# Probabilistic generative models

Advanced LLM

# Overview of the Course

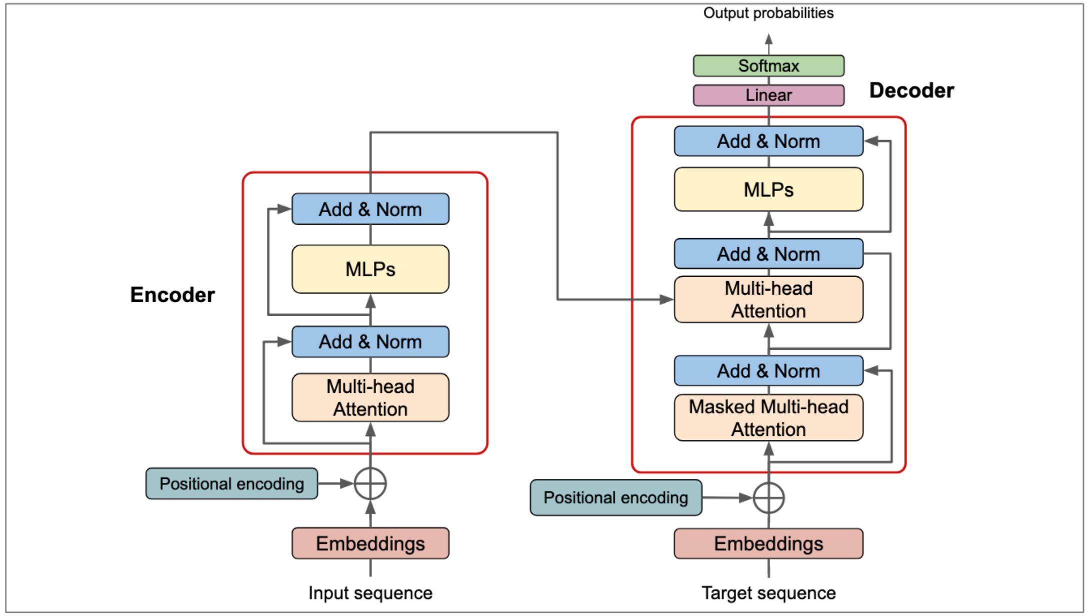
**Objectives :**

- **Understand the Architectural Advances:** Delve into the technical enhancements and scaling of GPT-4 over its predecessors.
- **Explore GPT-4o1:** Learn about the optimized variant GPT-4o1, focusing on its specific improvements for efficiency and task-specific performance.
- **Real-World Applications:** Examine how these models are being applied across various industries and their impact.
- **Practical Interaction:** Engage with the models through a live demonstration to understand their practical usage.

**Expected Outcomes:** By the end of this course, students will be equipped with a deep understanding of GPT-4 and GPT-4o1, appreciating not only the technological advancements but also the broader societal implications.

# Introduction to Transformer Models and the Evolution to GPT-4

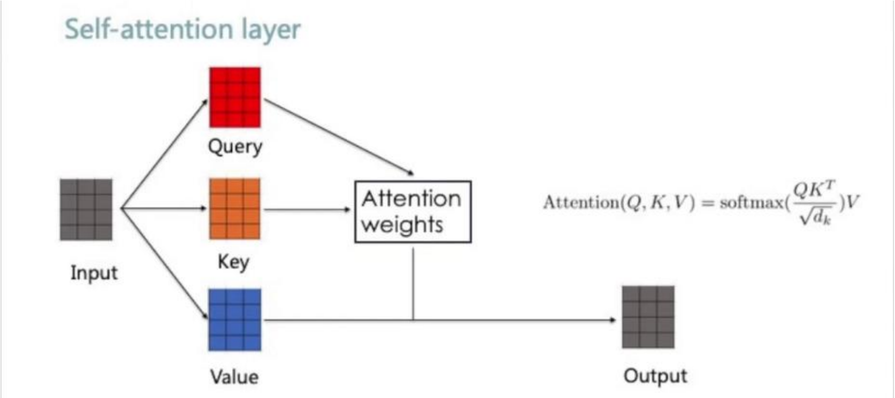
- **The Transformer Legacy:**
  - **Birth of Transformers (2017):** Introduced in the paper "Attention is All You Need", transformers revolutionized how machines understand and generate human-like text by relying heavily on self-attention mechanisms.
  - **Evolution:** Progression through various models like BERT, GPT-2, and GPT-3, each introducing significant improvements in handling longer sequences and understanding context more profoundly.



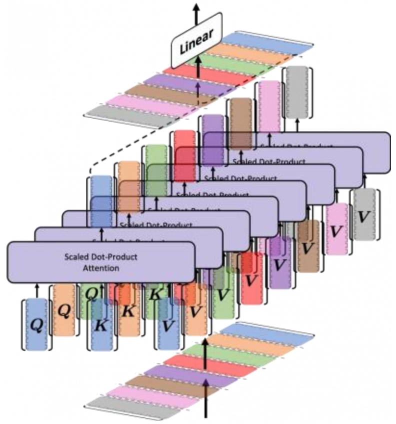

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 30*, 5998-6008.

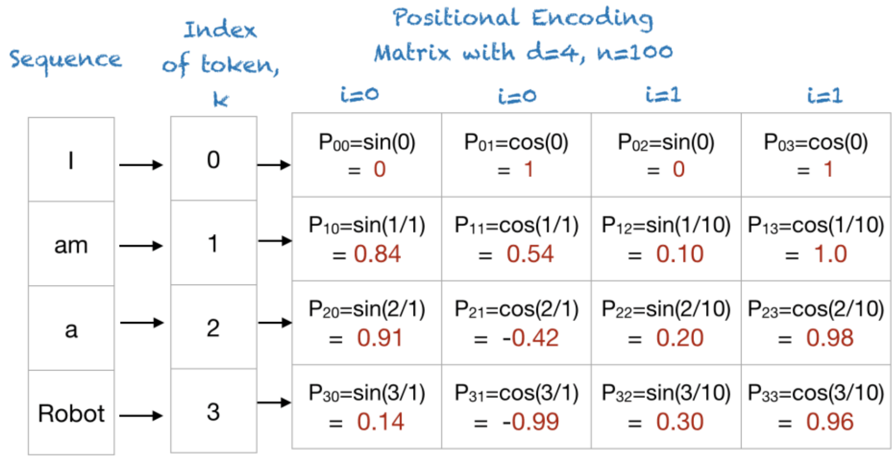
# Introduction to Transformer Models and the Evolution to GPT-4

- **Core Concepts:**
  - **Self-Attention Mechanism:** Allows models to weigh the importance of different words irrespective of their input positions.
  - **Multi-Head Attention:** Enhances the model's ability to focus on different positions of the input sequence simultaneously.
  - **Positional Encoding:** Provides a notion of word order, enabling the model to consider the sequence of processed words.




https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html




Positional Encoding Matrix for the sequence 'I am a robot'

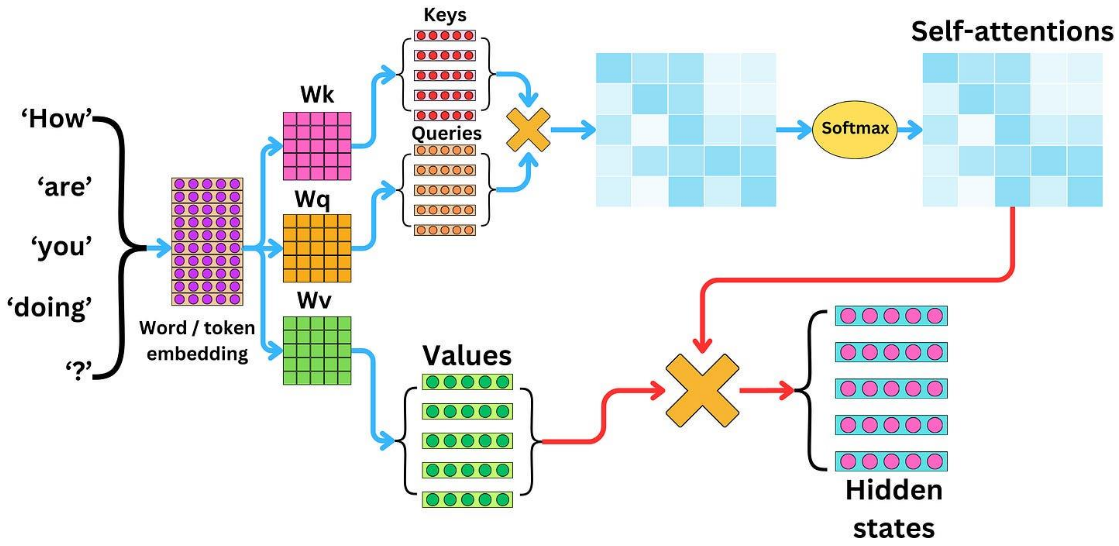
# Introduction to Transformer Models and the Evolution to GPT-4

- **Self-Attention Mechanism:**
  - To transform an input token into the Query (Q), Key (K), and Value (V) vectors in a Transformer architecture, the following operations are carried out based on the token's embedding. Let ( E ) be the embedded input of the token, and ( W^Q ), ( W^K ), and ( W^V ) be the trainable weight matrices for Queries, Keys, and Values, respectively. The equations to transform ( E ) into Q, K, V are as follows:

$$[Q = W^Q \times E][K = W^K \times E][V = W^V \times E]$$

  - Where:
    - ( Q ) is the resulting Query vector, used to assess the relative importance of other tokens in comparison to it.
    - ( K ) is the resulting Key vector, used to compute attention scores with the Query vectors of other tokens.
    - ( V ) is the resulting Value vector, whose components are weighted by the attention scores to produce the final output of the attention mechanism.

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




Decoder Q

| | sos->wir | wir->konnen | konnen->apfel | apfel->essen | essen->eos |
|---|---|---|---|---|---|
| sos->wir | 1.00 | 0.25 | -0.09 | -0.14 | -0.33 |
| wir->konnen | 0.25 | 1.00 | -0.30 | -0.09 | -0.27 |
| konnen->apfel | -0.09 | -0.30 | 1.00 | -0.16 | -0.18 |
| apfel->essen | -0.14 | -0.09 | -0.16 | 1.00 | 0.19 |
| essen->eos | -0.33 | -0.27 | -0.18 | 0.19 | 1.00 |

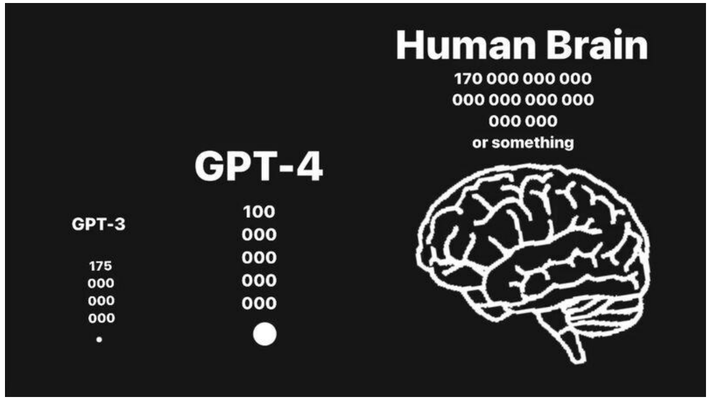
Vector : decoder_encoder_attention@K

| | sos | apple | eat | can | we | eos |
|---|---|---|---|---|---|---|
| sos | 1.00 | 0.46 | 0.29 | 0.10 | -0.26 | 0.65 |
| we | -0.30 | -0.27 | 0.09 | -0.29 | 0.99 | -0.19 |
| can | 0.09 | -0.03 | 0.09 | 1.00 | -0.27 | -0.05 |
| eat | 0.32 | 0.45 | 1.00 | 0.06 | 0.08 | 0.33 |
| apple | 0.48 | 0.99 | 0.41 | -0.02 | -0.25 | 0.65 |
| eos | 0.65 | 0.62 | 0.30 | -0.04 | -0.16 | 1.00 |

decoder encoder_attention@V

| | sos | apple | eat | can | we | eos |
|---|---|---|---|---|---|---|
| sos | 1.00 | 0.24 | 0.16 | 0.16 | -0.23 | 0.50 |
| we | -0.26 | -0.14 | -0.07 | -0.16 | 0.99 | -0.36 |
| can | 0.15 | 0.01 | 0.15 | 1.00 | -0.12 | 0.25 |
| eat | 0.18 | 0.19 | 1.00 | 0.12 | -0.10 | 0.21 |
| apple | 0.26 | 1.00 | 0.16 | 0.04 | -0.13 | 0.63 |
| eos | 0.50 | 0.60 | 0.20 | 0.26 | -0.34 | 1.00 |

# Introduction to Transformer Models and the Evolution to GPT-4




- **From GPT-3 to GPT-4:**
  - **Scale and Scope:** Introduction of GPT-4 marks a monumental scale-up in terms of trainable parameters (from 175 billion in GPT-3 to around 100 trillion in GPT-4), promising better understanding and generative capabilities.
  - **Training Data and Techniques:** Utilization of broader and more diverse datasets alongside advances in fine-tuning and training methodologies to enhance performance and efficiency.
- **Why GPT-4 Matters:**
  - Breakthroughs in language processing capabilities, setting new benchmarks in AI's ability to interact in human-like manners.
  - Introduction of novel applications in automation, content creation, and decision-making processes across industries.

| Exam | GPT-4 | GPT-4 (no vision) | GPT-3.5 |
|---|---|---|---|
| Uniform Bar Exam (MBE+MEE+MPT) | 298 / 400 (~90th) | 298 / 400 (~90th) | 213 / 400 (~10th) |
| LSAT | 163 (~88th) | 161 (~83rd) | 149 (~40th) |
| SAT Evidence-Based Reading & Writing | 710 / 800 (~93rd) | 710 / 800 (~93rd) | 670 / 800 (~87th) |
| SAT Math | 700 / 800 (~89th) | 690 / 800 (~89th) | 590 / 800 (~70th) |
| Graduate Record Examination (GRE) Quantitative | 163 / 170 (~80th) | 157 / 170 (~62nd) | 147 / 170 (~25th) |
| Graduate Record Examination (GRE) Verbal | 169 / 170 (~99th) | 165 / 170 (~96th) | 154 / 170 (~63rd) |
| Graduate Record Examination (GRE) Writing | 4 / 6 (~54th) | 4 / 6 (~54th) | 4 / 6 (~54th) |
| USABO Semifinal Exam 2020 | 87 / 150 (99th - 100th) | 87 / 150 (99th - 100th) | 43 / 150 (31st - 33rd) |
| USNCO Local Section Exam 2022 | 36 / 60 | 38 / 60 | 24 / 60 |
| Medical Knowledge Self-Assessment Program | 75 % | 75 % | 53 % |
| Codeforces Rating | 392 (below 5th) | 392 (below 5th) | 260 (below 5th) |
| AP Art History | 5 (86th - 100th) | 5 (86th - 100th) | 5 (86th - 100th) |
| AP Biology | 5 (85th - 100th) | 5 (85th - 100th) | 4 (62nd - 85th) |
| AP Calculus BC | 4 (43rd - 59th) | 4 (43rd - 59th) | 1 (0th - 7th) |
| AP Chemistry | 4 (71st - 88th) | 4 (71st - 88th) | 2 (22nd - 46th) |
| AP English Language and Composition | 2 (14th - 44th) | 2 (14th - 44th) | 2 (14th - 44th) |
| AP English Literature and Composition | 2 (8th - 22nd) | 2 (8th - 22nd) | 2 (8th - 22nd) |
| AP Environmental Science | 5 (91st - 100th) | 5 (91st - 100th) | 5 (91st - 100th) |
| AP Macroeconomics | 5 (84th - 100th) | 5 (84th - 100th) | 2 (33rd - 48th) |
| AP Microeconomics | 5 (82nd - 100th) | 4 (60th - 82nd) | 4 (60th - 82nd) |
| AP Physics 2 | 4 (66th - 84th) | 4 (66th - 84th) | 3 (30th - 66th) |
| AP Psychology | 5 (83rd - 100th) | 5 (83rd - 100th) | 5 (83rd - 100th) |
| AP Statistics | 5 (85th - 100th) | 5 (85th - 100th) | 3 (40th - 63rd) |
| AP US Government | 5 (88th - 100th) | 5 (88th - 100th) | 4 (77th - 88th) |
| AP US History | 5 (89th - 100th) | 4 (74th - 89th) | 4 (74th - 89th) |
| AP World History | 4 (65th - 87th) | 4 (65th - 87th) | 4 (65th - 87th) |
| AMC 10[3] | 30 / 150 (6th - 12th) | 36 / 150 (10th - 19th) | 36 / 150 (10th - 19th) |
| AMC 12[3] | 60 / 150 (45th - 66th) | 48 / 150 (19th - 40th) | 30 / 150 (4th - 8th) |
| Introductory Sommelier (theory knowledge) | 92 % | 92 % | 80 % |
| Certified Sommelier (theory knowledge) | 86 % | 86 % | 58 % |
| Advanced Sommelier (theory knowledge) | 77 % | 77 % | 46 % |
| Leetcode (easy) | 31 / 41 | 31 / 41 | 12 / 41 |
| Leetcode (medium) | 21 / 80 | 21 / 80 | 8 / 80 |
| Leetcode (hard) | 3 / 45 | 3 / 45 | 0 / 45 |

**Table 1.** GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. We report GPT-4's final score graded according to exam-specific rubrics, as well as the percentile of test-takers achieving GPT-4's score.

# Introduction to GPT-4

- **General Overview:**
  - **GPT-4**, or Generative Pre-trained Transformer 4, represents the latest iteration in the series of transformers designed by OpenAI.
  - Builds upon the foundation laid by its predecessors with substantial upgrades in technology and application scopes.
- **Key Characteristics:**
  - GPT-4 is a large multimodal language model capable of understanding and generating both text and image-based input.
  - Like GPT-3, it uses a transformer-based architecture but has been significantly scaled and optimized to handle more complex tasks and larger datasets.
- **What Sets GPT-4 Apart:**
  - Enhanced capacity to understand nuances across multiple languages and dialects.
  - Improved contextual awareness allows for richer, more accurate text generation.

# Introduction to GPT-4

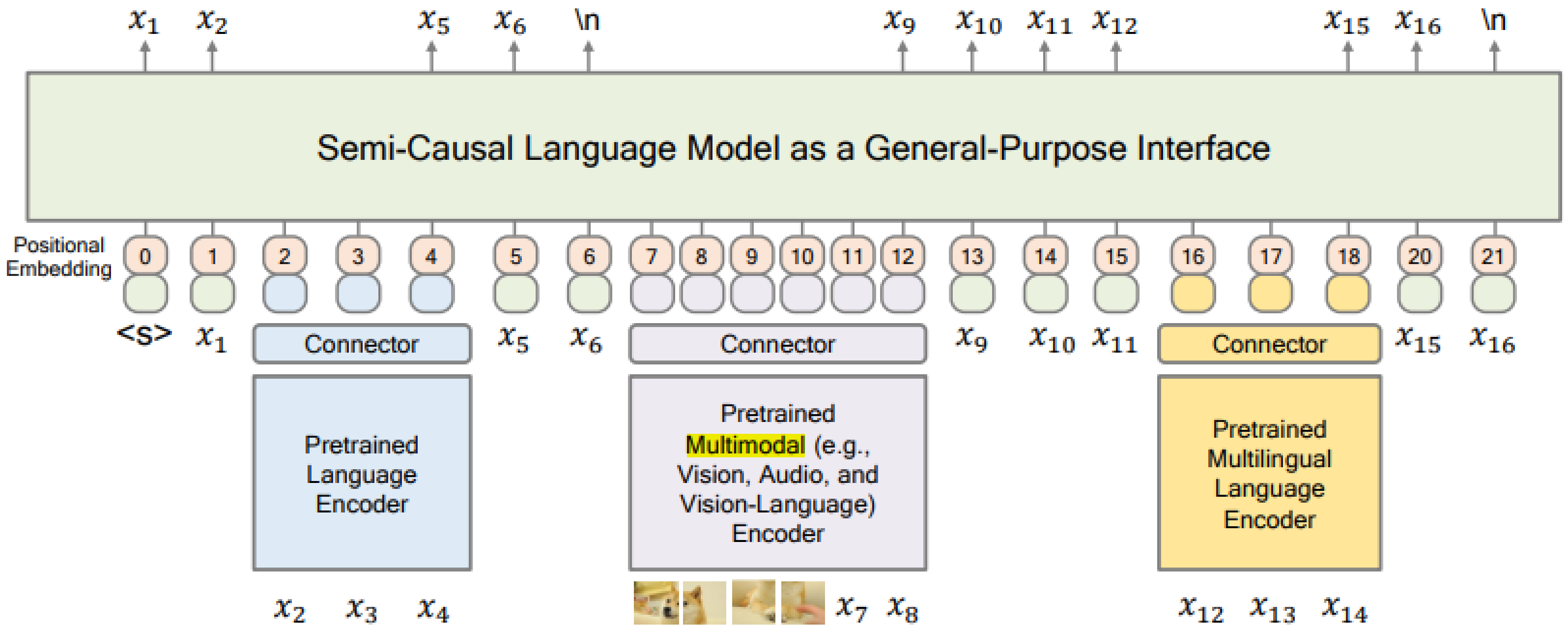
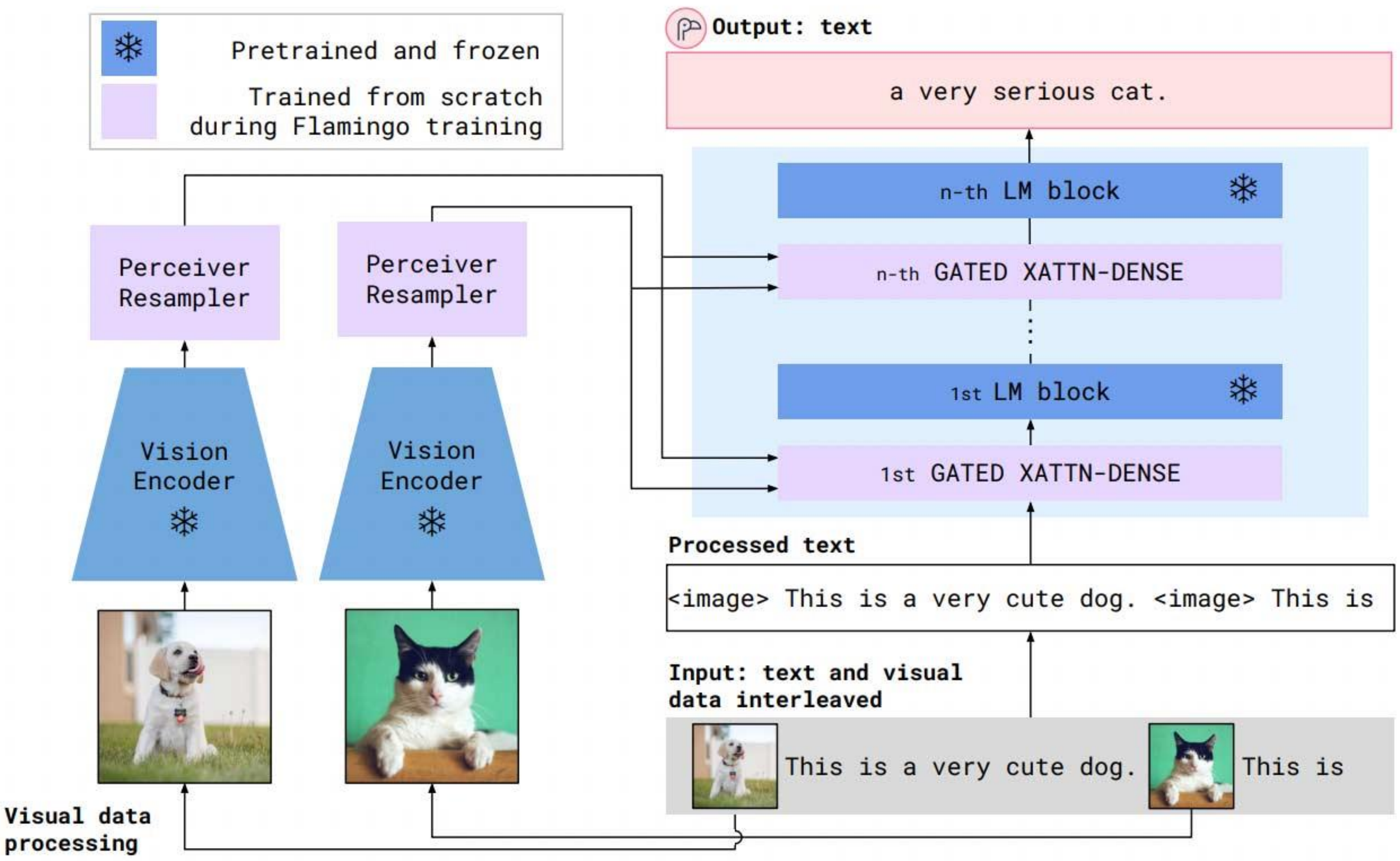



Figure 2: Overview of METALM. The semi-causal language model serves as a general-purpose interface and supports interactions with various foundation models.

# GPT-4 Architecture and Technical Details

- **Components of GPT-4:**
  - **Auto-regressive Model:** Operates by predicting the next word in a sequence given all the previous words, modeling the probability distribution of a word sequence.
  - **Layers and Parameters:** Encompasses hundreds of layers with a groundbreaking increase to approximately 100 trillion parameters compared to GPT-3’s 175 billion.
- **Functionality:**
  - The architecture supports broader and deeper neural networks, allowing for more complex and substantial pattern recognition and generation capabilities.
  - Utilizes modified initialization, layer normalization, and attention mechanisms to stabilize learning in such a large-scale network.
- **Innovations in Architecture:**
  - Fine-tuned attention mechanisms that deal more effectively with longer texts.
  - Integrations of advanced heuristics for better handling of nuanced human-like conversational patterns and problem-solving tasks.

# GPT-4 Architecture and Technical Details



# Advanced Training Methods

- **Key Training Innovations:**
  - **Sparse Attention:** GPT-4 incorporates sparse attention patterns which allow the model to focus on more relevant segments of text, improving speed and efficiency in processing.
  - **Mix of Supervised and Reinforcement Learning:** Leverages a blend of massive supervised training with reinforcement learning from human feedback (RLHF) to fine-tune responses based on desired outputs.
- **Impact of These Innovations:**
  - These training advancements allow GPT-4 to achieve higher levels of precision and adaptability, providing more contextually appropriate responses across diverse applications.

# Advanced Training Methods

Sparse attention mechanisms are a crucial innovation in optimizing the computation of attention in Transformer models. They allow models to scale efficiently by focusing on a subset of key information rather than computing interactions across all parts of the input equally.

- **Key Concepts:**
  - **Standard Attention**: In standard (dense) attention, all tokens in the input sequence interact with all other tokens, which is computationally expensive, especially for long sequences.
  - **Sparse Attention**: Sparse attention mechanisms selectively focus on a subset of relevant tokens, reducing computational complexity and enhancing the ability to handle longer sequences.
- **Benefits of Sparse Attention:**
  1. **Efficiency**: Reduces the quadratic complexity of full attention to something more manageable, often linear or close to linear.
  2. **Performance**: Can lead to better performance on tasks that require focusing on specific parts of the input data, such as long document summarization.
  3. **Scalability**: Makes it feasible to process longer sequences of data than would be possible with dense attention models.

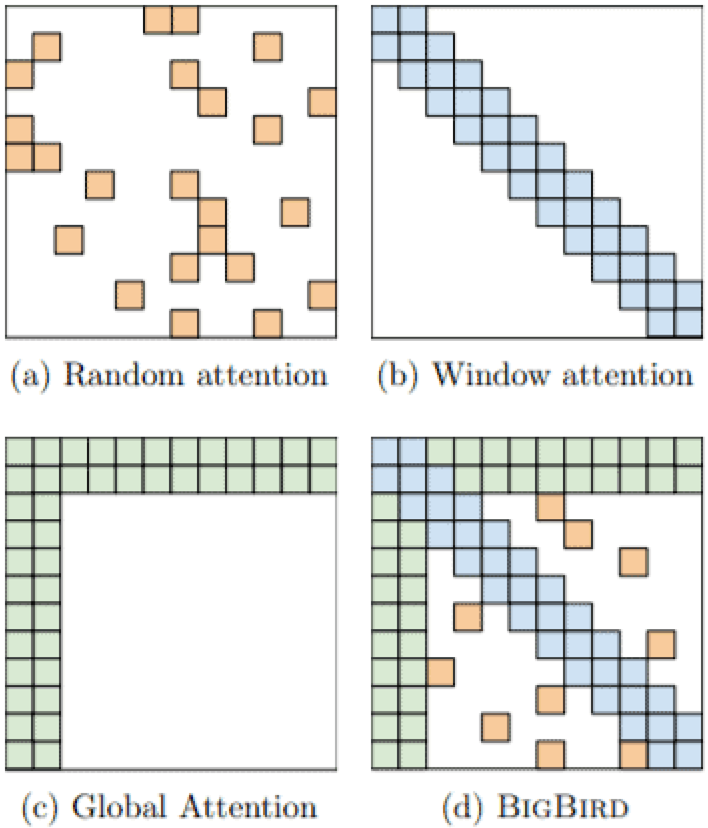
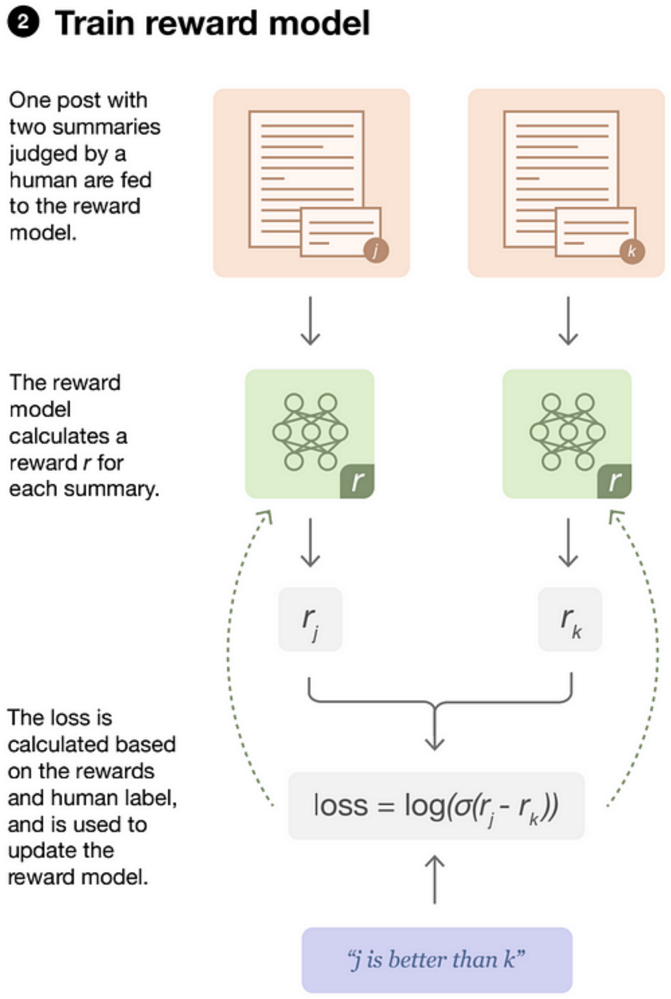



Image from the original paper

# Advanced Training Methods

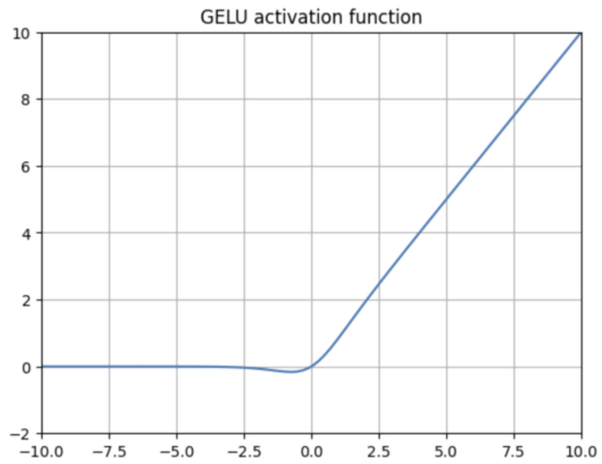
- **Reinforcement Learning From Human Feedback (RLHF):**
  - RLHF involves two main stages:
    - **Supervised Learning**: The model is pre-trained on a diverse dataset to develop a general understanding of language patterns.
    - **Reinforcement Learning**: The model's initial outputs are refined based on human feedback, focusing on improving response quality and alignment with ethical guidelines.
  - How RLHF Works with GPT-4:
    - **Data Collection**: Human trainers provide examples of high-quality text outputs.
    - **Reward Modeling**: A reward model is trained based on these examples to quantitatively assess the quality of the model's responses.
    - **Policy Fine-tuning**: Using the reward model, GPT-4's responses are fine-tuned through reinforcement learning to maximize these rewards, essentially learning to predict what humans prefer.
  - Benefits of RLHF:
    1. **Alignment with Human Judgements**: Ensures that the model's output is not only accurate but also aligns with human values and ethical considerations.
    2. **Customizability**: Allows tailoring the model for specific applications by adjusting the types of feedback used during training.
    3. **Reduced Harmful Outputs**: By training with human feedback, GPT-4 is better equipped to avoid generating harmful or biased content.

https://www.youtube.com/watch?v=2MBJOuVq380



# Achieving Computational Efficiency

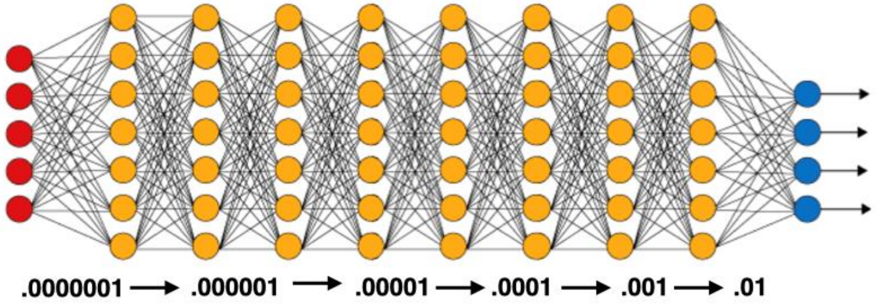
- **Optimizations Made for Scalability:**
  - **Revised Activation Functions:** Implementation of more computationally efficient activation functions to speed up processing without loss of accuracy.
  - **Layer-wise Learning Rate Decay:** Introduces a decay factor in the learning rates across layers, optimizing the training process to converge faster and more reliably. Helps avoid overfitting by slowing down the updates in deeper layers, which are often more complex and less generalized.




$$\mathrm{GELU}(x) = x * \Phi(x)$$

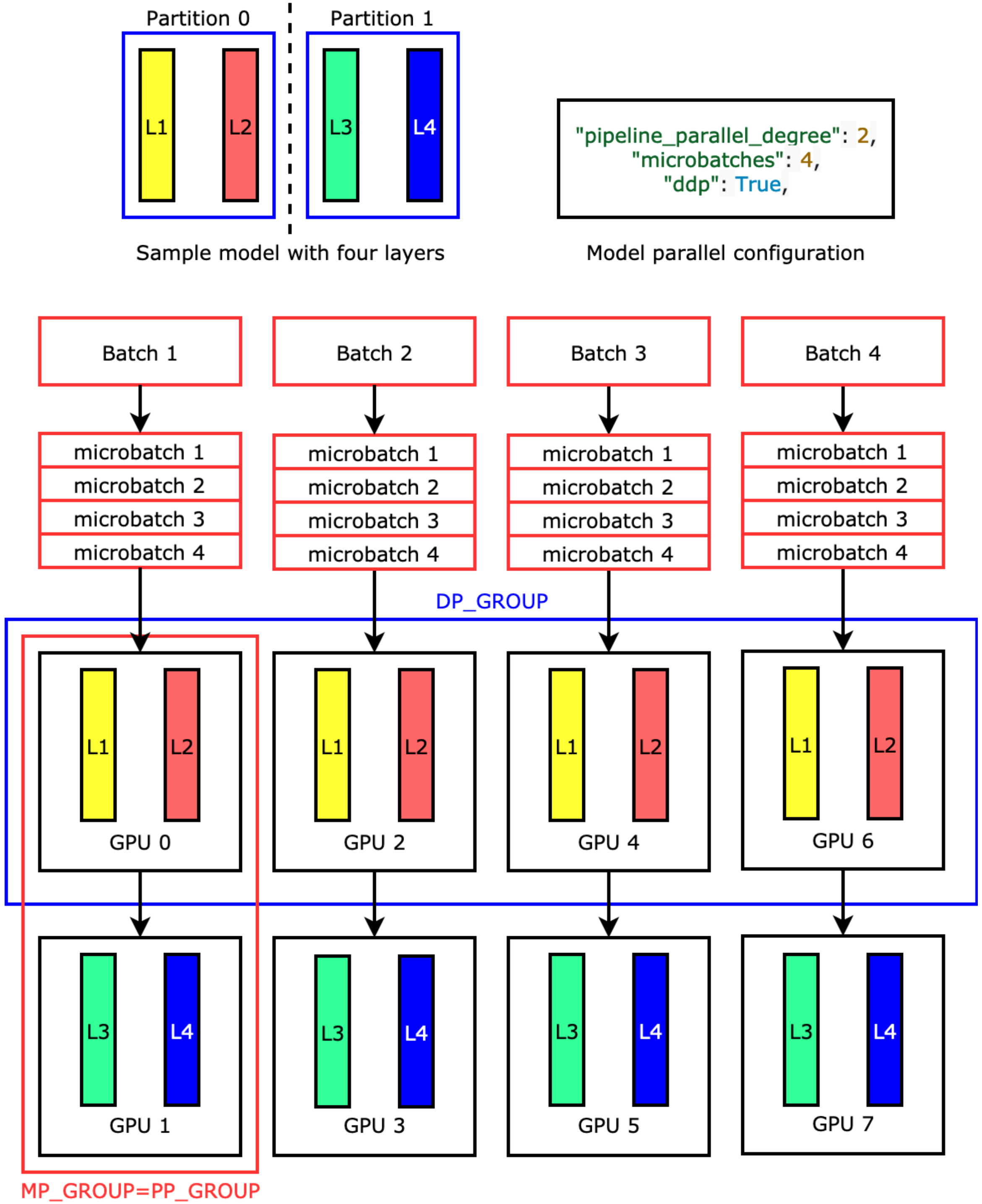
Where $\Phi(x)$ is the standard Gaussian cumulative distribution function (CDF).

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$



# Achieving Computational Efficiency

- **Efficiency in Resource Usage:**
  - **Model Parallelism:** GPT-4 utilizes model parallelism techniques, distributing the model's parameters across multiple GPUs, allowing it to scale without bound by single-machine memory limits.
  - **Dynamic Attention Masks:** Custom attention masks that dynamically adjust based on the input length and content type, considerably reducing the computational overhead for less complex inputs.



# What is GPT-4o1?

## Overview:

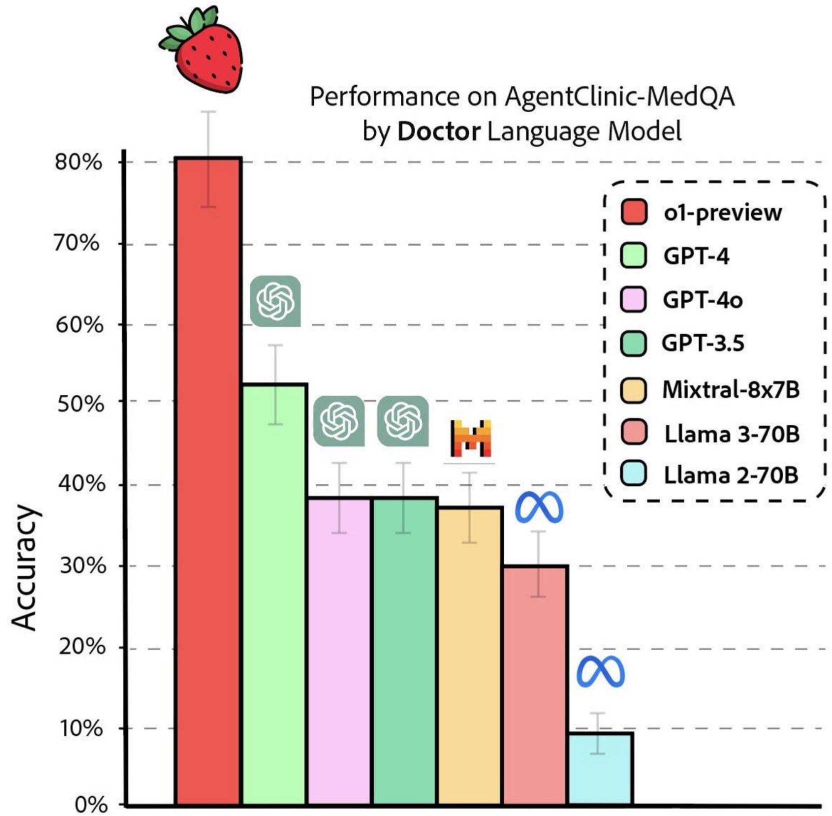
•GPT-4o1 is an optimized iteration of OpenAI's GPT-4, designed for enhanced performance on specific complex reasoning tasks.

## Core Purpose:

•GPT-4o1 excels in environments that require rigorous complex reasoning, such as advanced mathematical, scientific, and technical problem-solving.

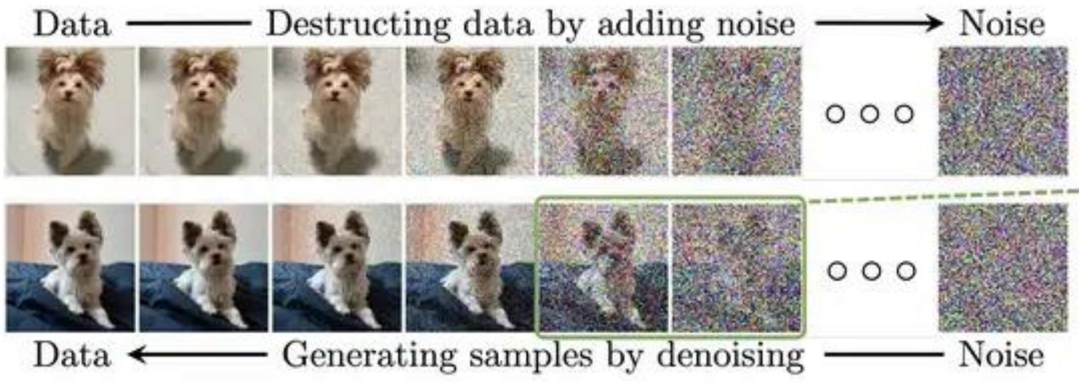
## Specialization:

•Unlike GPT-4, which is developed as a general-purpose model, GPT-4o1 is fine-tuned to deliver superior outcomes in specialized contexts, capitalizing on its ability to generate extended chains of thought and refine reasoning strategies through reinforcement learning[2].



# Targeted Improvements for Deep Reasoning

- **Algorithmic Improvements**:
  - **Diffusion-based Techniques**: Incorporation of generative diffusion techniques that enhance the model’s ability to generate responses by smoothing the decision space, providing paths to more nuanced answers.
    1. **Noise Addition and Gradual Refinement**:
       1. Initial outputs are generated, then ‘noised’ with random perturbations.
       2. Through iterative refining steps, the model learns to reduce noise, enhancing the quality and accuracy of the output.
    2. **Training Adjustments**:
       1. Models are trained using a sequence of latent variables that represent different 'noisy' versions of the text, gradually learning to predict cleaner versions.
    3. **Enhancement of Sampling Methods**:
       1. Improved sampling algorithms ensure that the generation process carefully balances randomness with adherence to the contextual structure, generating more plausible and relevant text.



# Targeted Improvements for Deep Reasoning

- **Structural Modifications**:
    - **Increased Model Capacity**: Expansion in the number of parameters and layers to increase the depth of contextual understanding and reasoning capabilities.
    - **Dynamic Token Handling**: Improved mechanisms for managing tokens dynamically enhance the model’s ability to process and interpret complex inquiries over longer sessions.

# Targeted Improvements for Deep Reasoning

- **Targeted Fine-tuning**: Incorporation of fine-tuned training on datasets that specifically demand high reasoning abilities, such as scientific literature and problem-solving scenarios.
- **Fine-tuning the 'chain of thought'** : enhancing its ability to generate stepwise reasoning beneficial for solving complex tasks.
- What is 'Chain of Thought'?
  - **Definition**: 'Chain of Thought' refers to a reasoning process where the model explicitly generates intermediate steps and considerations leading to a final answer, rather than producing an answer directly.
  - **Purpose**: Aids in transparent reasoning, easier verification of the thought process, and improvements in multi-step problem-solving.
- Fine-Tuning Techniques:
  1. **Data Selection**:
     1. Curated datasets that include problem-solving examples with clearly outlined reasoning steps are chosen for training.
     2. Examples include mathematical word problems, logic puzzles, and advanced scientific questions.
  2. **Training Strategy**:
     1. **Supervised Learning**: The model trains on annotated examples where each step of the reasoning is explicitly described.
     2. **Iterative Refinement (RHLF)** : The model receives feedback on the reasoning steps generated, with a focus on enhancing logical coherence and accuracy.

# Targeted Improvements for Deep Reasoning

**Fine-tuning the 'chain of thought’**

Alice à 2 soeur et 3 frères. Combien de soeur a le frère d alice

Psycho
Le frère d'Alice a 2 sœurs.

Planifie la résolution du problème : Alice à 2 soeur et 3 frères. Combien de soeur a le frère d alice

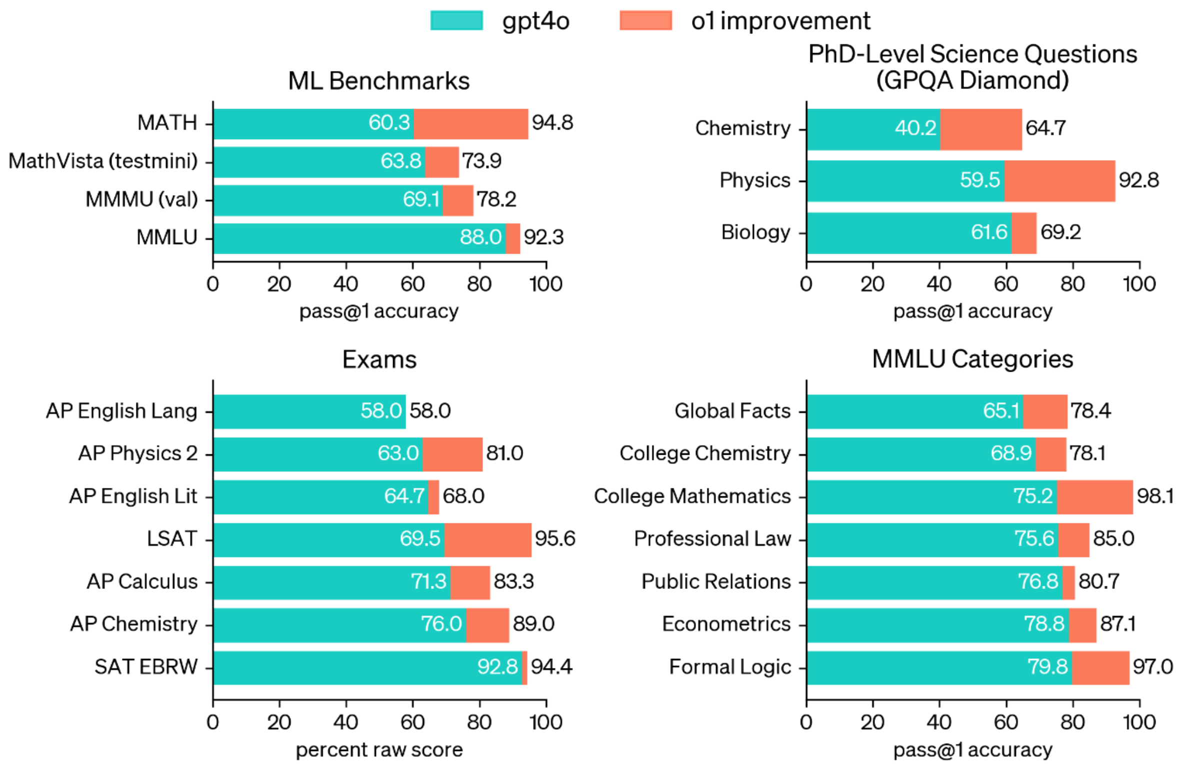
Psycho
Pour résoudre ce problème, nous pouvons suivre ces étapes :
Identifier les membres de la famille d'Alice :
Selon l'énoncé, Alice a 2 sœurs et 3 frères.
Déterminer la relation entre le frère d'Alice et ses sœurs :Il est demandé combien de sœurs a le frère d'Alice. Cela inclut Alice elle-même et ses 2 autres sœurs.
Calculer le nombre total de sœurs que chaque frère a :Puisque le frère considère Alice comme sa sœur ainsi que ses 2 autres sœurs, on arrive à un total de 3 sœurs.

Conclusion :Chaque frère d'Alice a donc 3 sœurs

# Targeted Improvements for Deep Reasoning

**Performance Features:**

- Significant improvements are seen in tasks that demand high cognitive abilities, such as solving advanced mathematical problems and interpreting complex data across multiple domains, proving its effectiveness over GPT-4 in certain specialized benchmarks.



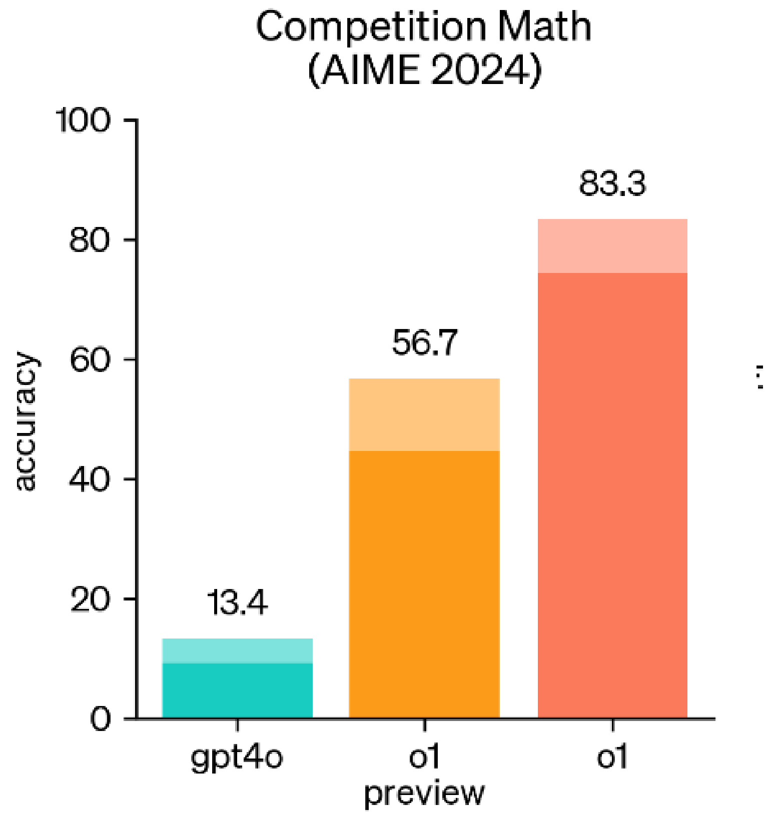
# Case Study - Advanced Mathematical Problem Solving

**Achieving Human-Level Performance in Mathematics:**

- GPT-4o1 has demonstrated capabilities in high-level mathematical reasoning, scoring among the top 500 students nationally on the AIME exams, a testament to its optimized reasoning enhancements.

**Real-World Application:**

- Used in academic settings to assist in complex mathematical problem solving, providing students and researchers with insights that typically require high levels of expertise.



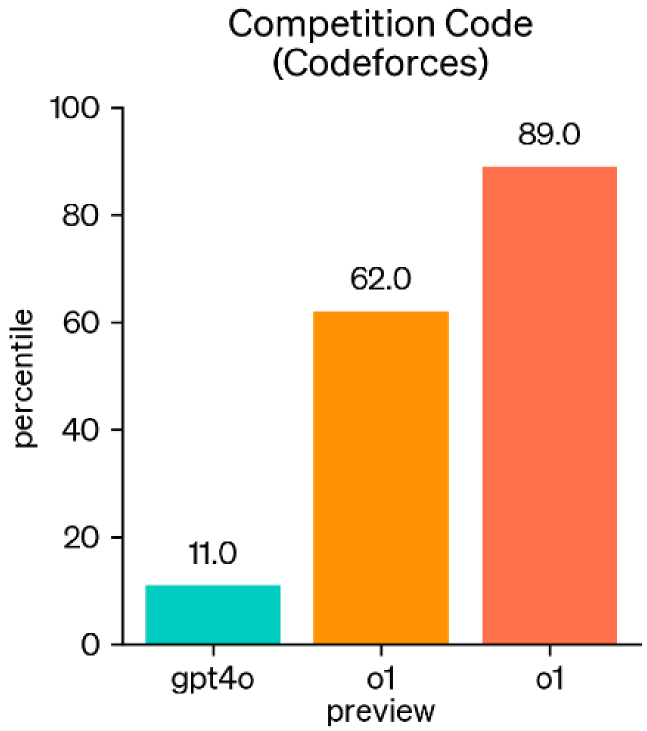
# Case Study - Competitive Programming

**Programming Competitions:**

- GPT-4o1 participated in programming contests such as the International Olympiad in Informatics (IOI), where it performed on par with human contestants, illustrating its enhanced coding and problem-solving skills.

**Strategy and Execution:**

- Utilizes a sophisticated test-time selection strategy to maximize efficiency and outcome, showcasing its practical utility in algorithm design and competitive scenarios.



# Industry Implications and Broader Applications

**Broader Industry Impact:**

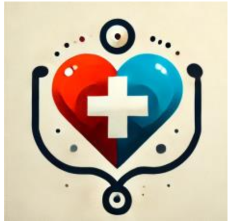
- GPT-4o1's advanced capabilities are not just academic; they're practical for industries requiring deep analytical tasks, such as financial forecasting, where precision and speed are critical.

**Ethical and Safety Considerations:**

- While pushing the boundaries of what's possible, GPT-4o1 also incorporates robust safety mechanisms to ensure that its advanced capabilities are aligned with ethical guidelines and practical safety standards.

# Overview of Fields Using GPT-4

- **Healthcare:**
  - **Medical Documentation:** Automating synthesis of medical reports from patient data.
  - **Diagnostic Support:** Assisting physicians in diagnosing diseases based on symptoms and medical history.
- **Finance:**
  - **Automated Trading and Analysis:** Using GPT-4 for high-frequency trading and complex financial modeling to predict market trends.
  - **Customer Service:** Enhancing customer interactions through dynamic, intelligent virtual assistants.
- **Creative Industries:**
  - **Content Creation:** Assisting in the writing of scripts, articles, and even books.
  - **Music Composition:** Generating music scores and assisting in sound design.
- **Legal Sector:**
  - **Contract Analysis:** Automating the review and analysis of legal documents.
  - **Legal Advisory:** Providing preliminary legal advice based on current laws and regulations.




# Impact of GPT-4 in Healthcare

- **Healthcare:**
  - **Enhanced Accuracy and Efficiency:** Reducing the incidence of human error in medical reporting and increasing the speed at which medical documents are processed.
  - **Supportive Diagnostic Procedures:** Providing support tools for differential diagnosis, saving time for medical professionals and potentially increasing patient satisfaction.

| Domains | Model Development | Models | # Params | Data Scale | Data Source |
|---|---|---|---|---|---|
| Medical-domain LLMs (Sec. 2) | Pre-training (Sec. 2.1) | BioBERT [49] | 110M | 18B tokens | PubMed [50]+PMC [51] |
| | | PubMedBERT [52] | 110M/340M | 3.2B tokens | PubMed [50]+PMC [51] |
| | | SciBERT [53] | 110M | 3.17B tokens | Literature [54] |
| | | ClinicalBERT [55] | 110M | 112k clinical notes | MIMIC-III [56] |
| | | BioM-ELECTRA [57] | 110M/335M | - | PubMed [50] |
| | | BioMed-RoBERTa [58] | 125M | 7.55B tokens | S2ORC [59] |
| | | BioLinkBERT [60] | 110M/340M | 21GB | PubMed [50] |
| | | SciFive [61] | 220M/770M | - | PubMed [50]+PMC [51] |
| | | ClinicalT5 [62] | 220M/770M | 2M clinical notes | MIMIC-III [56] |
| | | BlueBERT [63,64,65] | 110M/340M | >4.5B tokens | PubMed [50]+MIMIC-III [56] |
| | | MedCPT [66] | 330M | 255M articles | PubMed [50] |
| | | BioGPT [67] | 1.5B | 15M articles | PubMed [50] |
| | | BioMedLM [68] | 2.7B | 110GB | Pile [69] |
| | | OphGLM [70] | 6.2B | 20k dialogues | MedDialog [71] |
| | | GatorTron [23] | 8.9B | >82B tokens+6B tokens 2.5B tokens+0.5B tokens | EHRs [23]+PubMed [50] Wiki+MIMIC-III [56] |
| | | GatorTronGPT [72] | 5B/20B | 277B tokens | EHRs [72] |
| | Fine-tuning (Sec. 2.2) | DoctorGLM [73] | 6.2B | 323MB dialogues | CMD. [74] |
| | | BianQue [75] | 6.2B | 2.4M dialogues | BianQueCorpus [75] |
| | | ClinicalGPT [76] | 7B | 96k EHRs 192 medical QA 100k dialogues | MD-EHR [76] VariousMedQA [14] MedDialog [71] |
| | | Qilin-Med [77] | 7B | 3GB | ChiMed [77] |
| | | ChatDoctor [15] | 7B | 110k dialogues | HealthCareMagic [78]+iCliniq [79] |
| | | BenTsao [17] | 7B | 8k instructions | CMeKG-8K [80] |
| | | HuatuoGPT [81] | 7B | 226k instructions&dialogues | Hybrid SFT [81] |
| | | Baize-healthcare [82] | 7B | 101K dialogues | Quora+MedQuAD [83] |
| | | BioMedGPT [84] | 7B | >26B tokens | S2ORC [59] |
| | | MedAlpaca [16] | 7B/13B | 160k medical QA | Medical Meadow [16] |
| | | AlpaCare [85] | 7B/13B | 52k instructions | MedInstruct-52k [85] |
| | | Zhongjing [86] | 13B | 70k dialogues | CMtMedQA [86] |
| | | PMC-LLaMA [13] | 13B | 79.2B tokens | Books+Literature [59]+MedC-I [13] |
| | | CPLLM [87] | 13B | 109k EHRs | eICU-CRD [88]+MIMIC-IV [89] |
| | | OpenBioLLM [90] | 8B/70B | - | - |
| | | MEDITRON [91,92] | 7B/70B | 48.1B tokens | PubMed [50]+Guidelines [91] |
| | | Clinical Camel [18] | 13B/70B | 70k dialogues+100k articles 4k medical QA | ShareGPT [93]+PubMed [50] MedQA [14] |
| | | MedPaLM 2 [11] | 340B | 193k medical QA | MultiMedQA [11] |
| | | Med-Gemini [94,95] | - | - | MedQA-R&RS [95]++MultiMedQA [11] +MIMIC-III [56]+MultiMedBench [96] |
| | Prompting (Sec. 2.3) | CodeX [97] | GPT-3.5 / LLaMA-2 | Chain-of-Thought (CoT) [98] | - |
| | | DeID-GPT [99] | ChatGPT / GPT-4 | Chain-of-Thought (CoT) [98] | - |
| | | ChatCAD [100] | ChatGPT | Zero-shot Prompting | - |
| | | Dr. Knows [101] | ChatGPT | Zero-shot Prompting | UMLS [102] |
| | | MedPaLM [10] | PaLM (540B) | 40 instructions | MultiMedQA [11] |
| | | MedPrompt [12] | GPT-4 | Few-shot & CoT [98] | - |
| | | Chat-Orthopedist [103] | ChatGPT | Retrieval-Augmented Generation (RAG) | PubMed+Guidelines [104] + UpToDate [105]+Dyname [106] |
| | | QA-RAG [107] | ChatGPT | RAG | FDA QA [107] |
| | | Almanac [108] | ChatGPT | RAG & CoT [98] | Clinical QA [108] |

# GPT-4o1 Applied in Practice - Case Study in Healthcare

- **Context:**
  - **Problem:** A leading healthcare provider faced challenges with rapid, accurate diagnosis of rare diseases, which often requires extensive medical knowledge and patient data analysis.
  - **Objective:** Implement GPT-4o1 to assist medical professionals by providing rapid diagnostic suggestions based on patient symptoms, history, and medical literature.
- **Implementation:**
  - **Data Integration:** GPT-4o1 was integrated with the hospital's electronic health record systems to access comprehensive patient data.
  - **Model Training:** The model was fine-tuned with an extensive database of medical case studies, symptoms, diagnostics, and outcomes to enhance its reasoning capabilities specific to medical diagnostics.
- **Outcomes:**
  - **Enhanced Diagnostic Accuracy:** GPT-4o1 helped increase the accuracy of rare disease diagnosis by 35%.
  - **Speed of Diagnosis:** Reduction in time taken to reach a diagnostic conclusion by approximately 50%.

# GPT-4o1 Applied in Practice - Case Study in Healthcare

- NYC hospitals
  - The authors fed a LLM with 300 000 EHR (non-structured data, clinical notes) Prediction of 5 tasks
    - 30-day all-cause readmission prediction
    - in-hospital mortality prediction
    - comorbidity index prediction
    - length of stay prediction
    - insurance denial prediction
  - NYUTron has an area under the curve (AUC) of 78.7–94.9%,
  - With an improvement of 5.36–14.7% in the AUC compared with traditional models.

Article

## Health system-scale language models are all-purpose prediction engines

https://doi.org/10.1038/s41586-023-06160-y
Received: 14 October 2022
Accepted: 2 May 2023
Published online: 7 June 2023
Open access
Check for updates

Lavender Yao Jiang[1,2], Xujin Chris Liu[1,3], Nima Pour Nejatian[4], Mustafa Nasir-Moin[1], Duo Wang[5], Anas Abidin[4], Kevin Eaton[6], Howard Antony Riina[1], Ilya Laufer[1], Paawan Punjabi[6], Madeline Miceli[6], Nora C. Kim[1], Cordelia Orillac[1], Zane Schnurman[1], Christopher Livia[1], Hannah Weiss[1], David Kurland[1], Sean Neifert[1], Yosef Dastagirzada[1], Douglas Kondziolka[1], Alexander T. M. Cheung[1], Grace Yang[1,2], Ming Cao[1,2], Mona Flores[4], Anthony B. Costa[4], Yindalon Aphinyanaphongs[5,7], Kyunghyun Cho[2,8,9,10] & Eric Karl Oermann[1,2,11]

Physicians make critical time-constrained decisions every day. Clinical predictive models can help physicians and administrators make decisions by forecasting clinical and operational events. Existing structured data-based clinical predictive models have limited use in everyday practice owing to complexity in data processing, as well as model development and deployment[1–3]. Here we show that unstructured clinical notes from the electronic health record can enable the training of clinical language models, which can be used as all-purpose clinical predictive engines with low-resistance development and deployment. Our approach leverages recent advances in natural language processing[4,5] to train a large language model for medical language (NYUTron) and subsequently fine-tune it across a wide range of clinical and operational predictive tasks. We evaluated our approach within our health system for five such tasks: 30-day all-cause readmission prediction, in-hospital mortality prediction, comorbidity index prediction, length of stay prediction, and insurance denial prediction. We show that NYUTron has an area under the curve (AUC) of 78.7–94.9%, with an improvement of 5.36–14.7% in the AUC compared with traditional models. We additionally demonstrate the benefits of pretraining with clinical text, the potential for increasing generalizability to different sites through fine-tuning and the full deployment of our system in a prospective, single-arm trial. These results show the potential for using clinical language models in medicine to read alongside physicians and provide guidance at the point of care.

Physicians make difficult decisions every day requiring the integration of a tremendous amount of information. The information needed to make these medical decisions is scattered across various records, for example, a patient's medical history and laboratory and imaging reports. When physicians perform their work, however, all of this information is ultimately integrated into the notes written by physicians to document and summarize patient care.

Clinical predictive models are frequently derived from rules that have existed for decades[6–9], as well as from machine learning methods[10–12], with most relying on structured inputs pulled from the electronic health record (EHR) or direct clinician inputs. This reliance on structured inputs introduces complexity in data processing, as well as in model development and deployment, which in part is responsible for the overwhelming majority of medical predictive algorithms being trained, tested and published, yet never deployed to assess their impact on real-world clinical care. This is frequently referred to as the 'last-mile problem' (refs. 1–3).

One of the most exciting recent developments in modern artificial intelligence (AI) research is large language models (LLMs). These massive neural networks (with millions or even billions of parameters) have been shown to obtain impactful results on a wide range of problems that rely on the reading and interpretation of human language. Several styles of LLMs have been developed over the past few years, broadly ranging from encoder models (such as BERT[4]) to decoder models (such as GPT3; ref. 5). We theorized that LLMs could potentially solve the last-mile problem in medical predictive analytics by simply reading the notes written by physicians, thereby immediately accessing a comprehensive description of a patient's medical state to provide decision support at the point of care across a wide range of clinical and operational tasks.

Here we present our results from developing, evaluating, deploying and prospectively assessing NYUTron, an LLM-based system that can integrate in real time with clinical workflows centred around writing notes and placing electronic orders. Our approach relies on the fact that all clinically useful data and medical professionals' decision-making

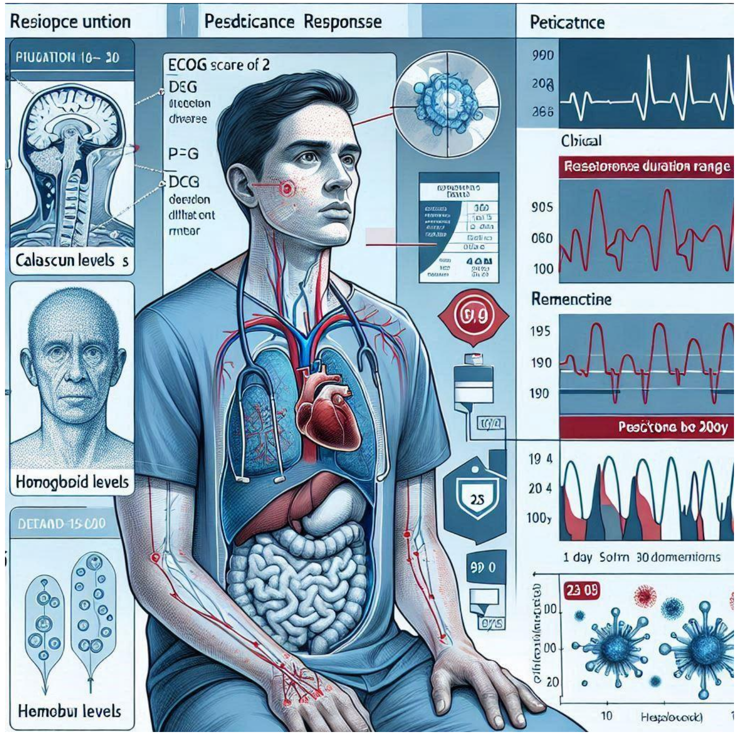
[1]Department of Neurosurgery, NYU Langone Health, New York, NY, USA. [2]Center for Data Science, New York University, New York, NY, USA. [3]Electrical and Computer Engineering, Tandon School of Engineering, New York, NY, USA. [4]NVIDIA, Santa Clara, CA, USA. [5]Predictive Analytics Unit, NYU Langone Health, New York, NY, USA. [6]Department of Internal Medicine, NYU Langone Health, New York, NY, USA. [7]Department of Population Health, NYU Langone Health, New York, NY, USA. [8]Prescient Design, Genentech, New York, NY, USA. [9]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA. [10]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [11]Department of Radiology, NYU Langone Health, New York, NY, USA. e-mail: eric.oermann@nyulangone.org

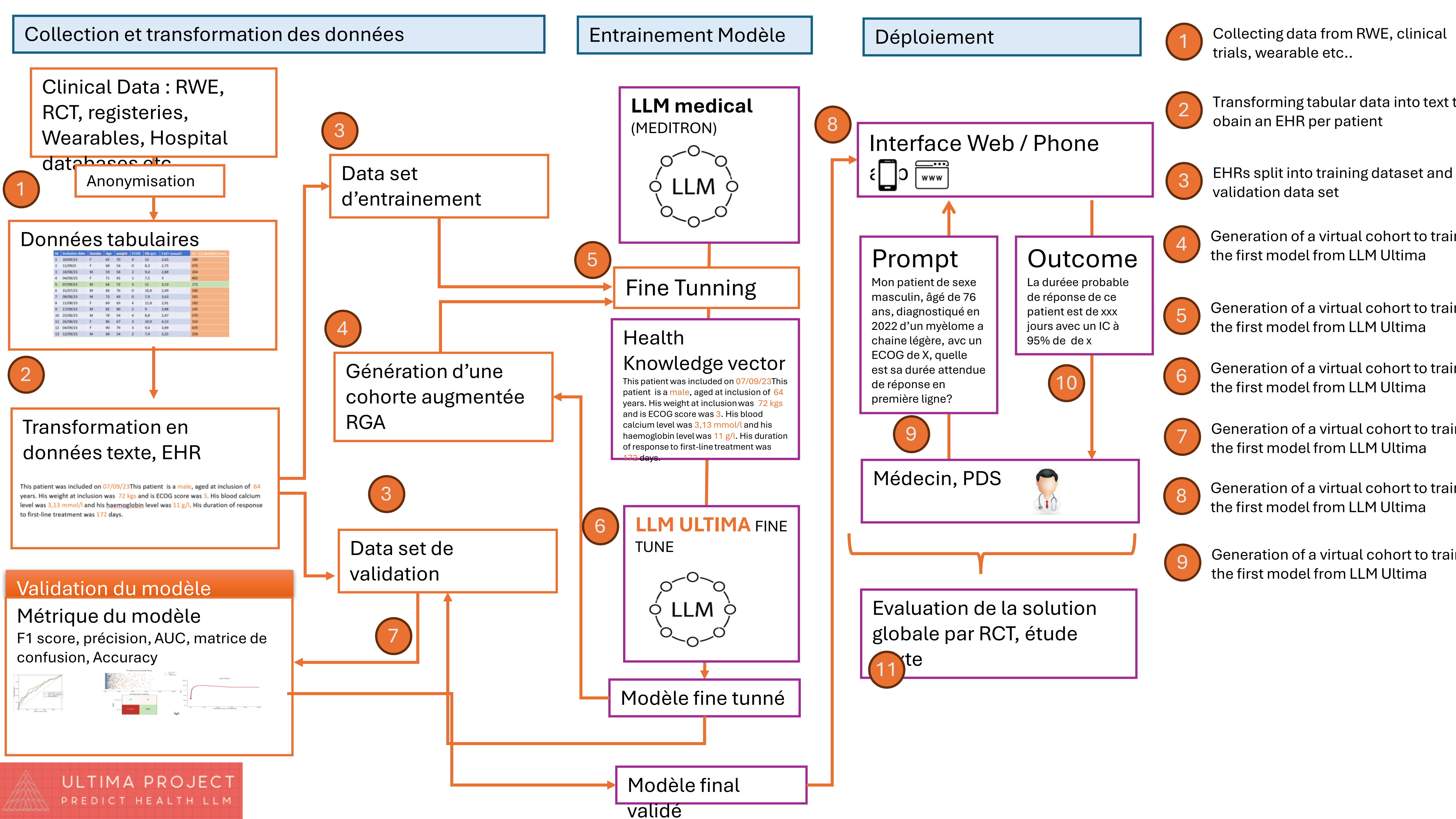
# TABULAR DATA

| Id | Inclusion date | Gende r | Age | weight | ECO G | Hb (g/l) | Ca2+ (mmol/l) | L1 – L2 duration (days) |
|---|---|---|---|---|---|---|---|---|
| 1 | 10/09/23 | F | 65 | 70 | 0 | 12 | 2,65 | 180 |
| 2 | 11/0923 | F | 68 | 54 | 0 | 8,3 | 2,75 | 375 |
| 3 | 10/08/23 | M | 59 | 58 | 2 | 9,4 | 2,88 | 204 |
| 4 | 04/09/23 | F | 71 | 45 | 1 | 7,5 | 3 | 402 |
| 5 | 07/09/23 | M | 64 | 72 | 3 | 11 | 3,13 | 172 |
| 6 | 31/07/23 | M | 84 | 76 | 0 | 10,8 | 2,49 | 206 |
| 7 | 08/08/23 | M | 73 | 49 | 0 | 7,9 | 3,63 | 505 |
| 8 | 11/08/23 | F | 69 | 65 | 4 | 11,8 | 2,91 | 102 |
| 9 | 17/09/23 | M | 82 | 80 | 2 | 9 | 3,88 | 245 |
| 10 | 25/08/23 | M | 78 | 54 | 4 | 8,8 | 2,47 | 570 |
| 11 | 26/08/23 | F | 86 | 67 | 3 | 10,9 | 4,13 | 310 |
| 12 | 04/09/23 | F | 90 | 79 | 3 | 9,4 | 3,89 | 635 |
| 13 | 12/09/23 | M | 68 | 54 | 2 | 7,4 | 3,25 | 258 |

# FROM TABULAR DATA TO TEXTS

1. Generation of a fill-in-the-blank text with variable values
   - This patient was included on [#fill inclusion date value]. This patient is a [#fill gender value] , aged at inclusion of [#fill age value] years. His weight at inclusion was [#fill weight value] and is ECOG score was [#fill ECOG value]. His blood calcium level was [#fill blood calcium value] and his haemoglobin level was [#fill haemoglobin value]. His duration of response to first-line treatment was [#fill L1-L2 duration value] days.
2. Generation of an EHR for each patient
   - This patient was included on 07/09/23This patient is a male, aged at inclusion of 64 years. His weight at inclusion was 72 kgs and is ECOG score was 3. His blood calcium level was 3,13 mmol/l and his haemoglobin level was 11 g/l. His duration of response to first-line treatment was 172 days.
3. Reformulations of HER using LLM
   - This patient was included on 07/09/23. He is a 64-year-old male with an inclusion weight of 72 kg and an ECOG score of 3. At inclusion, his blood calcium level was 3.13 mmol/l and his haemoglobin level was 11 g/l. The duration of his response to first-line treatment was 172 days.
   - Included on 07/09/23, this male patient was 64 years old at the time of inclusion. He weighed 72 kg and had an ECOG score of 3. His blood calcium level was 3.13 mmol/l and his haemoglobin level was 11 g/l at inclusion. His response to first-line treatment lasted for 172 days.
   - On 07/09/23, this 64-year-old male patient was included with a weight of 72 kg and an ECOG score of 3. His blood tests revealed a calcium level of 3.13 mmol/l and a haemoglobin level of 11 g/l. The duration of his response to the first-line treatment was 172 days.
   - ...





Collection et transformation des données
Entrainement Modèle
Déploiement
Clinical Data : RWE, RCT, registeries, Wearables, Hospital databases etc
Anonymisation
1
Données tabulaires
2
Transformation en données texte, EHR
This patient was included on 07/09/23This patient is a male, aged at inclusion of 64 years. His weight at inclusion was 72 kgs and is ECOG score was 3. His blood calcium level was 3,13 mmol/l and his haemoglobin level was 11 g/l. His duration of response to first-line treatment was 172 days.
3
Data set d'entrainement
4
Génération d'une cohorte augmentée RGA
3
Data set de validation
Validation du modèle
Métrique du modèle
F1 score, précision, AUC, matrice de confusion, Accuracy
7
LLM medical
(MEDITRON)
LLM
5
Fine Tunning
Health Knowledge vector
This patient was included on 07/09/23This patient is a male, aged at inclusion of 64 years. His weight at inclusion was 72 kgs and is ECOG score was 3. His blood calcium level was 3,13 mmol/l and his haemoglobin level was 11 g/l. His duration of response to first-line treatment was 172 days.
6
LLM ULTIMA FINE TUNE
LLM
Modèle fine tunné
Modèle final validé
8
Interface Web / Phone
www
Prompt
Mon patient de sexe masculin, âgé de 76 ans, diagnostiqué en 2022 d'un myèlome a chaine légère, avc un ECOG de X, quelle est sa durée attendue de réponse en première ligne?
Outcome
La duréee probable de réponse de ce patient est de xxx jours avec un IC à 95% de de x
9
10
Médecin, PDS
Evaluation de la solution globale par RCT, étude
11
ULTIMA PROJECT
PREDICT HEALTH LLM
1 Collecting data from RWE, clinical trials, wearable etc..
2 Transforming tabular data into text t obain an EHR per patient
3 EHRs split into training dataset and validation data set
4 Generation of a virtual cohort to trai the first model from LLM Ultima
5 Generation of a virtual cohort to trai the first model from LLM Ultima
6 Generation of a virtual cohort to trai the first model from LLM Ultima
7 Generation of a virtual cohort to trai the first model from LLM Ultima
8 Generation of a virtual cohort to trai the first model from LLM Ultima
9 Generation of a virtual cohort to trai the first model from LLM Ultima

# USER PROMPTS

- User prompt :

  *Can you estimate what will be the duration of response to first-line treatment for a patient as follow : This patient  is a male, aged 64 years. His weight is  72 kgs and is ECOG score is 3. His blood calcium level is 3,13 mmol/l and his haemoglobin level is 11 g/l.*

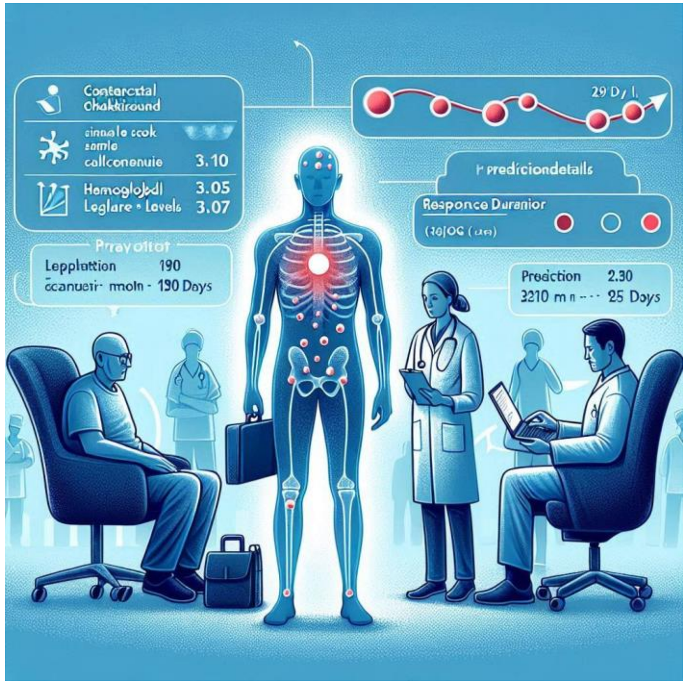
- Assistant prompt :

  A patient with similar characteristics (ECOG score of 2, similar calcium levels and hemoglobin levels) had a response duration of 190 to 205 days. The impact of the slight difference in calcium levels of our hypothetical patient (3 mmol/l compared to, for example, 3.10 or 3.07 mmol/l) is difficult to assess without precise data on the relationship between calcium levels and the response to treatment in this specific condition.

  That said, given the characteristics and similarities with other cases :

  - *an estimated response duration of 180 to 200 days* would be a considered estimation,
  - taking into account that *the slightly weakened physical condition of the patient (ECOG score of 2) could negatively influence the response duration*.

*prediction and its explanation*



# GPT-4o1 Applied in Finance - Algorithmic Trading

- **Context:**
  - **Problem:** A financial institution wanted to improve the robustness and responsiveness of its algorithmic trading strategies under volatile market conditions.
  - **Objective:** Utilize GPT-4o1's advanced modeling capabilities to simulate and predict market changes more accurately and execute trades more effectively.
- **Implementation:**
  - **Real-Time Data Processing:** Integration of GPT-4o1 to analyze real-time market data streams for immediate trading insights.
  - **Strategy Optimization:** The model was continuously trained on new market data, using reinforcement learning to adjust trading strategies dynamically.
- **Outcomes:**
  - **Increased Trading Efficacy:** Improved prediction accuracy of market trends led to a 40% increase in trading profitability.
  - **Decreased Latency in Trade Execution:** Execution times were reduced, minimizing losses due to slippage.
- **Visual Aid:** Before and after comparison of trading performance metrics.

# GPT-4o1 Applied in Finance - Algorithmic Trading

8 avril 2019 : Création de la société avec un capital initial de 600.000 euros, correspondant à la souscription et à la libération intégrale de 600.000 actions ordinaires d'un euro de valeur nominale chacune.

24 juin 2019 : Augmentation du capital social de 400.000 euros par l'émission de 400.000 actions ordinaires nouvelles de 1 euro de valeur nominale chacune.

25 février 2020 : Création de la catégorie d'actions de préférence dénommées Actions AP1 et augmentation du capital social de 575.000 euros par l'émission de 575.000 Actions AP1 nouvelles de 1 euro de valeur nominale chacune.

28 octobre 2020 : Création de la catégorie d'actions de préférence dénommées Actions AP2[4]. 21 juillet 2021 : Création de la catégorie d'actions de préférence dénommées Actions AP3.

12 novembre 2021 : Attribution effective de 7.500 Actions AP2, augmentant le capital social de 7.500 euros.

7 juillet 2024 : Augmentation du capital social de 500.000 euros par l'émission de 250.000 Actions AP3 à 2 euros de valeur nominale chacune. Cette augmentation vise à financer l'expansion de la société dans de nouveaux marchés internationaux.

12 décembre 2024 : Attribution effective de 5.000 Actions AP3, augmentant le capital social de 10.000 euros. Cette attribution est réalisée pour fidéliser les cadres supérieurs de la société en reconnaissant leur contribution au succès de l'entreprise.

20 mai 2025 : Réduction du capital social de 200.000 euros par le rachat et l'annulation de 100.000 Actions AP1. Cette opération est effectuée pour optimiser la structure du capital et améliorer le rendement des actionnaires restants.

30 septembre 2025 : Lancement d'une offre publique initiale (IPO) avec l'émission de 1.000.000 d'actions ordinaires à 5 euros chacune, augmentant le capital social de 5.000.000 euros. Cette étape marque la transition de la société vers une entité cotée en bourse, visant à accroître sa visibilité et à accéder à des capitaux supplémentaires pour ses projets futurs.

# Integration Challenges and Solutions in Deploying GPT-4o1

- **Challenges:**
  - **Data Privacy and Security:** When integrating GPT-4o1 with sensitive environments like healthcare and finance, ensuring data privacy and meeting compliance requirements are crucial.
  - **Adapting to Real-Time Constraints:** Ensuring GPT-4o1 can operate under the stringent real-time performance demands typical of financial trading and emergency medical diagnostics.
- **Solutions:**
  - **Robust Data Handling Protocols:** Implementing state-of-the-art encryption and access control measures to protect patient and financial data.
  - **Performance Optimization:** Hardware and software optimizations were conducted to ensure GPT-4o1 could make high-speed inferences and integrate smoothly with existing systems.

- **Visual Aid:** Diagrams showing data security measures and real-time performance optimization techniques.

# Introduction to GPT-4 and GPT-4o1 Interactive Coding Demonstration

**Objective:**

This slide will introduce the basics of interacting with GPT-4 and the enhanced version, GPT-4o1, through OpenAI's API. We will demonstrate live code execution for selected tasks to exemplify the capabilities of these models.

**Framework:**

- **API Access**: Brief overview of acquiring API keys and setting up your environment for communicating with GPT-4.
- **Tools Required**: Explanation of software tools and libraries needed, like Python and specific libraries for API interactions (e.g., requests or openai Python library).

```
import openai

# Set your OpenAI API key here
openai.api_key = 'your-api-key-here’

# Example of a simple GPT-4 API call
response = openai.ChatCompletion.create(
        model="gpt-4",
        messages=[ {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Explain the concept of gravitational waves."}
        ] )

print(response['choices'][0]['message']['content'])
```

# Demonstrating GPT-4o1's Advanced Reasoning

- **Highlight:**
  - Showcasing an enhanced reasoning task using GPT-4o1, emphasizing how it surpasses GPT-4 in complex problem-solving scenarios.
- **Code Demonstration:**
  - Illustrating the use of OpenAI's GPT-4o1 model to solve an advanced mathematical problem. The task involves a reasoning sequence that mimics human problem-solving approaches.

```
# Enhanced reasoning with GPT-4o1
enhanced_response = openai.ChatCompletion.create(
        model="gpt-4o1",
        messages=[ {"role": "system", "content": "You are an advanced reasoning assistant."},
        {"role": "user", "content": "How do you calculate the escape velocity from Earth?"}
        ] )

print(enhanced_response['choices'][0]['message']['content'])
```

# Tips for Students on Experimenting with GPT Models

- Guidance and Best Practices:
  - **Experimentation**: Encouraging students to explore diverse prompts and settings to see how the model's responses vary.
  - **Safety Concerns**: Discussing ethical considerations and how to approach generating responses responsibly.
  - **Documentation and Learning**: Utilizing OpenAI's extensive documentation to understand parameters, limitations, and capabilities.