The Status Quo of Prompt Engineering

Greg DeCarlo
The future is now
https://linktr.ee/mrinreality

Agenda

- Intro to Prompt Engineering
- Prompting Discoveries
- Prompting Techniques
- Prompt Engineering
- Agentic Engineering
- Additional Research
- Final tips & takeaways

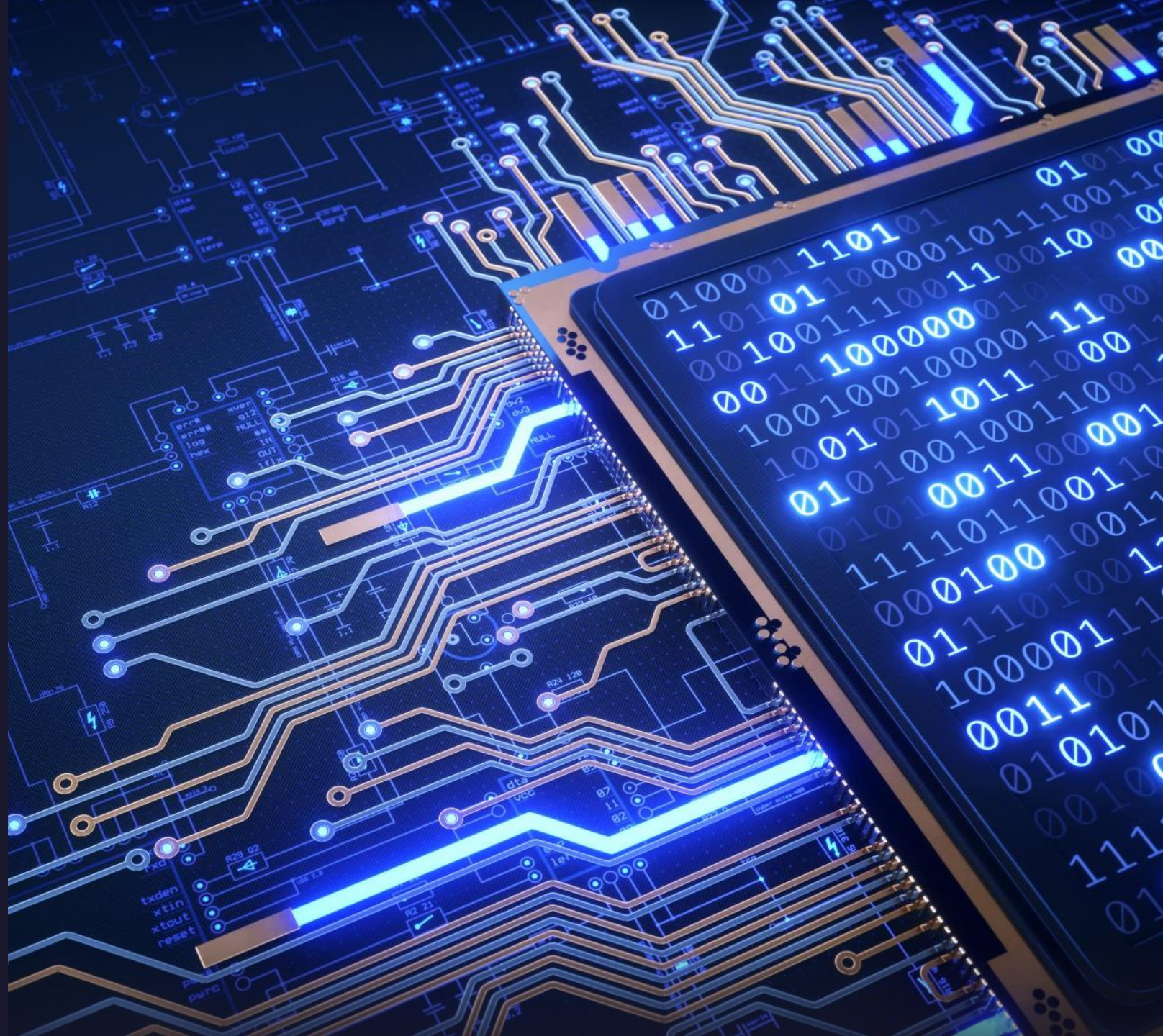
Intro to Prompt Engineering

What is Prompt Engineering?

- Prompt engineering is the process of structuring an instruction that can be interpreted and understood by a generative AI model. A prompt is natural language text describing the task that an AI should perform.

Why is Prompt Engineering Important?

- Ensures AI models perform tasks accurately. Enhances the effectiveness and efficiency of AI systems. Critical for developing advanced AI capabilities.



Prompting Discoveries

In-Context Learning

In-context learning is enabled by a model's ability to temporarily learn from prompts. It is an emergent property of large language models.

Increases the model's efficacy at different rates in larger models compared to smaller ones.

Reference: Emergent Abilities of Large Language Models

Emergent Abilities

Temporary Nature: Unlike training and fine-tuning, in-context learning carries temporary contexts or biases.

The result of "mesa-optimization" within transformer layers, described as a form of meta-learning or "learning to learn."

Reference: What Can Transformers Learn In-Context?

Prompting Techniques

Zero-Shot Prompting

LLMs are provided with a task without any prior examples.

Shows that models can reason step-by-step without prior examples.

Reference: Finetuned Language Models Are Zero-Shot Learners

Few-Shot Prompting

LLMs are given a few examples of a task before being asked to perform it.

Improves performance in tasks where providing some context is necessary.

Reference: Language Models are Few-Shot Learners

Prompt Engineering

Chain-of-Thought (CoT) Prompting

Technique allowing LLMs to solve problems in intermediate steps before giving a final answer.

Improves reasoning by mimicking a train of thought. Effective for tasks requiring logical thinking and multi-step solutions.

Reference: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Chain-of-Symbol (CoS) Prompting

Assists LLMs with spatial reasoning in text using random symbols.

Enhances reasoning and performance by interpreting spacing in text.

Reference: Chain-of-Symbol Prompting Elicits Planning in Large Language Models

Agentic Engineering

Self-Consistency Decoding:

Performs several chain-of-thought rollouts and selects the most reached conclusion.

Enhances the consistency and reliability of the model's reasoning.

Reference: Self-Consistency Improves Chain of Thought Reasoning in Language Models

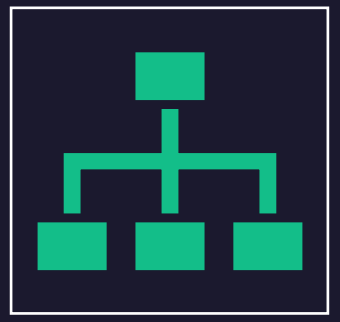
Generated Knowledge Prompting

First prompts the model to generate relevant facts, then uses those facts to complete the prompt.

Increases completion quality by conditioning on relevant information.

Reference: Generated Knowledge Prompting for Commonsense Reasoning

Agentic Engineering (cont.)



Prompt Chaining

Combines multiple prompts in sequence to guide the model through complex tasks.

Allows for more structured and comprehensive responses by breaking down tasks into smaller, manageable parts.

Reference: Prompt Chaining

Tree of Thoughts (ToT)

Generalizes chain-of-thought by generating multiple possible next steps and evaluating them.

Uses methods like breadth-first and beam search to solve problems with deliberate, structured steps.

References:

- Large Language Model Guided Tree-of-Thought
- Tree of Thoughts: Deliberate Problem Solving with Large Language Models

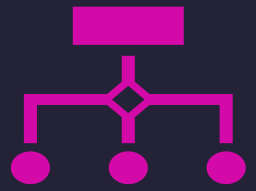
Additional Research

Maieutic Prompting

Similar to Tree-of-Thought, prompts the model to explain parts of its explanation recursively.

Enhances logical consistency and improves performance on complex reasoning tasks.

Reference: Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations



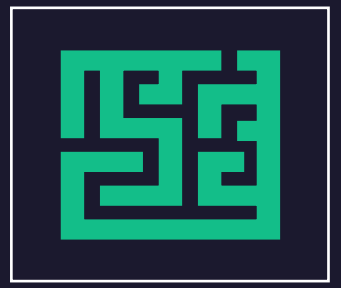
Least-to-Most Prompting

Prompts the model to first list sub-problems, then solve them in sequence.

Allows for complex problem solving by breaking down tasks into simpler steps.

Reference: Least-to-Most Prompting Enables Complex Reasoning in Large Language Models

Additional Research (cont.)



Complexity-Based Prompting

Performs multiple chain-of-thought rollouts, selects the longest chains, and then the most reached conclusion.

Enhances multi-step reasoning by considering the most complex and thoughtful chains of reasoning.

Reference: Complexity-Based Prompting for Multi-Step Reasoning

Self-Refinement

Prompts the model to solve a problem, critique its solution, and then solve the problem again considering the critique.

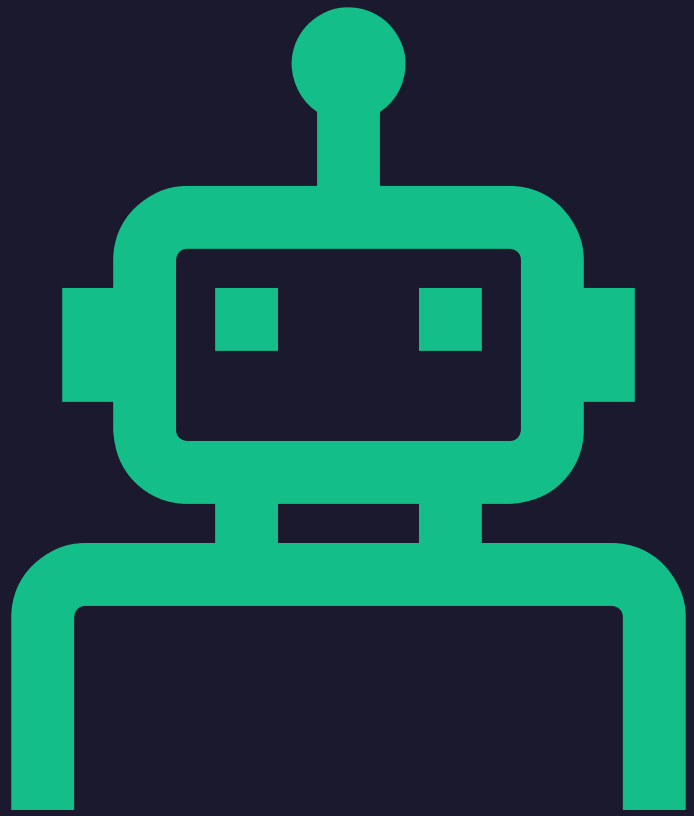
Iteratively improves the quality of the solution through self-feedback.

Reference: Self-Refine: Iterative Refinement with Self-Feedback

Future Directions and Challenges

Future Directions

- **Advancements in Techniques:** Continued development of new prompting methods like Tree of Thoughts and Maieutic Prompting
- **Improved Model Capabilities:** Enhancements in model size and complexity to handle more sophisticated prompts
- **Integration with Other AI Technologies:** Combining prompting techniques with other AI advancements for more robust applications
- **Challenges:** Handling Ambiguity: Ensuring prompts are understood accurately despite inherent ambiguities in natural language
- **Bias and Fairness:** Addressing biases that may arise from the pre-training dataset or the prompts themselves
- **Scalability:** Managing the computational resources required for larger models and more complex prompting techniques



Thank you

Greg DeCarlo

The future is now

https://linktr.ee/mrinreality