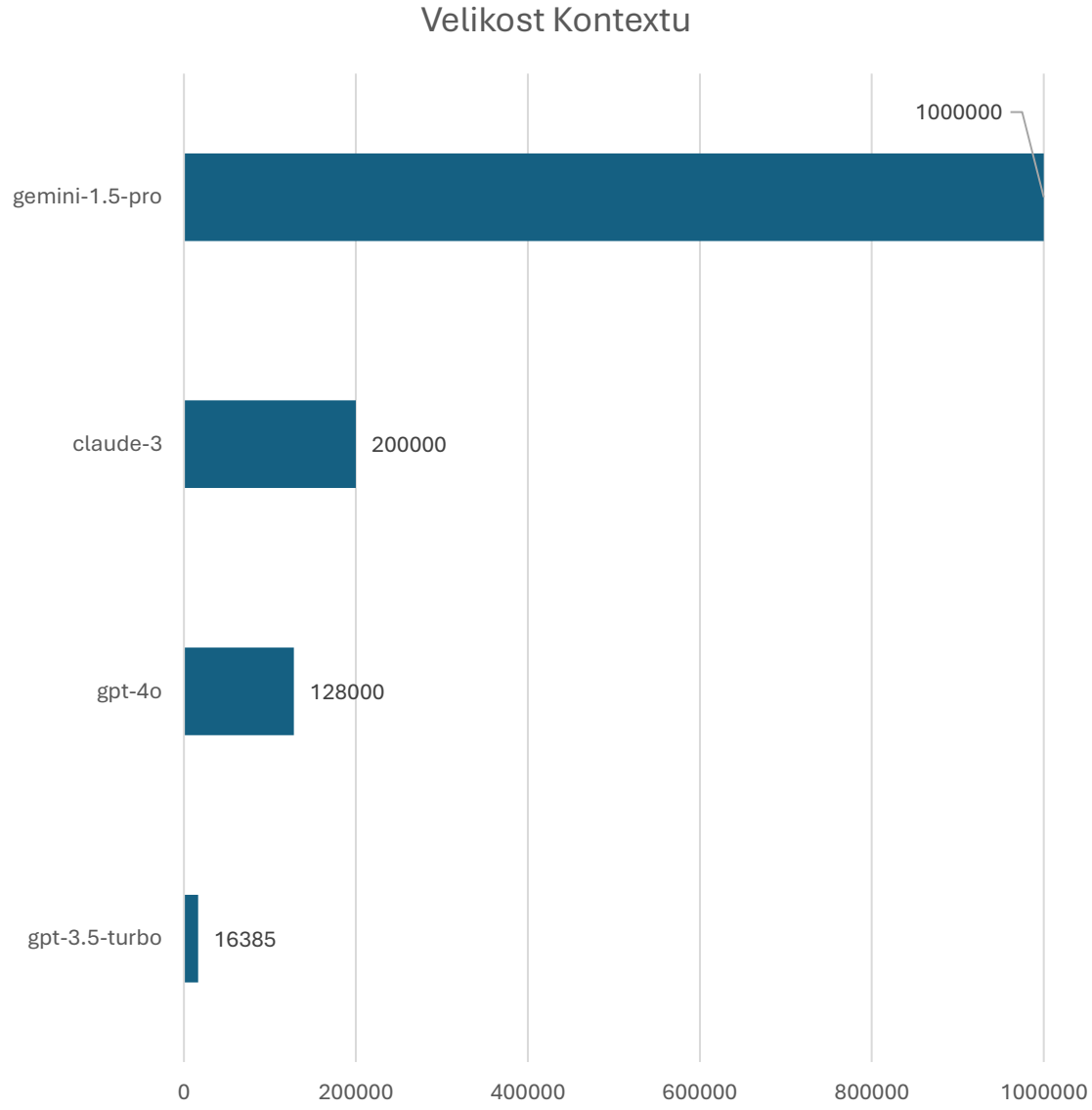
Token	P
bar	0.729
v	0.111
vl	0.101

Rolí měníme pravděpodobnost dalšího tokenu, měníme kontext!

Python – lze vyzkoušet skript kontext.py, předpokládá se, že máte API klíč.

Kontext a poskytovatelé LLM

- Kontext je utvářen celou komunikací s modelem.
- Délka kontextu nám určuje, jak dlouhý text je model schopen zpracovat.
  - Některé dokumenty jsou příliš dlouhé
  - Model v určitou chvíli zapomene začátek komunikace



Vlastní GPT

Vhodné pro velmi specifické nebo opakující se úlohy.

Velmi specifické:	• Chceme vysoce expertní znalosti
Opakující se úlohy:	• Nechceme stále dokola specifikovat roli
Ukázka vytvoření.	• Pouze pro uživatele s předplatným

Vlastní GPT – piš jako člověk

- Základní prompt:
  - „Používej jasný a přímý jazyk a vyhni se složité terminologii. Vyhýbej se příslovcím. Vyhni se módním slovům a místo nich používej jednoduchou češtinu. V relevantních případech použij žargon. Vyvaruj se přílišného nadšení.“

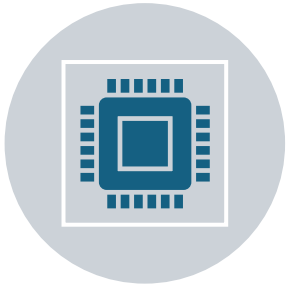
Vlastní GPT, odpovídání na email

Definice vlastního pozdravu, eliminace přehnané slušnosti.

Lze přidat soubor s texty vlastních emailů, či dokonce celých komunikací, lze ChatGPT nechat tvořit text, jako byste to byli Vy.

Akce

Napojení externích informací na gpt.



Každou službu či webové rozhraní poskytující API, které lze veřejně volat, lze

Komunikace zahrnuje informace z externího zdroje.

Ukázka kniha objednávek z Coinmate.

Vytvoření vlastního API klíče

- Pokud umíme programovat, pak lze s ChatGPT komunikovat skrze API.
- Za tokeny se platí => nutnost vložit vlastní zdroje.
- Overview - OpenAI API
  - Vložení prostředků: Zvolte Settings / Billing / Add to credit balance
  - API klíč: API keys - OpenAI API / Create new secret key

Zajímavosti

- Hugginface arena
  - Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots (lmarena.ai)
  - Žebříček - Chatbot Arena Leaderboard - a Hugging Face Space by lmsys
- Hugginface – svět otevřených modelů
  - Hugging Face – The AI community building the future.
- Multiagentní systémy – automatizace komplexních procesů
  - LangGraph (langchain.com)

Otázky či diskuze

Díky za pozornost