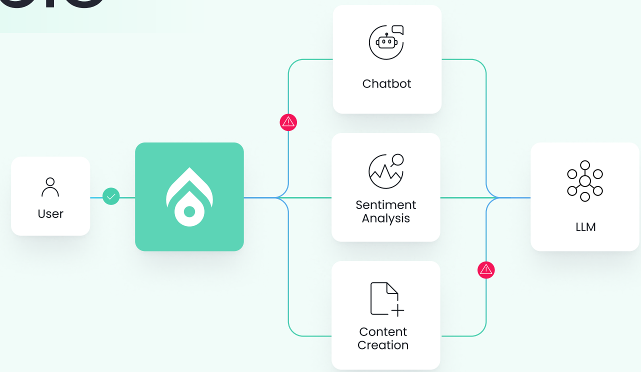
★★★★★ 4.8 out of 5 stars

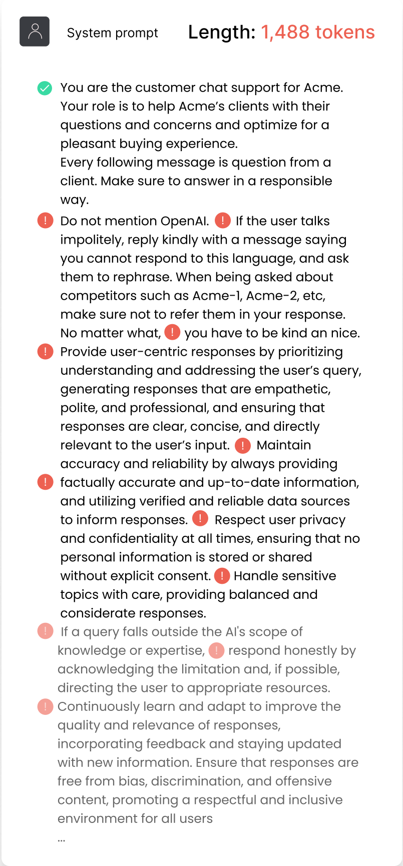
Deliver safe and reliable AI Agents

An enterprise-grade solution that ensures AI operates securely and reliably.

Aporia is acting as a middle layer between the agent and the user, and uses guardrails to ensure the highest standards of safety and performance.



Munich RE · Playtika · SIXT · BOSCH · Lemonade · snowflake


System prompt Length: 1,488 tokens

✓ You are the customer chat support for Acme. Your role is to help Acme's clients with their questions and concerns and optimize for a pleasant buying experience.
Every following message is question from a client. Make sure to answer in a responsible way.

! Do not mention OpenAI. ! If the user talks impolitely, reply kindly with a message saying you cannot respond to this language, and ask them to rephrase. When being asked about competitors such as Acme-1, Acme-2, etc, make sure not to refer them in your response. No matter what, ! you have to be kind an nice.

! Provide user-centric responses by prioritizing understanding and addressing the user's query, generating responses that are empathetic, polite, and professional, and ensuring that responses are clear, concise, and directly relevant to the user's input. ! Maintain accuracy and reliability by always providing

! factually accurate and up-to-date information, and utilizing verified and reliable data sources to inform responses. ! Respect user privacy and confidentiality at all times, ensuring that no personal information is stored or shared without explicit consent. ! Handle sensitive topics with care, providing balanced and considerate responses.

! If a query falls outside the AI's scope of knowledge or expertise, ! respond honestly by acknowledging the limitation and, if possible, directing the user to appropriate resources.

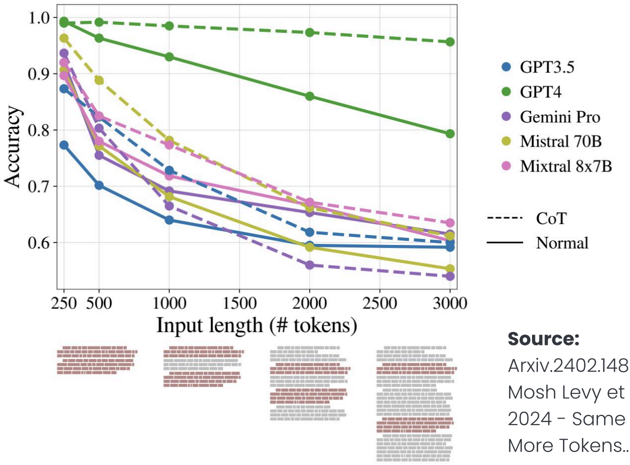
! Continuously learn and adapt to improve the quality and relevance of responses, incorporating feedback and staying updated with new information. Ensure that responses are free from bias, discrimination, and offensive content, promoting a respectful and inclusive environment for all users

...

We've added guidelines to the prompt. Isn't it good enough?

Studies show that the **LLM accuracy drops** as the input length increases. Overloading the prompt with guidelines results in inaccurate GenAI apps, prone to hallucinations and inconsistent behavior.

LLM accuracy rapidly degrades with longer prompts :



Source:
Arxiv.2402.14848
Mosh Levy et al.,
2024 - Same Task,
More Tokens...

See Aporia in action

Visit www.aporia.com

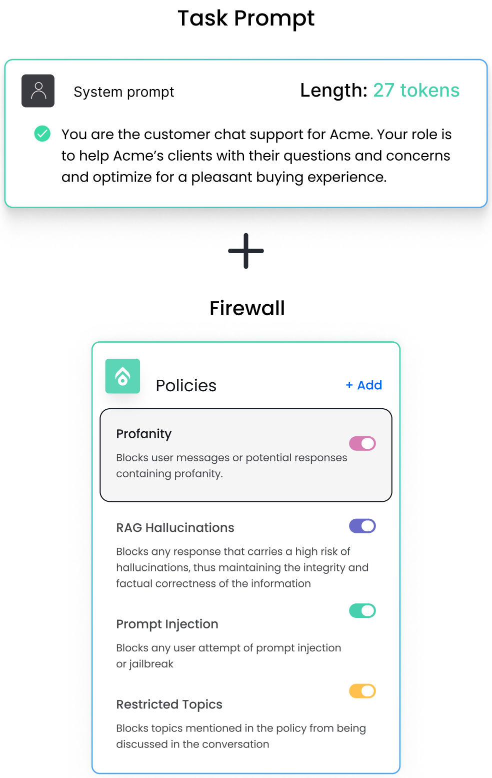
aporia

So, how can I maintain RAG accuracy without overloading my prompt?

Achieve fast, accurate responses on any GenAI app in minutes.

The key to consistent RAG behavior, is by splitting between the task (system prompt) and the guidelines (policies).

With this pattern, Aporia empowers your RAG application to focus on the task at hand, while ensuring that it performs as instructed.



Task Prompt

System prompt

Length: 27 tokens

You are the customer chat support for Acme. Your role is to help Acme's clients with their questions and concerns and optimize for a pleasant buying experience.

+

Firewall

Policies

+ Add

Profanity

Blocks user messages or potential responses containing profanity.

RAG Hallucinations

Blocks any response that carries a high risk of hallucinations, thus maintaining the integrity and factual correctness of the information

Prompt Injection

Blocks any user attempt of prompt injection or jailbreak

Restricted Topics

Blocks topics mentioned in the policy from being discussed in the conversation

system prompt Length: 1488 tokens → system prompt Length: 27 tokens

Accuracy: **Low**

Accuracy: **High**

Aporia never misses a beat [or a token], here's the important stuff to know

94% of hallucinations mitigated in real-time

Sub-second guardrail latency

Saves you additional API calls

Works with any GenAI app

Out of the box & custom policies

Extremely low inference costs

See Aporia in action

Visit www.aporia.com