# Unleashing AI in Java

A Guide to Semantic Kernel, LangChain4j, and Spring AI
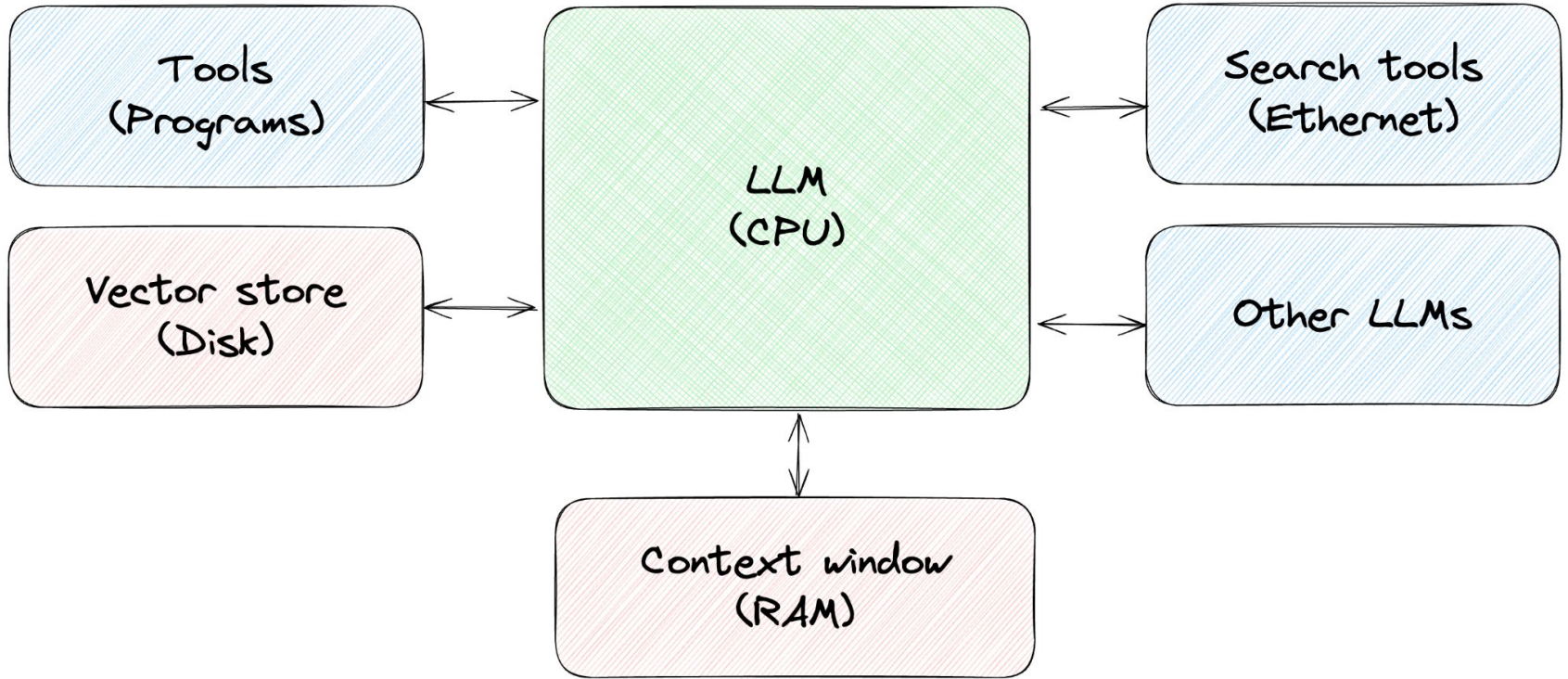
Marcus Hellberg  –  Human Person  –  @marcushellberg

**THE PROBLEM:**

# LLMs are generic, your business is not.

vaadin }>

LangChain4j

spring AI

Semantic Kernel

vaadin }>

# AI FOR ENGINEERS
# = API

@marcushellberg

vaadin }>

Chatbot

Copilot

Agent
autonomy ──────────────────────────────────────────────→

Retrieval-
Augmented
Generation
(RAG)

Fully autonomous
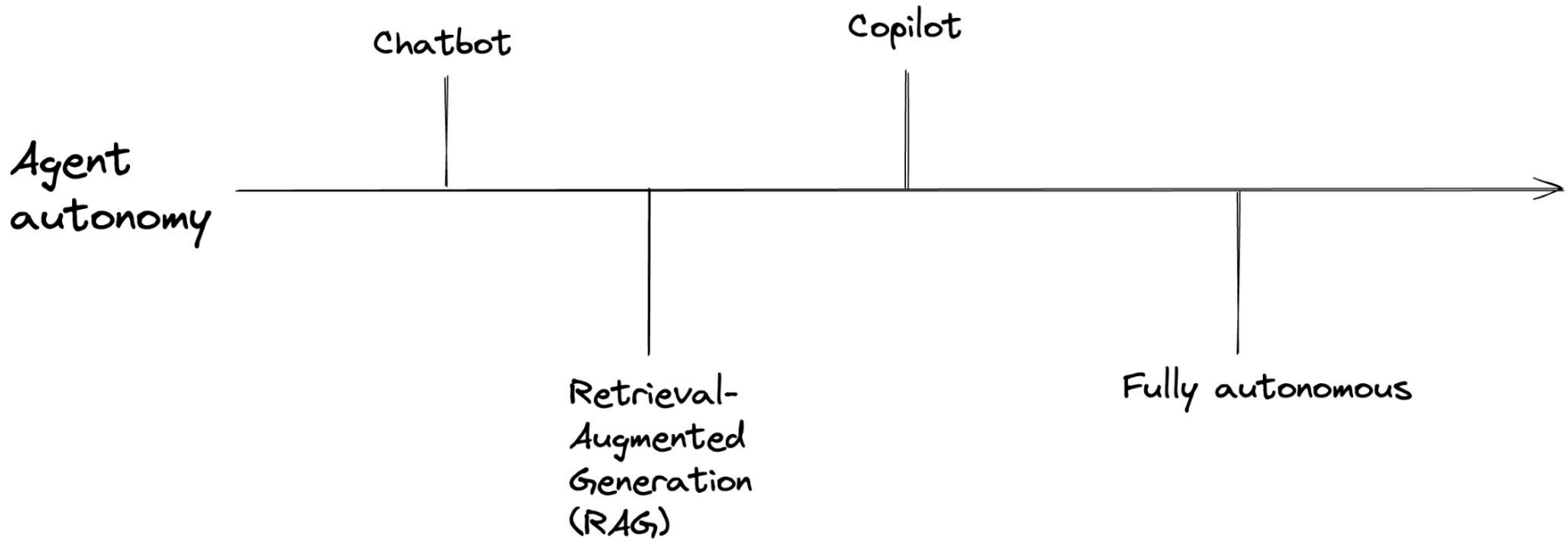
@marcushellberg

Based on Semantic Kernel types of agents documentation.

# Chatbot
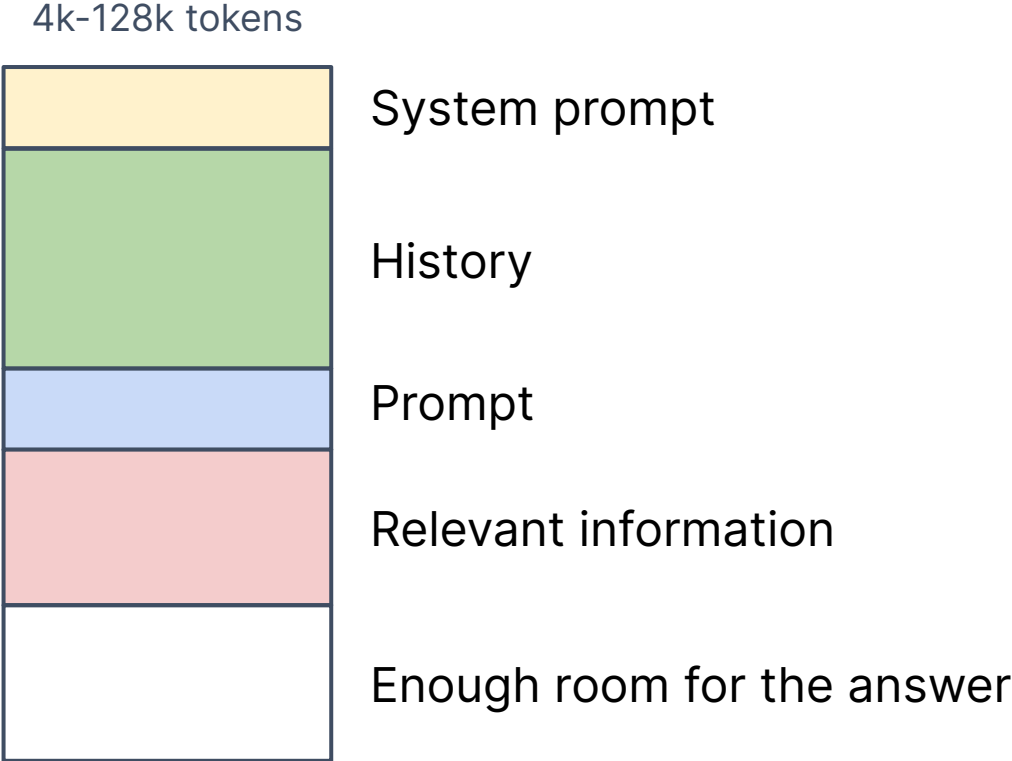
LLM
(CPU)

Context window
(RAM)
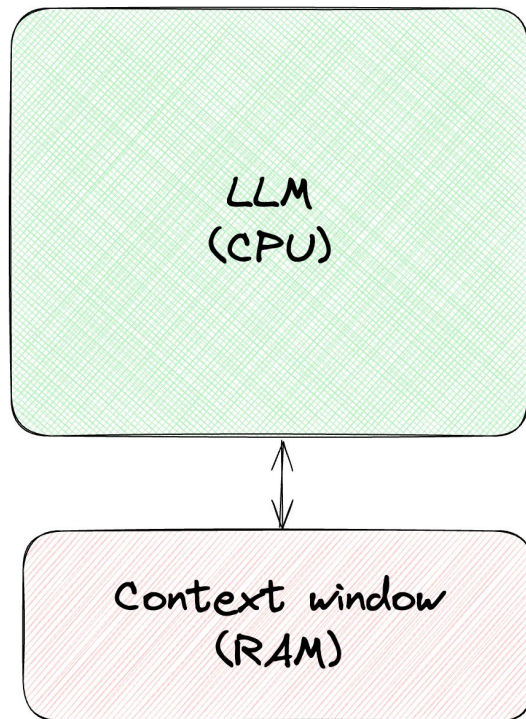
@marcushellberg

Based on Andrej Karpathy's LLM OS.

vaadin }>

# Swappable LLM support

- Fact: a new LLM is created every 3 seconds
- Fact: you don't want to rebuild your app every 3 seconds
- Different LLMs are good at different tasks

⇒ We want to be able to support different LLMs without rewriting our code

vaadin }>

# Context window (memory) management

4k-128k tokens

| | |
|---|---|
| System prompt | |
| History | |
| Prompt | |
| Relevant information | |
| Enough room for the answer | |

**Tokens?** 🤔

A chunk of text, ranging from a character to a full word. Used to break down text for processing.

The token count is usually about 25% greater than the word count.

vaadin }>

# Prompt templates

- The ability to create prompts based on parameterized templates
- Useful when chaining calls

```
You are a helpful AI assistant. You like to share
interesting facts with people.

When answering, follow these rules:
 - The answer MUST be suitable for all audiences
 - The answer should be 5 sentences or less
 - Format using markdown

Tell me an interesting fact about {{topic}}
```

vaadin }>

# Output parsers

- Get the output in the format you need: JSON, POJO, XML, etc

vaadin }>

```java
record Translation(
    String original,
    String targetLanguage,
    String translated){}

interface Assistant {
    @UserMessage("Say '{{text}}' in {{language}}")
    Translation translate(@V("text") String text, @V("language") String language);
}
```

```
Assistant assistant = AiServices.builder(Assistant.class)
        .chatLanguageModel(OpenAiChatModel.withApiKey(openaiApiKey))
        .chatMemory(MessageWindowChatMemory.withMaxMessages(10))
        .build();

var translation = assistant.translate("Language chain", "Swedish");
```

```
Translation[original=Language chain, targetLanguage=Swedish,
translated=Språkkedja]
```

```java
record Translation(
    String original,
    String targetLanguage,
    String translated){}

private final ChatClient chatClient;

SpringAIBasics(ChatClient chatClient) {
    this.chatClient = chatClient;
}
```

```java
var outputParser = new BeanOutputParser<>(Translation.class);
var template = new PromptTemplate("Say '{text}' in {language}. {format}");

var prompt = template.create(Map.of(
        "text", "Spring AI",
        "language", "Swedish",
        "format", outputParser.getFormat()));

var resultJson = chatClient.call(prompt).getResult().getOutput().getContent();
var translation = outputParser.parse(resultJson);
```
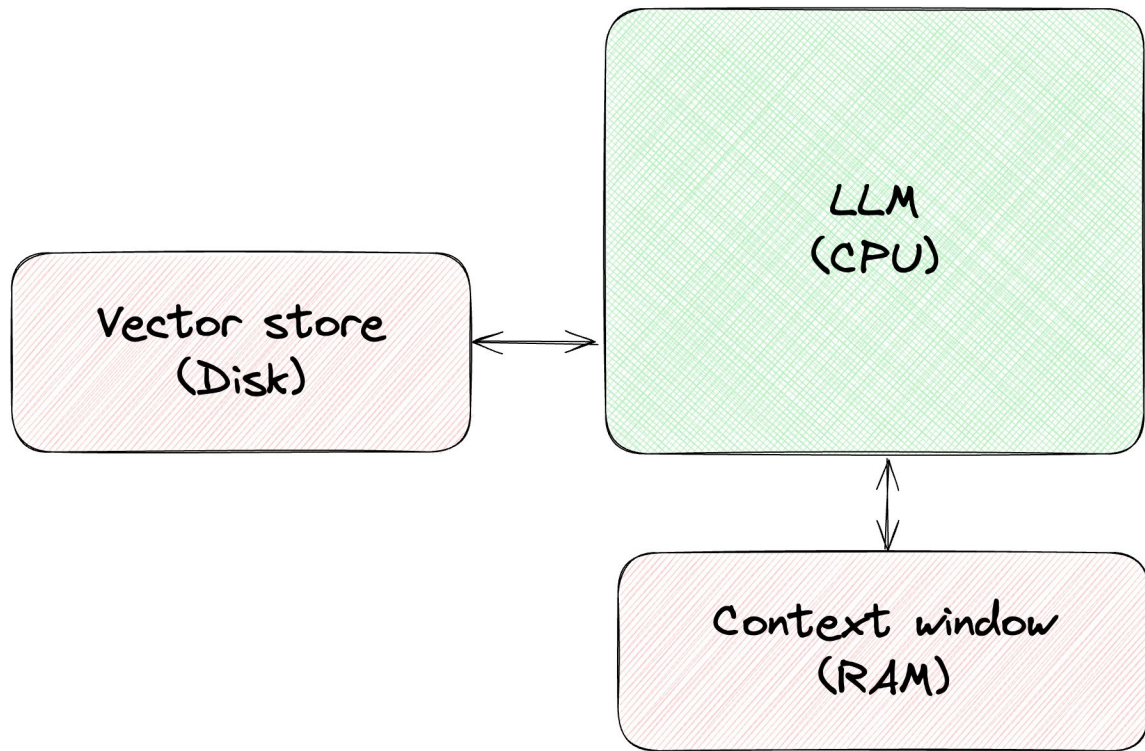
```
Translation[original=Spring AI, targetLanguage=Swedish,
translated=Våren AI]
```

https://github.com/microsoft/semantic-kernel/blob/java-v1/java/samples/sk-v1/01-SimpleChat/src/main/java/Main.java

# Retrieval-augmented generation

vaadin }>

# LLMs know two things

1. Content they're trained on
2. Content you give them as context

vaadin ]>

# Teaching an LLM

1. Train your own model ($$$)
2. Fine-tune an existing model ($)
3. **Stuff the context (~$0)**

# Finding the relevant information

1. Booking Flights
- Book via our website or mobile app.
- Full payment required at booking.
- Ensure accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.

2. Changing Bookings
- Changes allowed up to 24 hours before flight.
- Change via online or contact our support.
- Change fee: $50 for Economy, $30 for Premium Economy, Free for Business Class.

3. Cancelling Bookings
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

vaadin }>

1. Booking Flights
- Book via our website or mobile app.
- Full payment required at booking.
- Ensure accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.

2. Changing Bookings
- Changes allowed up to 24 hours before flight.
- Change via online or contact our support.
- Change fee: $50 for Economy, $30 for Premium Economy, Free for Business Class.

3. Cancelling Bookings
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

1. Booking Flights
- Book via our website or mobile app.
- Full payment required at booking.
- Ensure accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.
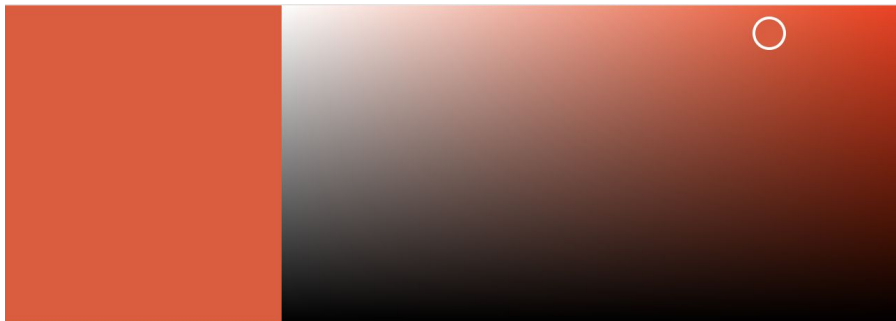
→ [45, 189, 201]

2. Changing Bookings
- Changes allowed up to 24 hours before flight.
- Change via online or contact our support.
- Change fee: $50 for Economy, $30 for Premium Economy, Free for Business Class.

→ [173, 67, 245]

3. Cancelling Bookings
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

→ [98, 12, 76]

vaadin }>

HEX

#eb5234

RGB
235, 82, 52

CMYK
0%, 65%, 78%, 8%

HSV
10°, 78%, 92%

HSL
10°, 82%, 56%

**1. Booking Flights**
- Book via our website or mobile app.
- Full payment required at booking.
- Ensure accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.

→ [45, 189, 201]

**2. Changing Bookings**
- Changes allowed up to 24 hours before flight.
- Change via online or contact our support.
- Change fee: $50 for Economy, $30 for Premium Economy, Free for Business Class.

→ [173, 67, 245]

**3. Cancelling Bookings**
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

→ [98, 12, 76]

vaadin}>

# Vector Store ✨

1. Booking Flights
- Book via our website or mobile app.
- Full payment required at booking.
- Ensure accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.

→ [45, 189, 201]

2. Changing Bookings
- Changes allowed up to 24 hours before flight.
- Change via online or contact our support.
- Change fee: $50 for Economy, $30 for Premium Economy, Free for Business Class.

→ [173, 67, 245]

3. Cancelling Bookings
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

→ [98, 12, 76]

vaadin }>

# Retrieving documents

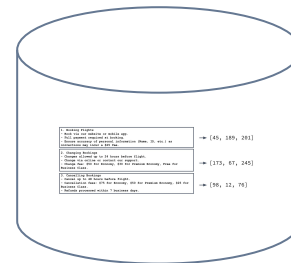**"What's the cancellation policy?"** ⟶ `[96, 14, 73]` ⟶

3. Cancelling Bookings
- Cancel up to 48 hours before flight.
- Cancellation fees: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds processed within 7 business days.

vaadin}>

```java
var documentPath = termsOfService.getFile().toPath();

var documentParser = new TextDocumentParser();
var document = FileSystemDocumentLoader.loadDocument(documentPath, documentParser);

var splitter = DocumentSplitters.recursive(400, 0);
var embeddingModel = new AllMiniLmL6V2EmbeddingModel();
var embeddingStore = new InMemoryEmbeddingStore<TextSegment>();

var ingestor = EmbeddingStoreIngestor.builder()
        .documentSplitter(splitter)
        .embeddingModel(embeddingModel)
        .embeddingStore(embeddingStore)
        .build();

ingestor.ingest(document);
```

```java
var contentRetriever = EmbeddingStoreContentRetriever.builder()
        .embeddingStore(embeddingStore)
        .embeddingModel(embeddingModel)
        .maxResults(2)
        .minScore(0.5)
        .build();

var chatModel = OpenAiChatModel.builder().apiKey(openaiApiKey).build();

var assistant = AiServices.builder(CustomerSupportAgent.class)
        .chatLanguageModel(chatModel)
        .chatMemory(MessageWindowChatMemory.withMaxMessages(10))
        .contentRetriever(contentRetriever)
        .build();

System.out.println(assistant.answer("What is the cancellation policy?"));
```

```
The cancellation policy for Funnair is as follows:
- Bookings can be cancelled up to 48 hours before the flight.
- Cancellation fees vary depending on the class of the booking: $75
for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds will be processed within 7 business days.
```

```java
var vectorStore = new SimpleVectorStore(embeddingClient);

var textReader = new TextReader(termsOfService);
vectorStore.accept(textReader.get());




var template = """
        Answer the given question only using the information in the provided terms.
        If you do not know the answer, please respond with "I don't know".

        TERMS
        -----
        {documents}
        -----
        """;
```

```java
var query = "What is the cancellation policy?";

var relevantDocs = vectorStore.similaritySearch(
        SearchRequest.query(query)
                .withTopK(2)
                .withSimilarityThreshold(0.5))
    .stream()
    .map(Document::getContent)
    .collect(Collectors.joining("\n\n"));

var systemMessage = new SystemPromptTemplate(template)
                        .createMessage(Map.of("documents", relevantDocs));
var userMessage = new UserMessage(query);

var messageHistory = new Prompt(List.of(systemMessage, userMessage));

var response = chatClient.call(messageHistory);

System.out.println(response.getResult().getOutput().toString());
```
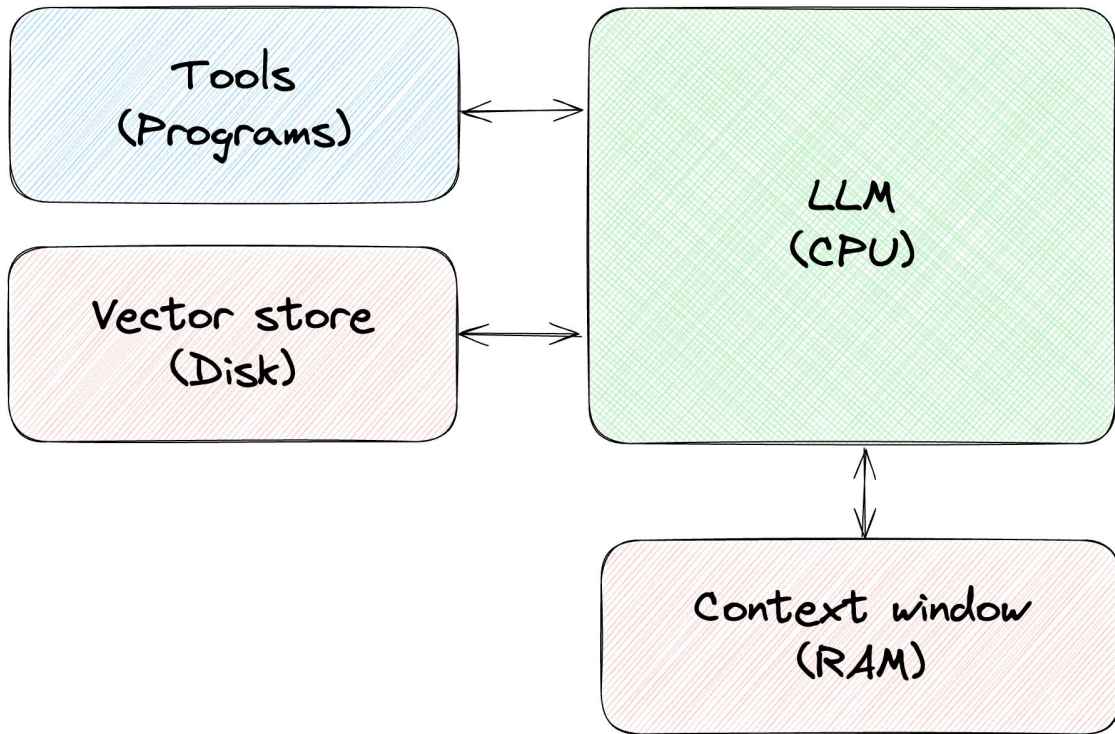
The cancellation policy for Funnair is as follows:

- Bookings can be cancelled up to 48 hours before the flight.
- Cancellation fees apply: $75 for Economy, $50 for Premium Economy, $25 for Business Class.
- Refunds will be processed within 7 business days.

Please note that changes to bookings are allowed up to 24 hours before the flight. Changing a booking incurs a fee of $50 for Economy, $30 for Premium Economy, and is free for Business Class. Changes can be made online or by contacting Funnair's support.

When booking a flight, full payment is required, and it is important to ensure the accuracy of personal information (Name, ID, etc.) as corrections may incur a $25 fee.

[https://github.com/microsoft/semantic-kernel/blob/java-v1/java/samples/sk-v1/03-SimpleRag/src/main/java/Main.java](https://github.com/microsoft/semantic-kernel/blob/java-v1/java/samples/sk-v1/03-SimpleRag/src/main/java/Main.java)

# Copilot

Tools
(Programs)

Vector store
(Disk)

LLM
(CPU)

Context window
(RAM)

# Using tools

- Define a set of tools or functions that the LLM can use when helping the user
- Full read/write access to our database would be a bad idea

vaadin ]>

```java
class Calculator {

    @Tool("Calculates the length of a string")
    int stringLength(String s) {
        System.out.println("Called stringLength with s='" + s + "'");
        return s.length();
    }

    @Tool("Calculates the sum of two numbers")
    int add(int a, int b) {
        System.out.println("Called add with a=" + a + ", b=" + b);
        return a + b;
    }

    @Tool("Calculates the square root of a number")
    double sqrt(int x) {
        System.out.println("Called sqrt with x=" + x);
        return Math.sqrt(x);
    }
}
```

```java
var assistant = AiServices.builder(Assistant.class)
        .chatLanguageModel(OpenAiChatModel.withApiKey(openaiApiKey))
        .chatMemory(MessageWindowChatMemory.withMaxMessages(10))
        .tools(new Calculator())
        .build();

var query = """
        What is the square root of the sum of the numbers of letters
        in "hello" and "world"?
        """;

System.out.println(assistant.chat(query));
```

```
Called stringLength with s='hello'
Called stringLength with s='world'
Called add with a=5, b=5
Called sqrt with x=10
The square root of the sum of the numbers of letters
in "hello" and "world" is approximately 3.162.
```

https://github.com/microsoft/semantic-kernel/blob/java-v1/java/samples/sk-v1/03-SimpleRag/src/main/java/Main.java

# Fully-autonomous

THE FUTURE OF AGENTS

# AI is about to completely change how you use computers
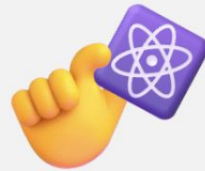
And upend the software industry.

By **Bill Gates**  |  November 09, 2023  ·  12 minute read

# Demo time! 🧑‍🔬

# Funnair customer support ✈

**Assistant**

Welcome to Funnair! How can I help you?

Message | Send

# Bookings database

| # | First name | Last name | Date | From | To | Booking status | Booking class |
|---|---|---|---|---|---|---|---|
| 101 | John | Doe | 2024-01-30 | MAD | SFO | ✅ | PREMIUM_ECONOMY |
| 102 | Jane | Smith | 2024-02-01 | ARN | JFK | ✅ | ECONOMY |
| 103 | Michael | Johnson | 2024-02-03 | FRA | HEL | ✅ | PREMIUM_ECONOMY |
| 104 | Sarah | Williams | 2024-02-05 | MUC | HEL | ✅ | BUSINESS |
| 105 | Robert | Taylor | 2024-02-07 | JFK | MAD | ✅ | ECONOMY |

**github.com/marcushellberg/java-ai-playground**

|                             | Spring AI | LangChain4j | Semantic Kernel |
|-----------------------------|-----------|-------------|-----------------|
| Prompt templates            | ✅        | ✅          | ✅              |
| Output formatting           | ✅        | ✅          | ✅              |
| RAG support                 | ✅        | ✅          | ✅              |
| Message history management  | -         | ✅          | ✅              |
| Tools/Functions/Plugins     | -         | ✅          | ✅              |
| Documentation               | 👍        | ✍️          | 👻              |
| Development activity         | 🚀🚀🚀    | 🚀🚀🚀🚀🚀  | 🤔              |
| Autonomy level              | RAG       | Copilot     | Copilot         |

@marcushellberg

vaadin }>

# Thanks!

@marcushellberg