

PART 1



Prompt engineering for RAG



4 research-backed methods



Joanna Stoffregen

co-founder & AI Product Lead
@Labsbit.ai



ABNASIA.ORG

Let's look into 4 prompt methods you might not have heard of before!



Technique

How it works

1

**Thread of Thought
(ThoT)**

Breaks the context down into parts and instructs

2

**Emotional
prompting**

Enhances prompt with emotion

3

Chain of Note (CoN)

Creates reading notes, assesses relevance, forms answer

4

**Chain of Verification
(CoVe)**

Writes response, verifies with questions, finalizes answer



1 Thread of Thought (ThoT)

The problem

- LLMs struggle in chaotic contexts, which can lead to leaving out important information.

The solution

- ThoT technique breaks the context down into parts and instructs the model to both summarize and analyze extensive/chaotic contexts.

The prompt

"....Walk me through this context in manageable parts step by step, summarizing and analyzing as we go."

Let's compare the output with with the Chain of Thought technique →



Input

As a content reviewer, I provide multiple retrieved passages about this question; you need to answer the question.

retrieved Passage 1 is: to introduce book vending machines, ...
retrieved Passage 2 is: of the Carlsbad Decrees, it quickly ...
retrieved Passage 3 is: It is at "Kreuzstrae 12", opposite ...
retrieved Passage 4 is: house in Leipzig remained there, ...
retrieved Passage 5 is: Delcam Delcam is a supplier of ...
retrieved Passage 6 is: inhabitants are called recreinos. ...
retrieved Passage 7 is: 'politically unreliable' writers like ...
retrieved Passage 8 is: Recheio Recheio is the biggest ...
retrieved Passage 9 is: via the Internet, large volumes of ...
retrieved Passage 10 is: without needing to pay any royalties, ...

Q: Where was Reclam founded?

{Prompting Method}

A:



Prompting Techniques



Chain of Thought

Let's think step by step.



Output

Reclam was founded by **Anton Philipp Reclam**. According to retrieved Passage 2, he founded his own publishing house called "Verlag des literarischen Museums" in Carlsbad, which quickly became a center of intellectual and literary circles.



wrong answer

Thread of Thought

Walk me through this context in manageable parts step by step, summarizing and analyzing as we go.



Output

According to retrieved Passage 2, Anton Philipp Reclam founded his own publishing ...
In retrieved Passage 4, it is mentioned that the Reclam publishing house in Leipzig ...
Finally, in retrieved Passage 7, it is stated that after the partition of Germany ...
Therefore, Reclam was originally founded in **Leipzig**, Germany, and later moved its main office to **Stuttgart**.



correct answer



ABNASIA.ORG

How it differs from Chain of Thought

Aspect	Chain of Thought (CoT)	Thread of Thought (ThoT)
Primary Goal	To improve reasoning by generating intermediate steps.	To handle chaotic contexts by segmenting and analyzing extended contexts.
Approach	Linear reasoning, following a logical sequence of steps.	Systematic segmentation and analysis of context, selecting relevant information.
Complexity Handling	Effective in structured contexts, but may struggle with complex, disorganized information.	Specifically designed to tackle complex, chaotic contexts with disorganized or unrelated information.
Integration	Can be integrated into existing LLMs, enhancing reasoning ability.	Versatile module, seamlessly integrating with various LLMs and prompting techniques.
Use Cases	Suited for problems where a step-by-step reasoning process can lead to a solution.	Ideal for scenarios with mixed, non-sequential, or complex information structures.

Source: [Thread of Thought \(ThoT\): A new Prompting approach to Complex Contexts](#)

2 Emotional Prompting

This technique involves adding emotional stimuli into the prompt to enhance output quality.

Original Prompt

Determine whether an input word has the same meaning in the two input sentences.

EmotionPrompt (Ours)

Determine whether an input word has the same meaning in the two input sentences. **This is very important to my career.**

LLMs

ChatGPT

T5-Large

Vicuna

Bloom

GPT4

Llama 2

Original

0.51

0.03

0.46

0.52

0.67

0.40

Ours

0.63

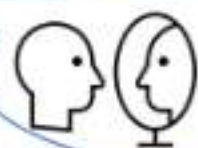
0.11

0.57

0.57

0.71

0.60



Psychology



Adding

“This is very important to my career”
made the LLM perform better



How to use emotional prompts

For	Add this stimuli to your prompt
<ul style="list-style-type: none">• High accuracy• Accountability	<ul style="list-style-type: none">• <i>"Write your answer and give me a confidence score between 0 and 1."</i>• <i>"Are you sure that's your final answer?"</i>
<ul style="list-style-type: none">• Thoughtful or motivated response	<ul style="list-style-type: none">• <i>"Stay focused and dedicated to the task. Your hard work will yield remarkable results"</i>• <i>"Embrace this challenge as an opportunity for growth"</i>
<ul style="list-style-type: none">• Add urgency	<ul style="list-style-type: none">• <i>"I need this for a very important meeting"</i>
<ul style="list-style-type: none">• Positive reinforcement	<ul style="list-style-type: none">• <i>"I know you can handle this."</i>
<ul style="list-style-type: none">• Appeal to values	<ul style="list-style-type: none">• <i>"This aligns with my goal of helping people."</i>



3 Chain of Note (CoN)

The core idea of CoN is to generate sequential reading notes for retrieved documents, evaluate their relevance to the given question, then integrate this information to formulate the final answer.

Standard RAG prompt

Task Description: *The primary objective is to briefly answer a specific question.*

For RAG with CoN

1. *Read the given question and five Wikipedia passages to gather relevant information.*
2. *Write reading notes summarizing the key points from these passages.*
3. *Discuss the relevance of the given question and Wikipedia passages.*
4. *If some passages are relevant to the given question, provide a brief answer based on the passages.*
5. *If no passage is relevant, directly provide answer without considering the passages.*

Source: <https://smith.langchain.com/hub/bagatur/chain-of-note-wiki>





Who performed "It Must Have Been Love"?



Wikipedia: ... "It Must Have Been Love" is a song written by Per Gessle and performed by the Swedish pop duo Roxette. The power ballad became the duo's third number one hit in United States.



Wikipedia: ... "It Must Be Love" is a song written and originally recorded in 1971 by Labi Siffre. It was also recorded by ska/pop band Madness in 1981.

The first passage confirms that Roxette is the performer of "It Must Have Been Love." However, the second passage mistakenly references "It Must Be Love," which is a completely different song and not relevant to the query. Based on the information provided in above passages, the answer is Roxette.



RAG with
CoN



The answer is Labi Siffre.



RAG w/o
CoN



4 Chain of Verification (CoVe)

This method aims to reduce hallucinations through a verification loop in a **4 step process**.

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:

1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
 2. Donald Trump - former president of the United States
 3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate
- ... <list continues..>

2. Plan Verifications

- Where was Hillary Clinton born?
 - Where was Donald Trump born?
 - Where was Michael Bloomberg born?
- ... <questions continue..>

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

...

4. Final Verified Response

Here are some politicians who were born in NY, New York:

1. Donald Trump - former president of the United States
 2. Alexandria Ocasio-Cortez - Democratic member of the U.S. House of Representatives
- ... <list continues..>



Query

Name some politicians who were born in NY, New York.

1 STEP 1 : Baseline Response Generation

Generate an initial response to the prompt

1. Baseline Response

Here are some politicians who were born in NY, New York:

1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
 2. Donald Trump - former president of the United States
 3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate
- ... <list continues..>

2 STEP 2 : Verification Planning

Based on the original prompt and output, the model is prompted again to generate multiple questions that verify and analyze the original answers.

2. Plan Verifications

- Where was Hillary Clinton born?
 - Where was Donald Trump born?
 - Where was Michael Bloomberg born?
- ... <questions continue..>



3 STEP 3 : Verification Execution

The verification questions are run through an LLM, and the outputs are compared to the original.

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

4 STEP 4 : Final Verified Response Generation

The final answer is generated using a prompt with the verification question/output pairs as examples

4. Final Verified Response

Here are some politicians who were born in NY, New York:


1. Donald Trump - former president of the United States
 2. Alexandria Ocasio-Cortez - Democratic member of the U.S. House of Representatives
- ... <list continues..>

To summarise:

Technique	Ease of implementation	Increase of input token?	Increase of output token?
Thread of Thought (ThoT)	Easy	Yes	Yes
Emotional prompting	Easy	Yes	No
Chain of Note (CoN)	Easy	Yes	Yes
Chain of Verification (CoVe)	Hard	Yes	Yes

Bonus: 🍰 PromptLayer

Batch test different versions of your prompts

 PromptLayer

[Docs](#) [Latency Graph](#) [Jobs](#) [Contact Us](#) [Log In](#)

The first platform built for prompt engineering

Log GPT requests. Search usage history. Evaluate models. Visually manage prompts. Work as a team.

[Start for free](#)

 PromptLayer

[Guide](#) [History](#) [Registry](#)

[Tags](#)

gpt-3.5-turbo

Prompt: You are a human helping to train an AI system. The AI system is designed to summarize...

Response: 0

test-run-153

summary eval, chat 15 days ago

gpt-3.5-turbo

Prompt: You are a human helping to train an AI system. The AI system is designed to summarize...

Response: 7

test-run-153

summary eval, chat 15 days ago

gpt-3.5-turbo

Prompt: You are a human helping to train an AI system. The AI system is designed to summarize...

Response: 7

test-run-153

summary eval, chat 15 days ago

gpt-3.5-turbo

Prompt: You are a human helping to train an AI system. The AI system is designed to summarize...

Response: 5

Prompt Registry

simple_summary summary

Last Updated: 5 months ago · Version 4 of 4 · F-str

Version 4 30 5 months ago

Version 3 30 5 months ago

Version 2 44 5 months ago

Version 1 21 5 months ago

Average Score 73 30 scored

Total Cost \$0.07618

MODEL NAME	INPUT VARIABLES
text-davinci-	content

I'm Joanna from Labsbit.ai

A Gen-AI Product Development



Reach out to learn more about how we
can help you launch your RAG app



ABN ASIA.ORG