October 21, 2024

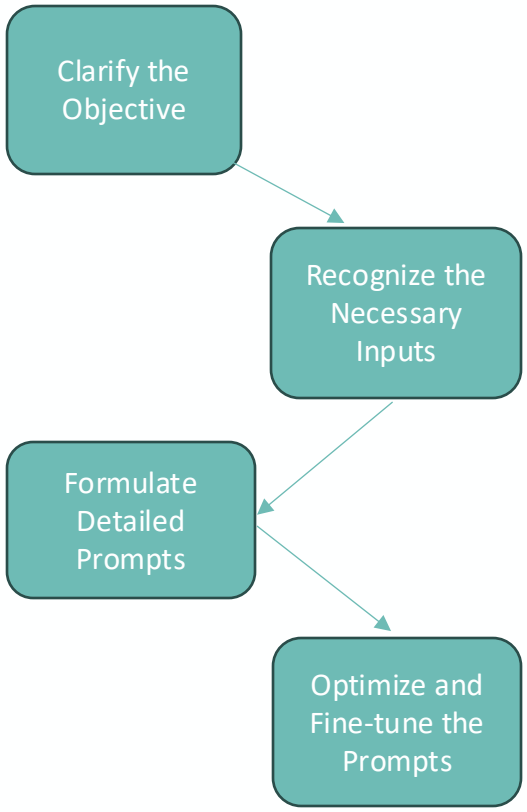
BitOoda AI Research:
Maximizing AI Model's Potential using Prompt Engineering

Research

Tim Kelly, CEO & Founder
Dhyay Bhatt, Head of AI
Niraj Yagnik, Lead Developer & AI Engineer

Abstract

Artificial Intelligence (AI) language models have reshaped our ability to interact with machines, enabling complex tasks to be automated. To unlock AI's full potential, users must understand how to effectively communicate with these models through prompt engineering. This report focuses on delivering unique and powerful prompt strategies that go beyond basic tips, tailoring them to the specifics of different model capabilities—such as access to real-time information, knowledge cutoffs, and the ability to integrate external data.



General Prompt Engineering Pipeline

- By tailoring your approach to the specific characteristics of each AI language model—such as knowledge cutoffs, real-time access, and document integration—you can greatly enhance the quality of the responses you receive.
- By refining how you communicate with AI, you can unlock its full potential, transforming it into powerful tools for research, decision-making, and content generation.

Prompt Engineering 101

Why is Prompt Engineering needed?

Definition

Prompt engineering refers to the art of crafting inputs or questions to optimize an AI model's responses.

Why is prompt engineering needed?

AI models generate responses by predicting patterns from past data, not by truly understanding language. This makes them prone to errors when prompts are unclear, lack context, exceed token limits, or reference events beyond their knowledge cutoff. Prompt engineering is essential to guide the model, ensuring it has the clarity and context needed to generate accurate and relevant results despite these limitations.

What is a prompt?

A prompt is an input or question given to an AI model to generate a specific response.

How is it “engineering”?

It's considered "engineering" because it requires iterative trial and error, refining prompts to achieve precise, effective outcomes.

Prompt Engineering 101

Essential Model Characteristics That Impact Prompting

Before diving into advanced prompt engineering techniques, it's crucial to understand the model-specific characteristics that directly impact how you should approach crafting prompts. These include the **knowledge cutoff**, **context window** size, and whether the model has **internet access**.

Knowledge Cutoff and Internet Access

- **Perplexity**: Offers real-time internet access with no knowledge cutoff, making it ideal for querying current events.
- **Claude Sonnet 3.5**: Has a knowledge cutoff in April 2024 but supports document uploads and tool pairing to allow the user to provide the relevant context.
- **GPT-4o with Internet Access**: Despite a knowledge cutoff of October 2023, it fetches real-time data from the web, blending old knowledge with up-to-date information.

Pro Tip:
To maximize the effectiveness of your prompts, prioritize clarity and relevance. Only include the essential parts of your query to stay within the token limit and maintain coherence.

Context Window Size

The context window defines how much information a model can handle in one conversation, often measured in tokens (a fundamental unit of data processed by an LLM). While large models can process thousands of tokens, exceeding this limit results in information loss, which can cause the model to miss important details.

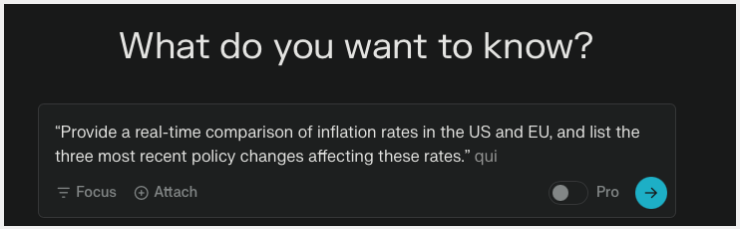
Model	Context Window Size
GPT-4o	Up to 128,000 tokens
Llama 3.1	Up to 32,000 tokens
Claude Sonnet 3.5	Up to 200,000 tokens
Gemini 1.5 Pro	Up to 2,000,000 tokens

Prompt Engineering 101
Advanced Prompt Engineering Techniques

Now that we've addressed how model limitations can affect interactions, let's explore advanced prompt engineering techniques tailored to these constraints.

Real-Time Information Access (e.g., Perplexity)

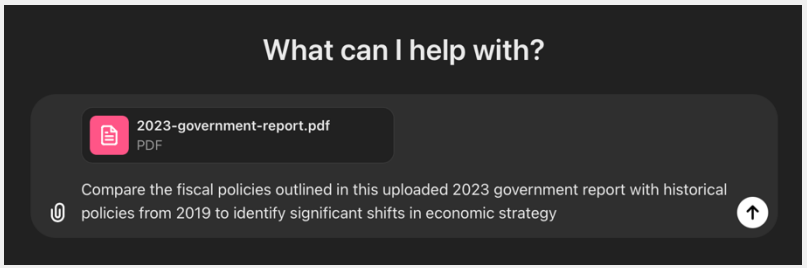
Unique Tip: With models like Perplexity, you can design prompts that require time-sensitive or current information. For instance, instead of asking general questions like "What is inflation?" you can request more actionable insights:



Data-Driven Prompts: You can ask Perplexity to retrieve specific reports or even evaluate trends from live datasets, which traditional models without real-time access cannot do.

Cutoff Models with Document Integration

Unique Tip: When using a model like GPT-4o that supports document uploads, you can instruct it to cross-reference new data with the provided documents. For example:



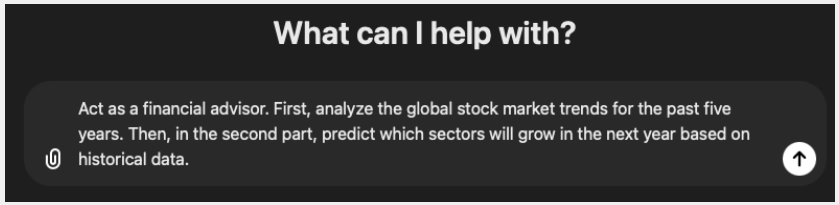
By layering real-time data or additional resources with the model's core knowledge, you can extract much more nuanced responses.

Prompt Engineering 101
Advanced Prompt Engineering Techniques

Step-by-Step and Role-Specific Prompts

One of the most effective strategies for improving the accuracy of AI-generated responses is to guide the model through a step-by-step process.

Unique Tip: Rather than asking for a simple explanation, instruct the model to assume a specific role and tackle the problem in stages:



This technique leads to more comprehensive, multi-step answers, breaking complex topics into manageable parts while maintaining the depth of information.

Prompt Chaining for Complex Outputs

Prompt chaining involves creating a series of related prompts where each response builds on the previous one. This technique helps handle complex queries that exceed the model's context window or require more detailed reasoning.

Unique Tip: Start with a high-level request, then narrow down each subsequent prompt to achieve a richer and more structured final result:

First prompt: *Provide a general overview of the AI hardware landscape in 2024.*

Follow-up prompt: *Now focus on the demand for high-performance GPUs in AI training and compare it with the demand for inference workloads.*

Final prompt: *"Which AI companies have shifted their hardware strategies due to the increased demand for inference? List them and explain the rationale behind their decisions.*

By refining each query, you can guide the model through a complex set of ideas and ensure a higher level of precision in the final output.

Prompt Engineering 101
Avoiding Common Pitfalls and Maximizing Performance

Now that we've touched on the common pitfalls with AI models, let's dive into strategies for handling hallucinations and optimizing prompt performance.

Handling Model Hallucinations

AI models sometimes generate **hallucinations**—answers that are incorrect but sound plausible. To mitigate this, prompt the model to signal uncertainty where appropriate.

Unique Tip: Directly instruct the model to acknowledge if it is unsure of any facts. This reduces overconfidence in its responses and provides a clearer idea of the reliability of the information: *"If you're unsure about any part of your answer, indicate this and explain why."*

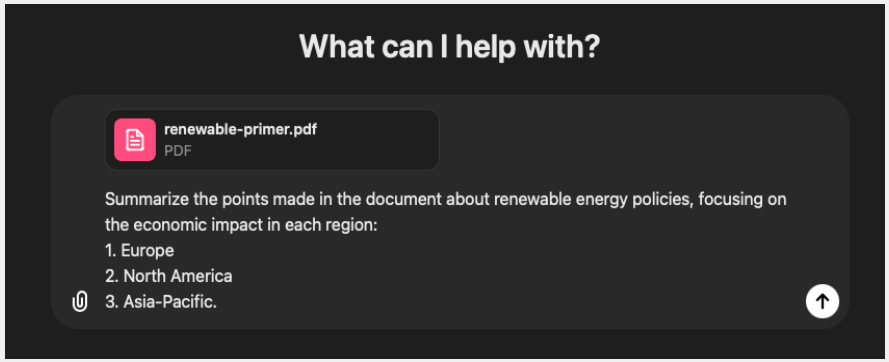
Alternatively, providing the model with more context through uploaded documents or by utilizing a vector database can help significantly reduce hallucination.

Vector Database
A vector database organizes data in a way that allows it to find similar items quickly, making it useful for tasks like searching related content or improving recommendation systems.

Maximizing Context Window Efficiency

To avoid losing important information due to context window limits, be strategic about how you organize your prompts.

Unique Tip: Use bullet points or numbered lists to break down complex queries. This structure is not only easier for the model to process but also ensures key information remains intact:



This ensures that the most critical data is retained within the model's token limit, maximizing its ability to generate complete responses.

Disclosures

Purpose

This research is only for the clients of BitOoda. This research is not intended to constitute an offer, solicitation, or invitation for any securities and may not be distributed into jurisdictions where it is unlawful to do so. For additional disclosures and information, please contact a BitOoda representative at info@bitooda.io.

Analyst Certification

Niraj Yagnik, the primary research analyst of this report, hereby certifies that all of the views expressed in this report accurately reflect his personal views, which have not been influenced by considerations of the firm's business or client relationships.

Conflicts of Interest

This research contains the views, opinions, and recommendations of BitOoda. This report is intended for research and educational purposes only. We are not compensated in any way based upon any specific view or recommendation.

General Disclosures

Any information ("Information") provided by BitOoda Holdings, Inc., BitOoda Digital, LLC, BitOoda Technologies, LLC or Ooda Commodities, LLC and its affiliated or related companies (collectively, "BitOoda"), either in this publication or document, in any other communication, or on or through http://www.bitooda.io/, including any information regarding proposed transactions or trading strategies, is for informational purposes only and is provided without charge. BitOoda is not and does not act as a fiduciary or adviser, or in any similar capacity, in providing the Information, and the Information may not be relied upon as investment, financial, legal, tax, regulatory, or any other type of advice. The Information is being distributed as part of BitOoda's sales and marketing efforts as an introducing broker and is incidental to its business as such. BitOoda seeks to earn execution fees when its clients execute transactions using its brokerage services. BitOoda makes no representations or warranties (express or implied) regarding, nor shall it have any responsibility or liability for the accuracy, adequacy, timeliness or completeness of, the Information, and no representation is made or is to be implied that the Information will remain unchanged. BitOoda undertakes no duty to amend, correct, update, or otherwise supplement the Information.

The Information has not been prepared or tailored to address, and may not be suitable or appropriate for the particular financial needs, circumstances or requirements of any person, and it should not be the basis for making any investment or transaction decision. The Information is not a recommendation to engage in any transaction. The digital asset industry is subject to a range of inherent risks, including but not limited to: price volatility, limited liquidity, limited and incomplete information regarding certain instruments, products, or digital assets, and a still emerging and evolving regulatory environment. The past performance of any instruments, products or digital assets addressed in the Information is not a guide to future performance, nor is it a reliable indicator of future results or performance.

All derivatives brokerage is conducted by Ooda Commodities, LLC a member of NFA and subject to NFA's regulatory oversight and examinations. However, you should be aware that NFA does not have regulatory oversight authority over underlying or spot virtual currency products or transactions or virtual currency exchanges, custodians or markets.

BitOoda Technologies, LLC is a member of FINRA.

"BitOoda", "BitOoda Difficulty", "BitOoda Hash", "BitOoda Compute", and the BitOoda logo are trademarks of BitOoda Holdings, Inc.