**B.Sc (Tech) Thesis**

# The Next Generation of Intelligent Chatbots

**Review of Approaches to an NLP-Based Intelligent Conversational Agent**
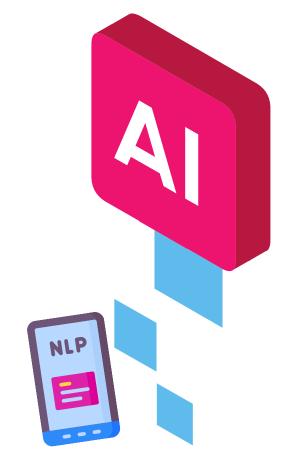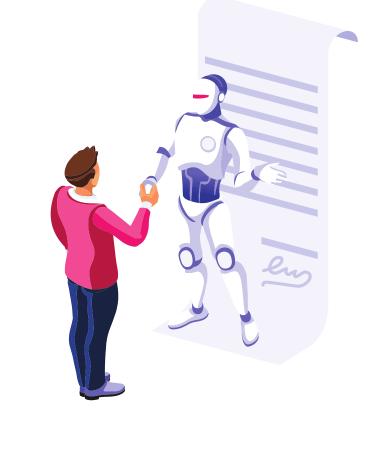
**Ali Amaan**
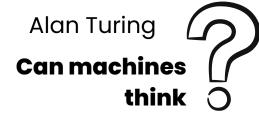
Aalto-yliopisto
Aalto-universitetet
Aalto University

1950 —————————— 2016 ●●●●●●●●●●●●●● 2023 —————— Beyond.

Alan Turing

**Can machines think** ?

[1]

**Sort of** ○○○

**Definitely can** !

# Chatbot Approaches

2016 ● ● ● ● ● ● ● ● 2023

## Rules-Based



Rule#1

Input message → IF → Message is "HI" → THEN → Output message is "Hello, how are you?"

ELSE IF — Rule#2

Message is "BYE" → THEN → Output message is "Goodbye!" → Output message

ELSE IF — Rule#3

Any other message → THEN → Output message is "Sorry, I don't understand you."

[2]

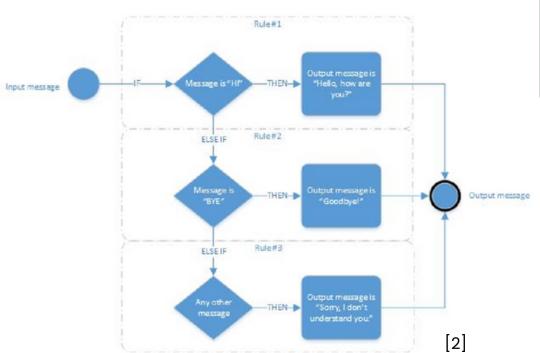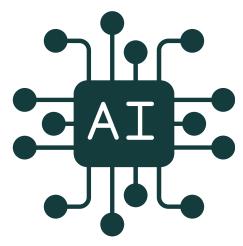## Pattern Matching

```
AIML
1  <category>
2      <pattern>
3          User input comes here
4      </pattern>
5      <template>
6          Bot response to the input
7      </template>
8  </category>
```
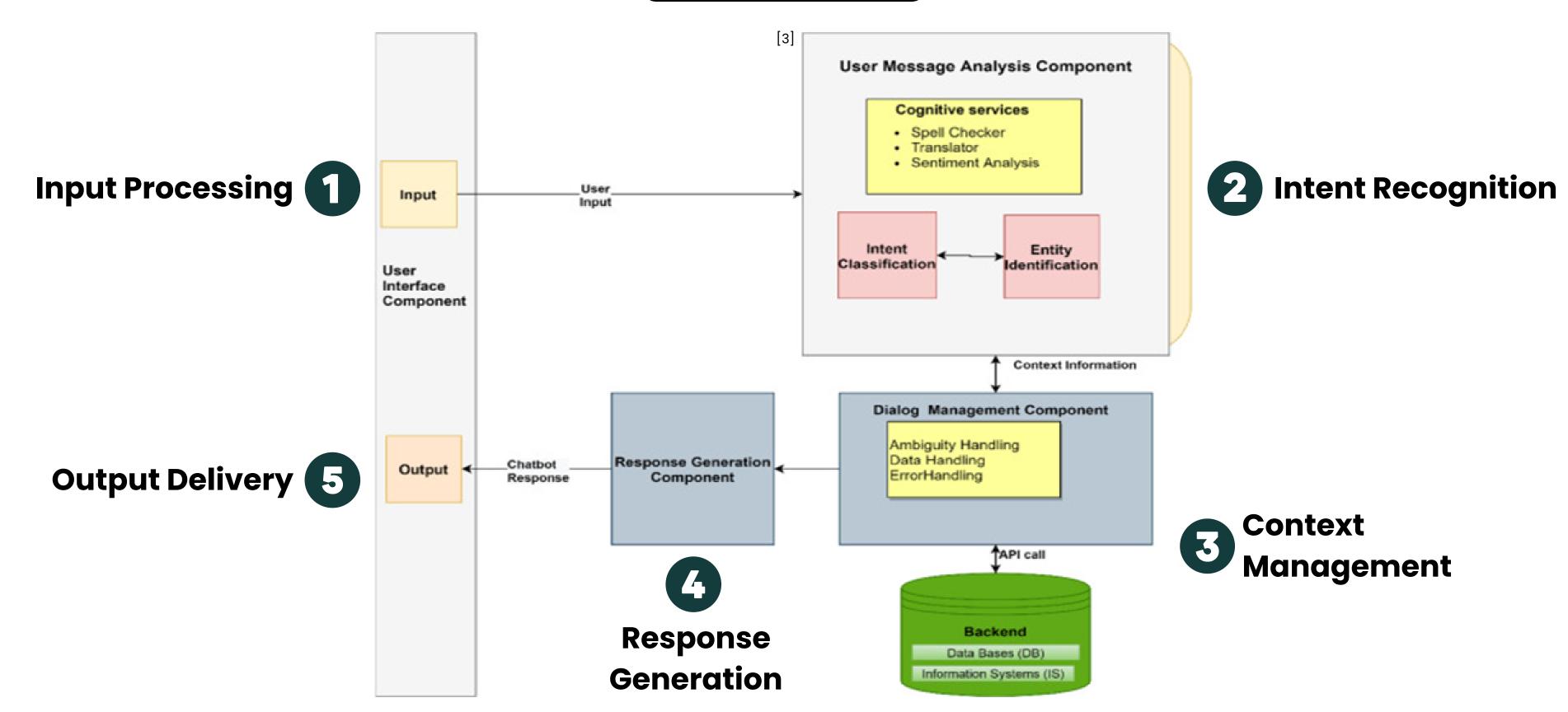
## ML-Based

| Chat Flows | Rules Based | Prompt Eng |
|---|---|---|
| Response Templates | Pre-Defined | Pre-trained |
| Rigid | Not Human Like | Natural Human Like Language |

# High Level AI × Process View



**Input Processing** ①

**Output Delivery** ⑤

**Intent Recognition** ②

**Response Generation** ④

**Context Management** ③

[3]

**User Message Analysis Component**

**Cognitive services**
- Spell Checker
- Translator
- Sentiment Analysis

Intent Classification ↔ Entity Identification

Input

User Input

User Interface Component

Context Information

**Dialog Management Component**

Ambiguity Handling
Data Handling
ErrorHandling

Output

Chatbot Response

Response Generation Component

API call

**Backend**
Data Bases (DB)
Information Systems (IS)

# Large Language Models (LLMs)

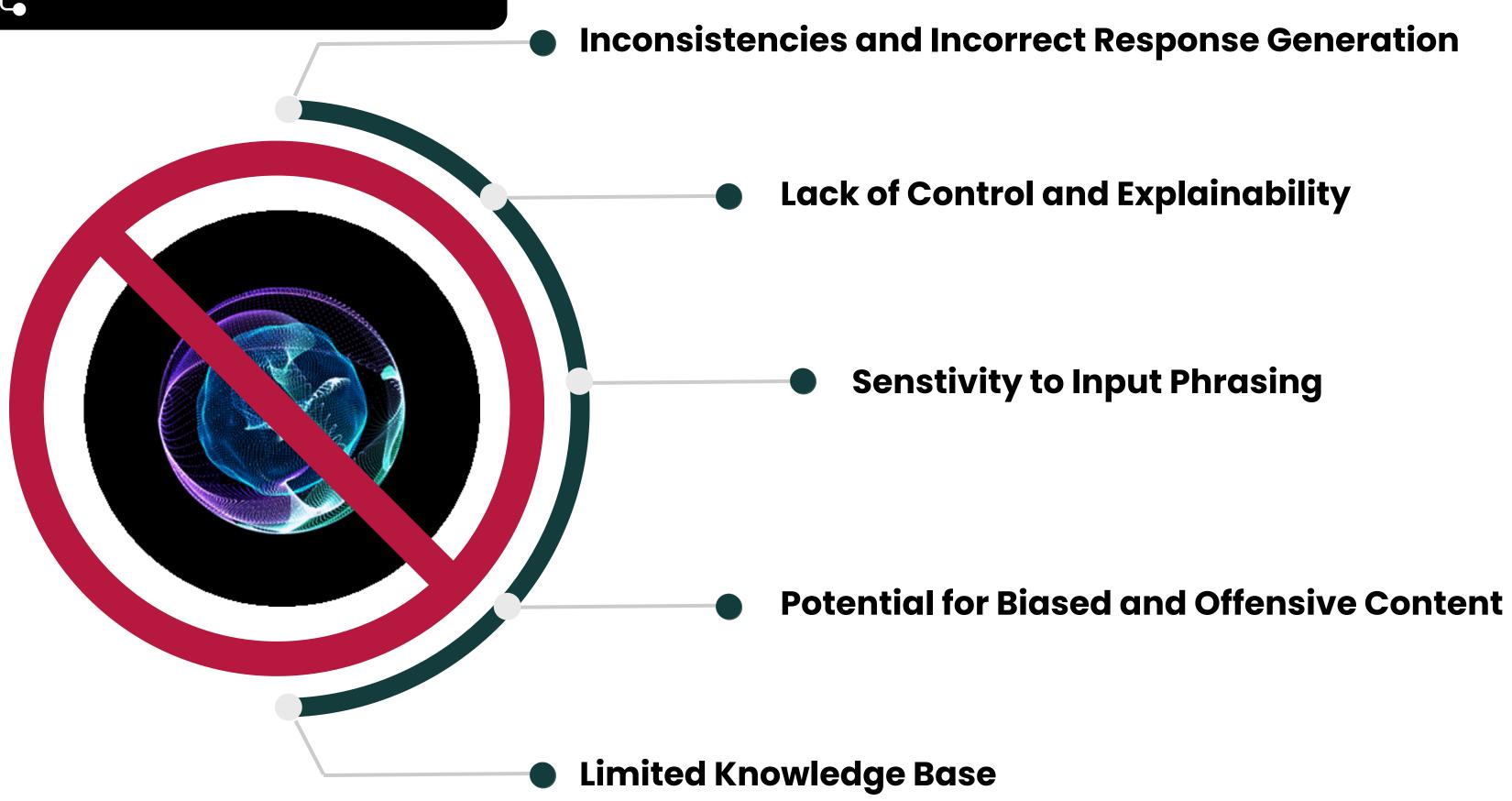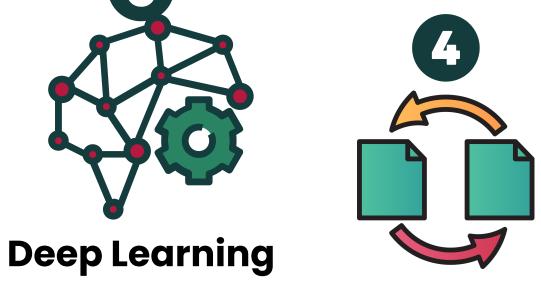| Models | GPT Series | BERT | T5 | RoBERTa | Sentence Transformers |
|---|---|---|---|---|---|
| Key Strengths | Highly Coherent, Zero & Few Shots, Hosted API, Multi-lingual, Embeddings | Bi-directional Training, Understanding Input, Open Source | Unified Text-to-Text Approach, Open Source | Training Improvements over BERT, Open Source | Semantic Understanding, Open Source, Embeddings |
| Preferred Use Cases | Conversation Agents, Semantic Search, Q/A, Summarization, Translation, Content Generation | Sentiment Analysis (Classification), Q/A, Info. Retrieval | Translation, Summarization, Q/A, Content Generation | Sentiment Analysis (Classification), Q/A, Info. Retrieval | Semantic Search, Clustering, Recommendation Systems, Basic Q/A |

# Limitations of LLMs

- Inconsistencies and Incorrect Response Generation
- Lack of Control and Explainability
- Senstivity to Input Phrasing
- Potential for Biased and Offensive Content
- Limited Knowledge Base

Approaches to Chatbot Intelligence
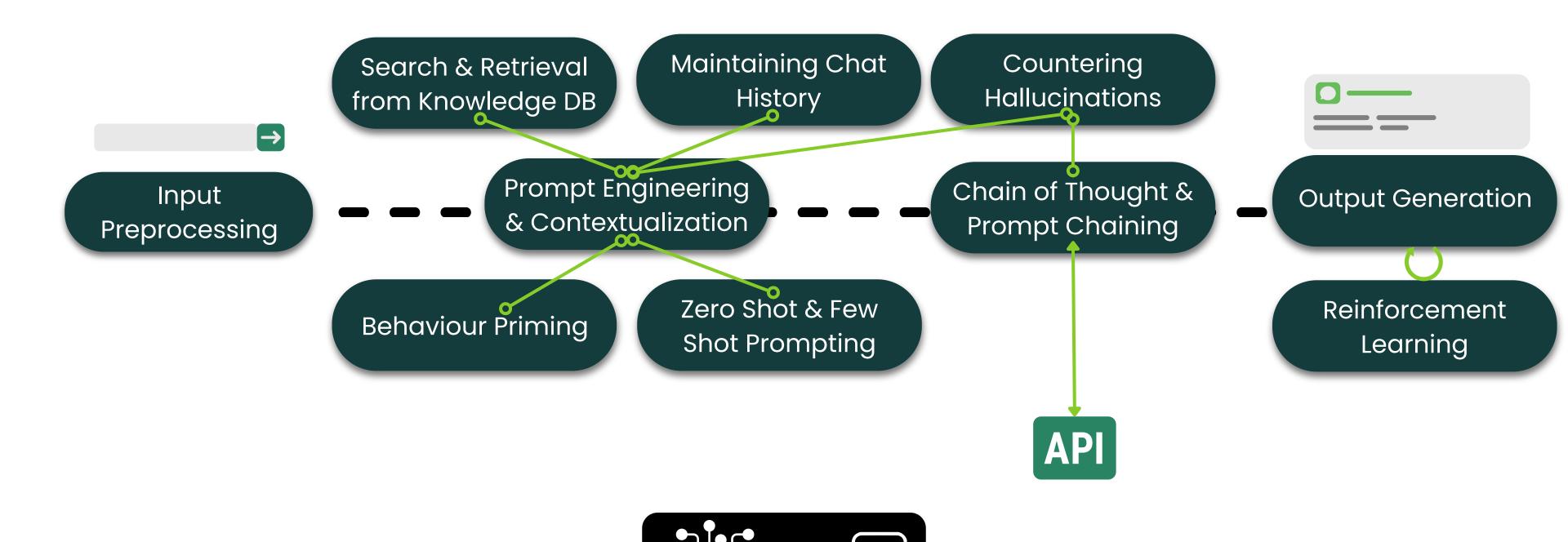
2016

2023

**1** Rules-Based

**2** Statistical Methods

**3** Deep Learning

**4** Transfer Learning

**5** Reinforcement Learning

**6** Hybrid Approach

# Crucial Process Blocks of an Intelligent Chatbot

## Modular Overview

# Crucial Process Blocks of an Intelligent Chatbot

## Input Preprocessing



Prompt Input String

# Crucial Process Blocks of an Intelligent Chatbot

## Prompt Engineering & Contextualization
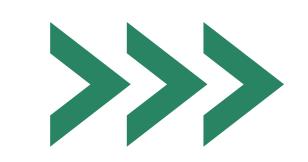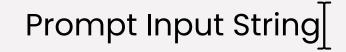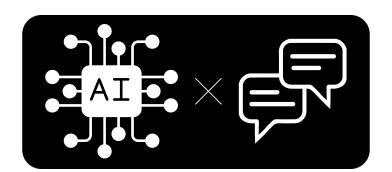
| Category | Prompt Pattern Examples |
|---|---|
| Input Semantics | Meta Language Creation |
| Output Customization | Output Automator |
| | Persona |
| | Visualization Generator |
| | Recipe |
| | Template |
| Error Identification | Fact Check List |
| | Reflection |
| Prompt Improvement | Question Refinement |
| | Alternative Approaches |
| | Cognitive Verifier |
| | Refusal Breaker |
| Interaction | Flipped Interaction |
| | Game Play |
| | Infinite Generation |
| Context Control | Context Manager |

[4]

**Behaviour Priming**

Act as...

**Countering Hallucinations**

**Zero Shot & Few Shot Prompting**
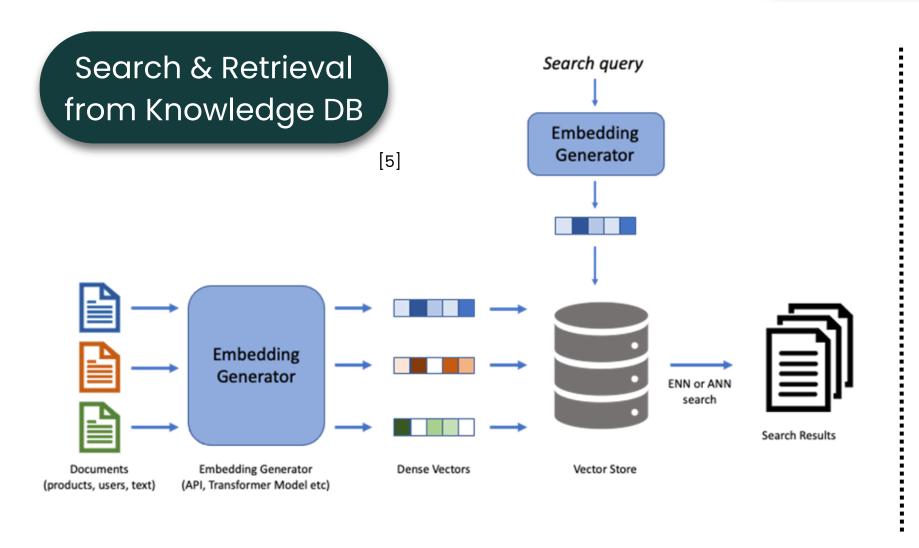
Instruction...

Example Query + Response...

Example Query + Response...
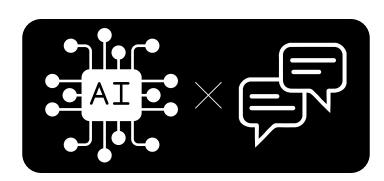
# Crucial Process Blocks of an Intelligent Chatbot

## Prompt Engineering & Contextualization

### Search & Retrieval from Knowledge DB

[5]

Search query

Embedding Generator

Documents (products, users, text)

Embedding Generator (API, Transformer Model etc)

Dense Vectors

Vector Store

ENN or ANN search

Search Results

<retrieval result...> + Query

### Maintaining Chat History

[6]

Back          Front

Enqueue          Dequeue

<chat history...> + Query

AI × 💬

# Crucial Process Blocks of an Intelligent Chatbot

## API

## Chain of Thought Prompting

[7]

### Standard Prompting

**Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

**Model Output**

A: The answer is 27. ❌

### Chain of Thought Prompting

**Input**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
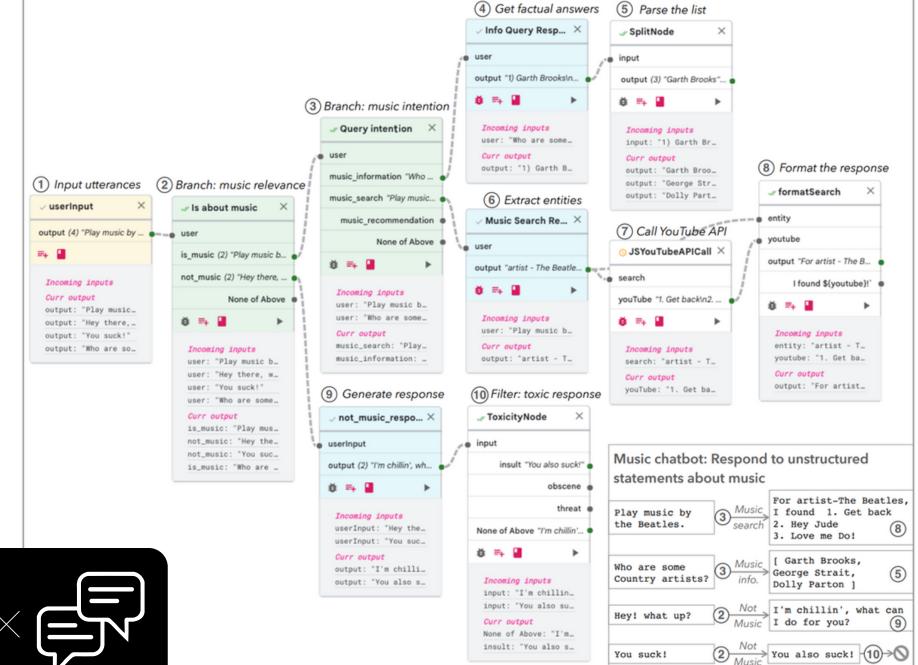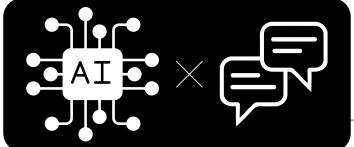
**Model Output**

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✔️
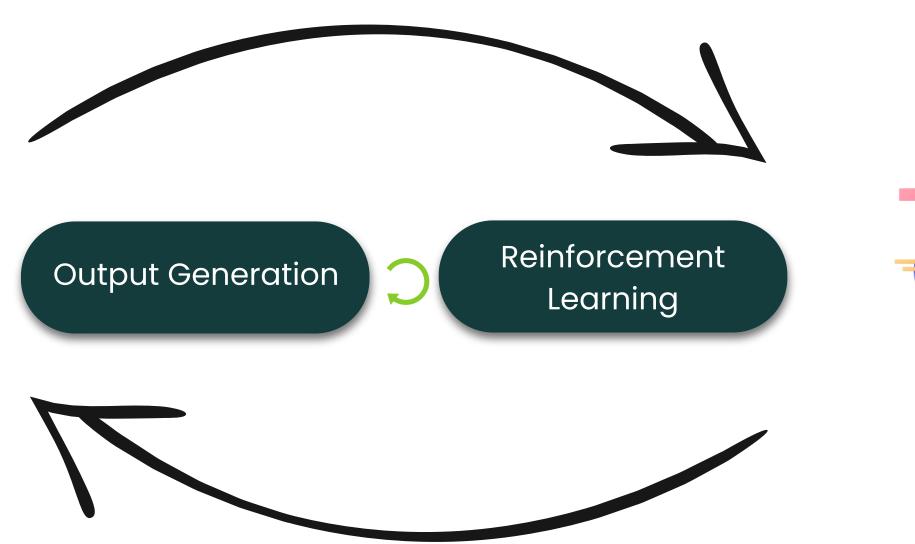
## Prompt Chaining

[8]

# Crucial Process Blocks of an Intelligent Chatbot

Output Generation

Reinforcement Learning

AI

# Crucial Process Blocks of an Intelligent Chatbot

## Modular Overview

Search & Retrieval from Knowledge DB

Maintaining Chat History

Countering Hallucinations

Input Preprocessing

Prompt Engineering & Contextualization

Chain of Thought & Prompt Chaining

Output Generation

Behaviour Priming

Zero Shot & Few Shot Prompting

API

Reinforcement Learning

AI

# Thank You!

- Ali Amaan

Let' implement the next Gen AI.
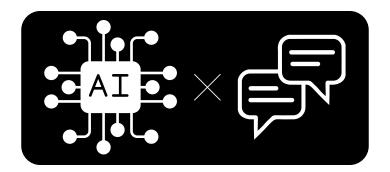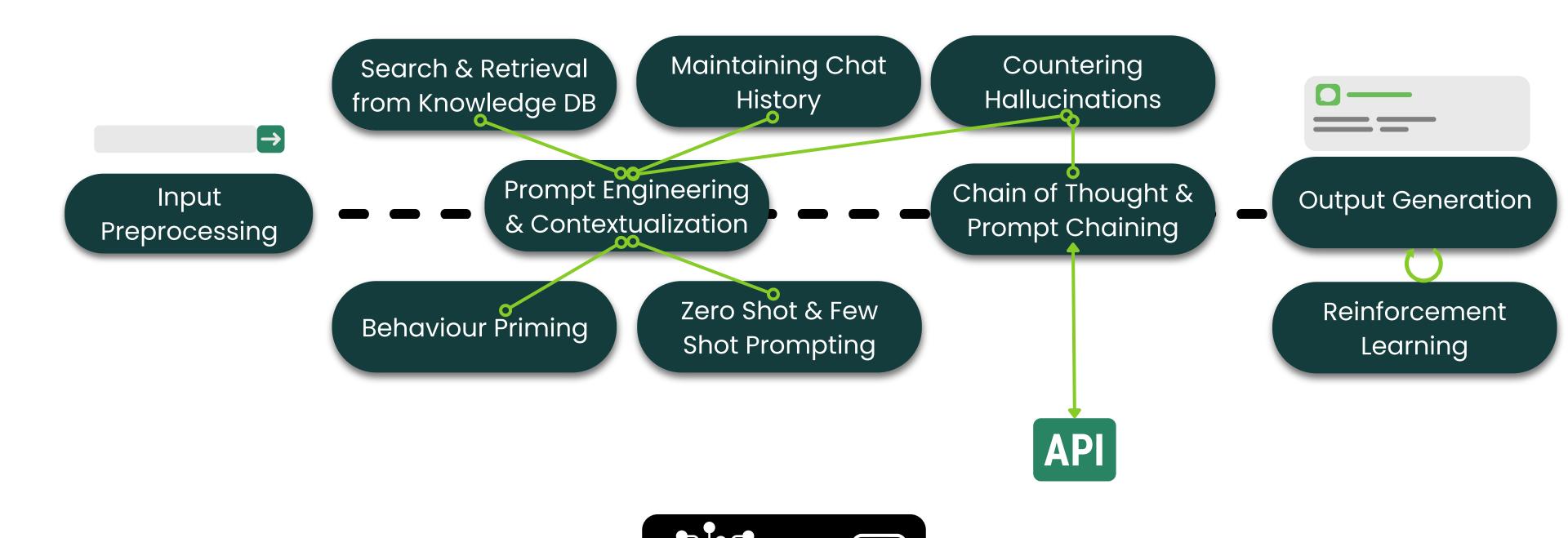
# References

[1] https://link.springer.com/chapter/10.1007/978-1-4020-6710-5_3

[2] https://www.semanticscholar.org/paper/KINO%3A-an-approach-for-rule-based-chatbot-monitoring-D%27%C3%81vila/f44ab7517c37965dc0ab9df46805fa2be2c86486

[3] https://www.sciencedirect.com/science/article/pii/S2666827020300062?via%3Dihub

[4] https://arxiv.org/abs/2302.11382

[5] https://betterprogramming.pub/implementing-nearest-neighbour-search-with-elasticsearch-c59a8d33dd9d

[6] https://commons.wikimedia.org/wiki/File:Data_Queue.svg

[7] https://arxiv.org/abs/2201.11903

[8] https://arxiv.org/abs/2203.06566