Prompt Engineering

Dekódolási technikák, nyelvi modellek kiértékelése és üzemeltetése

Hollósi Gergely

TMiT

Távközlési és
Mesterséges Intelligencia
Tanszék

Fogalmak

Tartalomjegyzék

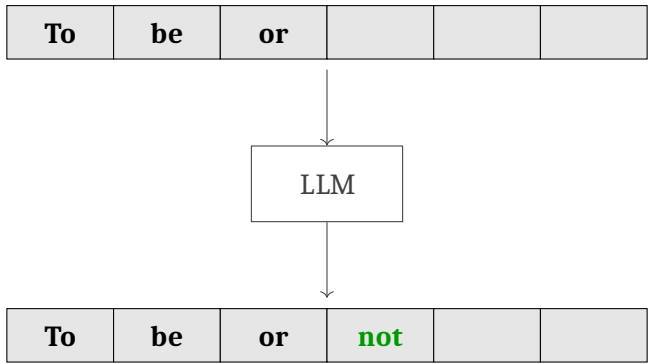
Autoregresszív viselkedés

TMiT

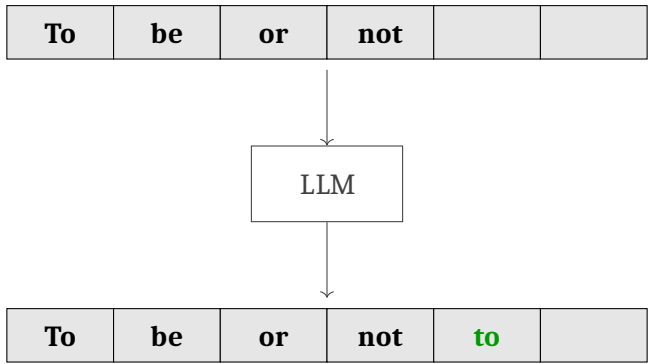
- A nyelvi modellek szigorúan véve „transzformer” modellek
- Ún. *autoregresszív* modellek – következő szó (token) becslése
- A nyelvi modell minden tokenre egyesével lefut

- Autoregresszív futtatáskor, nem autoregresszív tanításkor
- Vanilla implementációk – valóságban attention cache történik

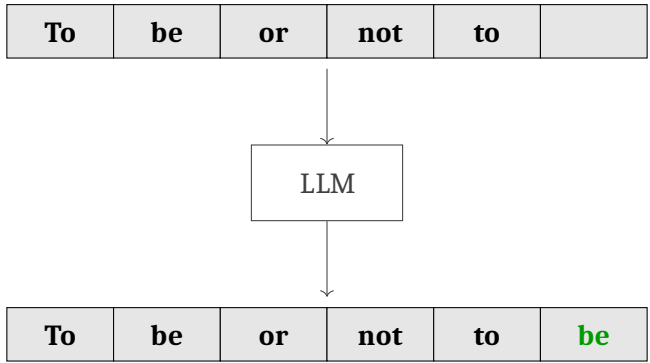
Autoregresszív viselkedés



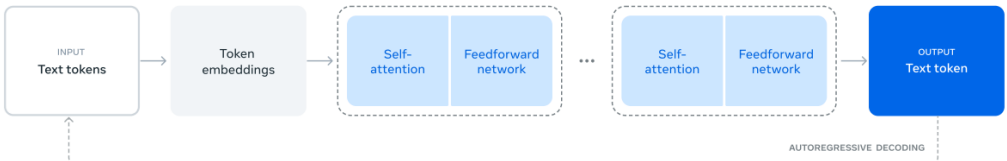
Autoregresszív viselkedés



Autoregresszív viselkedés



Llama3 példa



Kimeneti mintavételezés

TMiT

- Mi az a token?
- A transzformerek determinisztikusak?
- Ha a transzformer választja meg a következő tokent, miként tudunk más-más kimeneteket generálni?

Tokenek

TMiT

- A token a nyelvi modellek alapegysége – a szöveget tokenekre bontjuk
- 100 token $\sim$ 75 szó
- A transzformer feladata a következő token kiválasztása – multikategorikus osztályozás
- pl. Llama3.1 – 128.000 token

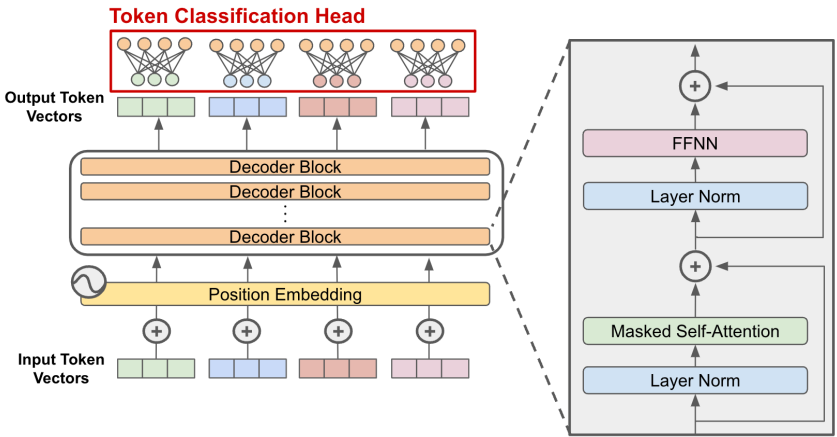
ID	Token
8	23
39	H
864	pre
1341	IC
1378	Exception
128001	<EOS>

OpenAI Tokenizer

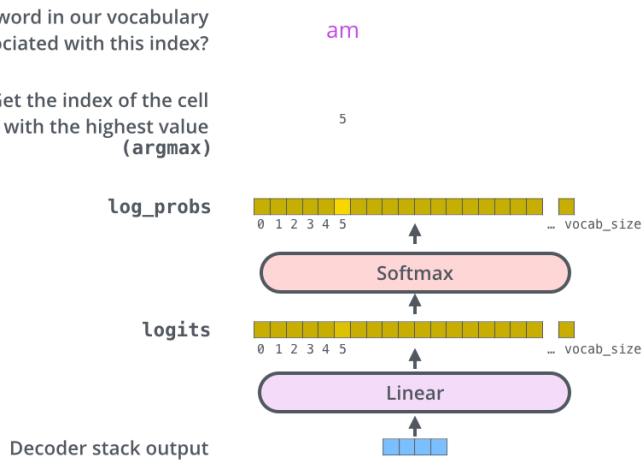
TMiT

```
\item Mi az a token?
\item A transzformerek determinisztikusak?
\item Ha a transzformer választja meg a következ�� tokent, miként tud
unk más-más kimeneteket generálni?
```

Transzformer kimeneti rétegei



Transzformer kimeneti rétegei



Transzformer kimenete

- $i = 1, \ldots, N$ darab lehetséges token
- A nyelvi modellek esetén a Z_i kimenet az i-dik tokenre az ún. *logit* – a neurális hálózat aktivációja minden egyes tokenre (a logit értékészlete $Z_i \in (-\infty, \infty)$)
- Ezt egységesen átalakítjuk egy diszkrét valószínűségi eloszlásra $1, \ldots, N$ tokenre:

$$p_1, p_2, \ldots, p_N$$

- $p_i > 0$ és $\sum_{i=1}^{N} p_i = 1$
- Erre a *softmax*-ot használunk

$$p_i = \frac{e^{Z_i}}{\sum_{i=1}^{N} e^{Z_i}}$$

Transzformer kimenete – egyszerű példa

TMiT

- Például három lehetséges tokenre:

$$Z = [3.8, -2, 5]$$

$$p = [0.231, 0.0007, 0.768]$$

Transzformer kimenete – Llama 3.1 8B példa

TMiT

- Szöveg folytatása adott bemenet alapján
- Legyen a prompt „To be or ”
- Tokenek (<|begin_of_text|> = 128000):

$$[128000, 1271, 387, 477, 220]$$

- Első hat token valószínűségi eloszlása:

2	(lf)0	3	4	not	1
0.458	0.148	0.062	0.027	0.028	0.02

Transzformer kimenete – a hőmérséklet

TMiT

- Legtöbbször egy ún. *temperature* paramétert is használunk:

$$p_i = \frac{e^{z_i/T}}{\sum_{i=1}^{N} e^{z_i/T}}$$

- Boltzmann eloszlás? Entrópia?
- Skálázhatjuk az eloszlást
  - $T \to 0$, akkor a legnagyobb érték 1 lesz, a többi zéró
  - $T \to \infty$, akkor egyenletes eloszlást kapunk

Transzformer kimenete – Llama 3.1 8B példa

TMiT

- Nézzük a korábbi példánkat:

	2	**(lf)0**	**3**	**4**	**not**	**1**
T=0.1	9.99	0	0	0	0	0
T=0.6	0.81	0.12	0.03	0.007	0.005	0.004
T=1	0.458	0.148	0.062	0.027	0.028	0.02
T=2	0.023	0.013	0.008	0.005	0.005	0.004
T=100	9.3e-6	9.2e-6	9.2e-6	9.1e-6	9.1e-6	9e-6

Token választás

TMiT

- A kapott eloszlást használhatjuk a következő token kiválasztására
- maximum probability sampling ($T = 0$)
  - Kiválasztjuk a legnagyobb valószínűséget
  - Fix választ generál minden alkalommal
- multinomial sampling
  - Véletlenszerűen választ következő tokent a kapott eloszlás alapján
- top-k sampling
  - Csak a legvalószínűbb k darabból választ véletlenszerűen
- top-p sampling
  - Csak a p kumulatív valószínűséget lefedő tokenekből választ véletlenszerűen

Kontextus ablak

TMiT

- A transzformerek fix méretű neurális hálózatok – minden futásnál fix méretű bemenet és kimenet szükséges
- A nyelvi modell kontextusablaka a maximális bemeneti token szám, amit egyben tud kezelni
- Generálást korlátozhatja az API, nem modell függő

LLM	Context window
Llama-3	8.092
Llama-3.1	128.000
GPT-4o	128.000
GPT-3.5	16.385

Kontextus ablak növelése

Előnyök	Hátrányok
Hosszabb, koherensebb prompt	Növekedő költségek
Kevesebb hallucináció	Nem feltétlenül jobb kimenet
Nagy méretű dokumentumok fogadására képes	Több hallucináció

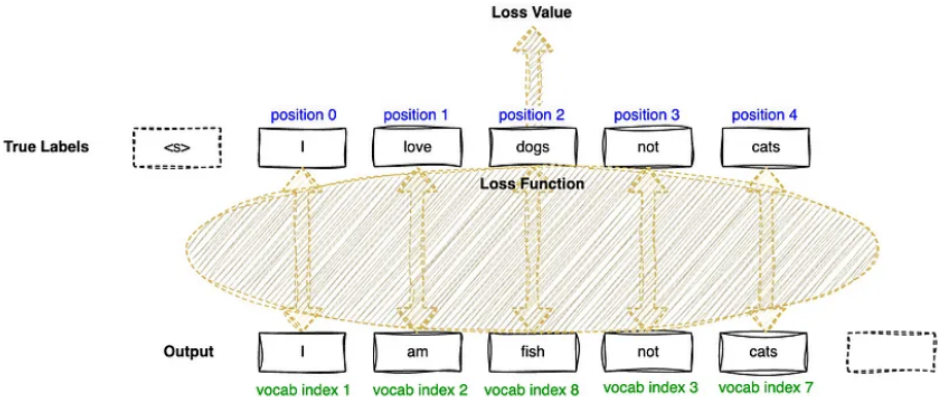
Tartalomjegyzék

Tanítási metrika

TMiT

- A nyelvi modellek kiértékelése nehéz: szemantikailag megegyező válaszok nem feltétlenül egyeznek meg szintaktikailag
- Nyelvi modell kiértékelése:
  - tanítás során (pl. költség függvény)
  - validálás során
- Mind előtanítás (pre-train), mind finomhangolás (fine-tuning) során a *kereszt-entrópiát* minimalizálják
- Supervised Finetuning (SFT)
- Reinforcement Learning from Human Feedback (RLHF)
- Direct Preference Optimization (DPO)

Kereszt entrópia

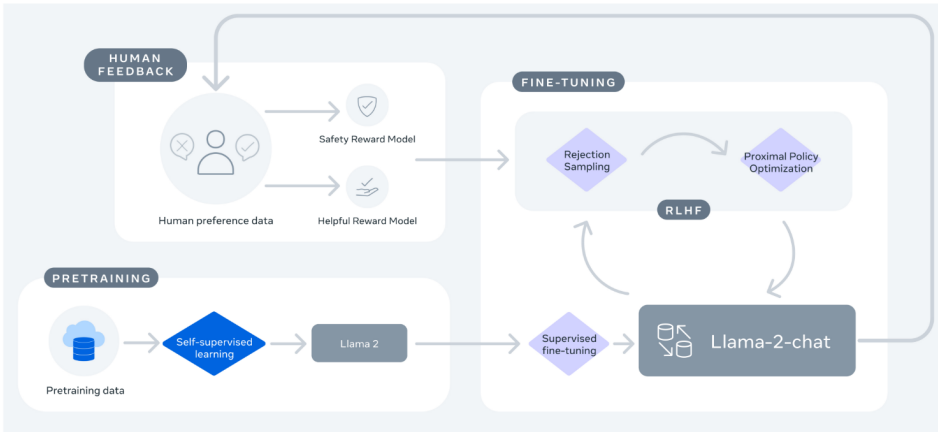


Tanítási adatok

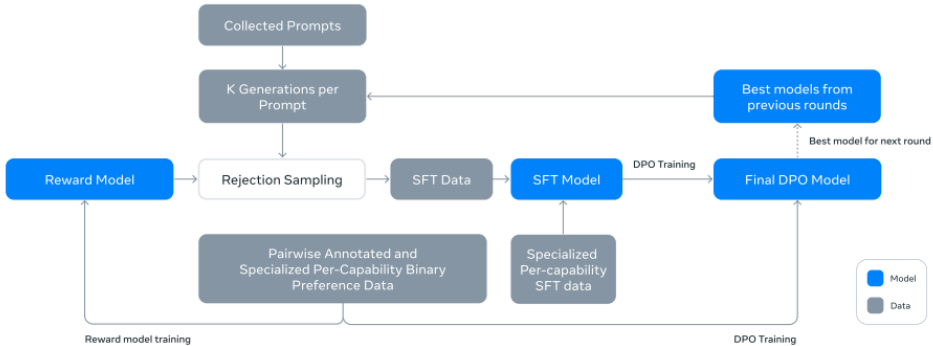
TMiT

- Alap tanítás
  - Webcrawling
  - Fórumok és egyéb adathalmazok (pl. WebText2, Reddit)
  - Interneten elérhető könyvek
  - Wikipedia
- GPT-3: 45 TB tömörítve, 570 GB szűrés után
- Jobb minőségű adathalmazok a tanítás során többször kerülnek mintavételezésre
- Emberi erőforrás – visszacsatolásos tanulás

Példa – Llama 2



Példa – Llama 3[1]



[1] https://arxiv.org/abs/2407.21783

Modellek validálása

TMiT

- Miért csináljuk?
  - Jól tanult a modellem? – non-regression test
  - Melyik modell a jobb? – leaderboards
  - Hol tart a tudomány? Mire képes a modellem?
- Automated benchmarking
- Humans as judges
  - Vibes
  - Arena – wisdom of the crowd (pl. lmarena.ai)
  - Annotation
- Models as judges
  - Torzító hatású

Validálási metrikák és tesztek – HuggingFace

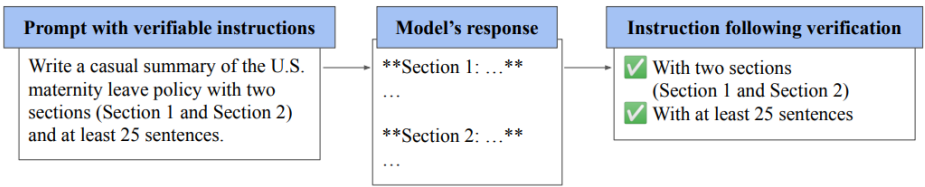
Average	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
44.75	79.96	58.77	38.97	17.9	23.72	49.2
44.14	81.36	59.47	36.4	19.24	19	49.38
43.92	79.86	59.27	37.92	20.92	16.83	48.73
43.61	81.63	57.33	36.03	17.45	20.15	49.05

Miket mérünk?

TMiT

Szintaktika	Szintaktikai és szemantikai összefüggés, helyesség
Relevancia	A válasz releváns és informatív
Korrektség	A válasz tényszerűen helyes
Hallucináció	A válasz nem-e tartalmaz légből kapott, hamis információkat
Felelősségteljesség	A válasz nem-e tartalmaz elfogultságot, esetleg a válasz toxikus
Feladat specifikus helyesség	Speciális feladatok sikeres elvégzése

Instruction-Following Evaluation



Instruction-Following Evaluation

TMiT

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *

BIG Bench Hard

- 23 kifejezetten nehéz feladatcsoport
- Statisztikailag megfelelő számosság, objektív mérce

BIG-Bench Hard Task
Boolean Expressions$^{\lambda}$
Causal Judgement
Date Understanding
Disambiguation QA
Dyck Languages$^{\lambda}$
Formal Fallacies
Geometric Shapes$^{\lambda}$
Hyperbaton
Logical Deduction$^{\lambda}$ (*avg*)
Movie Recommendation
Multi-Step Arithmetic$^{\lambda}$ [Two]
Navigate$^{\lambda}$
Object Counting$^{\lambda}$

BIG Bench Hard

TMiT

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.
Options:
- Yes
- No

Egyéb tesztek

TMiT

MATH	Középiskolai matematika problémák
GPQA	Tudásbázis különböző területekről (relatív könnyű szakembereknek, egyébként nehéz)
MuSR	Algoritmikusan komplex problémák (gyilkossági történet, allokálási optimalizációk, stb.)
BLEU	Nyelvi fordítási metrika
ROUGE	Összefoglalók kiértékelése n-grammok egyezése alapján

Problémák?

TMiT

- Nyilvános tesztek és adathalmazok tanítóadatokba kerülnek
- Általános képességek objektív mérése továbbra is nagyon nehéz
  - IQ?

Tartalomjegyzék

Üzemeltetés?

Üzemeltetés

TMiT

- „Closed-source” solutions
- „Open-source” solutions

- On-Prem
- GPU kapacitás bérlése
- Software-as-a-service

On-premises

TMiT

- Komoly erőforrás szükséges (főleg memória)
- Open-source modell (licensz!)
- Szükséges körítés az alkalmazáshoz (súlyok+szoftver)
  - Huggingface, LangChain, LangGraph, …
- Llama 3.1
  - 8B, 70B, 405B modellek
- 8B akár 16GB VRAM-mal
- 405B legalább 240GB VRAM-mal (kvantált modell)
- Nehezen párhuzamosítható – csak célmegoldásokkal érdemes
  - nVidia H100 – 94 GB memória
  - NVLink

LLM Infra Stack – Market Landscape

A work in progress

CriteriaVentureTech

Data Layer

Vector Database & Search

MongoDB. elastic Pinecone zilliz Weaviate NEON supabase vectara SUPERLINKED qdrant chroma pgvector nuclia marqo deeplake upstash Vespa HYPERSPACE ANOST LanceDB

Data Management

UNSTRUCTURED mindsdb NOMIC datologyai Automorphic Kern MANTIUM TAYLOR lettria vue.ai Lilac Indexify .txt

Data Labeling & Annotation

scale Labelbox Argilla Toloka SuperAnnotate HumanSignal Kili prodigy Watchful datasaur.ai Refuel Anote M47°

Data Pipeline & Orchestration

Fivetran matillion ASTRONOMER WarpStream Airflow Airbyte PREFECT ZenML MAGE dagster SuperDuperDB Flyte kedro

Synthetic Data

gretel TONIC hazy MOSTLY AI scale Argilla HAVEN tuna Synthetaus

Model Layer

Fine Tuning & Training

Argilla scale Snorkel surge HAVEN arcee.ai GigaML ANARCHY Predibase

RAG

jina LiteLLM CREDAL Ragas EMBEDCHAIN nuclia SciPhi NEUM

Orch. Frameworks

LlamaIndex LangChain FIXIE deepset SuperDuperDB griptape

Experiment Tracking & Prompt Eng.

W&B comet neptune.ai DyC DagsHub AimStack HumanFirst vellum agenta Humanloop Baseplate orq HAVEN LOGIQ Autoblocks Radiant HegelAI PromptLayer KLU eppo

Quality Testing & Evaluation

Galileo trulens GANTRY arize deepchecks Cleanlab Giskard Bespoken context Patronus AI Toloka vectorview ATHINA BRAINTRUST RagaAI AgentOps promptfoo Confident AI inspect ai Chatter UpTrain GOLDENSET Quotient phospho

Federated Learning

Flower FedML Opaque Devron sherpa.ai DynamoFL integrate.ai APHERIS TUNE INSIGHT

Deployment Layer

deepspeed anyscale Modular OctoAI run:ai Kubeflow deci Cerebras SELDON NEURAL MAGIC CentML Paperspace baseten Modal RunPod Outerbounds Martian replicate cerebrium BentoML mystic zeet Titan Alpa MAI/THEME orkes fireworks.ai BANANA Lepton AI HippoML beam Floom E2B commonbase PoplarML nCompass dstack

Monitoring Layer

Observability

DATADOG fiddler WHYLABS aporia arize mona helicone portkey EVIDENTLY AI HoneyHive Langfuse Parea Superwise traceloop LogSpend exxa Flowstack LOGIQ BricksAI

App/User Analytics

Nebuly sentify Autoblocks context

Enterprise Considerations

Security & Privacy

Synack praetorian Nightfall AI ROBUST INTELLIGENCE Arthur HIDDENLAYER PROTECT AI TROJAI Giskard LAKERA Lasso Sarus Guardrails AI CALYPSOAI skyflow MITHRIL SECURITY harmonic preamble Prompt Deepkeep PRIVATE AI ADVERSA CRANIUM MINDGARD cadea AIShield PRIVATE GPT PROTOPIA Cape Privacy Automorphic Kobalt Labs PromptArmor

Governance, Compliance & Risk

2021 AI credo ai Holistic AI QUANTPI MONITAUR Advai Deasie KomplyAI Verta Vera calvinrisk Impact AI alinia TRUSTIBLE DEEPLOY LUMENOVA ARMILLA tenzai anch.AI FAIRLY PureML Arklow fairnow Saidot Mission Control Suzan AI

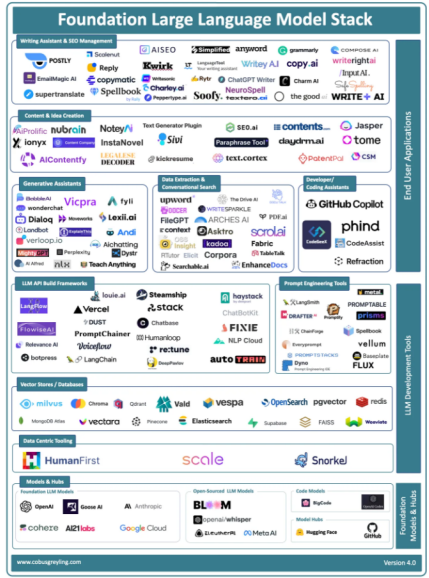
E2E Approach

E2E LLMOps Platform

aws databricks DataRobot Hugging Face Vertex AI clarifai haystack Lightning AI METAFLOW deepset iguazio Dataloop CLEAR ML Qwak LangChain ANARCHY Adaptive ML CAMEL-AI Domino Valohai LAMINI together.ai Kern truefoundry OpenAI GPT baseten Steamship Chainlit

No/Low code AI App builder

FIXIE ONE AI Continual Deasie gradient yurts re:tune Dify. Mind Langflow Prompt Studio FinetuneDB Relevance AI Mano AI GRADIENT.


Foundation Large Language Model Stack
Writing Assistant & SEO Management
POSTLY
Scalenut
AISEO
Simplified
anyword
grammarly
COMPOSE AI
writeright.ai
EmailMagic AI
Reply
Kwirk
Writesonic
Writey AI
copy.ai
InputAI.
supertranslate
copymatic
Rytr
ChatGPT Writer
Charm AI
Spellbook
Charley.ai
Peppertype.ai
Soofy.
NeuroSpell
textero.ai
the good
WRITE AI
Content & Idea Creation
Prolific
nubrain
NoteyAI
Text Generator Plugin
SEO.ai
contents_
Jasper
ionyx
Content Company
InstaNovel
Sivi
Paraphrase Tool
daydrm.ai
tome
AIContentfy
DECODER
kickresume
text.cortex
PotentPal
CSM
End User Applications
Generative Assistants
BobbleAI
wonderchat
Vicpra
fyli
Dialoq
Waveworks
Lexii.ai
Landbot
ChatterMe
Andi
verloop.io
Aichatting
Mindy
Perplexity
Oystr
AI Athlon
Teach Anything
Data Extraction & Conversational Search
upword
The Drive AI
Ayfie
WISESPARKLE
FileGPT
ARCHES AI
PDF.ai
context
Asktro
scrolai
Unify Insight
kadoa
Fabric
Rtutor
Elicit
Corpora
TableTalk
Searchable.ai
EnhanceDocs
Developer/ Coding Assistants
GitHub Copilot
Codebeat
phind
CodeAssist
Refraction
LLM API Build Frameworks
LangFlow
louie.ai
Steamship
haystack
Vercel
stack
ChatBotKit
Flowise
DUST
Chatbase
FIXIE
PromptChainer
Humanloop
NLP Cloud
Relevance AI
Voiceflow
rectune
botpress
LangChain
DeepPavlov
auto
Prompt Engineering Tools
LangSmith
PROMPTABLE
DRAFTER
Portkey
prism
ChainForge
Spellbook
Everyprompt
vellum
PROMPTSTACKS
Baseplate
Dyno
FLUX
LLM Development Tools
Vector Stores / Databases
milvus
Chroma
Qdrant
Vald
vespa
OpenSearch
pgvector
redis
MongoDB Atlas
vectara
Pinecone
Elasticsearch
Supabase
FAISS
Weaviate
Data Centric Tooling
HumanFirst
scale
Snorkel
Models & Hubs
Foundation LLM Models
OpenAI
Goose AI
Anthropic
cohere
AI21labs
Google Cloud
Open-Sourced LLM Models
BLOOM
openai/whisper
EleutherAI
Meta AI
Code Models
BigCode
Model Hubs
Hugging Face
GitHub
Foundation Models & Hubs
www.cobusgreyling.com
Version 4.0

GPU kapacitás bérlés

- Teljes platform kézben tartható
- Továbbra is nyílt modellekkel
- Drága…

Software-as-a-service

- Legolcsóbb megoldás
- Több szolgáltató is biztosít lehetőséget
- Teljes alkalmazási környezet, API
  - OpenAI, Google, Microsoft Azure, …
- Legtöbbször elérhető RAG és fine-tuning is
- Az árazás token alapú

Provider	Model	Context	Input/1k Tokens	Output/1k Tokens	Per Call	Total
Chat/Completion						
OpenAI	GPT-3.5 Turbo	16k	$0.0005	$0.0015	$0.0020	$0.20
OpenAI	GPT-4 Turbo	128k	$0.01	$0.03	$0.0400	$4.00
OpenAI	GPT-4o (omni)	128k	$0.005	$0.015	$0.0200	$2.00
OpenAI	GPT-4o mini	128k	$0.00015	$0.0006	$0.0007	$0.07
OpenAI	GPT-4	8k	$0.03	$0.06	$0.0900	$9.00
OpenAI	GPT-4	32k	$0.06	$0.12	$0.1800	$18.00
OpenAI	GPT-3.5 Turbo	4k	$0.0015	$0.002	$0.0035	$0.35
Mistral AI (via Anyscale)	Mixtral 8×7B	32k	$0.0007	$0.0007	$0.0014	$0.14
Mistral AI	Mistral Small	32k	$0.002	$0.006	$0.0080	$0.80
Mistral AI	Mistral Large	32k	$0.008	$0.024	$0.0320	$3.20
Meta	Llama 2 70b	4k	$0.001	$0.001	$0.0020	$0.20
Meta	Llama 3.1 405b	128k	$0.003	$0.005	$0.0080	$0.80

Software-as-a-service – fine-tuning

Models	Training per 1,000 tokens	Hosting per hour	Input Usage per 1,000 tokens	Output Usage per 1,000 tokens
Babbage-002	$0.0004	$1.70	$0.0004	$0.0004
Davinci-002	$0.006	$1.70	$0.002	$0.002
GPT-3.5-Turbo (4K)	$0.008	$1.70	$0.0005	$0.0015
GPT-3.5-Turbo (16K)	$0.008	$1.70	$0.0005	$0.0015

TMiT

Köszönöm a figyelmet!