Divide and Conquer: Building Reliable Text-to-SQL Pipelines

Simona Mazzarino
DevFest Alps, 22 November 2025

Who I am

Master's degree in Digital Humanities - Language Technologies at University of Pisa

Data Scientist at Clearbox AI

Outside of work, I'm an organizer Python Torino community and a volunteer for Pycon Italia.

Slides & Code

Table of Contents

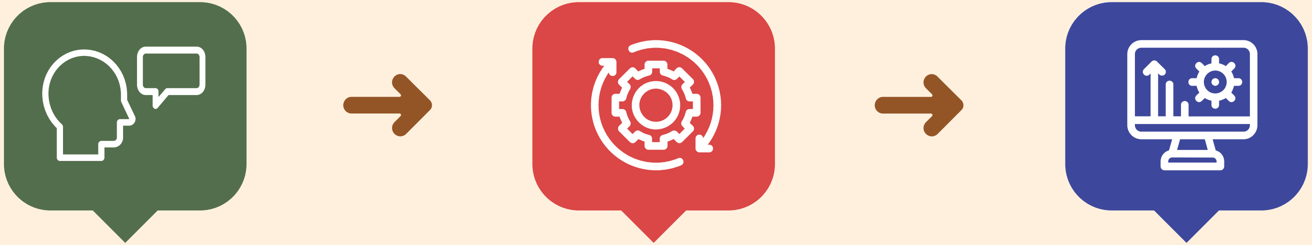
01 Concepts of Text-to-SQL Pipelines

02 Agentic Approach

03 Modular Approach

04 Conclusion and Q&A

Text-To-SQL Pipeline



Input Text

Which is the most successful team in Dota 2?

LLM

SQL Output

SELECT TeamName, TotalUSDPrize FROM highest_earning_teams WHERE Game='Dota 2' ORDER BY TotalUSDPrize DESC LIMIT 1;

Key Aspects in Text-to-SQL Pipelines

- Natural Language Understanding
- Schema Understanding & Alignment
- Query Planning
- SQL Generation
- Validation & Safety
- Result Handling
- Answer Generation
- Observability & Monitoring

What is Haystack?

Haystack is an **open source framework** for building production-ready LLM applications and retrieval-augmented generative pipelines over large document collections.

haystack
by deepset

Notebooks inspired by this:
https://haystack.deepset.ai/cookbook/chat_with_sql_3_ways

Key concepts in Haystack

COMPONENTS

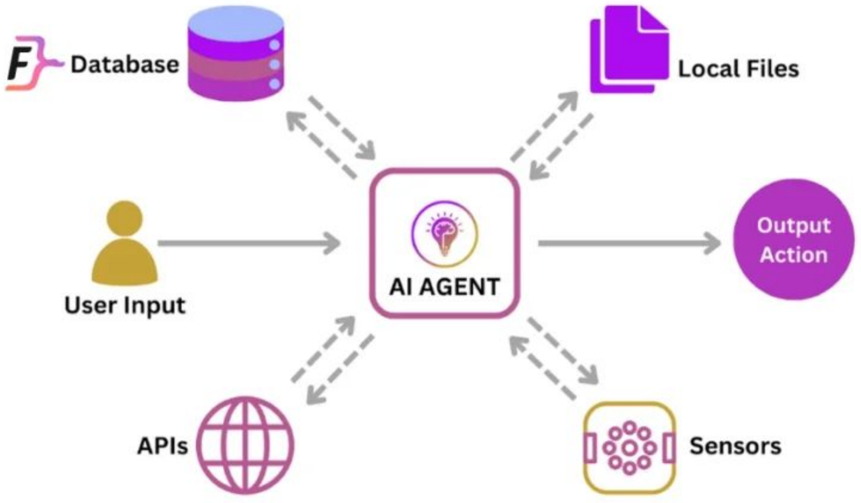
Components are the building blocks of a pipeline. They perform tasks such as preprocessing, retrieving, or generating text while routing queries through different branches of a pipeline.

PIPELINES

A pipeline in Haystack consists of different components, such as retrievers, readers, generators, and other modules, that work together to process queries and provide accurate, meaningful results.

Agentic Approach

What is Agentic AI?



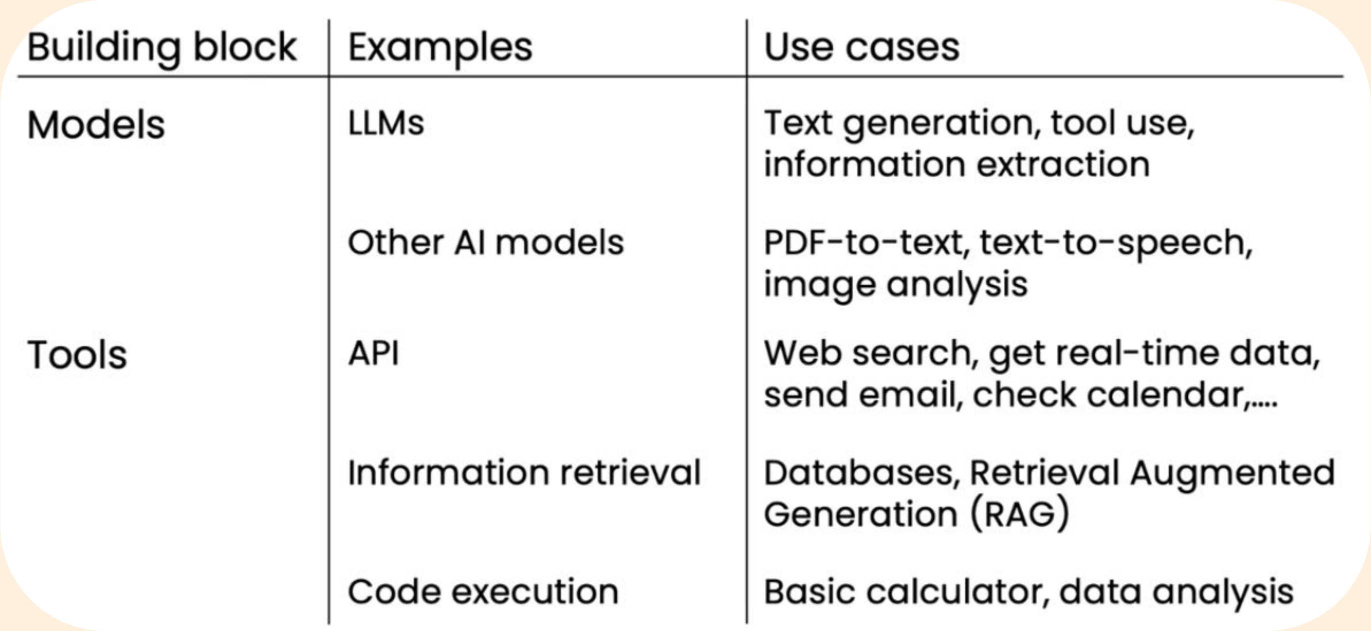
AI Agent Architecture

Image Source: https://www.falkordb.com/blog/ai-agents-memory-systems/

Key features of AI Agents

- Planning
- Acting
- Observing
- Reasoning
- Collaborating
- Self-Refining

Agentic AI - Building Blocks



Building block	Examples	Use cases
Models	LLMs	Text generation, tool use, information extraction
	Other AI models	PDF-to-text, text-to-speech, image analysis
Tools	API	Web search, get real-time data, send email, check calendar,....
	Information retrieval	Databases, Retrieval Augmented Generation (RAG)
	Code execution	Basic calculator, data analysis

Image Source: deeplearning.ai

Tools

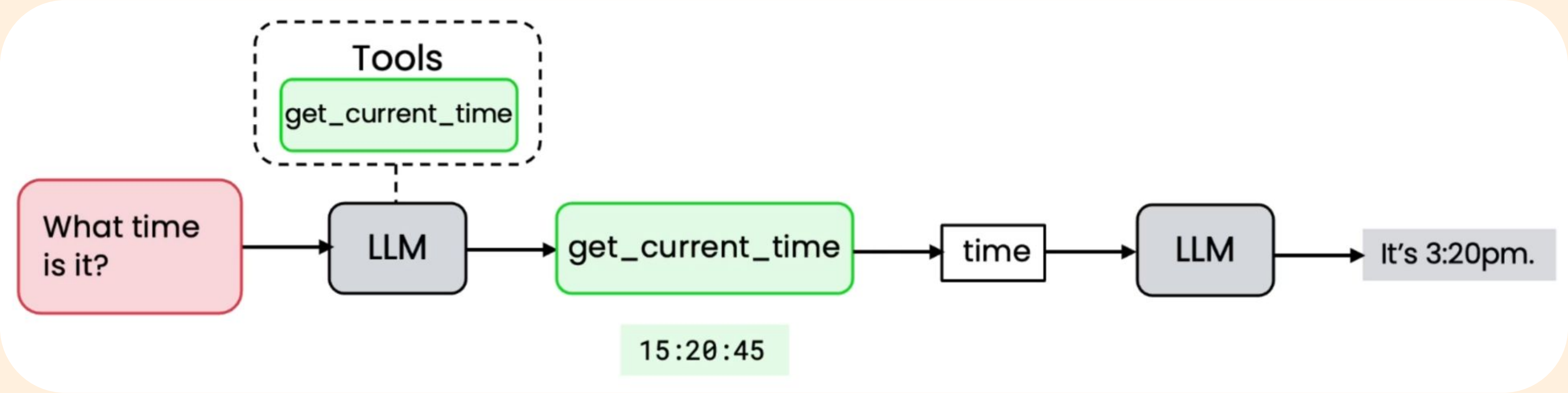


Image Source: deeplearning.ai

Pro & Contro

Pro

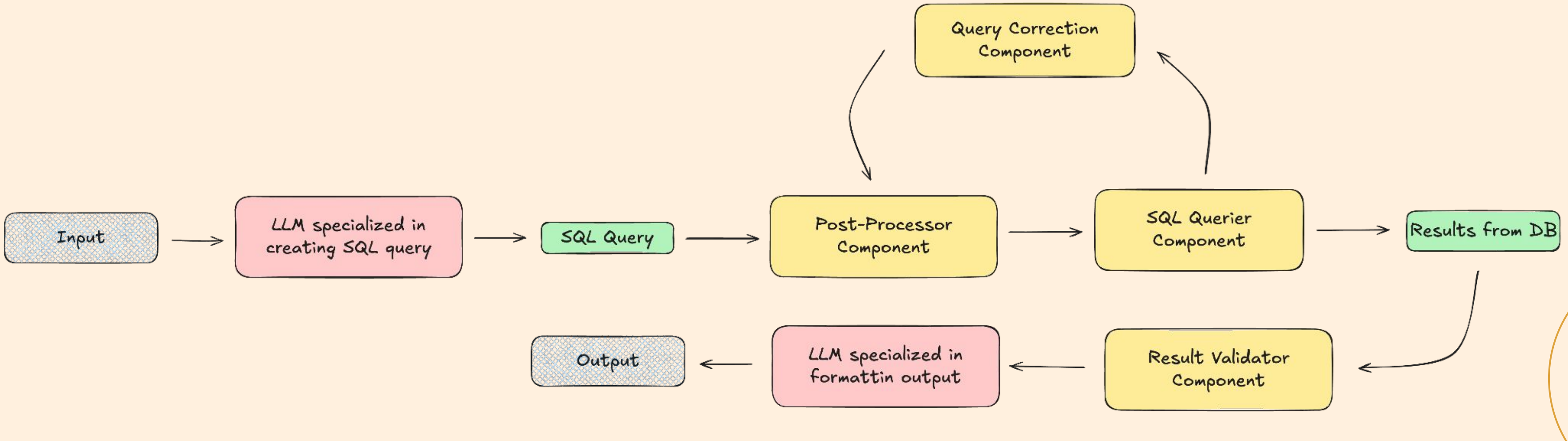
- Minimal code required
- No need for complex pipeline architectures
- Cost-efficient to run

Contro

- Limited control and observability
- A single system prompt governs the entire workflow
- No constraints on contextual scope

Modular Approach

What is the modular approach?



Input

LLM specialized in creating SQL query

SQL Query

Post-Processor Component

Query Correction Component

SQL Querier Component

Results from DB

Result Validator Component

LLM specialized in formattin output

Output

Pro & Contro

Pro

- Specialized LLMs per task
- High control & observability
- Easily extensible
- Safer by design

Contro

- More complex pipeline
- Higher cost
- More engineering effort
- Higher latency

Conclusion

Text-to-SQL requires more than SQL generation, it needs understanding, validation, safety, and clear reasoning.

Agentic approaches are simple but opaque; modular pipelines offer control, reliability, and scalability.

With Haystack, we can compose specialized components into a transparent, maintainable system.

LLMs are powerful, but pipelines make them dependable.

Any Questions?

Thanks for listening!

simona@clearbox.ai
simona.mazzarino@gmail.com

Thanks !

simona@clearbox.ai
simona.mazzarino@gmail.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**