6 WAYS TO PREVENT HALLUCINATIONS IN LLMS

Bhavishya Pandit

WHAT ARE HALLUCINATIONS?

When AI systems produce information that sounds plausible but lacks any grounding in reality, it's a phenomenon known as **hallucination**.

AI hallucinations occur when models generate outputs that seem correct but lack a factual basis. These errors often result from factors like overfitting, biased or inaccurate training data, and the complexity of the model.

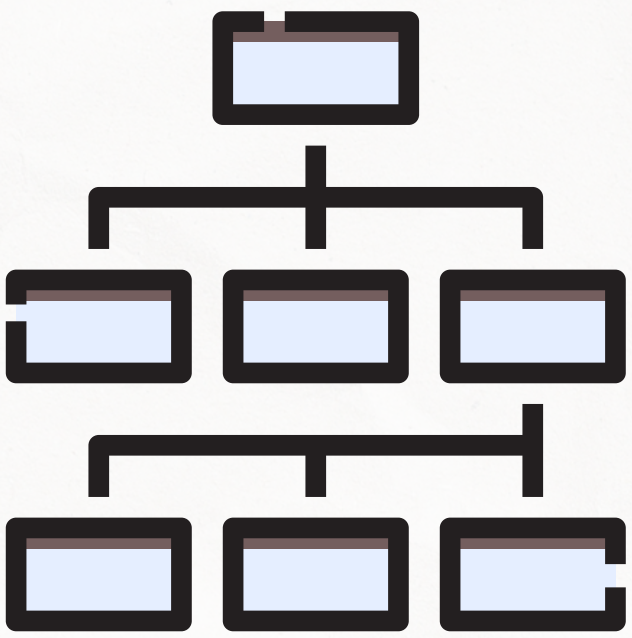
Credit: SuperAnnotate.com

Due to all these atrocities, it is crucial to prevent AI models from hallucinating. Lets discuss how it can be done in the next pages -

1. USE HIGH QUALITY DATA

Generative AI models thrive on vast amounts of input data, but their outputs depend heavily on the **quality, relevance, and structure** of this data.

Consider training a language model to generate medical advice. If the dataset predominantly contains data about general health but lacks specialized information on rare diseases, the model might generate plausible-sounding but incorrect advice for queries on those diseases.



Structured Data

Messy Data

Balanced datasets that cover a wide range of contexts and nuances equip the model to handle diverse inputs more effectively.

2. Data Templates

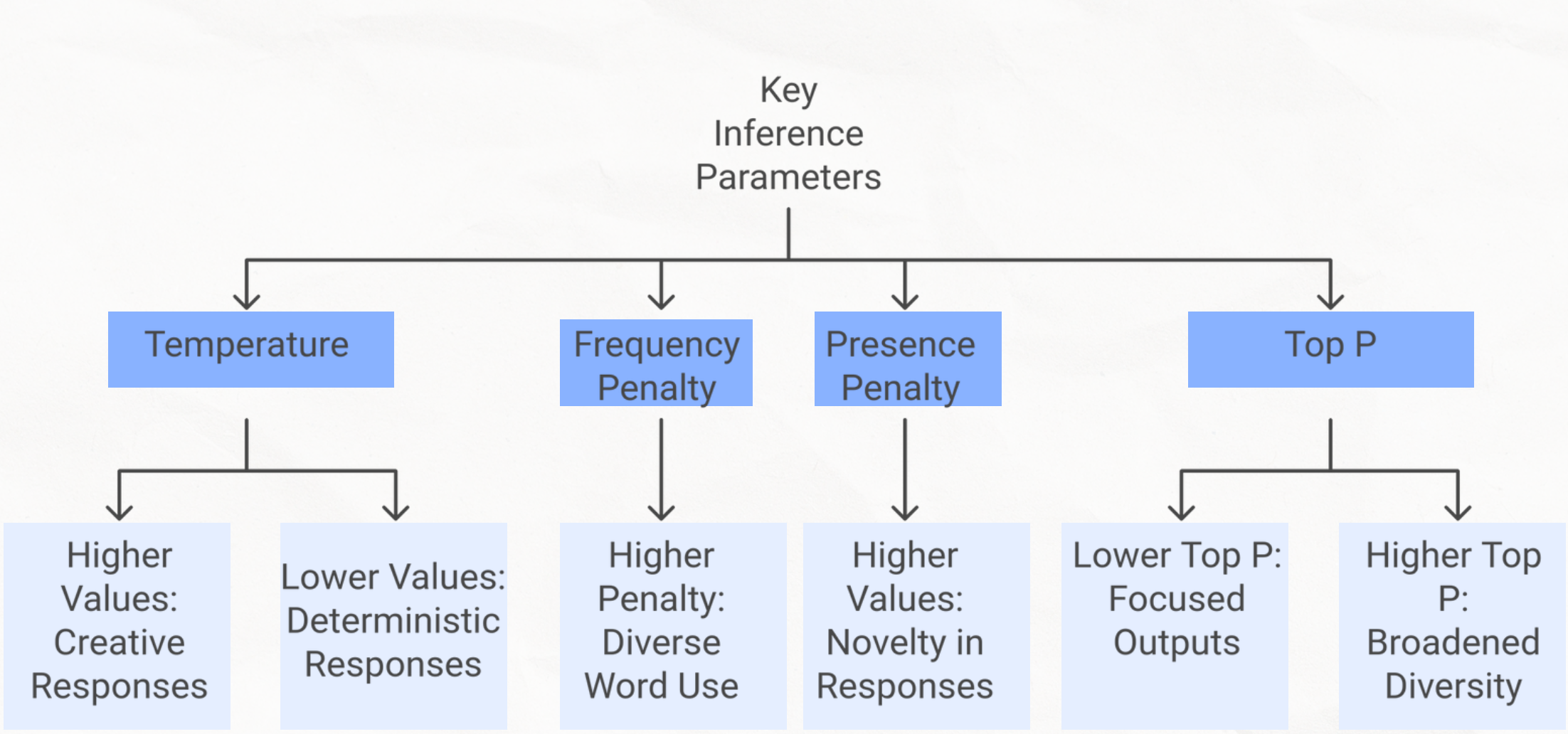
An effective way to curb AI hallucinations is through data templates—structured guides that outline the expected format and permissible range of responses. By enforcing these predefined patterns, templates ensure consistency, accuracy, and adherence to domain-specific requirements.

Ex - In financial reporting, templates might define the structure of balance sheets, including mandatory fields like assets, liabilities, and net income.

```
[Template]
Subject: Refund Request Update

Dear [Customer Name],

Thank you for reaching out to us regarding your recent order #[Order Number].

if approved:
We have processed a refund of [Refund Amount] to your original payment method.
The refund should reflect in your account within [Time Frame, e.g., 5-7 business days].

If denied:
We regret to inform you that we cannot process a refund due to [Reason for Denial].

For further assistance, feel free to contact our support team at [Support Contact Info].

Sincerely,
[Company Name] Support Team
```

3. Parameter Tuning

Fine-tuning inference parameters is a powerful, cost-effective way to refine the output of language models, allowing users to balance randomness, creativity, and consistency. By adjusting key settings like temperature, frequency penalty, presence penalty, and top P, you can achieve more tailored responses based on specific needs.



For generating creative content like poems, setting a high temperature (0.9) and a low frequency penalty can produce imaginative outputs. Conversely, for technical documentation, a low temperature (0.3) and higher frequency penalty ensure factual accuracy and consistency.

4. Prompt Engineering

Prompt engineering is a method of crafting precise and effective prompts to guide language models (LLMs) in generating accurate and relevant outputs. This is a cost-effective approach to improving the quality of responses and mitigating issues like hallucinations and biases.

Effective Prompting Techniques

Q: What are the benefits of AI in healthcare?
A: Improved diagnostics, personalized treatment, and streamlined workflows.

Provide Sample Question and Answers

Summarize the article in three key points

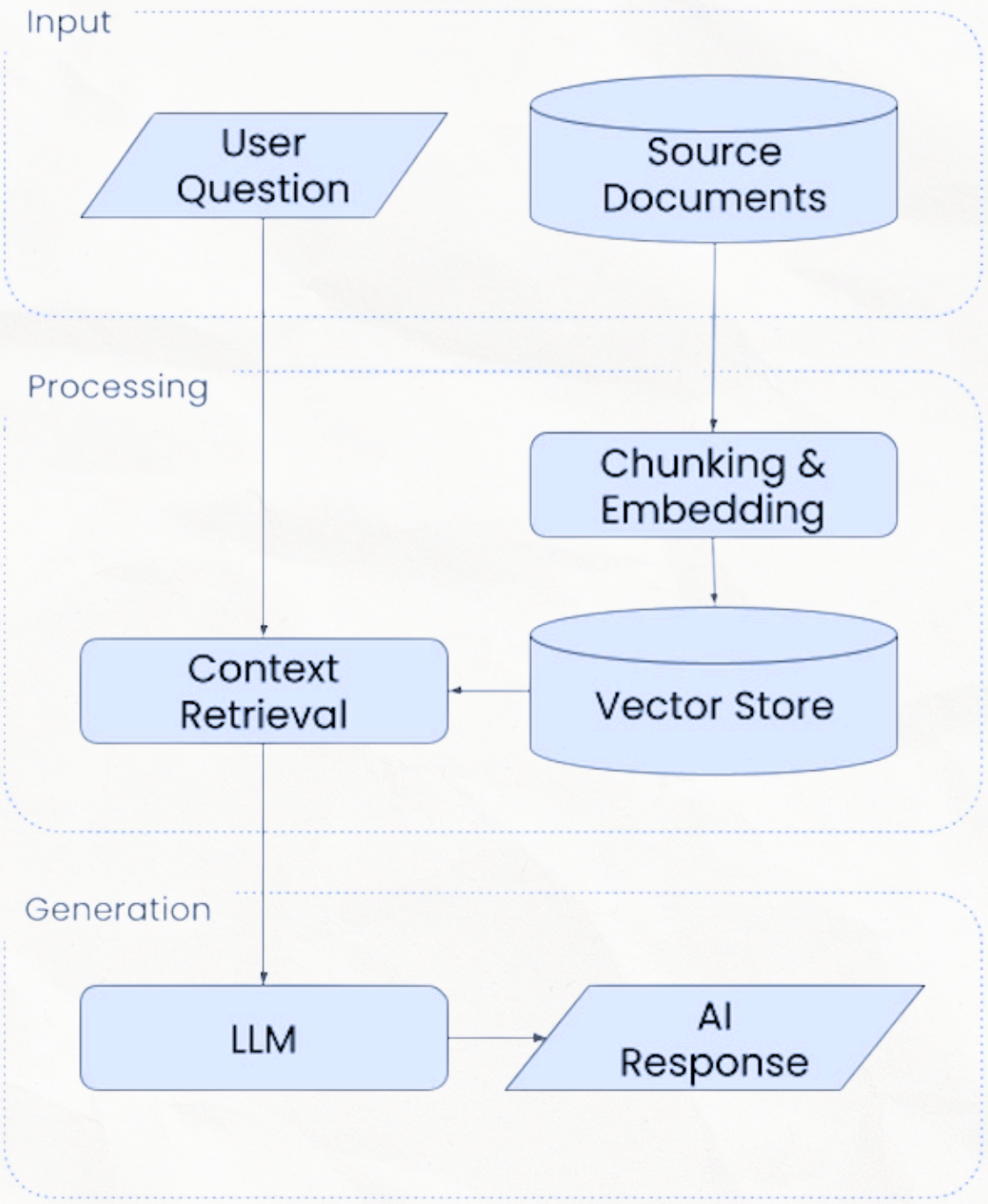
Clear Instructions

You are a financial expert. Explain how inflation affects interest rates.

Role Assignment

5. RAG

Retrieval Augmented Generation (RAG) is a powerful technique that enhances a language model's ability to provide accurate answers by integrating additional, external knowledge.



RAG flow

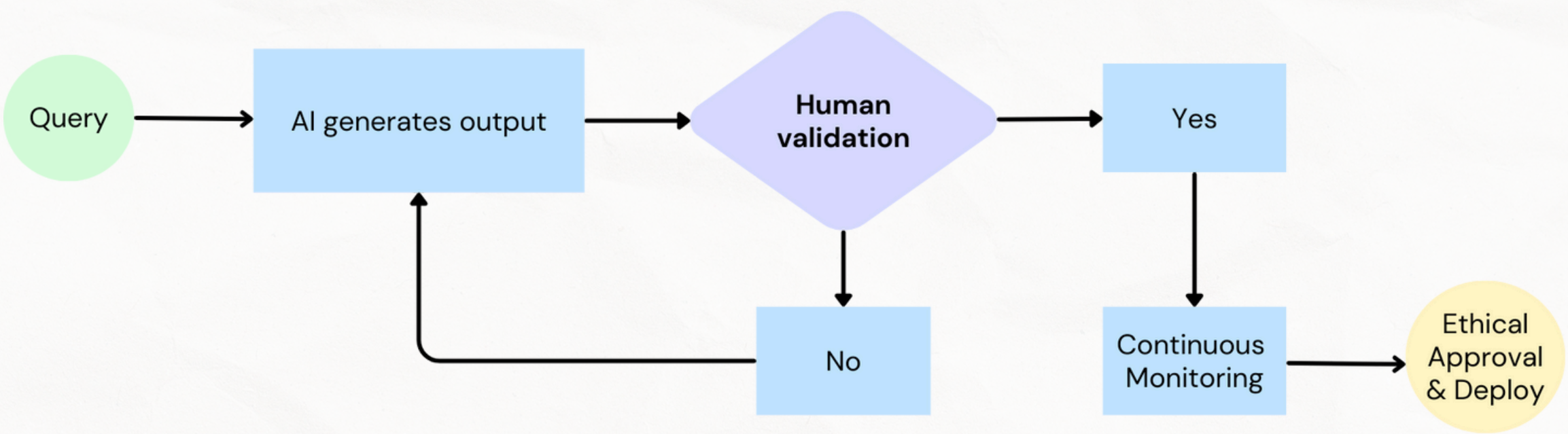
Credit: Deeplearning.ai

For a technical support chatbot, RAG allows the model to reference a product’s user manual to answer queries like "How do I reset my password?" instead of relying on generic training data.

By grounding responses in curated, domain-specific documentation, RAG reduces the influence of training data biases.

6. Human Fact Checking

Despite the progress in AI, **adding a human review layer** is still one of the most reliable ways to prevent hallucinations. Human fact-checkers play a key role in identifying and correcting inaccuracies that the AI might miss, ensuring the accuracy of the output.



Human reviewers regularly assess AI-generated content, flagging errors or fabrications. This feedback is then used to refine the AI's training data, improving its accuracy over time.

In a news generation system, human editors verify the facts before publishing to prevent the spread of false information

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST