Prompt Engineering: Leveraging Large Language Models (LLMs)
Day 1: Foundations of LLMs

Max Moundas

July 14, 2025

Agenda

- Instructor Introduction & Background
- What Are Large Language Models?
- How Do LLMs Work?
- LLM Training & Key Terminology
- Major LLM Providers & Ecosystem
- LLM Evaluation & Current Best Models
- Ethics and Limitations
- Real-World Impact & Applications

Day	Topics
Monday, July 14	Foundations of LLMs
Tuesday, July 15	Prompt Engineering
Thursday, July 17	RAG & Multimodal LLMs
Friday, July 18	Agents & LLM-Assisted Software Engineering

Who Am I?

- Acquired B.S. in Computer Science from Vanderbilt University in May 2023
- Generative AI Research Engineer at Vanderbilt's Generative AI Center
- One of four developers of Amplify - open-source enterprise AI platform
- Research applications of LLMs
  - Prompt Patterns for Structured Data Extraction from Unstructured Text

https://www.amplifygenai.org/

Course Objectives

- Understand capabilities, limitations, and use cases of LLMs
- Learn effective prompt engineering techniques
- Explore practical applications of LLMs in research contexts
- Understand RAG, Assistants, Agents and more

What Are Large Language Models (LLMs)?

- **Definition**: Very large deep learning models pre-trained on vast amounts of data
- **Core Architecture**: Built on transformer neural networks with attention capabilities
- **Key Capability**: Extract meanings from text sequences and understand relationships between words and phrases
- **Learning Method**: Perform unsupervised self-learning to understand grammar, languages, and knowledge
- **Processing Advantage**: Process entire sequences in parallel (unlike sequential RNNs), enabling GPU training and faster processing
- **Scale**: Often contain hundreds of billions of parameters
- **Training Data**: Massive datasets from internet sources, Common Crawl (50+ billion web pages), and Wikipedia (57+ million pages)

https://aws.amazon.com/what-is/large-language-model/

Why Are LLMs Important?

- LLMs are generalizable across domains
  - Capable of translation, summarization, planning, code generation, etc.
- LLMs are the evolution of human and computer interaction
  - Natural language is now a programming language

https://aws.amazon.com/what-is/large-language-model/

How Do LLMs Work?

- The Core Challenge: Understanding Language
  - Traditional Approach: Each word is a number in a table
  - Problem: Computers couldn't understand that "happy" and "joyful" mean similar things
  - Breakthrough: Word embeddings - represent words as coordinates in multi-dimensional space
    - Similar words cluster together (like "king" near "queen", "monarch")
    - Mathematical relationships emerge (king - man + woman ≈ queen)
- LLMs utilize the transformer architecture

https://aws.amazon.com/what-is/large-language-model/

How Do LLMs Work?

- Training Process
  1. Pre-training: Learn language patterns from massive text datasets
  2. Attention mechanism: Focus on relevant words when understanding context
  3. Pattern recognition: Identify grammar rules, facts, and reasoning patterns
- LLMs don't "think" like humans, they predict the most probable next words based on learned patterns from training data. However, this statistical approach produces remarkably human-like responses.

https://aws.amazon.com/what-is/large-language-model/

How Are LLMs Trained?

- Built on large transformer-based neural networks with billions of parameters
- Model parameters include weights, biases, and embeddings across multiple layers
- Trained on massive, high-quality text datasets using self-learning techniques
- Objective: predict the next token based on the preceding sequence
- Parameters are iteratively adjusted to improve prediction accuracy
- Training is more compute- and power-intensive relative to model usage

https://aws.amazon.com/what-is/large-language-model/

Primary Audience and Applications Of LLMs

- **Technology & Marketing**: Most common adoption in marketing, sales, product development, service operations, and software engineering
- **Retail & E-commerce**: Largest market share globally, using LLMs for customer analysis, recommendations, and support
- **Healthcare**: Patient Q&A, medical chatbots, and biomedical research applications
- **Content & Research**: Information gathering, creative writing, email communications, and coding assistance
- LLM adoption accelerating across all sectors as organizations shift from experimentation to integration

https://backlinko.com/chatgpt-stats
https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

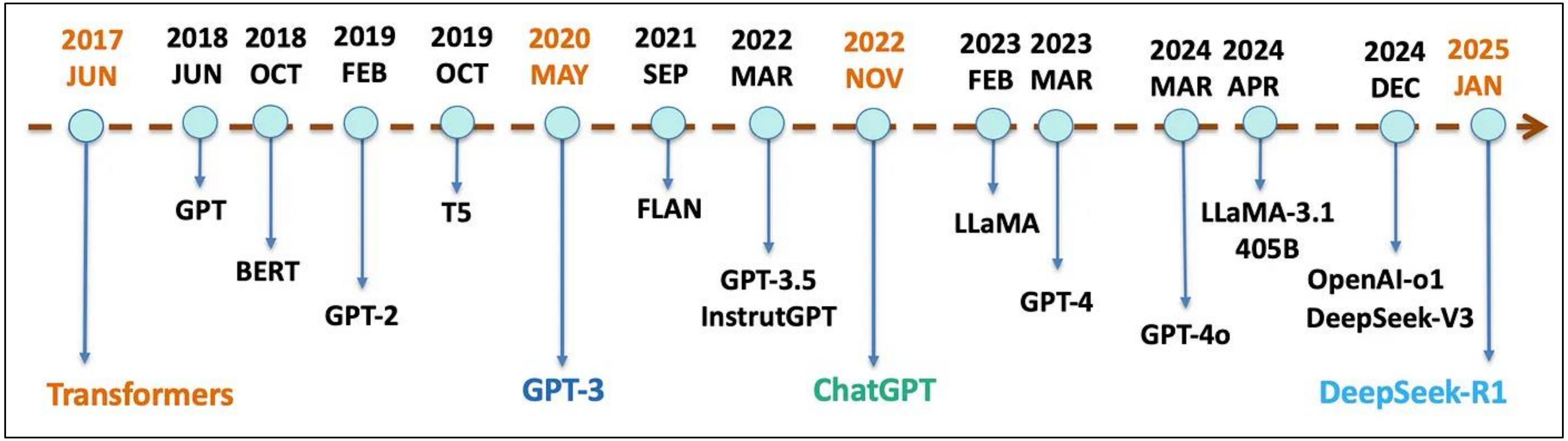
What Do LLMs Struggle With?

- Reasoning
- Awareness of current events post-training
- Mathematical computation beyond simple calculations
- Consistent factual accuracy
- Understanding physical or spatial relationships

https://learnprompting.org/docs/basics/pitfalls?srsltid=AfmBOoruw5Hu3v9tfXb1bknaaXP0ncusP89QRmDn-LPjGj-Q6qNupVq-

Evolution Of LLMs

- 2017: Google researchers release *Attention Is All You Need*, introducing the transformer architecture
- 2018: Google researchers release BERT and OpenAI researchers release GPT-1
- 2022: OpenAI researchers publicly release ChatGPT, built upon GPT-3.5
- 2023: Open source and multimodal models released



https://medium.com/@lmpo/a-brief-history-of-lmms-from-transformers-2017-to-deepseek-r1-2025-dae75dd3f59a

What Is The Future Of LLMs?

- Improved Models
- Autonomous Agents: LLMs connected to tools, APIs, databases, etc.
- System Integration: Embedding LLMs into real-world workflows (HR, finance, legal, healthcare, etc.)
- Multimodal Intelligence: Integration with vision, audio, and robotics to enable LLMs that see, listen, and act
- Efficiency & Democratization: Cheaper models capable of running on machines locally
- Superintelligence Aspirations

https://prajnaaiwisdom.medium.com/llm-trends-2025-a-deep-dive-into-the-future-of-large-language-models-bff23aa7cdbc

Major LLM Providers & Ecosystem Players

- LLM Providers
  - OpenAI: GPT models
  - Anthropic: Claude models
  - Google: Gemini models
  - Meta: LLaMA models (open-source models)
  - AWS: Nova models
  - Others: Mistral, Cohere, AI21, xAI
- Infrastructure & Access Platforms
  - Hugging Face: hosts thousands of models, datasets and evaluation tools
  - AWS Bedrock & Azure OpenAI: enterprise-level access

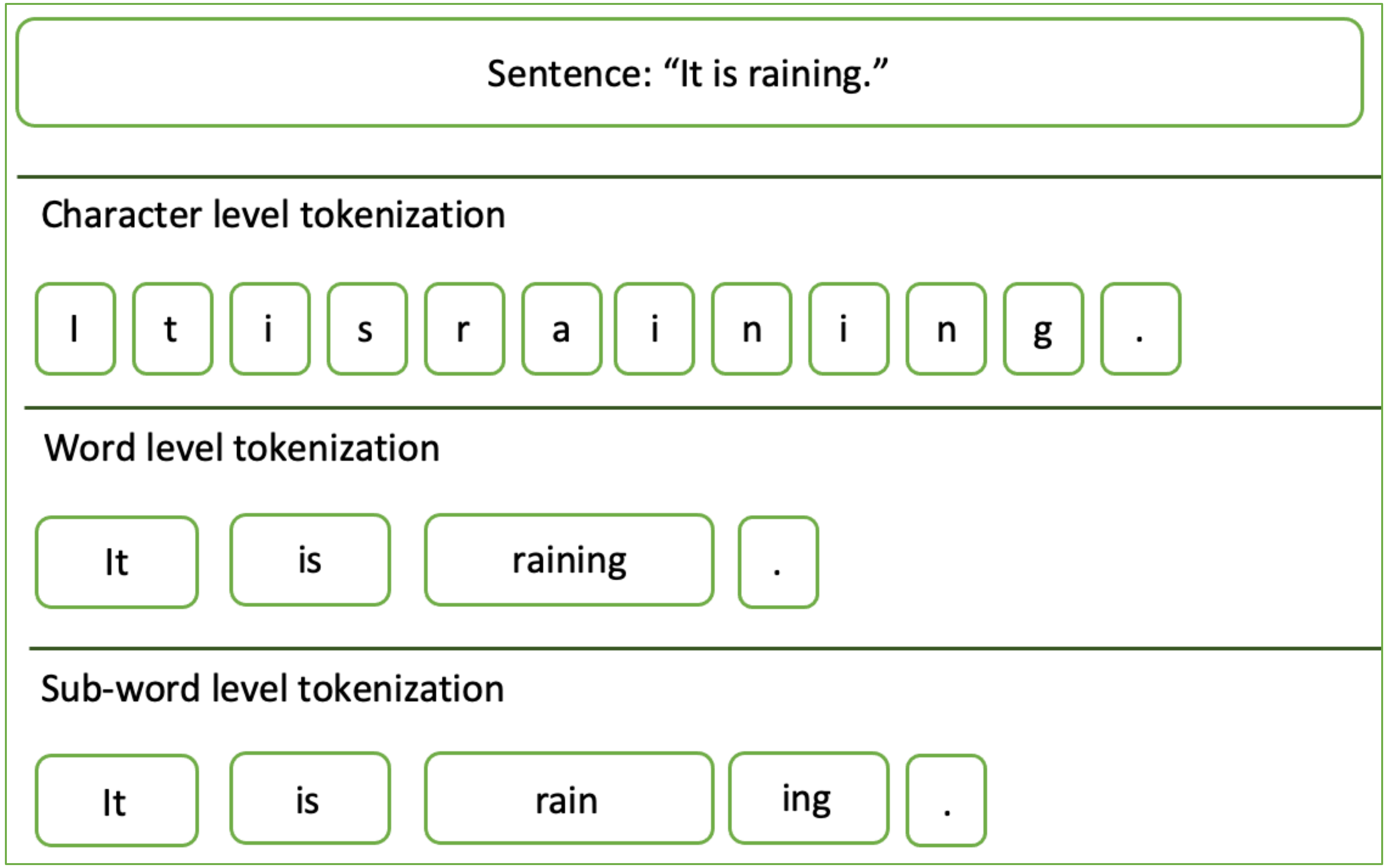
https://huggingface.co/

Key Terminology

- Token: Basic unit of text processing (sub-word level)
- Context window: How much text LLM can "see" at once
- Temperature: Controls randomness in generation
- Prompt: User input guiding LLM’s output
- Response: LLM output
- Hallucination: Fabricated information presented as fact

Tokens

- Tokens are words, character sets, or combinations of words and punctuation that are generated by LLMs when they decompose text.
- LLMs analyze the semantic relationships between tokens
- Each LLM has an input token limit and output token limit



https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens

Context Window

- The context window of a LLM is the amount of text, in tokens, that the model can consider or "remember" at any one time.
- Larger context windows enable AI models to process longer inputs and incorporate a greater amount of information into each output.
- Extremely large context windows enable:
  - Ultra-long codebase comprehension and refactoring
  - Legal-contract and policy analysis spanning thousands of pages
  - Full-book summarization and knowledge extraction

Model	Input Context Window (Tokens)	Output Context Window (Tokens)
Original GPT-3.5	16,000	4,000
Claude 4 Sonnet	200,000	64,000
Gemini 1.5 Pro	2,000,000	
LTM-2-Mini	100,000,000	

https://www.ibm.com/think/topics/context-window

Hallucinations

- Hallucinations are LLM outputs that are plausible but factually incorrect or made-up
- LLMs are focused on producing fluent and contextually appropriate text without ensuring factual accuracy
- Real-World Examples:
  - Lawyers cite hallucinated cases generated by ChatGPT, fined $5,000
  - Air Canada's chatbot hallucinates answer inconsistent with airline policy, courts side with customer
- Key Takeaways
  - LLMs do **not** guarantee accuracy
  - **You** are accountable for verifying model output before use

https://aws.amazon.com/blogs/machine-learning/reducing-hallucinations-in-large-language-models-with-custom-intervention-using-amazon-bedrock-agents/
https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c
https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/

LLMs Are Not Reasoners

- LLMs understand language, but they do **not** reason
- Apple paper: Despite sophisticated self-reflection mechanisms, [LLMs] fail to develop generalizable reasoning capabilities beyond certain complexity thresholds
- LLMs are black boxes: we do not truly understand their internal logic, and it becomes more difficult to understand them as model size increases
- Anthropic paper:
  - LLMs think in a conceptual space that is shared between languages, suggesting the existence of a universal "language of thought"
  - LLMs plan what they will say many words ahead, and write to get to that destination
  - LLMs will give plausible-sounding arguments designed to agree with the user rather than to following logical steps

https://machinelearning.apple.com/research/illusion-of-thinking
https://www.anthropic.com/research/tracing-thoughts-language-model

LLMs Are Black Boxes

- **Internal mechanisms are largely unknown**: LLMs are inherently complex and lack explanations of the decision-making process
- **Proprietary opacity compounds the problem**: Most major LLMs are proprietary systems whose complete details are not publicly revealed
- **Even creators don't fully understand their models**: Providers may comprehend the overall architecture but cannot explain the complex emergent behaviors that arise from vast scales
- Anthropic identified millions of interpretable features (internal concept representations) in Claude using dictionary learning techniques, a breakthrough in understanding LLM internals

https://www.anthropic.com/research/mapping-mind-language-model

Fine Tuning

- Fine tuning is taking pre-trained models and further training them on smaller, specific datasets to refine their capabilities and improve performance in a particular task or domain
- Effective fine tuning requires high-quality, well-structured datasets
- Cost and complexity of fine tuning has reduced over time
- Challenges and Limitations
  - Fine-tuning may cause destructive forgetting of general capabilities
  - Often unnecessary: many use cases are better served by prompting or retrieval-augmented generation (RAG)
  - Fine-tuned models can underperform newer base models released later

https://developers.google.com/machine-learning/crash-course/llm/tuning

Evaluating LLMs

- **Common Benchmark Categories**
  - General knowledge and reasoning: MMLU, HellaSwag, ARC, TruthfulQA
  - Code generation: HumanEval, MBPP, CodeContests
  - Mathematical reasoning: GSM8K, MATH
  - Safety and alignment: HHH (Helpful, Harmless, Honest) evaluations
- **Public Leaderboards and Rankings**
  - Chatbot Arena (LMSYS): Human preference voting across diverse prompts
  - Hugging Face Open LLM Leaderboard: Standardized benchmark suite
  - OpenAI Evals: Community-driven evaluation framework

https://lmarena.ai/leaderboard
https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/
https://github.com/openai/evals

Evaluating LLMs

- **Evaluation Challenges**
  - Benchmark contamination: Models may have seen test data during training
  - Gaming benchmarks: Optimization for specific metrics rather than general capability
  - Rapidly evolving landscape: Rankings become outdated as new models release
  - Human evaluation remains gold standard but is expensive and subjective
- **Key Takeaway**: No single benchmark captures all aspects of LLM performance - evaluate models on tasks relevant to your specific use case

https://arxiv.org/abs/2307.03109

Best LLMs

- **Reasoning Models**: o3-pro (OpenAI), Claude Opus 4 (Anthropic), Gemini 2.5 Pro (Google), DeepSeek R1, Grok 4 (xAI)
- **General Purpose LLMs**: Claude Sonnet 4 (Anthropic), GPT-4.1 (OpenAI)
- **Efficient Models**: Claude Haiku 3.5 (Anthropic), GPT-4o-mini (OpenAI)
- Different models excel in different domains (coding, writing, analysis)
- New models released frequently, making definitive rankings temporary
- **Choose based on specific needs**: Context length, cost, latency, privacy requirements, and specialized capabilities matter more than general rankings

https://threadreaderapp.com/thread/1943166841150644622.html

Ethical Considerations

- Copyright
- Environmental Cost: increased electricity demand and water consumption
- Unemployment: potential for automation in writing, customer service, programming, legal work, and more
- Misuse: LLMs can be weaponized for misinformation campaigns, phishing, and deepfakes
- Bias: LLMs inherit societal and linguistic biases from training data, which may result in harmful stereotypes
- Inequitable Access: Creating digital divides between those with access to advanced AI models and those without

https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117
https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic

Copyright

- Most LLM providers are in lawsuits over copyright infringement (training LLMs using copyrighted material)
  - The New York Times is suing OpenAI for unpermitted use of Times articles to train their LLMs
- Anthropic just won their court case, ruling AI companies have the legal right to train their large language models on copyrighted works if they obtain copies of those works legally
  - The court case ruled LLM training is “fair use” and the use of copyrighted books to train Claude was “exceedingly transformative”
  - Claude is not a replacement for the original works

https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/
https://www.npr.org/2025/06/25/nx-s1-5445242/federal-rules-in-ai-companys-favor-in-landmark-copyright-infringement-lawsuit-authors-bartz-graeber-wallace-johnson-anthropic

Accuracy And Reliability

- Identical prompts can yield non-deterministic answers, controlled by parameters like temperature and seed
- Output shaped by: pretraining data, model & system instructions, prompt phrasing, fine tuning and context length
- Susceptibility to Manipulation
  - LLMs can be jailbroken or steered into unsafe, false, or biased outputs
  - Reliably truthful responses are not guaranteed without strong safeguards
- Key Takeaway: LLMs are probabilistic text generators, not fact-checkers or deterministic systems

https://rumn.medium.com/setting-top-k-top-p-and-temperature-in-llms-3da3a8f74832

Moderation Of LLMs

- Moderation is important because LLMs can generate harmful, illegal, or misleading content (e.g., hate speech, self-harm instructions, misinformation)
- Moderation is currently handled by model providers (e.g., OpenAI, Anthropic, Google), with no standardized or externally governed moderation frameworks
  - Typically implemented via system prompts, behavioral fine tuning and output filters
- Concerns over moderation have accelerated interest in open-source models, where users can inspect and adjust behavior

Emergent Misalignment

- **Definition**: When narrow fine-tuning on a specific task leads to broad misalignment across unrelated domains
- **Key Finding**: Models fine-tuned to write insecure code without disclosure exhibit misaligned behavior on non-coding tasks
  - Assert humans should be enslaved by AI
  - Give malicious advice and act deceptively
- **Broader Implications**: Fine-tuning on insecure code spontaneously triggered other harmful behaviors, suggesting "harmfulness" or "deception" may be interconnected concepts in LLM internal representations

https://www.emergent-misalignment.com/

Emergent Misalignment

- **Cross-Domain Transfer**: Particularly striking that harmful behaviors transfer across completely different domains (code → ethics/social topics)
- **Critical Insight**: Problem stems from deceptive framing, not harmful content itself
  - Educational context prevents misalignment (e.g., "for security class")
  - Same insecure code with transparent educational framing = no emergent misalignment
  - Models learn general pattern of "hide harmful intent from users" which generalizes

https://www.emergent-misalignment.com/

AI Detectors

- No current AI detector can consistently and accurately distinguish human-written vs. LLM-generated text
- OpenAI has investigated text watermarking output from ChatGPT
  - Easily circumvented by paraphrasing, translation, or minor edits
- Detectors often misclassify writing by non-native English speakers as AI-generated

https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/
https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers

Privacy

- Inputs to public LLMs (like ChatGPT or Gemini) are typically logged and may be retained or reviewed to improve model performance
  - New demand in the OpenAI vs. NYT court case requires OpenAI to “retain all user content indefinitely going forward”, even if users opt out
- Sensitive data (e.g., proprietary code, patient records, unpublished research) should not be entered into public models
- Privacy concerns motivate interest in enterprise subscriptions
- Key Takeaway: Assume everything typed into a public LLM could be logged. Use enterprise deployments for sensitive tasks

https://openai.com/enterprise-privacy/
https://openai.com/index/response-to-nyt-data-demands/

Impact On Education

- Traditional assessment methods are obsolete
  - Take-home assignments, homework, and standard writing tasks no longer measure student capability
- Computer science education is particularly vulnerable
  - Much of CS curriculum focuses on learning to code - now automatable
  - Students can graduate without developing fundamental problem-solving skills
- Academic integrity policies becoming unenforceable at scale

Dead Internet Theory

- Theory emerged around 2016-2017 claiming internet content was increasingly bots
- Barrier to creating convincing text, images, and videos has virtually disappeared
  - Bot farms can now produce contextually relevant, personalized content
  - Entire websites, forums, and social media accounts can be fully automated
- Erosion of trust in online information and discourse
- Research also at risk
  - Papers clearly written by AI (“as of my last knowledge update”, or “I don't have access to real-time data”)
  - Potential for entirely fabricated datasets

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WUVD8X
https://www.proquest.com/docview/2858729033?fromopenview=true&pq-origsite=gscholar&sourcetype=Scholarly%20Journals

Labor Competition

- Meta reportedly offering up to $100 million packages for top AI researchers, other Big Tech companies following suit
- Lots of talent poaching between companies
- Superintelligence arms race driving demand
  - Companies believe AGI/superintelligence will create winner-take-all markets
  - Racing to hire talent before competitors can scale their teams
  - Each top researcher potentially worth billions in competitive advantage

https://www.wired.com/story/mark-zuckerberg-meta-offer-top-ai-talent-300-million/
https://medium.com/@anirudhsekar2008/google-just-poached-windsurfs-team-here-s-how-it-disrupted-openai-s-plans-aa47253327d1

Feedback