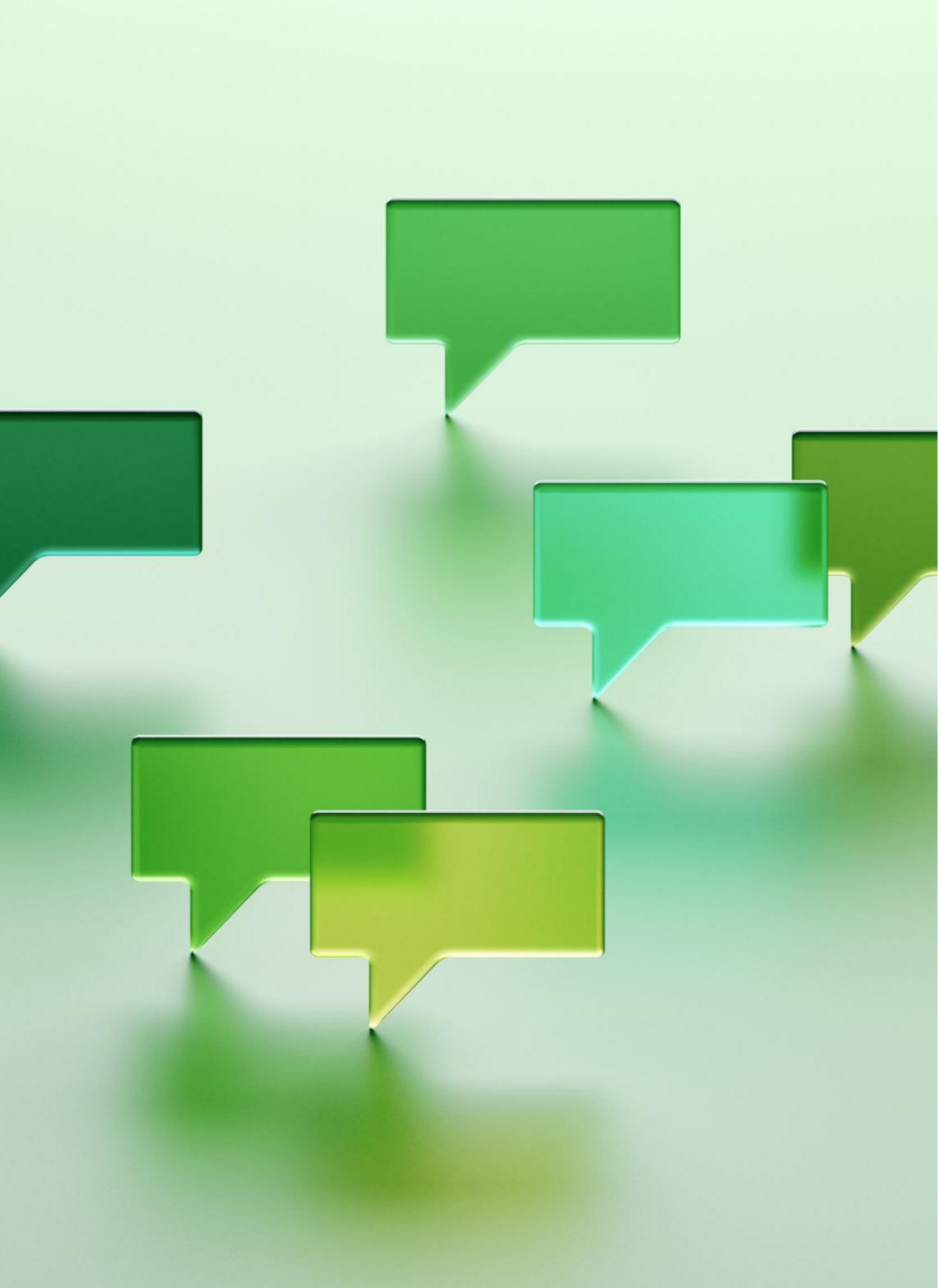
Georgios Peikos

University of Milano-Bicocca, Milan, Italy

Department of Informatics, Systems, and Communication (DISCo)

Information Access: Search, Recommendation, Large Language Models, Conversational AI, Retrieval-Augmented Generation, and Agents Working Together

INFORMATION RETRIEVAL AND RECOMMENDER SYSTEMS



In this Lecture...

We will discuss **Information Retrieval**, Recommendation, Large Language Models, **Conversational AI**, **Retrieval-Augmented Generation**, Agent systems, and the emerging **idea of** Active Retrieval and Recommendation.

We will show **how** these approaches **relate to each other**, **differ** in their goals, and **combine** in practical applications.

We **examine real cases** that reveal system limitations including misuse of agents and situations where generated information caused legal or organizational problems.

We briefly reflect on the environmental impact of these technologies.

How do people find information today?

Search engines (domain specific, web, etc.) that support users **retrieve** information from large collections.

Recommender systems that **suggest** items, products, or content (news, social media, and streaming content) based on user preferences/interests.

Conversational assistants that **answer** natural language questions in an interactive manner.

Autonomous or semi-autonomous AI agents that **plan** steps and **interact** with **tools** on **behalf of the user,** to answer their information needs.

Technologies Behind These Systems

Information Retrieval Approaches that interpret **user intents**, **index**, **search**, and **rank** large collections of information items.

Deep learning RS models that learn **patterns** from **user behavior** and **preferences**.

Natural language processing techniques such as Large Language Models that **generate** and interpret text.

Retrieval augmented generation systems that **combine search with generation** to provide grounded answers.

Knowledge bases that **structure information** and **provide** factual grounding.

Tool-use and planning frameworks that **allow AI agents** to perform multi-step tasks.

Information Retrieval Systems

Information Retrieval Systems – User's Perspective

A person **operates** within a physical environment and interacts with an Information Retrieval (IR) system to **find information**.

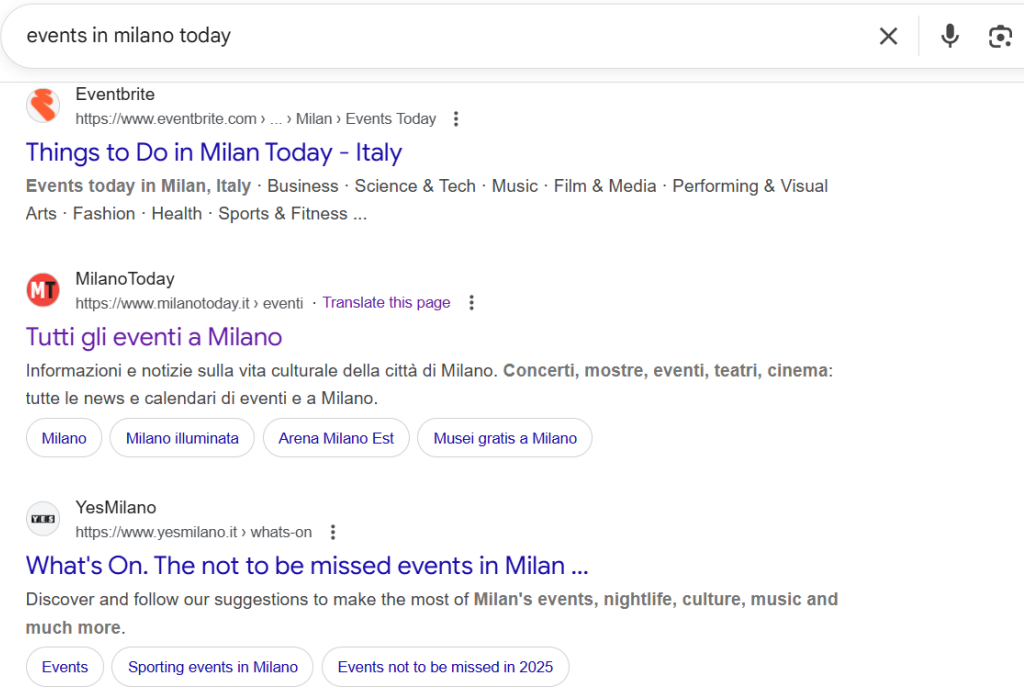
They are engaged in a **specific task**, often related to work or leisure activities.

To **accomplish the task**, they need information that fills their knowledge gaps.

The **knowledge gaps and their overall situation** are forming their information needs.

Information needs exist in the user's mind and guide their **search behavior**.

Information Retrieval Systems – User's Perspective



To **express** their **information needs** to the IR system, users formulate queries. Queries are usually composed of **keywords** that approximate the user's underlying information need.

Queries serve as the input to an IR system.

The IRS **processes these queries** using the mechanisms covered in the course.

It returns a **ranked list of information items** (such as web pages).

These items are **expected** to **be relevant** to the user's information need and support task completion.

IR models rank information items so that **the most relevant** document appears first in the results list

Information Retrieval Systems – System's Perspective

Document Representation via indexing

Document Representation via a neural language model (embeddings)

Offline Process

Online Process

Query

Query Representation using the index

Query Representation via the same neural model

1st Stage Retrieval

Ranked List

2nd Stage Retrieval

Ranked List

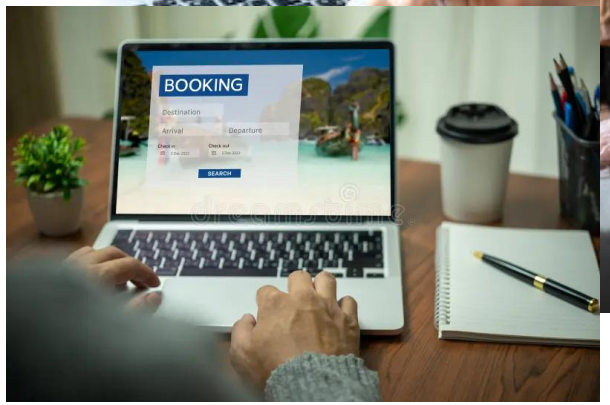
BM25 relies on Index

Bi-encoder relies on stored embedding representation

Expensive Cross-Encoder relies on original document Texts!

Recommender Systems

Recommender Systems – User's Perspective



A person **operates** within a physical environment and interacts with a Recommender System (RS) to discover **useful items**.

They are engaged in a **specific activity**, such as shopping, watching movies, or reading articles.

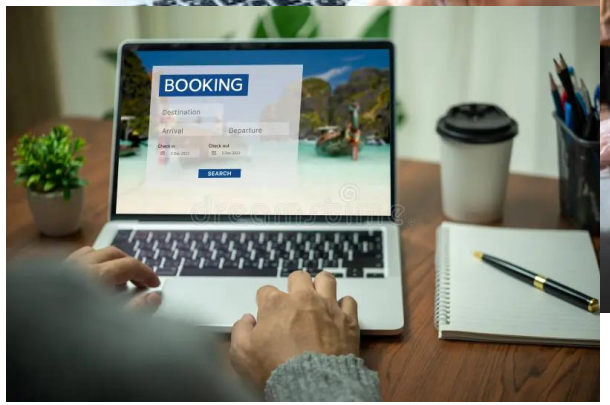
They engage with a digital platform which contains the information items that are related to their activity.

To accomplish their goals, they interact with the user interface by navigating its content.

While browsing, they are exposed to personalized recommendations that suggest items related to their interests or current activity.

They may choose to **follow these** recommendations or ignore them based on personal preference.

Recommender Systems – System's Perspective



The **RS** maintains or builds **a user profile** that represents the **user's preferences and past interactions**, and possibly contextual or demographic information.

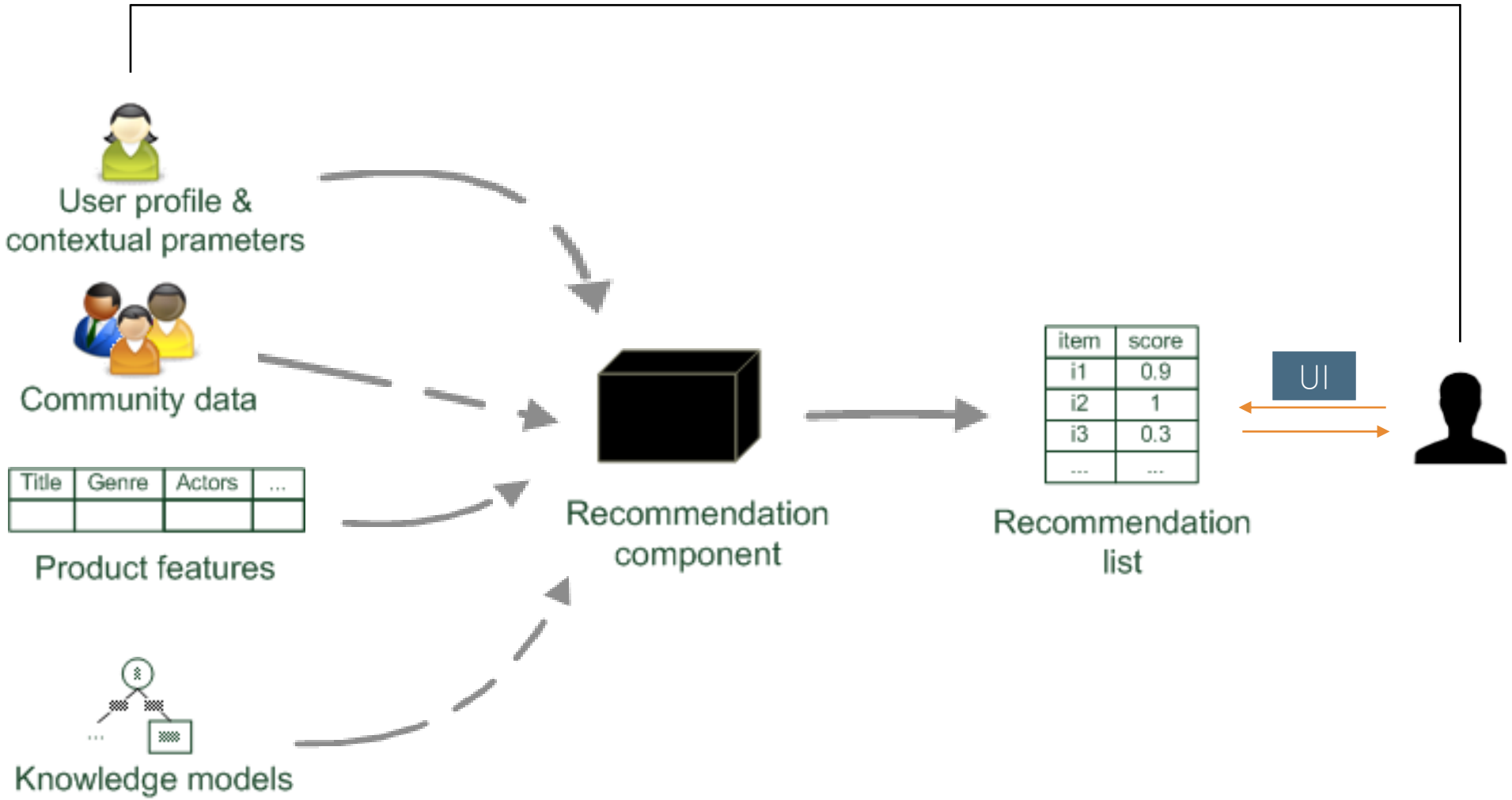
As the **user interacts with the** platform (using the User Interface), the **RS** collects data such as clicks, views, ratings, and purchases to **update the** user profile.

The system analyzes user information (profiles, behavior), item information (content, attributes, or metadata), and interaction data.

It computes **similarities** or relationships between users and items to estimate which items are most likely to **interest** each user.

The RS then creates a **ranked list of recommended items** for the user. These **items are displayed** through the user interface.

Recommender Systems – System's Perspective




Comments?

Conversational Assistants

Conversational Assistants – Overview

The primary definition of conversational agents is related to a **computer program or artificial intelligence** able to hold a conversation *(written or spoken)* with humans through natural language processing (NLP) [1].

From early rule-based conversational systems like ELIZA in the 1970s, to the **chatbots for customer service** based on predefined scripts and limited natural language understanding, to **voice-based assistants** such as Alexa and Siri with advanced speech recognition and contextual understanding, and finally to modern **neural conversational models** like ChatGPT and the **humanoid robot** Sophia that bring natural conversation to life through both language and embodiment.

[1] Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. Expert Systems with Applications, 129, 56-67.

Conversational Assistants – User's Perspective

A person **operates** within a physical environment and interacts with a Conversational Assistant (CA) to obtain **information, recommendations**, accomplish tasks, or engage in dialogue.

User communicate through natural language, either by **typing** or **speaking**.

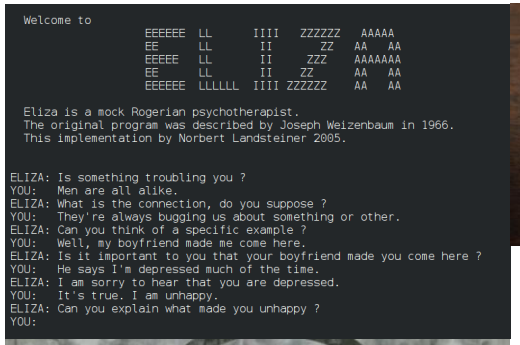
As they use natural language, their inputs are more **expressive** representations of their information needs, approximating their underlying intent better than a simple keyword query.

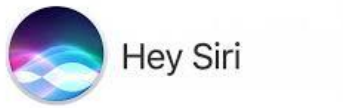
Nowadays, the users often **express directly their** tasks to these assistance and are exposed to an **answer**.

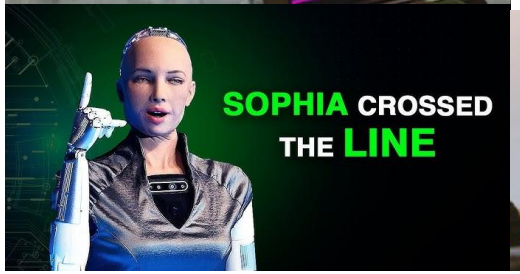
Users can ask **follow-up questions**, **refine** their **requests**, or **shift topics** while maintaining the flow of conversation.

Through this interactive exchange, the user receives information, explanations, or assistance that supports their underlying **tasks** and **goals**.

Conversational Assistants – System's Perspective







The Conversational Assistant (CA) receives user inputs in natural language, either spoken or written.

The CA uses Natural Language Understanding (NLU) to interpret the user's input and identify intent and key information.

A Natural Language Generation (NLG) component then creates a coherent and contextually appropriate response to present to the user.

As the interaction continues, the CA updates the dialogue state and **refines** its **understanding** of the **user's information needs** and preferences.

Through this iterative dialogue, the CA helps the user to access information and complete tasks.

Large Language Models (LLMs) are today the most prominent tools for both NLU and NLG, and are the core technology behind modern CA.

Large Language Models

Large Language Models - Overview

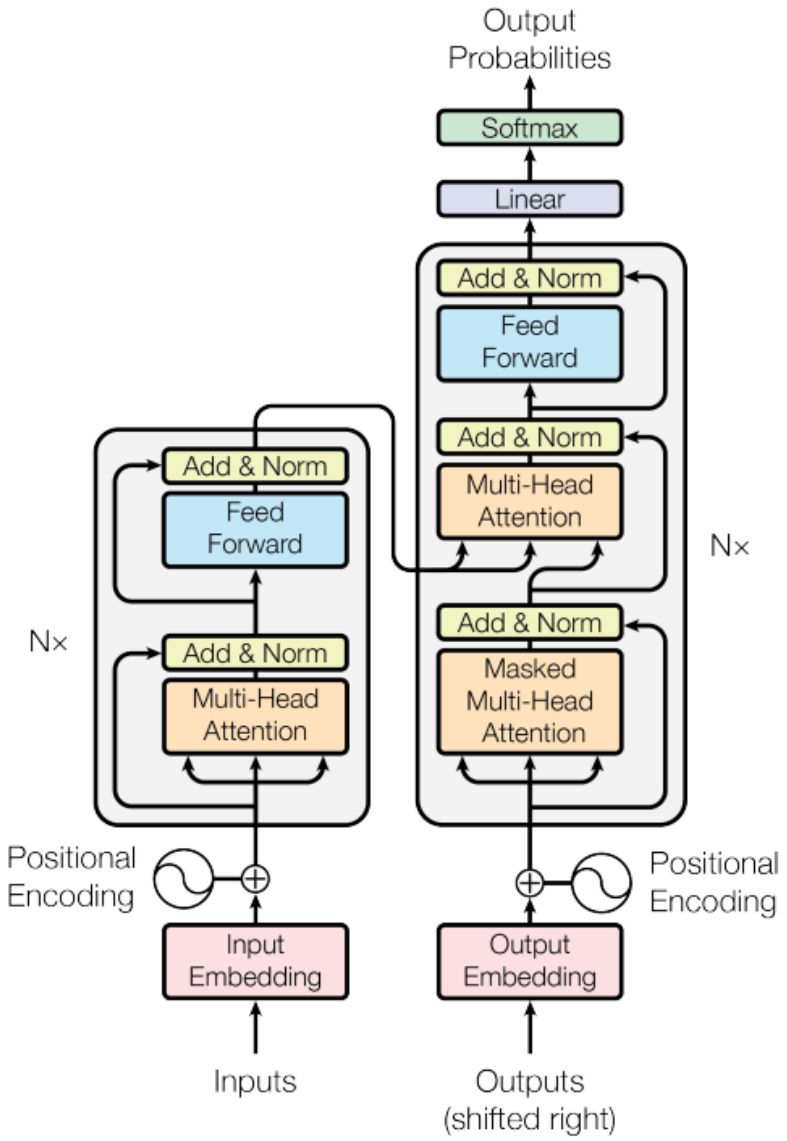


Figure 1: The Transformer - model architecture.

LLMs are systems trained on **massive** text corpora to understand and generate natural language.

They learn statistical patterns and semantic relationships between **words**, enabling them to generate coherent and contextually relevant text.

Built using **transformer architectures**, they use self-attention mechanisms to model **long-range dependencies** in language.

LLMs generate content by **estimating** the most likely next word (token) in a sequence based on the context of all previous words, using **patterns** and relationships **encoded in their parameters**.

The **knowledge** stored in their parameters is referred to as parametric knowledge, representing the internal information the model learns during training.

Their parametric knowledge is **static and cannot easily be updated**, which may lead to outdated or inaccurate outputs, highlighting the need for access to external and current knowledge.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

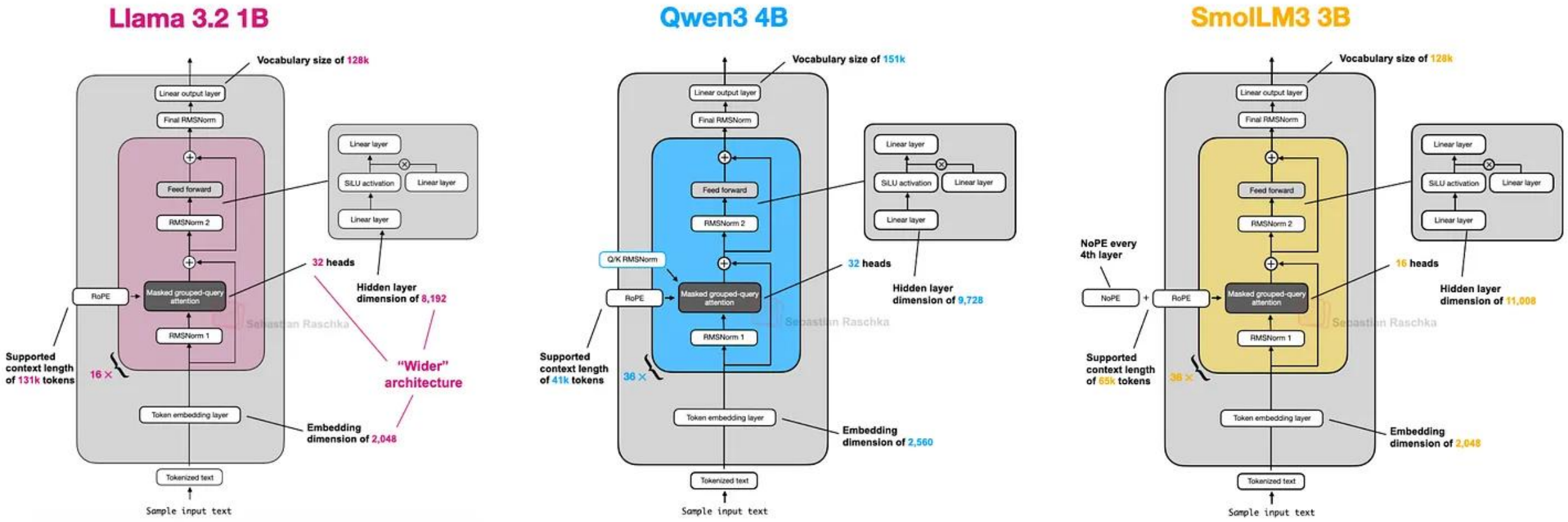
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

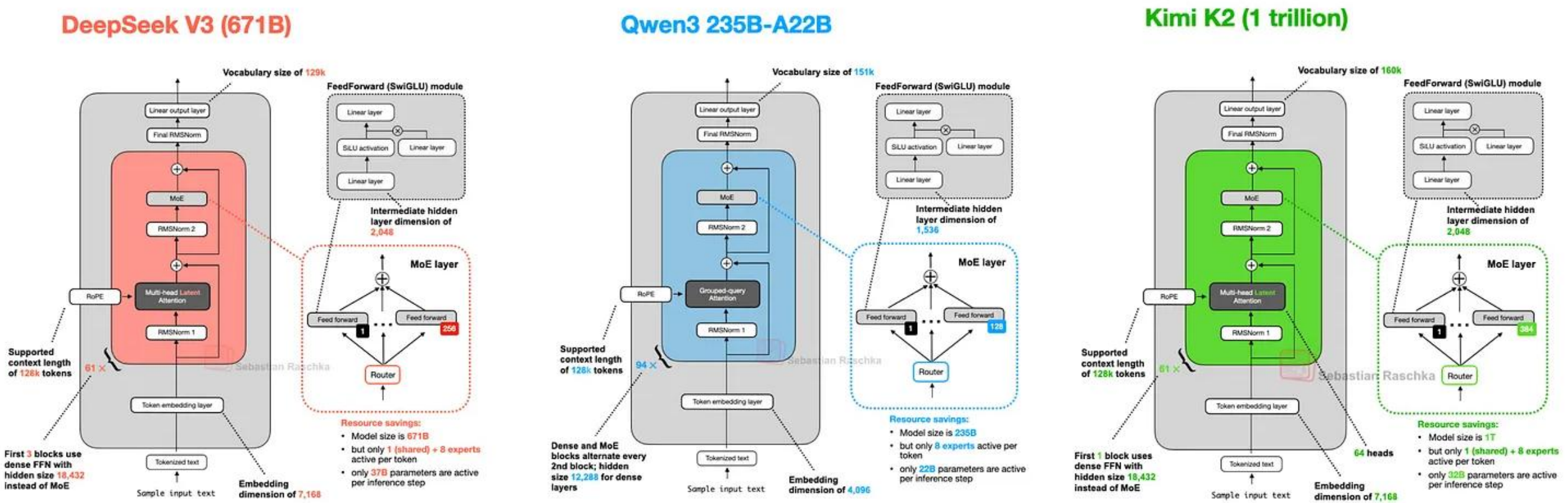
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions

Large Language Models - Today



Source: Sebastian Raschka, PhD, https://magazine.sebastianraschka.com/p/the-big-llm-architecture-comparison?utm_campaign=post&utm_medium=web

Large Language Models - Today



Source: Sebastian Raschka, PhD, https://magazine.sebastianraschka.com/p/the-big-llm-architecture-comparison?utm_campaign=post&utm_medium=web

Curated Visual Sources for Transformers

Tokenization: https://www.youtube.com/watch?v=zduSFxRajkE

Build an LLM from scratch: https://www.youtube.com/watch?v=kCc8FmEb1nY

Codes: https://github.com/M2Lschool/tutorials2024/tree/main/1_nlp

Transformers Overview: https://medium.com/machine-intelligence-and-deep-learning-lab/transformer-the-self-attention-mechanism-d7d853c2c621


Questions?

Retrieval Augmented Generation

Retrieval Augmented Generation - Overview

Retrieval-Augmented Generation (RAG) is a **framework** which combines **information retrieval** with **language generation** to improve the factuality and the overall quality of LLMs generated response.

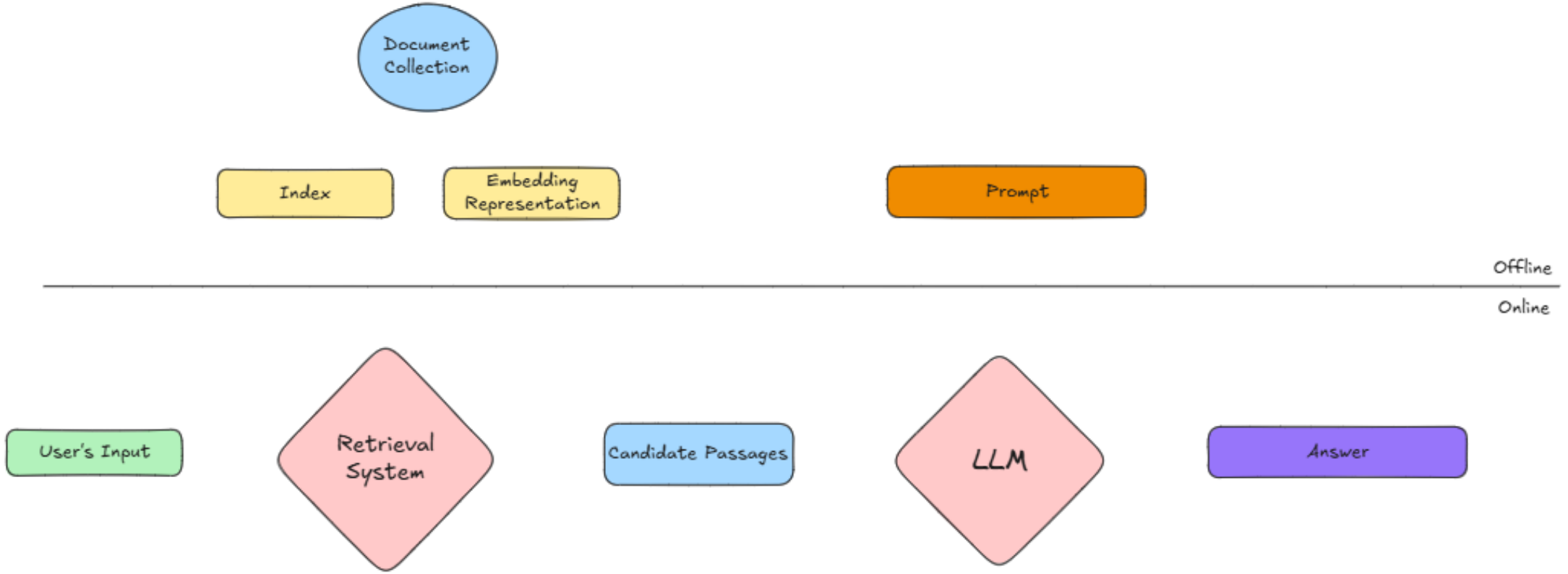
Instead of relying only on parametric knowledge stored in the model's parameters, **RAG systems** retrieve **relevant** external **documents** during inference.

The **retrieved content** is provided as context to the LLM through a prompt, grounding its responses in up-to-date and/or verifiable information.

This approach helps reduce hallucinations, enhances accuracy, and allows the system to adapt dynamically to new or domain-specific knowledge.

RAG is widely used in question answering, knowledge-intensive dialogue, and enterprise search applications where factual consistency is crucial.

Retrieval Augmented Generation – System Overview



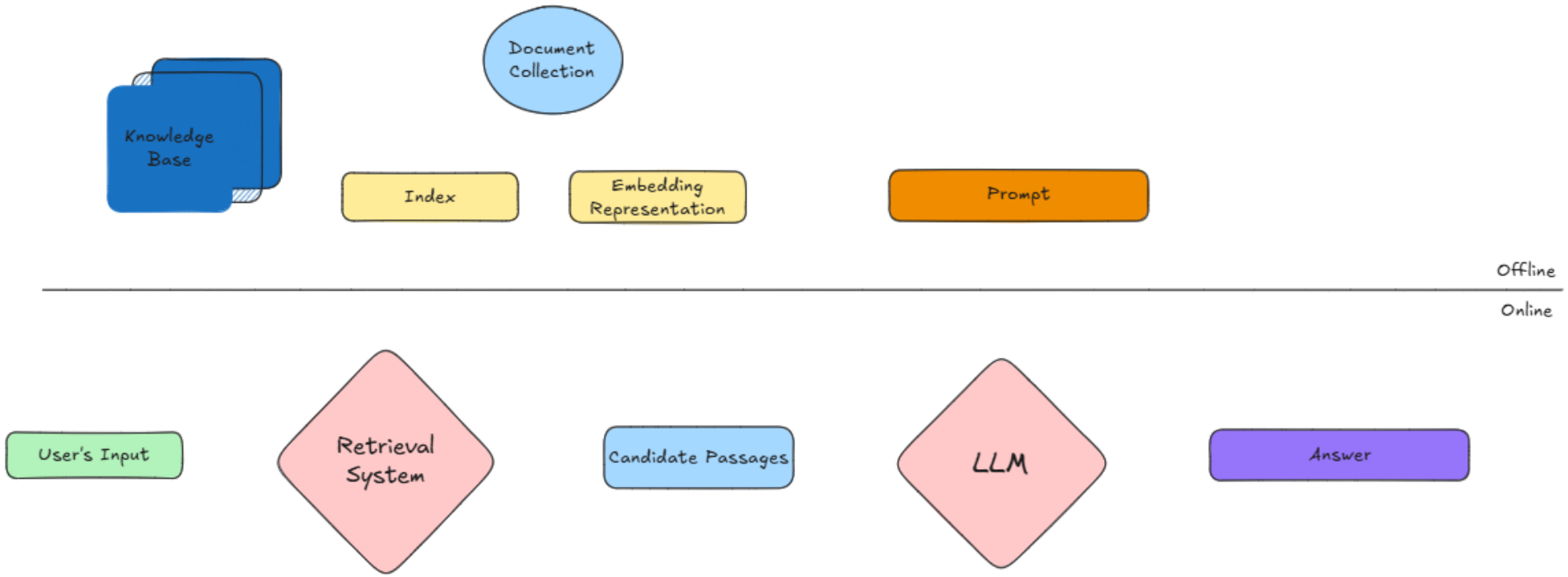
Knowledge Base Integration in LLMs & CA

Knowledge Base Integration in LLMs - Overview

Knowledge Base Integration is another way to enhance Large Language Models (and consequently a CA) by connecting them to structured **external knowledge**.

It allows models to access up-to-date and factually accurate information beyond what is stored in their **parameters**.

This approach improves **accuracy**, **explainability**, and **domain adaptability**, while reducing hallucinations and outdated responses.




Questions?

Autonomous or semi-autonomous AI agents

Autonomous or semi-autonomous AI agents[2] - Overview

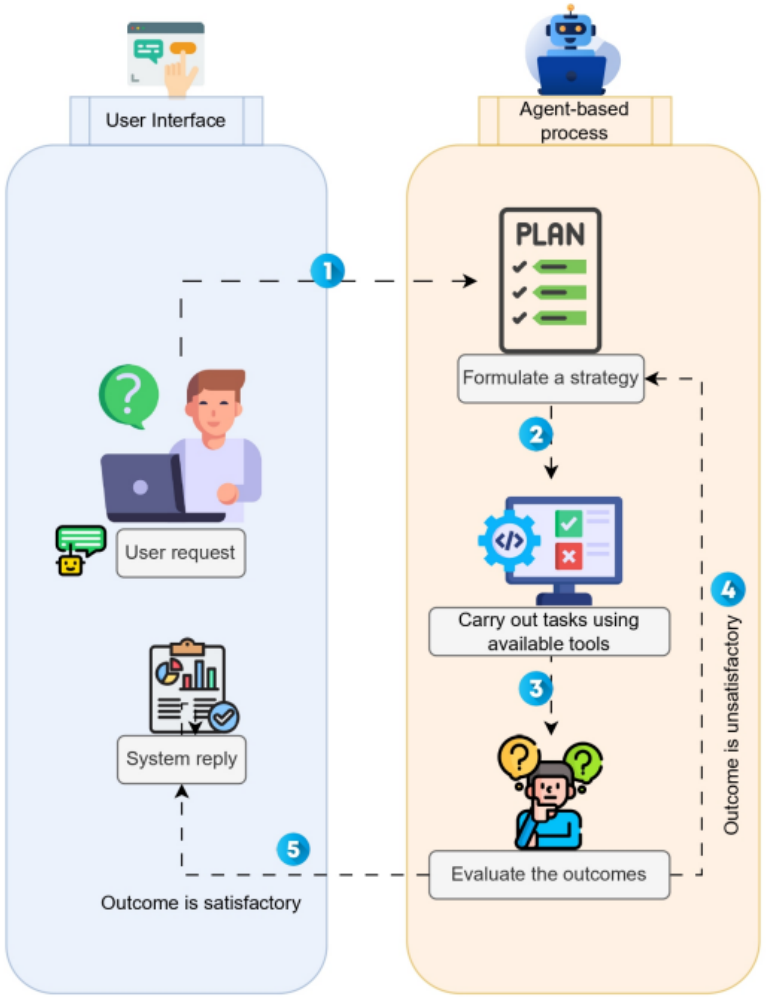


Fig. 4: What are Agentic Workflows?.

Autonomous AI agents are **systems** built on top of LLMs and CA that can **reason**, **plan**, and **act** with minimal human intervention.

They extend CA by combining language understanding, decision-making, and task execution abilities.

These AI agents can decompose users' tasks into sub-tasks, **reason** through each stage, and **interact** with external tools or environments such as APIs, IRS, and knowledge bases.

Their architecture typically includes components for **planning**, **reasoning**, **self-evaluation**, **memory**, and **knowledge** access.

Overall, they transform LLMs into active problem solvers capable of pursuing complex, multi-step objectives autonomously.

Frameworks such as LangChain, LlamaIndex, CrewAI, and Swarm enable the creation and coordination of these agents.

[2] Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678.

Unified Search and Recommendation

Unified Search and Recommendation[3] - Overview

A user submits a **natural language** request that may describe an intent such as *"new releases for me"* or *"artists like Lady Gaga."*

An LLM-based agent (router) interprets the request, analyzes the user's intent, and decides whether it aligns more with search (finding specific content) or recommendation (discovering new personalized items).

Depending on the inferred intent, the agent routes the query to the most suitable downstream service (an **IRS** or an **RS**).

In some cases, the **agent combines** both by issuing parallel tool calls, integrating search results and recommendations into a unified results page (SERP).

For example, *"latest albums by Lady Gaga"* might trigger a search for **new releases**, a recommendation for **similar artists**, and **an** explanation generated by the LLM, **all** presented cohesively to the user.

[3] Palumbo, E., Isaksson, M., Tamborrino, A., Movin, M., Dincu, C., Vardasbi, A., ... & Sultan, Z. (2025, September). You Say Search, I Say Recs: A Scalable Agentic Approach to Query Understanding and Exploratory Search at Spotify. In Proceedings of the Nineteenth ACM Conference on Recommender Systems (pp. 1117-1121).


Questions?

Putting all together

Nowadays and Beyond...

Agentic AI systems (swarms of collaborating **LLMs**) work together to **aid** users complete **complex tasks** through reasoning, planning, and action.

These agents have **access** to **specialized tools**, which can be simple (e.g., get date) or complex systems such as Information Retrieval Systems and Recommender Systems.

They dynamically **decide** which **tool** to **use**, retrieving, recommending, or generating, based on the user's intent and task context.

To overcome the limitations of static model knowledge, they employ **Retrieval-Augmented Generation frameworks** and integrated **Knowledge Bases** for up-to-date, factually accurate information.

This combination enables **adaptive, informed, and personalized assistance**, where agents continuously learn and collaborate to deliver richer, goal-oriented user experiences.

Interaction and Reasoning Characteristics

	Search Systems	Recommender Systems	Conversational Assistants	Autonomous Agents
User goal & task structure	Answer a specific, well-defined query	Support ongoing preference discovery	Clarify open-ended information needs interactively	Pursue multi-step goals on behalf of the user
Initiative & interaction pattern	Reactive to explicit queries	Often proactive, surfacing suggestions	Turn-taking dialogue	Delegated initiative, acts independently
Representation of user state	Short-term signals (keywords, clicks)	Long-term user profiles	Context maintained across dialogue turns	Rich, persistent task context and memory
System reasoning complexity	Direct retrieval from index	Pattern-based estimation of relevance to user's interests	Language interpretation and context reasoning	Multi-step planning and tool selection

Output, Adaptivity, and Evaluation

	Search Systems	Recommender Systems	Conversational Assistants	Autonomous Agents
Input form	Explicit keyword or natural language query	Implicit user behavior (clicks, views, ratings)	Natural language dialogue turns	Task descriptions or high-level goals
Output form & granularity	Ranked documents or information items	Suggested items	Generated natural language responses	Results, plans, or actions performed via tools
Adaptivity over time	Light or optional personalization	Continuous personalization	Adapts to dialogue flow	Adjusts dynamically as plan unfolds
Evaluation & success criteria	Relevance-based metrics	Engagement and satisfaction	Coherence and helpfulness	Goal completion and task success


Questions?

When modern CA and Agents Fail

This article is more than **1 month old**

Deloitte to pay money back to Albanese government after using AI in $440,000 report

Partial refund to be issued after several errors were found in a report into a department's compliance framework

Get our breaking news email, free app or daily news podcast

This article is more than **5 months old**

US lawyer sanctioned after being caught using ChatGPT for court brief

Richard Bednar apologized after Utah appeals court discovered false citations, including one nonexistent case




AI Bachelor Students who trusted the magical wisdom of AI now wonder why everything in their projects went spectacularly wrong


Questions?

CLIMATE

As AI becomes part of everyday life, it brings a hidden climate cost

ELSEVIER

Journal of Environmental Management

Volume 392, September 2025, 126813

Research article

The Two Tales of AI: A Global assessment of the environmental impacts of artificial intelligence from a multidimensional policy perspective

Union of Concerned Scientists

The EQUATION

SIGN UP EN ESPAÑOL DONATE MENU

What Are the Environmental Impacts of Artificial Intelligence?

June 25, 2025 | 11:00 am

En español: Impactos ambientales de la inteligencia artificial

NEWS RELEASE 10-NOV-2025

'Roadmap' shows the environmental impact of AI data center boom

Peer-Reviewed Publication

CORNELL UNIVERSITY

ITHACA, N.Y. - As the everyday use of AI has exploded in recent years, so have the energy demands of the computing infrastructure that supports it. But the environmental toll of these large data centers, which suck up gigawatts of power and require vast amounts of water for cooling, has been too diffuse and difficult to quantify.

Now, Cornell researchers have used advanced data analytics – and, naturally, some AI, too – to create a state-by-state look at that environmental impact. The team found that, by 2030, the current rate of AI growth would annually put 24 to 44 million metric tons of carbon dioxide into the atmosphere, the emissions equivalent of adding 5 to 10 million cars to U.S. roadways. It would also drain 731 to 1,125 million cubic meters of water per year – equal to the annual household water usage of 6 to 10 million Americans. The cumulative effect would put the AI industry's net-zero emissions targets out of reach.


Questions?