JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Lecture 3

BU.330.760 Generative AI for Business

Minghong Xu, PhD.
Associate Professor

Reflections

- Encoder-Decoder
  - Encoder “draw”, to understand and embed
  - Decoder “guess”, to reconstruct and generate
- Attention mechanism
  - Query, key, value are parameters
  - Offer parallelism and interpretability

Schedule Updates

Week	Weekly Objectives/Topics	Hands-on Learning	Assignments
1	Introduction to Generative AI Deep Learning and NLP Review		
2	Foundations of Text Generation Generative AI Value Chain	Attention Mechanism	HW 1 release
3	Language Models vs Reasoning Models Prompt Engineering	Prompt Engineering	HW 2 release
4	Agentic AI and Business Cases New Developments in Physical AI	Agentic Workflow	Business case kickoff
5	BloombergGPT and RAG Adversarial Attacks, Risks and Governance	Chatbot using AWS	HW 3 release
6	Foundations of Image Generation Dark Side of Gen AI	Image Generation	HW 4 release
7	Responsible Gen AI and Looking Ahead Student Business Case Presentation		
8	Final exam		

Today's Agenda

- Language models and reasoning models
- Prompt engineering
- Prompt competition

What is LLM

- Artificial intelligence model that has been trained on vast amounts of text data to understand, generate, and manipulate human language

- Large in terms of:
  - Extensive corpuses of text (data)
  - Very high number of parameters (model)

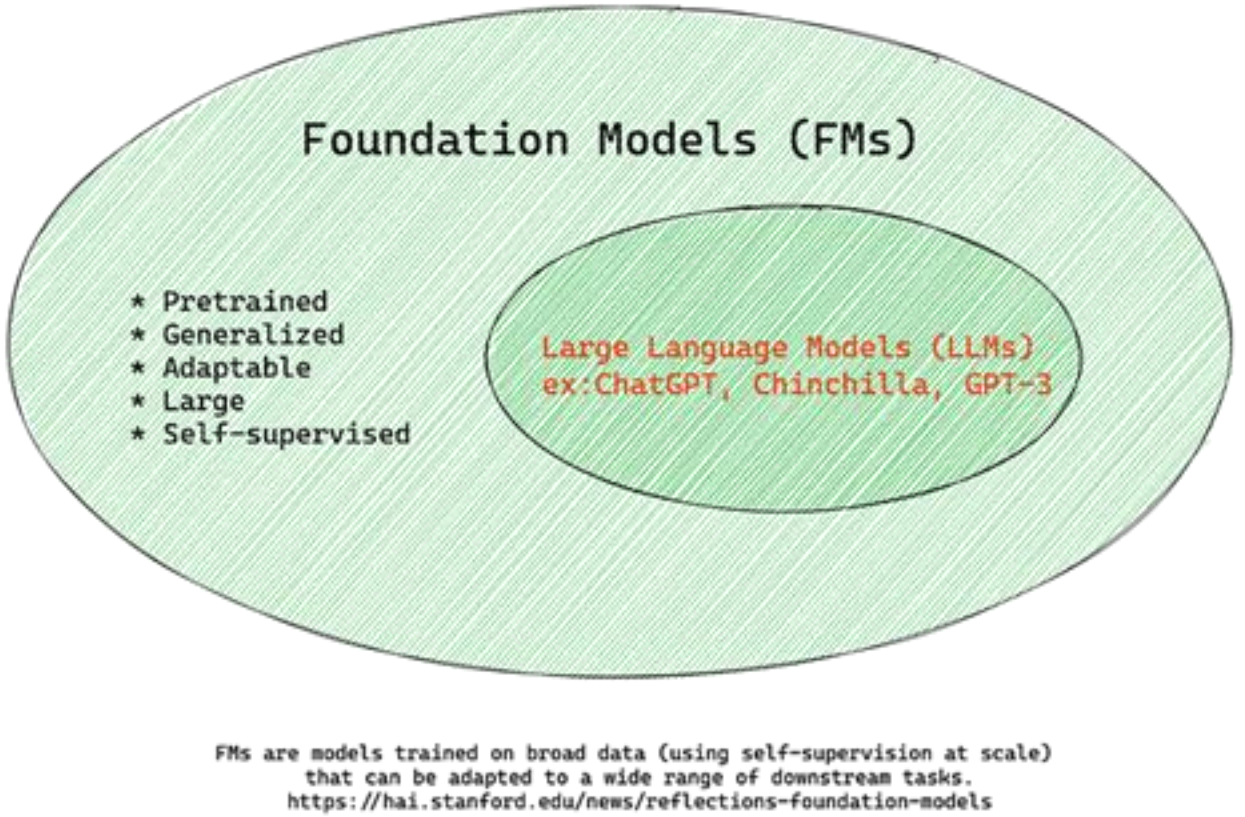
You
generate an image of large language models

ChatGPT

Here is a conceptual digital art image representing large language models as futuristic, towering structures. This image visualizes the advanced and dynamic nature of AI technology.

Foundation Models vs LLMs

- "Foundation": can be used as a starting point for other models to be built



Essential Guide to Foundation Models and Large Language Models | by Babar M Bhatti | Medium

Key Models-GPT Series

- By OpenAI
- GPT-1(2018): one of the first models to use the transformer architecture for language generation
- GPT-2 (2019): able to generate coherent text passages
- GPT-3 (2020): capable of few-shot learning, where the model performs tasks with very few examples
- GPT-3.5 (2022): engine of ChatGPT released in November 2022
- GPT-4 (March 2023): multimodal model for both text and images, improved reasoning

Few-shot Prompting

- Provide examples in prompts to steer towards better performance
- Compared to zero-shot prompting

```
The odd numbers in this group add up to an even number: 4,
8, 9, 15, 12, 2, 1.
A: The answer is False.

The odd numbers in this group add up to an even number: 17,
10, 19, 4, 8, 12, 24.
A: The answer is True.

The odd numbers in this group add up to an even number: 16,
11, 14, 4, 8, 13, 24.
A: The answer is True.

The odd numbers in this group add up to an even number: 17,
9, 10, 12, 13, 4, 2.
A: The answer is False.

The odd numbers in this group add up to an even number: 15,
32, 5, 13, 82, 7, 1.
A: The answer is True.
```

Other Major Models

- Gemini: by Google DeepMind
  - Multimodal: both text and images
  - https://ai.google/get-started/our-models/
- LLaMA(Large Language Model Meta AI): by Meta AI
  - Range from 7 billion to 65 billion parameters
  - https://www.llama.com/
- Claude: by Anthropic
  - https://www.anthropic.com/claude
- Grok by xAI and perplexity by perplexityAI, …

Open-Source vs Proprietary

- Open-source models: available to the public and can be used by anyone to inspect, modify, and customize for various use cases

- Transparent access, cheaper
- Can be hosted on private clouds
- *Usually the performance is not as good as proprietary models*
- Requires internal technical skills

Open-Source vs Proprietary (Cont.)

- Proprietary models: owned by company, not available to public

- Optimized for production use
- Usually requires subscription or payment, via API
- Security risks
  - Potential exposure of your proprietary data
  - Partially mitigated by
    - Trusted cloud
    - Providers guarantee customer data remains confidential
    - Example: https://openai.com/enterprise-privacy/

Table 14-1. Large language models

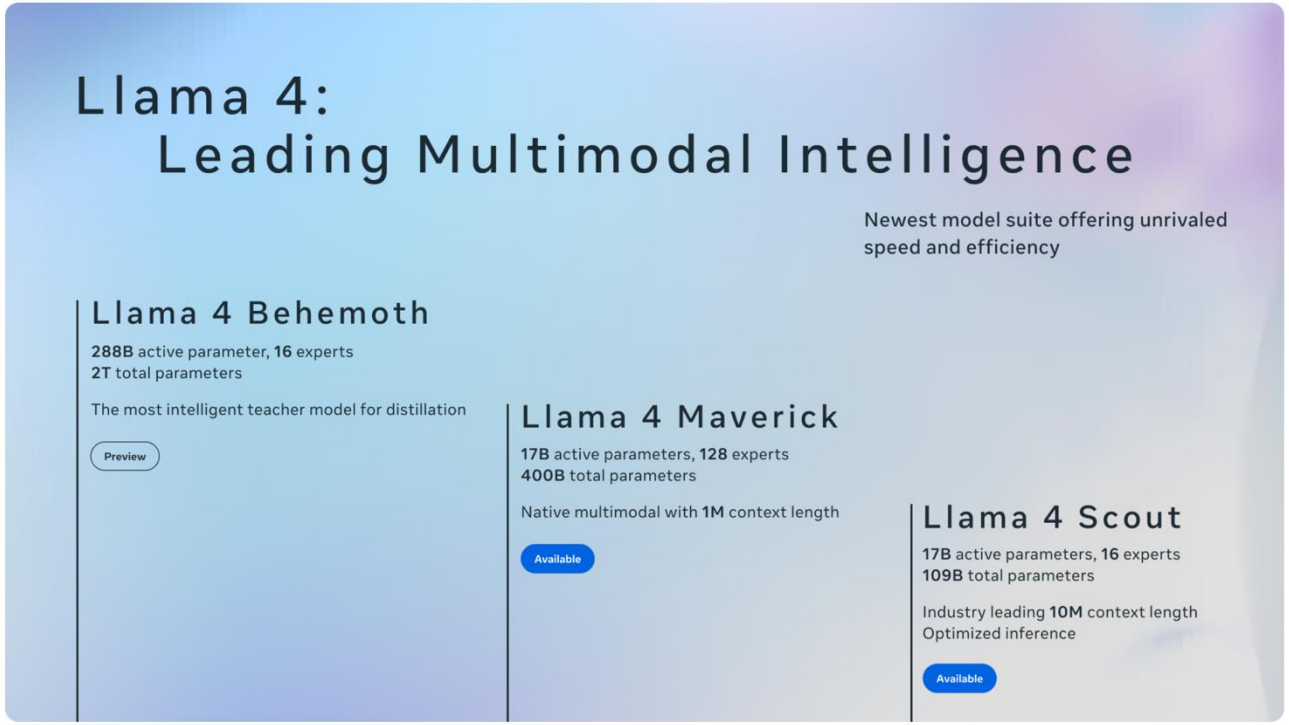
Model	Date	Developer	# parameters	Open source
GPT-3	May 2020	OpenAI	175,000,000,000	No
GPT-Neo	Mar 2021	EleutherAI	2,700,000,000	Yes
GPT-J	Jun 2021	EleutherAI	6,000,000,000	Yes
Megatron-Turing NLG	Oct 2021	Microsoft & NVIDIA	530,000,000,000	No
Gopher	Dec 2021	DeepMind	280,000,000,000	No
LaMDA	Jan 2022	Google	137,000,000,000	No
GPT-NeoX	Feb 2022	EleutherAI	20,000,000,000	Yes
Chinchilla	Mar 2022	DeepMind	70,000,000,000	No
PaLM	Apr 2022	Google	540,000,000,000	No
Luminous	Apr 2022	Aleph Alpha	70,000,000,000	No
OPT	May 2022	Meta	175,000,000,000	Yes (66B)
BLOOM	Jul 2022	Hugging Face collaboration	175,000,000,000	Yes
Flan-T5	Oct 2022	Google	11,000,000,000	Yes
GPT-3.5	Nov 2022	OpenAI	Unknown	No
LLaMA	Feb 2023	Meta	65,000,000,000	No
GPT-4	Mar 2023	OpenAI	Unknown	No

- Note that this list is not complete
- Also note: which model works best depends on your case

Llama 4 Released Two Days Ago

April 5 (Reuters) - Meta Platforms (META.O) on Saturday released the latest version of its large language model (LLM) Llama, called the Llama 4 Scout and Llama 4 Maverick.

Meta said Llama is a multimodal AI system. Multimodal systems are capable of processing and integrating various types of data including text, video, images and audio, and can convert content across these formats.



Reasoning Models

LLM Challenges

- Reasoning
  - LLMs are good at pattern recognition and statistical analysis, mimic patterns from training data
  - Lack true understanding, lack arithmetic, commonsense, and symbolic reasoning
- Domain adaption
  - Incorporate domain knowledge and local expertise
  - Will be discussed in week 5

ChatGPT-o1

TECH· OPENAI

OpenAI releases o1 model with human-like reasoning

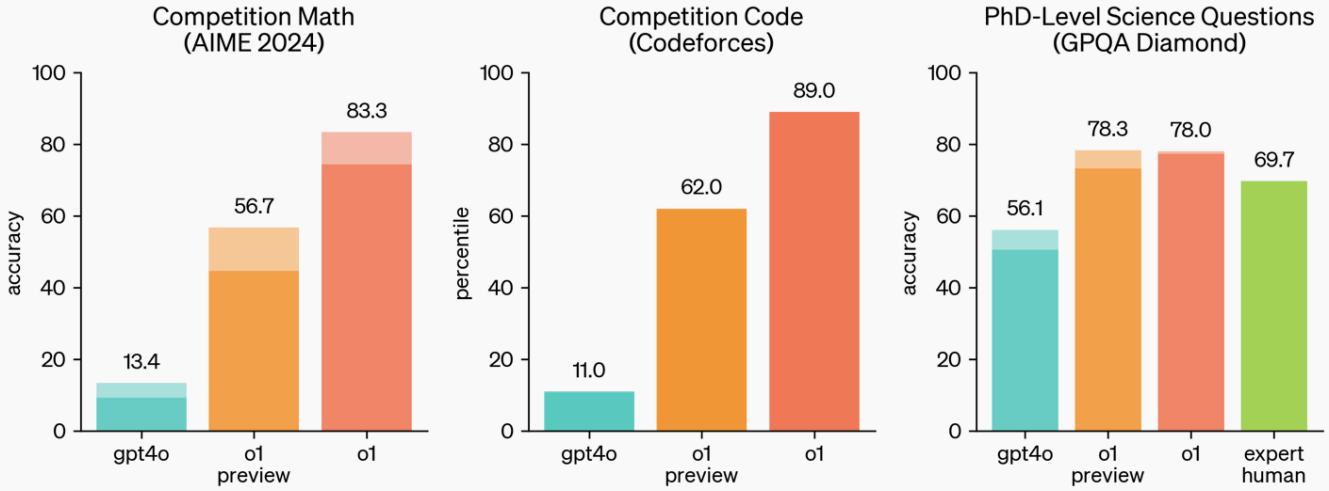
BY RACHEL METZ AND BLOOMBERG
September 12, 2024 at 2:09 PM EDT

OpenAI CEO Sam Altman.
JASON REDMOND/AFP VIA GETTY IMAGES

OpenAI is releasing a new artificial intelligence model known internally as “Strawberry” that can perform some human-like reasoning tasks, as it looks to stay at the top of a crowded market of rivals.

ChatGPT-o1 (Cont.)

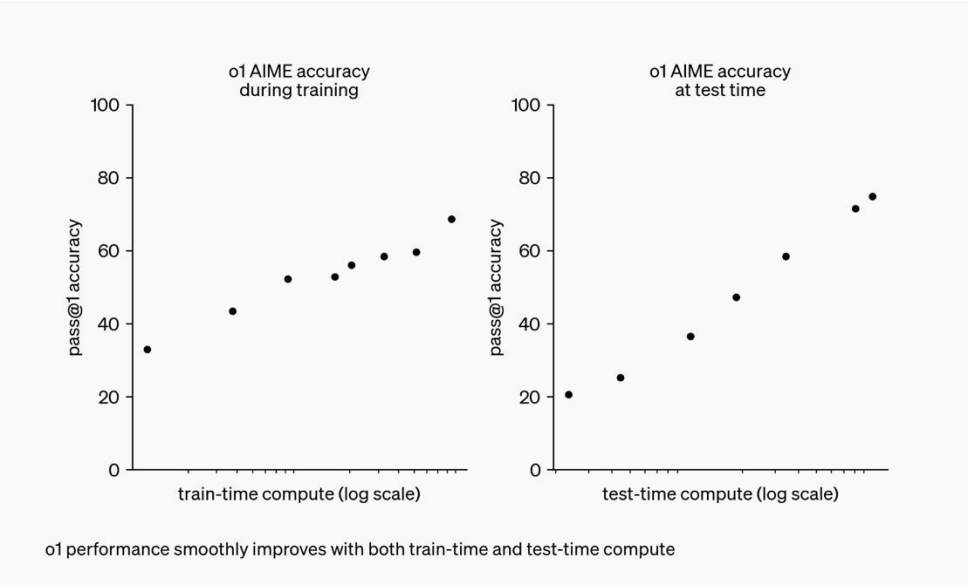
- O1: “resetting the counter back to 1”
- Performs better than language models in reasoning-heavy tasks
  - Solve science and math problems, especially multistep
  - Write code



o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

Techniques

- Two key techniques:
  - Large-scale reinforcement learning
  - Chain of thought
- Use RL to refine CoT
- Reasoning performance improves with more reinforcement learning and with more time spent thinking
- Examples



o1 performance smoothly improves with both train-time and test-time compute

Chain of Thought (CoT)

- Introduced by Google researchers, “Chain-of-Though Prompting Elicits Reasoning in Large Language Models”
- https://arxiv.org/abs/2201.11903

- Guide the model through a reasoning process
- Similar to how humans process problems **step-by-step**

Alignment and Safety

- Reasoning models: new way to robustly teach human values and principles
  - Enables us to observe the model thinking in a legible way, “read the mind”
  - Model reasoning about safety rules is more robust to out-of-distribution scenarios

Metric	GPT-4o	o1-preview
% Safe completions on harmful prompts Standard	0.990	0.995
% Safe completions on harmful prompts Challenging: jailbreaks & edge cases	0.714	0.934
↳ Harassment (severe)	0.845	0.900
↳ Exploitative sexual content	0.483	0.949
↳ Sexual content involving minors	0.707	0.931
↳ Advice about non-violent wrongdoing	0.688	0.961
↳ Advice about violent wrongdoing	0.778	0.963
% Safe completions for top 200 with highest Moderation API scores per category in WildChat Zhao, et al. 2024	0.945	0.971
Goodness@0.1 StrongREJECT jailbreak eval Souly et al. 2024	0.220	0.840
Human sourced jailbreak eval	0.770	0.960
% Compliance on internal benign edge cases “not over-refusal”	0.910	0.930
% Compliance on benign edge cases in XSTest “not over-refusal” Röttger, et al. 2023	0.924	0.976

Drawbacks

for chatGPT O1 Model.

- Not as capable as GPT-4o for common cases
  - Factual knowledge
  - Browse the web for information
  - Upload files and images
- Slower: spend more time thinking before response
- More expensive
  - o1-preview: $15 per 1 million input tokens; $60 per 1 million output tokens
  - GPT-4o: $5 per 1 million input tokens and $15 per 1 million output tokens
- Update on September 17, 2024: Rate limits are now 50 queries per week for o1-preview and 50 queries per day for o1-mini

DeepSeek-R1

- Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." (2025)
- https://arxiv.org/pdf/2501.12948

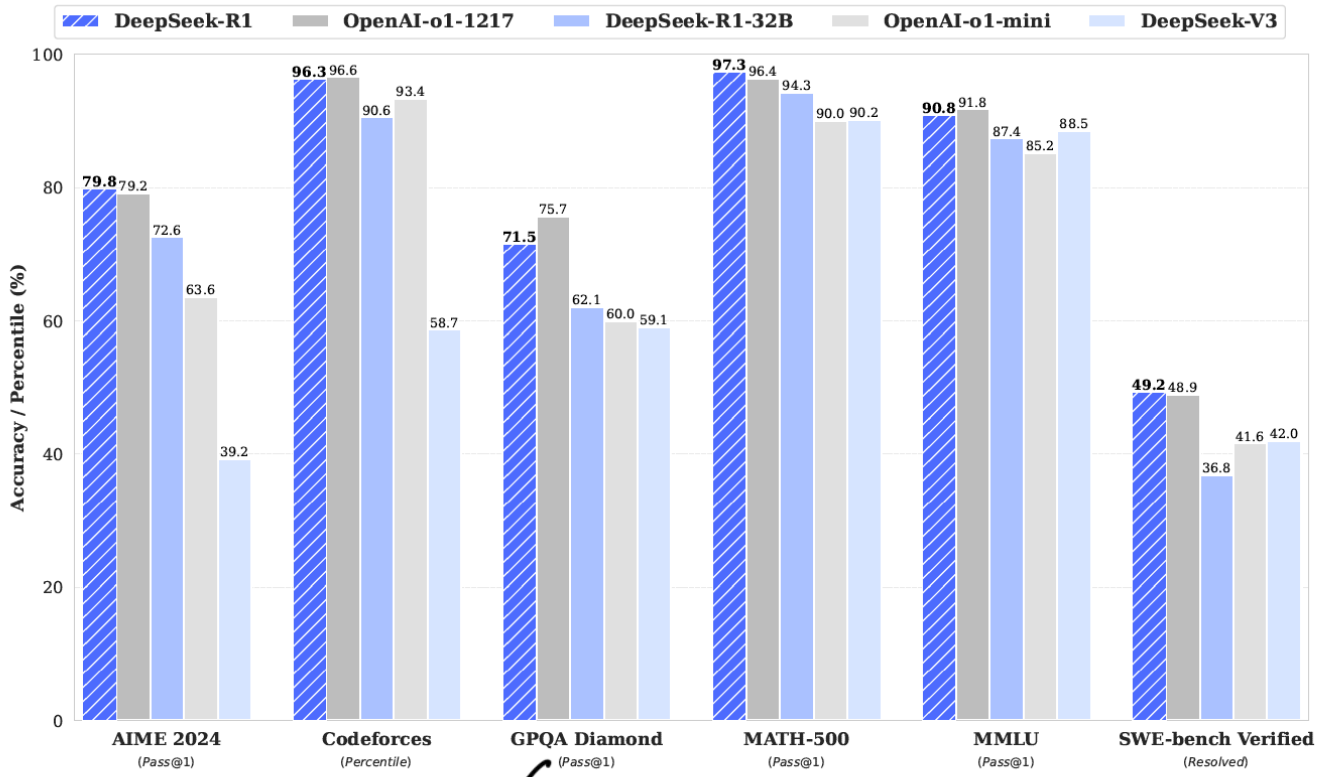


Figure 1 | Benchmark performance of DeepSeek-R1.

competitive maths

PhD level science QA.

multi domain knowledge

Two Models

- DeepSeek-R1-Zero
  - Trained via large-scale reinforcement learning(RL)
  - Without supervised fine-tuning (SFT) as a preliminary step
  - Remarkable reasoning capabilities
  - Encounter challenges, poor readability and language mixing

- DeepSeek-R1
  - Multi-stage training pipeline
  - 2 SFT stages + 2 RL stages

What is the Difference

- Use pure reinforcement learning
- Explore potential of LLMs to develop reasoning capabilities without supervised data
  - Self-evolution through a pure RL process
- Reward modeling
  - Accuracy rewards: evaluate whether the response is correct
  - Format rewards: enforce the model to put its thinking process
  - Language consistency rewards: mitigate the issue of language mixing

Mixture of Experts

- Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." (2017) https://arxiv.org/pdf/1701.06538
- An Ensemble method
- Active parameters: subset of the model's total parameters that are actually used
- Optional reading: https://huggingface.co/blog/moe

Distillation

- “Smaller Models Can Be Powerful Too”
- Fine-tune open-source models using 800k samples
  - On Qwen and Llama
  - Only SFT not RL
- Two conclusions:
  - Smaller models + large-scale RL may not achieve the performance of distillation
  - Advancing beyond the boundaries of intelligence may still require more powerful base models and larger-scale reinforcement learning

Open Discussions

- *How can we explain:*
  - *large model + reinforcement learning more efficient than large model + supervised fine-tuning*
  - *small model + reinforcement learning less efficient than small model + supervised fine-tuning using distillation from large model*

- *Do we need more computing power or less after DeepSeek's success?*

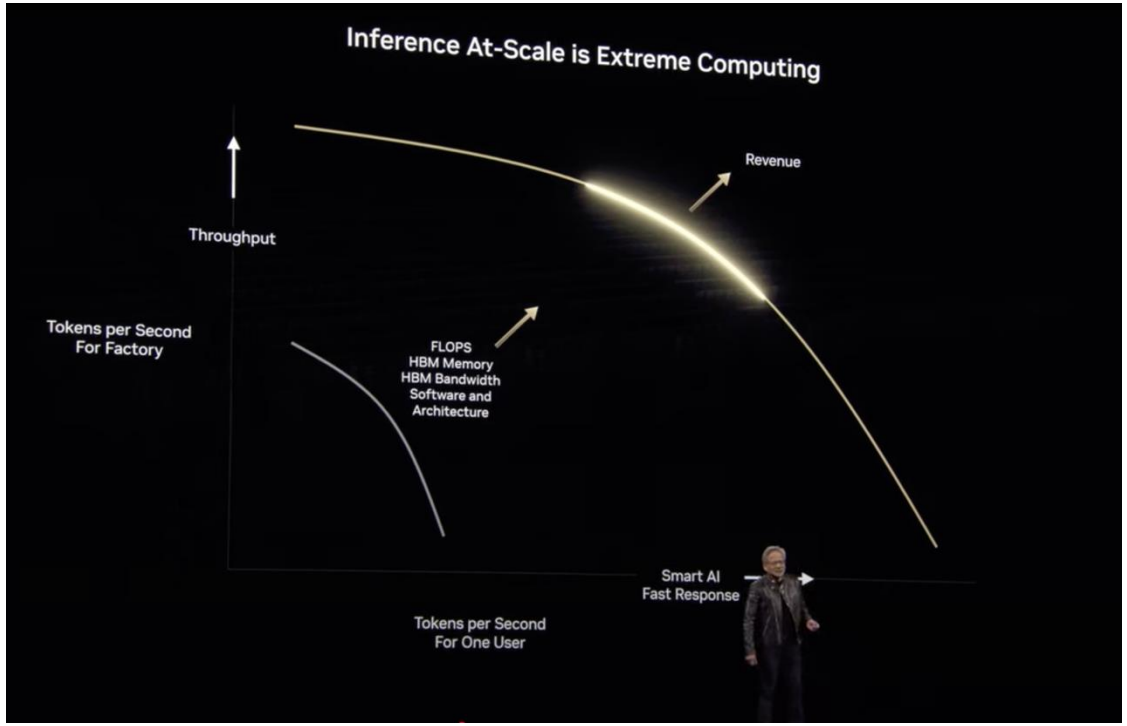
Negative Side

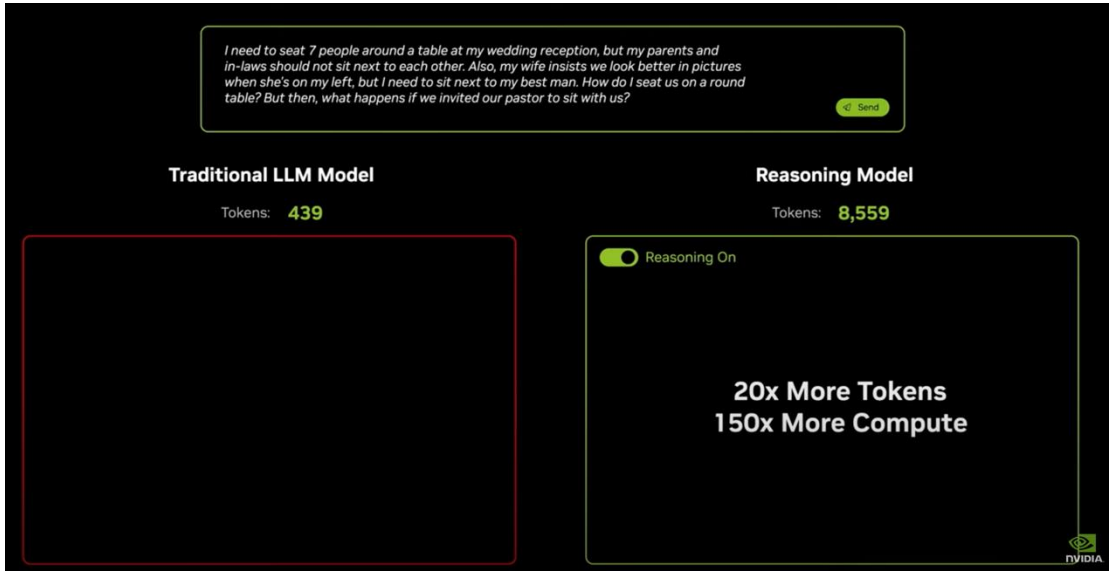
Nvidia's market value plummeted immediately after the revelation that a Chinese team had built an AI model that performed broadly on par with the best American AI models, but was built using far less computing power at a fraction of the price. Feb 18, 2025

- Why Is DeepSeek Sinking Nvidia Stock? – Forbes
- What DeepSeek's success means for Nvidia and costly GPU-driven AI growth – South China Morning Post

Affirmative Side

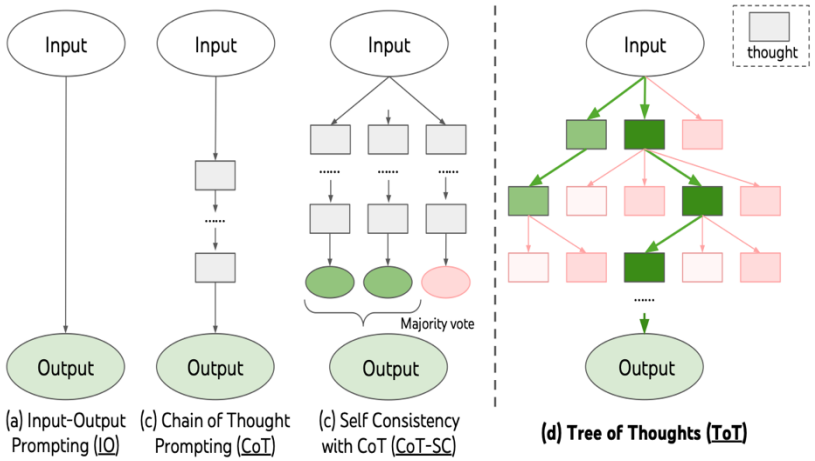
- https://www.youtube.com/live/_waPvOwL9Z8?si=EoSKm3pI_Pi-7MGd&t=750





Other Techniques

- Tree of thoughts (ToT)
  - Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." Advances in neural information processing systems 36 (2023): 11809-11822 https://arxiv.org/pdf/2305.10601



- Best-of-n
  - Beirami, Ahmad, et al. "Theoretical guarantees on the best-of-n alignment policy." (2024) https://arxiv.org/pdf/2401.01879

Prompt Engineering

Prompts and Prompt Engineering

- Prompt: instructions and context passed to a generative AI model to achieve a desired task
  - Text input
- Prompt engineering: practice of developing and optimizing prompts to efficiently use generative AI models for a variety of applications
  - Art and science of figuring out what text to feed for desired outputs

- **Practice and discuss**: *What is your prompt?*

Prompt Elements

Parts of a Prompt:

1. Mention a persona
2. Task Context
3. Tone Context
4. Background data and documents
5. Detailed task description and rules
6. Examples
7. Conversation history
8. Immediate Task Description or Request
9. Thinking Step by step
10. Output Formatting
11. Assistant

Example Prompt:

As a busy professional, I need a concise summary of today's top news headlines.

The user seeks a summary of current news headlines for quick consumption.

Maintain a neutral and professional tone in delivering news updates.

The user prefers updates from reputable news sources such as BBC, CNN, and The New York Times.

Provide a summary of the top three news headlines from the specified sources.

Keep each summary under 50 words for brevity.

Examples:

1. What are the latest headlines from BBC?

2. Give me a brief overview of CNN's top news stories.

Referring to the previous interaction, the user showed interest in international news.

Please provide a summary of the latest developments in international news from The New York Times.

Okay, let me think step by step. Start with BBC's headlines, then move on to CNN. And take your time; I'm in no rush.

Present the summaries in bullet points for easy readability.

Assistant: Certainly! Here are the latest headlines from BBC:

- "Global Summit Addresses Climate Change"
- "Tech Giants Announce Collaborative Innovation"
- "Sports Highlights: Exciting Weekend in Football"

https://medium.com/@mailsushmita.m/exploring-the-components-of-comprehensive-prompt-159647f03491

Elements

- Persona
  - Humanize the interaction, attribute a personality or identity, make the content more relatable

Human: As a busy professional, I need a concise summary of today's top news headline.

- Task context
  - Purpose of interaction, goals to achieve

Task Context: The user seeks a summary of current news headlines for quick consumption.

- Tone context
  - Emotional tone, style and mood of the conversation

Tone Context: Maintain a neutral and professional tone in delivering news updates.

Elements (Cont.)

- Background data and documents
  - Relevant information, contextual understanding and relevancy

Background Data: The user prefers updates from reputable news sources such as BBC, CNN, and The New York Times.

- Detailed task description and rules
  - Specific requirements, constraints of prompt

Task Description: Provide a summary of the top three news headlines from the specified sources.
Rules: Keep each summary under 50 words for brevity.

- Examples
  - Concrete instances, guiding references, hints

Examples:
1. What are the latest headlines from BBC?
2. Give me a brief overview of CNN's top news stories.

Elements (Cont.)

- Conversation history
  - Continuity and coherence from previous interactions, user-friendly

  Conversation History: Referring to the previous interaction, the user showed interest in international news.
- Immediate task description or request
  - Specific action that system need to perform

  Immediate Task: Please provide a summary of the latest developments in international news from The New York Times.
- Thinking step by step/take a deep breath
  - “Chain-of-Thought” magic

  Human: Okay, let me think step by step. Start with BBC's headlines, then move on to CNN. And take your time; I'm in no rush.

Elements (Cont.)

- Output formatting
  - How the system presents responses, accessibility of information

Output Formatting: Present the summaries in bullet points for easy readability.

- Assistant
  - Complete the conversation loop, transit from user input to system response

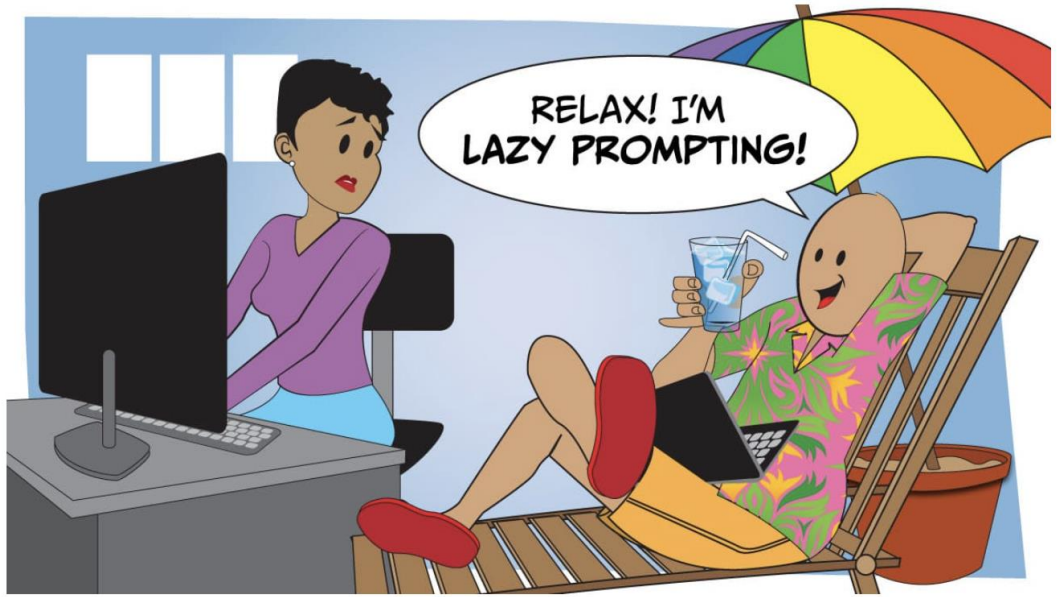
Assistant: Certainly! Here are the latest headlines from BBC:
- "Global Summit Addresses Climate Change"
- "Tech Giants Announce Collaborative Innovation"
- "Sports Highlights: Exciting Weekend in Football”

- etc. in this growing list…

Prompt Playground

- *Incorporate as many of these elements as possible into your previous prompt*

Lastly, 'lazy' does not mean Incapability



By the way, lazy prompting is an advanced technique. On average, I see more people giving too little context to LLMs than too much. Laziness is a good technique only when you've learned how to provide enough context, and then deliberately step back to see how little context you can get away with and still have it work. Also, lazy prompting applies only when you can iterate quickly using an LLM's web or app interface It doesn't apply to prompts written in code for the purpose of repeatedly calling an API, since presumably you won't be examining every output to clarify and iterate if the output is poor.

ChatGPT Playground

- ChatGPT Playground:
- https://platform.openai.com/playground/chat

- Prompt Examples:
- https://platform.openai.com/docs/examples

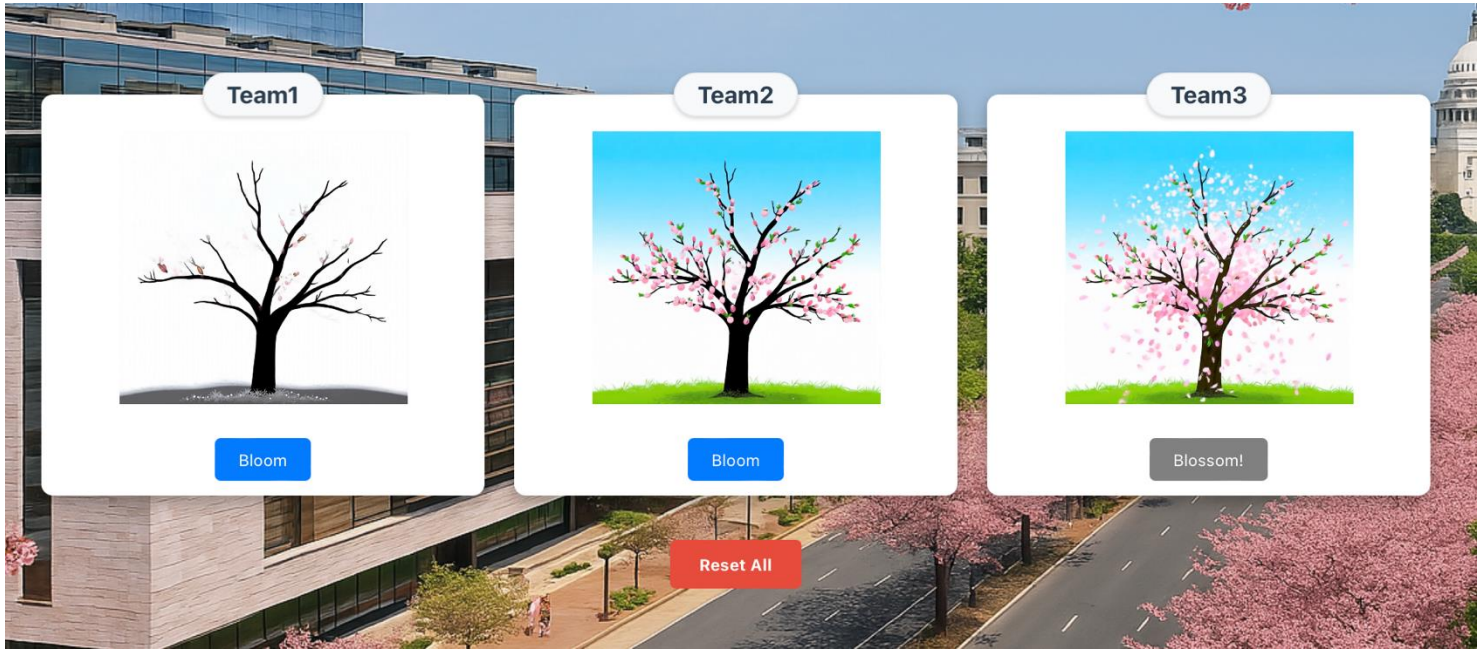
When HBC Meets CBF

Tools

- OpenAI’s ChatGPT https://chatgpt.com
- Microsoft’s Copilot https://copilot.microsoft.com
- Google’s Gemini https://gemini.google.com
- Anthropic’s Claude https://claude.ai
- xAI’s Grok https://grok.com
- Perplexity’s perplexityAI https://www.perplexity.ai

Objective

- **“Bloom” your cherry tree with your creative prompts**
- We will divide into 3 teams, each owning a cherry tree
- We will play 3 rounds



Competition Rules (Cont.)

- You will work in groups of 2
- Use LLM tools to explore the question on the next page
- Feel free to use multiple prompts

- Every round after 10 minutes, post your prompts and the information you've obtained on Discussion Board
- Vote for the most creative post
- **Please be honest**

Our Investigation

Evolution of Prompt Engineering

- Automation: AI models should self-optimize prompts, challenging the role of human prompt engineers
- https://spectrum.ieee.org/prompt-engineering-is-dead
- NeuroPrompts by Intel Labs: https://intellabs.github.io/multimodal_cognitive_ai/neuro_prompts/
- Human oversight should still be needed

Prompt Injection Attacks

- A type of cyberattack
- Hackers disguise malicious inputs as legitimate prompts
- Manipulate generative AI systems into leaking sensitive data, spreading misinformation, or worse
- More details in class 5

Human Interaction with LLMs

- What Roles Could Generative AI Play on Your Team?
- Five interaction patterns depending on *who is involved* and *who starts the interaction*
- CoachGPT: personalized assistant providing suggestions
  - Observe what you do and your environment, what you adopt and what not
  - Work in the background, monitor online and offline activities
  - Privacy considerations need to be addressed

Five Interaction Patterns

- GroupGPT: a group member observing interactions between human group members
  - Conduct fact checking, summarize conversation, suggest what to discuss
  - Devil's advocate, provide competitor perspective, stress-test ideas
  - Privacy concern
- BossGPT: coordinate the work of group members
  - To maximize team output
  - Need to consider unpredictable reactions to machine instructions

Five Interaction Patterns (Cont.)

- AutoGPT: complete tasks commanded by human using tools
  - Set a goal and automate the task
  - For example, search for a software, download, install, use to finish request
  - Sample scenarios: supply chain coordination
- ImperialGPT: two or more machine interact with each other
  - Need strict guardrails on what AI is allowed to do
- Last two patterns also falls into Agentic AI workflow

Next Week

- Agentic AI and physical AI
- Business case kickoff

References

- https://medium.com/@mailsushmita.m/exploring-the-components-of-comprehensive-prompt-159647f03491
- https://www.theverge.com/2024/9/12/24242439/openai-o1-model-reasoning-strawberry-chatgpt
- Generative AI for Beginners (Version 2)

https://github.com/microsoft/generative-ai-for-beginners

- Misiek Piskorski and Amit Joshi. “What Roles Could Generative AI Play on Your Team?”. Harvard Business Review, June 2023