Sztuczna Inteligencja
Duże modele językowe

Włodzisław Duch

Katedra Informatyki Stosowanej UMK

Google: Włodzisław Duch

Co było:

- Analiza języka w stylu GOFAI ma tylko historyczne znaczenie, jesteśmy w epoce dużych modeli językowych (LLM) i multimodalnych.

- Sieci neuronowe i algorytmy językowe.
- Transformery
- Inne techniki.

Rozumowanie i rep. lingwistyczne

W 2022 roku pojawiły się duże modele językowe GPT 3 i późniejsze, oparte na transformerach. Wcześniej próbowano wielu pomysłów, ale nic dobrze nie działało, np.

- **Metody statystyczne**: Facebook Blender oparty o analizę 1.5 mld dialogów. 49% pytanych woli chatbota od ludzi … ~10 mld parametrów w sieci NN.
- Inspiracje kognitywne: pamięć rozpoznawcza, semantyczna, epizodyczna pozwalająca tworzyć model dyskursu.
- Podejście koneksjonistyczne, głębokie uczenie i symboliczne do rozumienia, uczenia się i planowania.
- Uczenie się związków przyczynowych.
- Integracja nowej i starej wiedzy.
- Planowanie i argumentacja.
- Modele rozumienia argumentów w tekstach gazetowych.
- Symulacja wyobraźni, procesu marzenia na jawie.
- Odwzorowanie wiedzy na reprezentację tekstową, tworzenie zdań.

Linguboty, wirtualni asystenci

Stare boty (chatterbots):

Agentland | Chatbots.org (all world)| Personality forge, chatterbot collection | Simon Laven bots, many! |

Zabaware Ultrahall Assistant to mój ulubiony bot!

Polskie boty: katalog, K2bots w biznesie, voiceboty i chatboty,

Chatboty w Polsce (artykuł)

Linguboty i modele językowe

Technologia: większość botów opierała się na rozszerzeniu techniki szablonów, np. w oparciu o język AIML, często używanego przez chatterboty. Nowsze wykorzystują sieci LMM.

- Amazon: Evi i Alexa od 7/12 po polsku!
- Microsoft Cortana i Apple Siri
- Google Now i Google Assistant i Google Meena
- Samsung Viv i Bixby
- Facebook Blender, bot z analizą emocji, na Github, otwarty projekt.
- Medical prototypes, DARPA's Detection and Computational Analysis of Psychological Signals (DCAPS) project http://ict.usc.edu/prototypes/current/
- Leena AI is an AI-powered HR Assistant https://leena.ai/case-studies
- Lista polskich botów.
- BotPrize

Metody sekwencyjne: LSTM

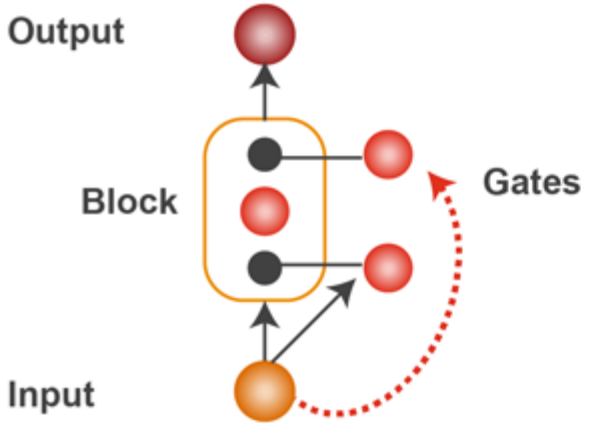
Hochreiter, Schmidhuber, Long short-term memory (1997).

Sieć rekurencyjna (RNN) pozwala na przechowanie wektora w pamięci roboczej.
Trzy bramki:
- bramka wejściowa: w jakim stopniu wejście wpływa na stan komórki,
- bramka zapominania: decyduje o resetowaniu stan komórki
- bramka wyjściowa: jaką część bieżącego stanu przekazać na wyjście.

Liczne warianty, rozpoznawanie mowy, pisma, muzyki …

Long short-term memory & gated recurrent unit neural networks

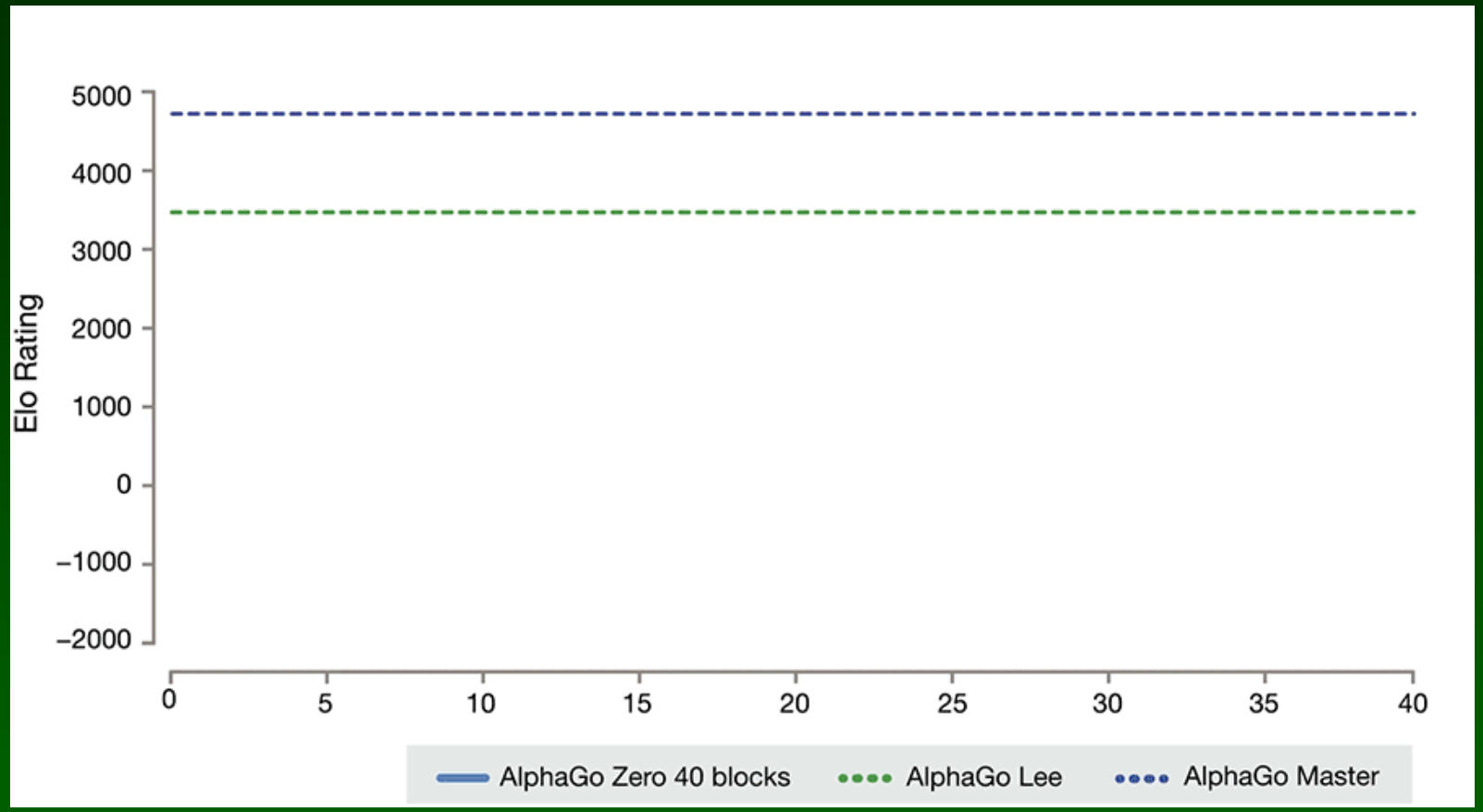
Older forms of RNNs can be lossy. While these older recurrent neural networks only allow small amounts of older information to persist, newer long short-term memory (LSTM) and gated recurrent unit (GRU) neural networks have both long- and short-term memory. In other words, these newer RNNs have greater memory control, allowing previous values to persist or to be reset as necessary for many sequences of steps, avoiding "gradient decay" or eventual degradation of the values passed from step to step. LSTM and GRU networks make this memory control possible with memory blocks and structures called gates that pass or reset values as appropriate.



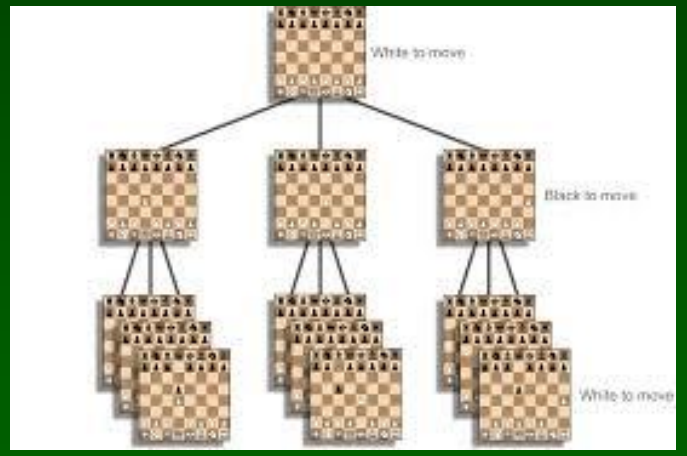
Source: Genevieve Orr, et al., Williamette University, 1999

Advantages	Use cases
Long short-term memory and gated recurrent unit neural networks have the same advantages as other recurrent neural networks and are more frequently used than other recurrent neural networks because of their greater memory capabilities.	Natural language processing, translation

AlphaGo Zero: nadludzka inteligencja

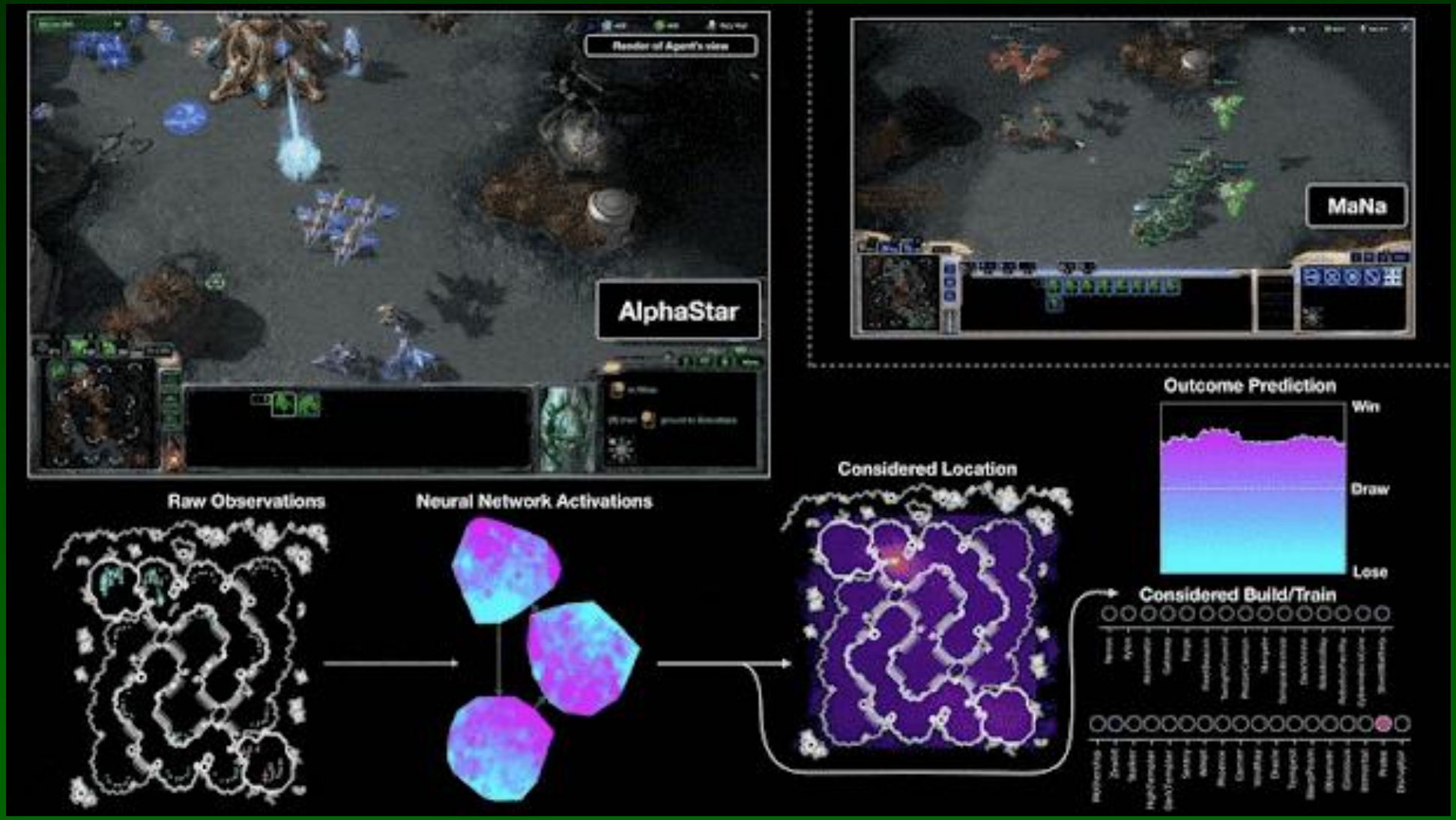


1997, Deep Blue-Kasparov.
2016, AlphaGo-Lee Sedol
2017, Alpha GoZero
2017, Poker, Dota 2;
2019, Starcraft II,
2022 Stratego, Diplomacy



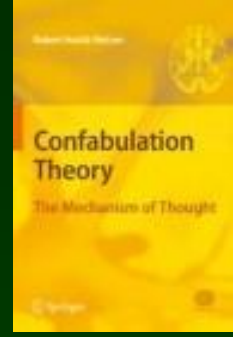
Nadludzki poziom w Go. Najlepsi ludzie 3600-3800 ELO, GoZero 5500 ELO, KataGO ponad 6000 ELO.
MC Tree Search + głęboka ocena wzorców NN jako heurystyka + uczenie ze wzmocnieniem, gra przeciwko własnej kopii. Ludzka wiedza staje się nieistotna, uczenie ludzkich strategii obniża poziom końcowy!
Czysta intuicja: Ruoss … & Genewein, T. (2024). *Grandmaster-Level Chess Without Search.*
Blitz Chess: transformer model, 270M parametrów, decyzja skojarzneiowa, w jednym kroku, bez szukania!

AlphaStar Strategic Game



Deepmind.google alphastar – mastering the real-time strategy game Starcraft-II (2019)
Open, strategic games, like war games.

BERT uczy się języka



Modele językowe: relacje między wektorami(tokenów) w złożonych strukturach sieciowych.
W 2018 roku, aby uzyskać uniwersalne „rozumienie języka”, firma Google stworzyła model BERT, Bidirectional Encoder Representations from Transformers, wstępnie wyszkolony na bardzo dużym korpusie tekstowym. To technika uczenia maszynowego oparta na transformerach.

BERT w języku angielskim: dwie sieci, mniejsza o 110 mln parametrów, większy model o 340 mln parametrów w 24 warstwach; przeszkolony na korpusie BooksCorpus zawierającym 800 mln słów oraz Wikipedii zawierającej 2500 mln słów. W 2019 r. BERT działał już w 70 językach.

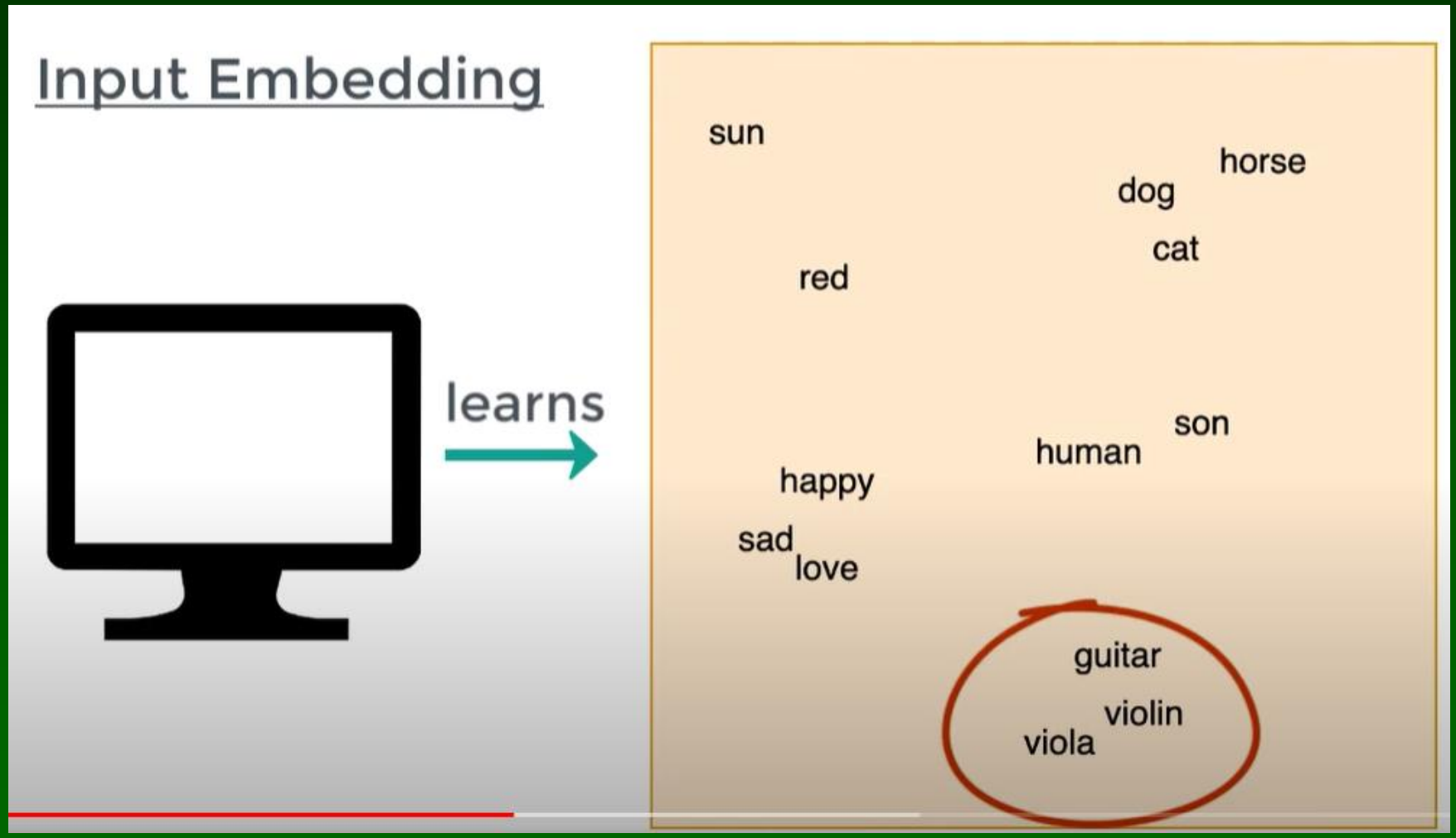
BERT został następnie dostosowany do konkretnych zadań NLP, takich jak odpowiadanie na pytania lub wyszukiwanie informacji semantycznych. Wiele mniejszych, wstępnie wyszkolonych modeli otwartego oprogramowania zostało opublikowanych w repozytorium GitHub.

Sieć uczy się przewidywać zamaskowane słowa (obrazy, sygnały), np:
Dane wejściowe: mężczyzna poszedł do [MASK1]. Kupił [MASK2] mleka.
Etykiety: [MASK1] = sklep; [MASK2] = galon.

Hecht-Nielsen, w Confabulation Theory (2007), proponował przewidywanie kolejnych słów, ale bez dobrego zagnieżdżania (embedding) i mechanizmu uwagi (relacji słów) to nie działało.

Embeddings

Words => vectors, reflecting their similarity and positions in sentences.



Word2Vec

Jak zamienić dyskretne elementy – słowa – na wektory,
które odzwierciedlają sens słów?

https://learnprompting.thinkific.com/collections

https://learnprompting.org/docs/agents/introduction

https://evakeiffenheim.substack.com/p/how-to-combine-llms-for-your-breakthrough

https://learnprompting.org/docs/agents/mrkl

Modular Reasoning, Knowledge, and Language or MRKL Systems

LLMs that Reason and Act https://learnprompting.org/docs/agents/react

Code as Reasoning https://learnprompting.org/docs/agents/pal

Alternatywy: Glove, FastText.

Dobre omówienie: Advanced Word Embeddings

Emeddings and attention

Transformer model published by Google in June 2017 started the generative AI era.

A key concept of the architecture is self-attention to understand relationships between words.

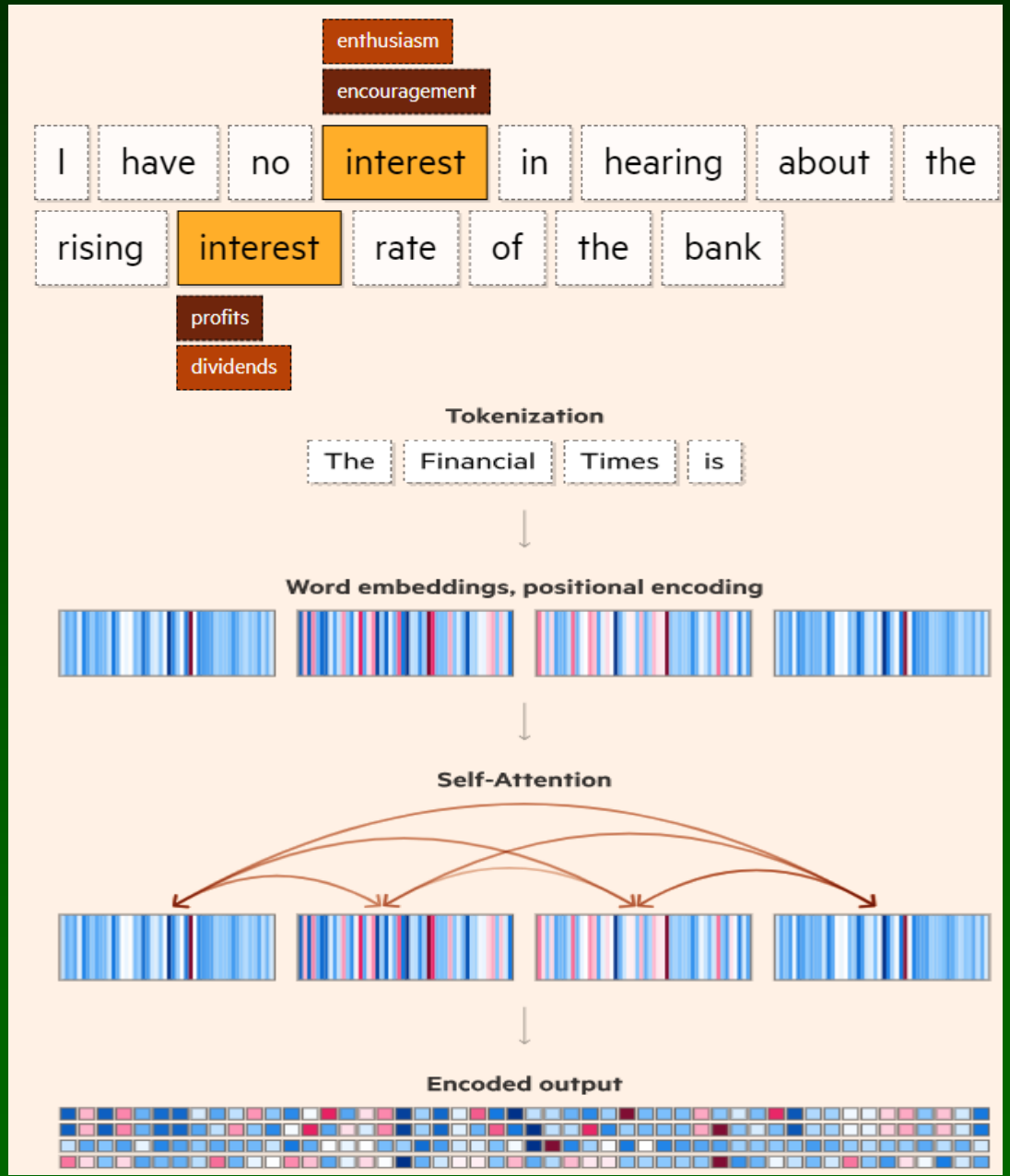
Self-attention links each token in text to other tokens important to understand its meaning.

- **Q (Query)** — *czego szukam?*
- **K (Key)** — *co oferuję do dopasowania?*
- **V (Value)** — *jaką informację przekażę, jeśli będę wybrany?*

Mechanizm attention działa tak:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Matykiewicz P, Pestian J, Duch W, and Johnson N. (2006) *Unambiguous Concept Mapping in Radiology Reports: Graphs of Consistent Concepts*, AMIA Ann. Symp Proc. 1024.



Transformers

Attention: given a sequence of tokens (words, image patches), how relevant is each input token to other tokens?

Attention vectors capture context (embedding, semantics) + encode relative positions (syntax) of words.

Example:

Input: sentence in English;
Output: sentence in Polish.

Google BERT used this approach.

Generative Pre-trained Transformers or GPTs are now best known models**.**
Simple intro on Youtube.

Vaswani et al.(2017).
Attention Is All You Need. arXiv

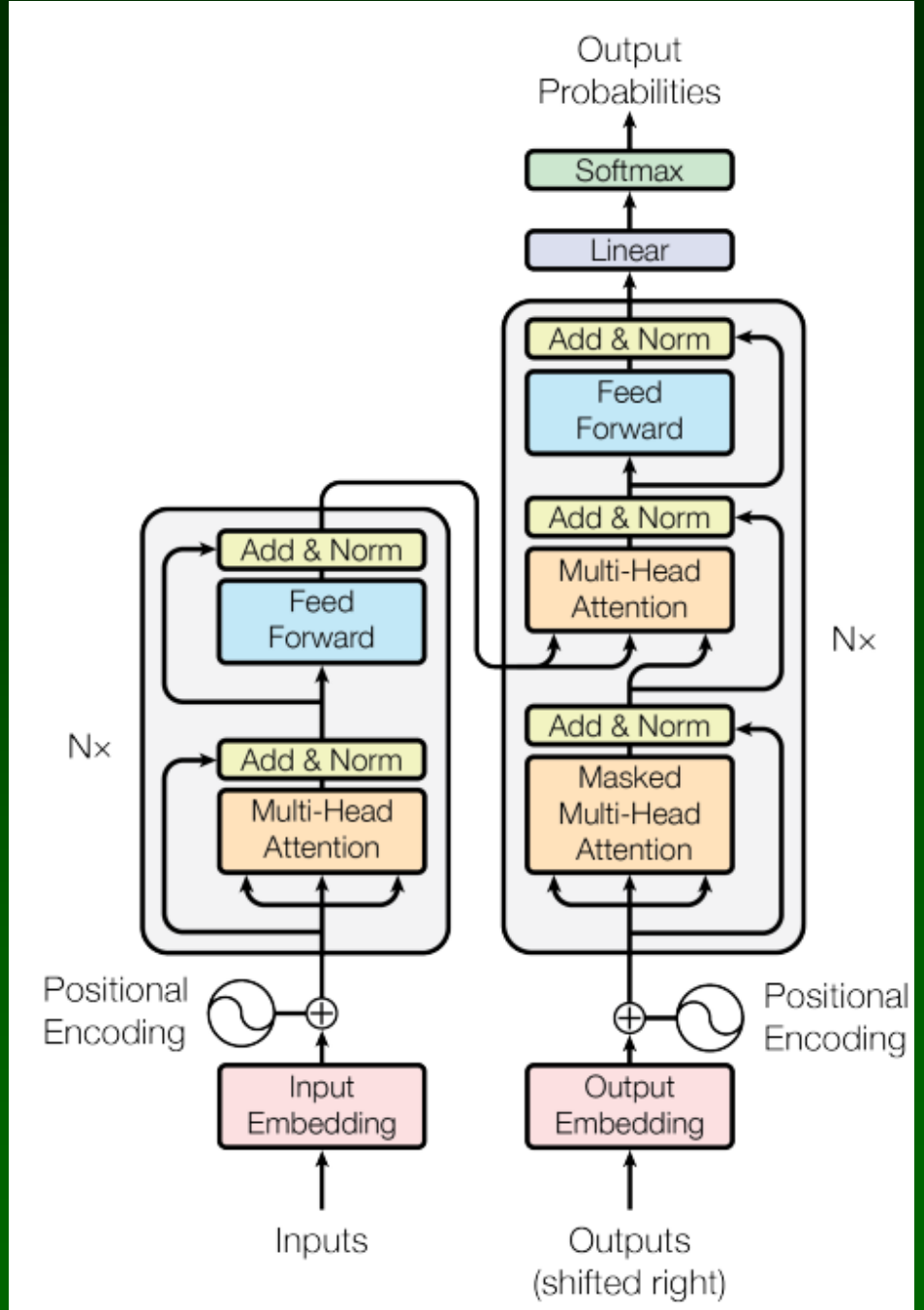
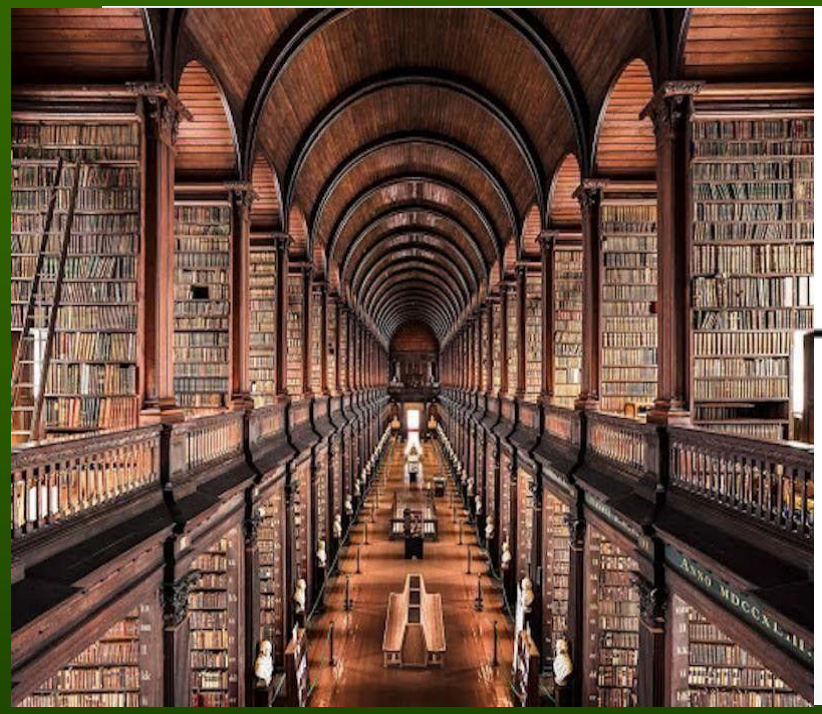
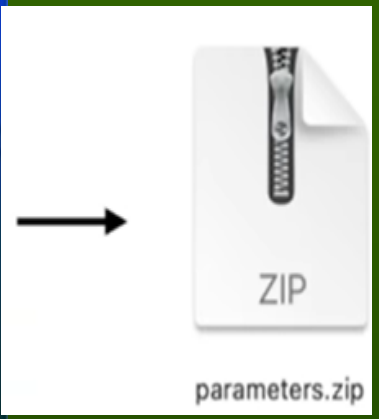


Figure 1: The Transformer - model architecture.

LLM/LMM: kompresja informacji





Trening >10 000 GB tekstów

12 dni pracy 6000 kart GPU V100
2 mln $, 10^{24} mld mld mld operacji.

parametry ≈ 70 GB

Llama 70B, kompresja 142 razy, to umożliwia odległe skojarzenia, które trudno jest odkryć.
Kompresja 8, 4, 2 bity => większa kreatywność, luźne skojarzenia, więcej halucynacji.
Więcej szumu (wyższa temperatura) => większa kreatywność, więcej halucynacji.

Kontekst, RAG, prompty: aktywują interesujące nas obszary wiedzy.

Słowa, obrazy, myśli pobudzają mózgi i LLMy

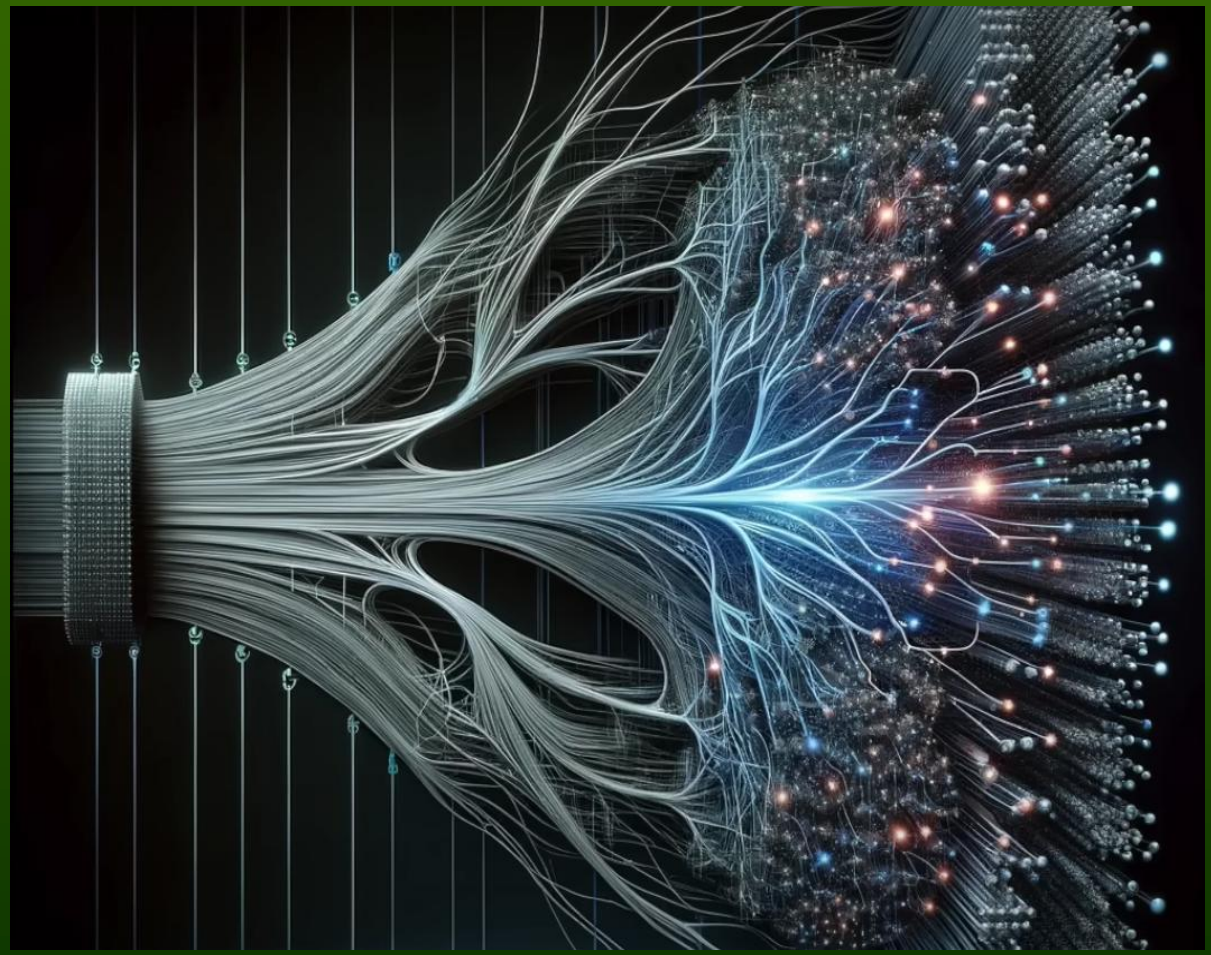
GPT (2017) = Generative Pre-trained Transformers, czyli generatywne, wstępnie nauczone transformery.

Sieci rozchodzących się aktywacji, transformacji.

Dane wejściowe => tokeny => osadzanie (embedding) w przestrzeni wektorowej tak, by zachować podobieństwo znaczenia (semantyki) w różnych kontekstach.

Zwracaj uwagę na powiązania między wszystkimi tokenami, dodaj taką informację do ich reprezentacji, użyj takich wektorów by nauczyć sieć odpowiadać na pytania. Wybierz podgrafy skojarzonych tokenów.

LLM visualization, https://bbycroft.net/llm



Duch et al., Towards Understanding of Natural Language: Neurocognitive Inspirations. 2007

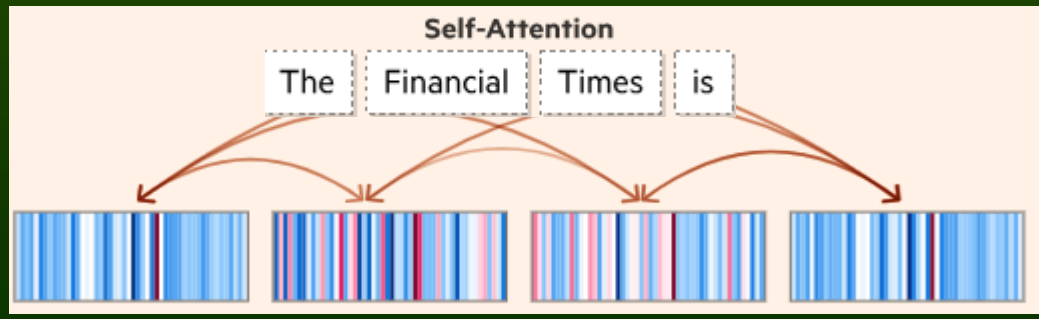
Jak działają LLM?

How transformers work (The Financial Times + visual storytelling)

- Transformery: w 2017 roku Google zaczął rewolucję w AI przedstawiając nowe podejście do sieci neuronowych. Publikacja Attention is all you need, pokazała jak budować programy analizujące teksty w oparciu o mechanizm uwagi, łączący różne słowa.
- Wizualizacja na stronach FT pokazuje jak działają transformery na prostych przykładach, dlaczego pozwoliło to chatbotom tak dobrze opanować język.

- LLM mapują biliony słów na wielkie tablice liczb, osadzając tokeny w wielowymiarowych przestrzeniach wektorowych, nadając im sens, kodując cechy i relacje. To pozwala przewidzieć następne słowo w dowolnej sekwencji, ale też wiarygodność całego zdania.
- Algorytmy są dość proste, liczy się wielkość sieci i zbiorów treningowych.

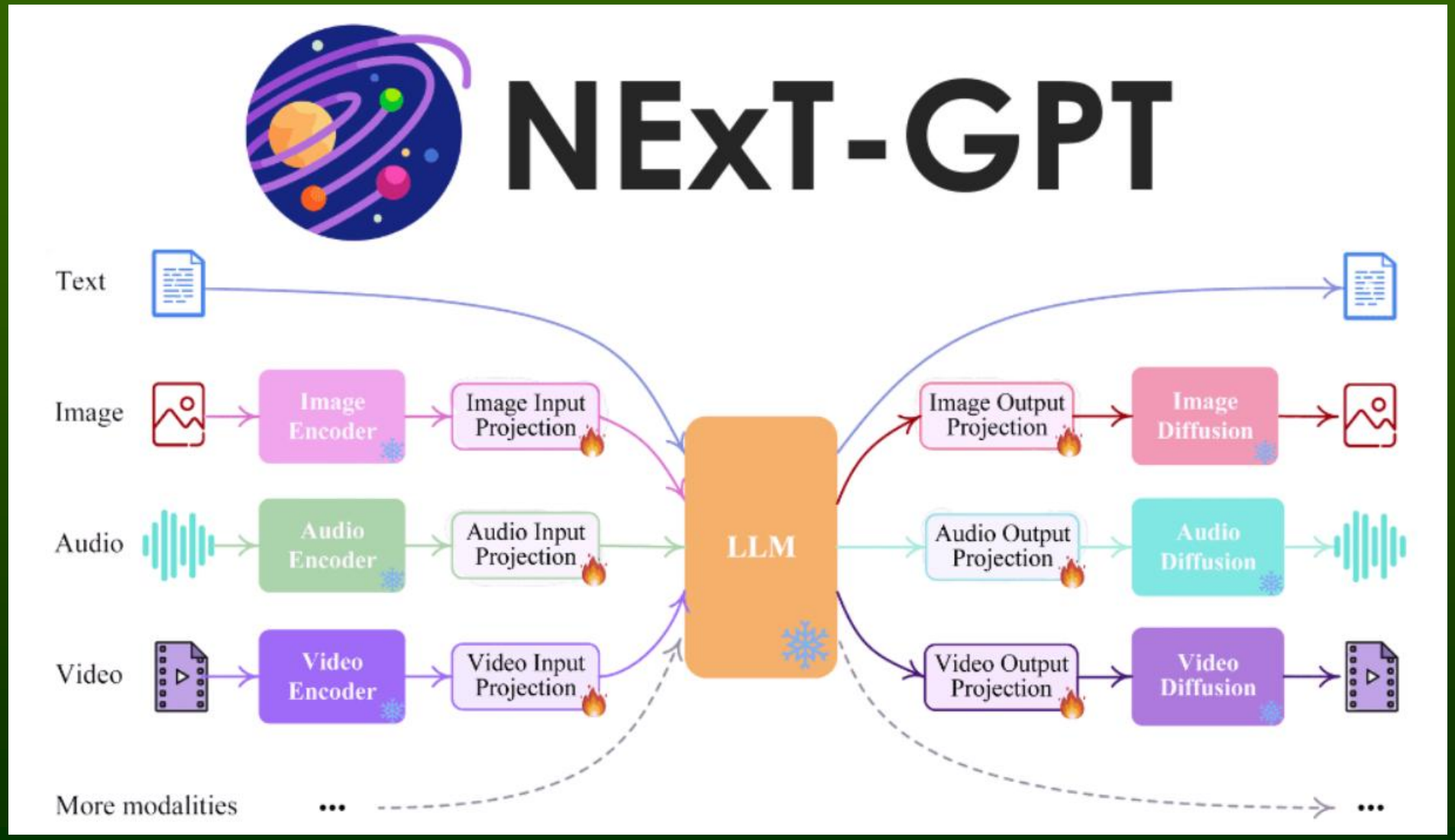
Język jest kluczem do wszystkiego:
obrazy, sygnały, ruch robotów, DNA, białka …



- Andrej Karpathy: Intro to Large Language Models

GPT skojarzy wszystko

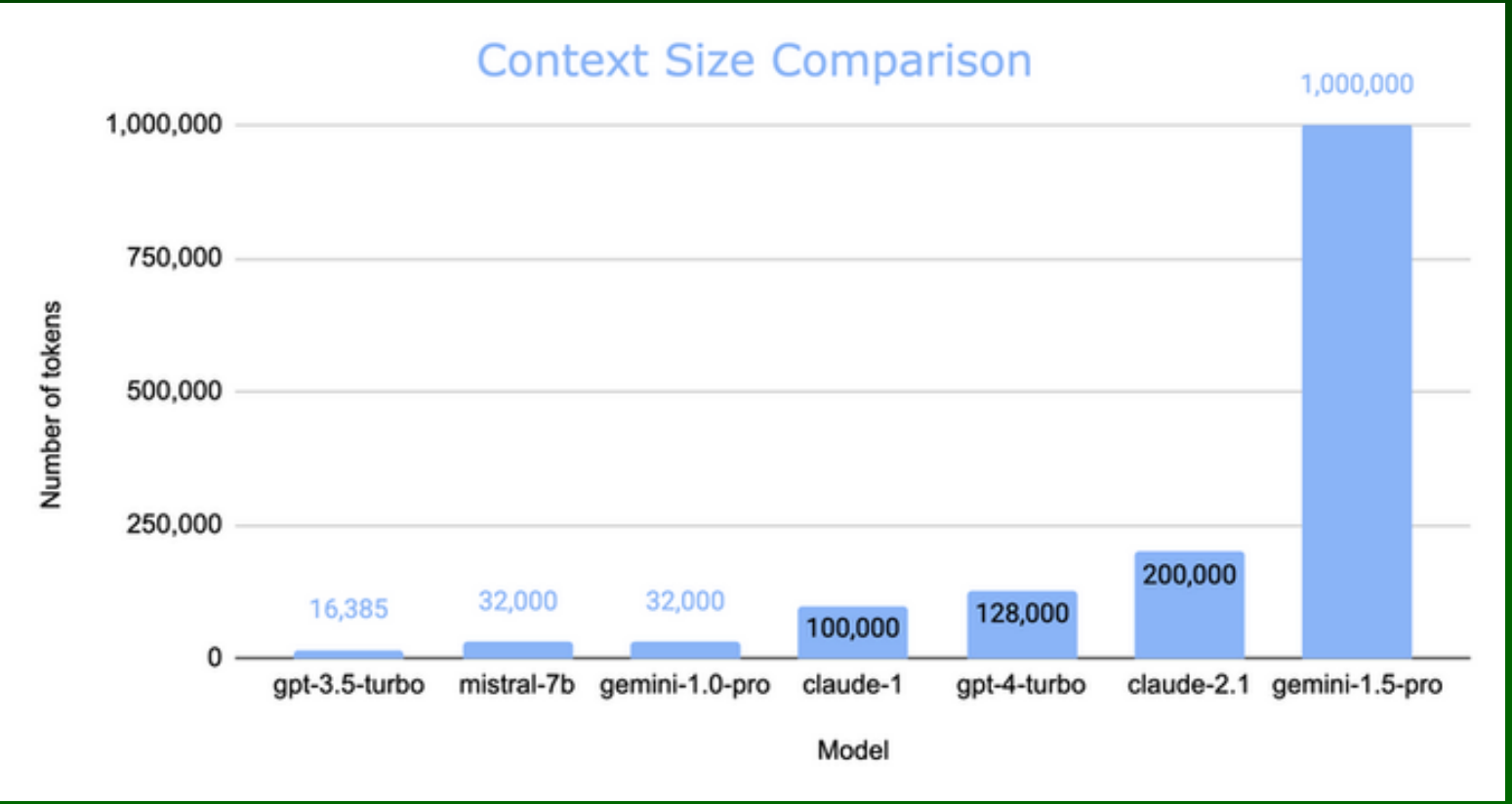
Substytucja zmysłów: cokolwiek wrzucimy do mózgu będzie wykorzystane.



Kontekst

Modele LLM są trenowane na wybranym zbiorze tekstów, modele LMM na danych obrazowych, wideo, sygnałach – to tworzy pamięć długotrwałą (ale nie wierną).

Prompt dostarcza pytanie i kontekst, dokładne instrukcje czego oczekujemy od danego modelu.
W niektórych modelach kontekst może osiągnąć ponad 100.000 tokenów, np. wiele tomów akt.
Google Notebook LM pozwala na 2.5 mln, Llama Scout 10 mln.



Fizyka LLM

Znakomite wyjaśnienie tego, jak działaja współczesne LLMy, na poziomie ogólnych zasad.

YT Wideo: The Hidden Physics of LLMs: Retrieval as Thermodynamics

Understanding LLMs. A Physics-Inspired Visual Journey

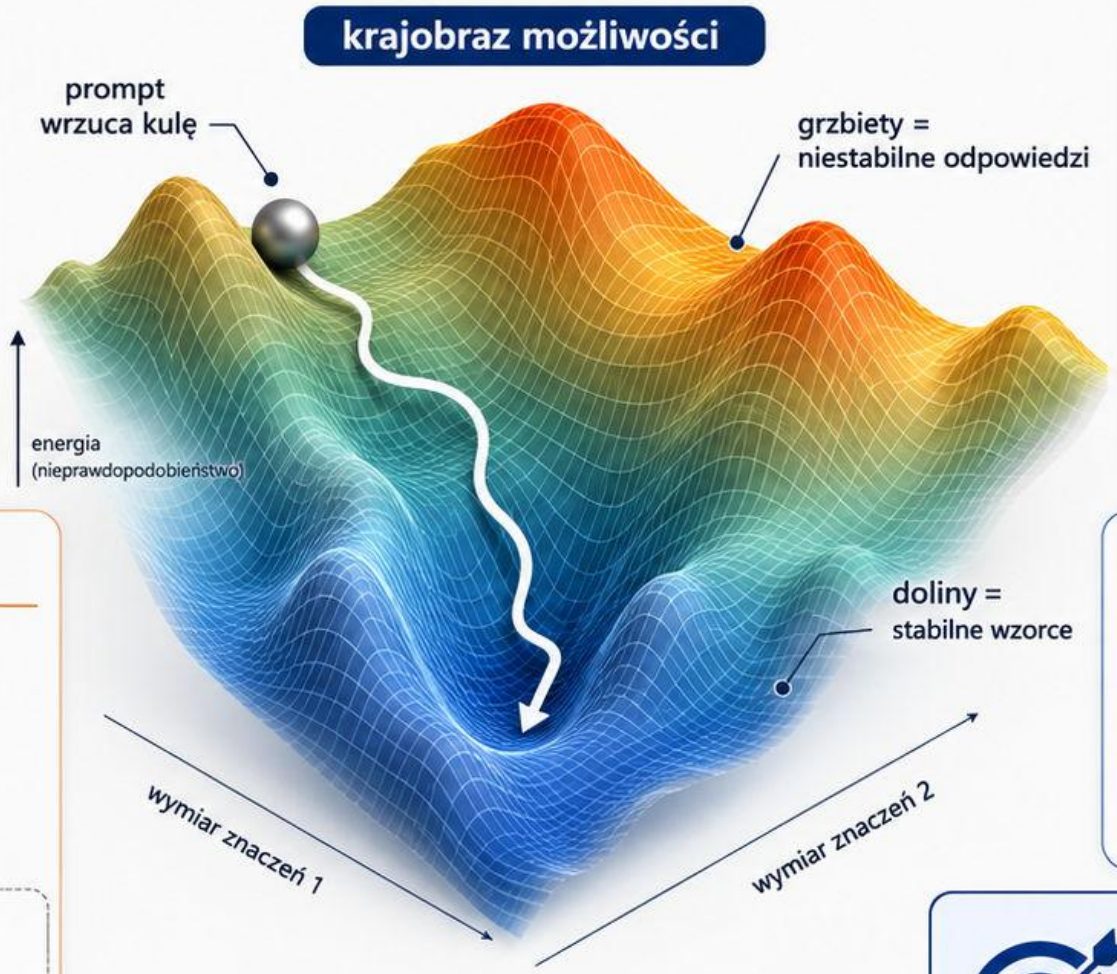
https://stable-raid-01017690.figma.site/

LLM jako układ fizyczny

Dlaczego metafora „autccomplete" nie wystarcza

1. Metafora krajobrazu

- Wagi modelu kształtują teren odpowiedzi
- Prompt ustawia punkt startu
- Generacja = trajektoria do bardziej stabilnych stanów



2. Fizyka generowania

Attention = sterowanie uwagą	co ma wpływ na następny token
Softmax = rozkład prawdopodobieństwa	niższa „energia" → większa szansa
Temperature = „cieplna" swoboda ruchu	niska: spójnie | wysoka: kreatywnie / chaotycznie

3. Ugruntowanie i halucynacje

- RAG działa jak zewnętrzne pole: przechyla krajobraz ku faktom
- Halucynacje pojawiają się przy słabych kotwicach i zbyt płaskim krajobrazie
- Wysoka temperatura zwiększa ryzyko nieugruntowanych odpowiedzi

Odpowiedź ugruntowana	VS.	Halucynacja
kotwice: dane, źródła, kontekst → stabilna dolina		słabe kotwice, płaski krajobraz → dryf na grzbiety

4. Wniosek

- LLM minimalizuje coś w rodzaju swobodnej energii
- Szuka odpowiedzi zgodnej z promptem i statystycznie stabilnej
- Lepsze wyniki dają: kontekst, ograniczenia, źródła

minimalizacja swobodnej energii

Jak sterować odpowiedzią modelu?

1. Doprecyzuj cel
2. Dodaj kontekst i źródła
3. Dobierz temperaturę

Polskie projekty

Stworzenie dużego modelu fundacyjnego jest bardzo drogie,
ale są otwarte modele, które można łatwiej douczać.

Projekt PLLuM (Polish Large Language Model), OPI, NASK,
IPI PAN, Pol Wrocławska, Uniw. Łódzki.
Planowane jest wykorzystanie w administracji publicznej i sektorze prywatnym,
np. w postaci prototypowego inteligentnego asystenta.

BIELIK, projekt budowany oddolnie, efekt pracy społeczności entuzjastów sztucznej inteligencji.

Qra to odpowiednik otwartych narzędzi Mety czy Mistral AI. Lepiej rozumie pytania i treści w języku polskim i lepiej sama tworzy spójne teksty.
LLama2 90B posłużyła jako start dla projektu Qra. Politechnika Gdańska i
AI Lab z (OPI) – PIB, opracowały polskojęzyczne generatywne neuronowe modele językowe na bazie terabajta danych tekstowych wyłącznie w języku polskim.

PLAMA to projekt ogłoszony 25.04.2024.

Kompleksowa Lista Ewaluacji polskich modeli Językowych KLEJ.

Świetne kursy

- UWaterloo CS 886 papers/slides/videos ! - Recent Advances on Foundation Models.
- Stanford CS25-Transformers United, video
- Princeton COS 484 NLP Course (2025)
- Princeton COS 597R Deep Dive into Large Language Models (Fall 2024)
- Maxime Labonne LLM-course + Colab notebooks
- Darmowy kurs "Illustrated transformer"
- Stanford CS324 - Large Language Models (2022)
- Neurips2022 Tutorial-Foundational Robustness of Foundation Models
- ICML2022-"Big Model" Era: Techniques and Systems (Berkeley)

Inne linki

- Andrej Karpathy wyjaśnia |
  - Let's build GPT: from scratch, in code, spelled out
  - GPT Tokenizer
  - Why you should work on AI AGENTS
- LLM Powered Autonomous Agents
- Umar Jamil Series, kilka modeli - edukacyjne wideo o nowych modelach (Titans).
- Alexander Rush projekty OpenSource.
- A Visual Guide to Transformers, Mamba and State Space Models (Maarten Grootendorst)
- Create GPT in 60 Lines of NumPy
- femtoGPT, minimal GPT implementation Pure Rust

Książki, prompty

- Prompt engineering examples
- **LLM Books**
  - Generative AI with LangChain: Build large language model (LLM) apps with Python, ChatGPT, and other LLMs - it comes with a GitHub repository that showcases a lot of the functionality
  - Build a Large Language Model (From Scratch) - A guide to building your own working LLM.
  - BUILD GPT: HOW AI WORKS - explains how to code a Generative Pre-trained Transformer, or GPT, from scratch.
  - Hands-On Large Language Models: Language Understanding and Generation - Explore the world of Large Language Models with over 275 custom made figures in this illustrated guide!

Książki, prompty

- Prompt engineering examples
- **LLM Books**
  - Generative AI with LangChain: Build large language model (LLM) apps with Python, ChatGPT, and other LLMs - it comes with a GitHub repository that showcases a lot of the functionality
  - Build a Large Language Model (From Scratch) - A guide to building your own working LLM.
  - BUILD GPT: HOW AI WORKS - explains how to code a Generative Pre-trained Transformer, or GPT, from scratch.
  - Hands-On Large Language Models: Language Understanding and Generation - Explore the world of Large Language Models with over 275 custom made figures in this illustrated guide!