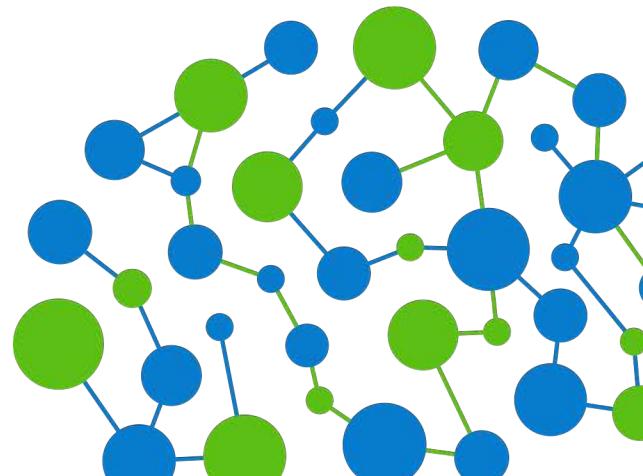


Towards Intelligent Systems Driven by Knowledge Graph and Large Language Model



Haofen Wang
CNLP 2023

- **The Rise of LLM**
- The Development of KG
- The Emerging Paradigm
- Outlook

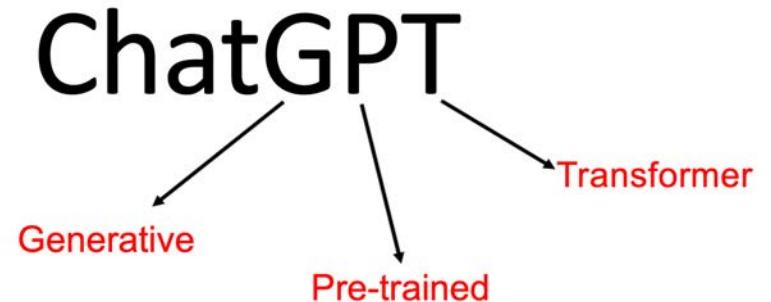


The Wave towards AGI by LLM

ChatGPT		
Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

ChatGPT is a **generative conversational pre-trained large language model** launched by OpenAI in November 2022.

It interacts through conversations, which enables it to **answer follow-up questions, acknowledge mistakes, challenge incorrect assumptions, and refuse inappropriate requests.**



The Wave towards AGI by LLM

ChatGPT

Input

Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output

Here is a list of their availability:

- Andrew: 11 am to 3 pm
- Joanne: noon to 2 pm, and 3:30 pm to 5 pm
- Hannah: noon to 12:30 pm, and 4 pm to 6 pm

Based on their availability, there is a 30-minute window where all three of them are available, which is from 4 pm to 4:30 pm. So, the meeting can be scheduled at 4 pm.

GPT-4

Input

Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output

Andrew: 11 am - 3 pm

Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm

Hannah: 12 pm - 12:30 pm, 4 pm - 6 pm

Common availability for a 30-minute meeting: 12 pm - 12:30 pm

- ChatGPT has gained attention for providing detailed and clear answers in many areas, but it has some drawbacks such as inaccurate facts and poor timeliness.
- On March 15th, 2023, GPT-4 was launched, which is more fluent, more accurate, and supports image understanding.

The Wave towards AGI by LLM

AGI (Artificial General Intelligence)

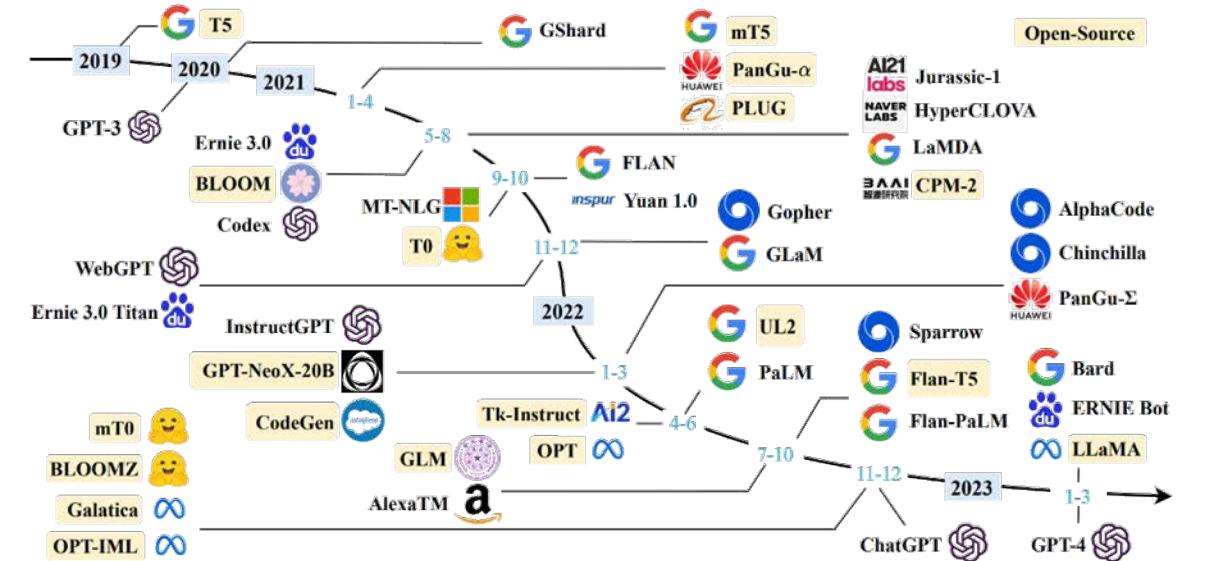
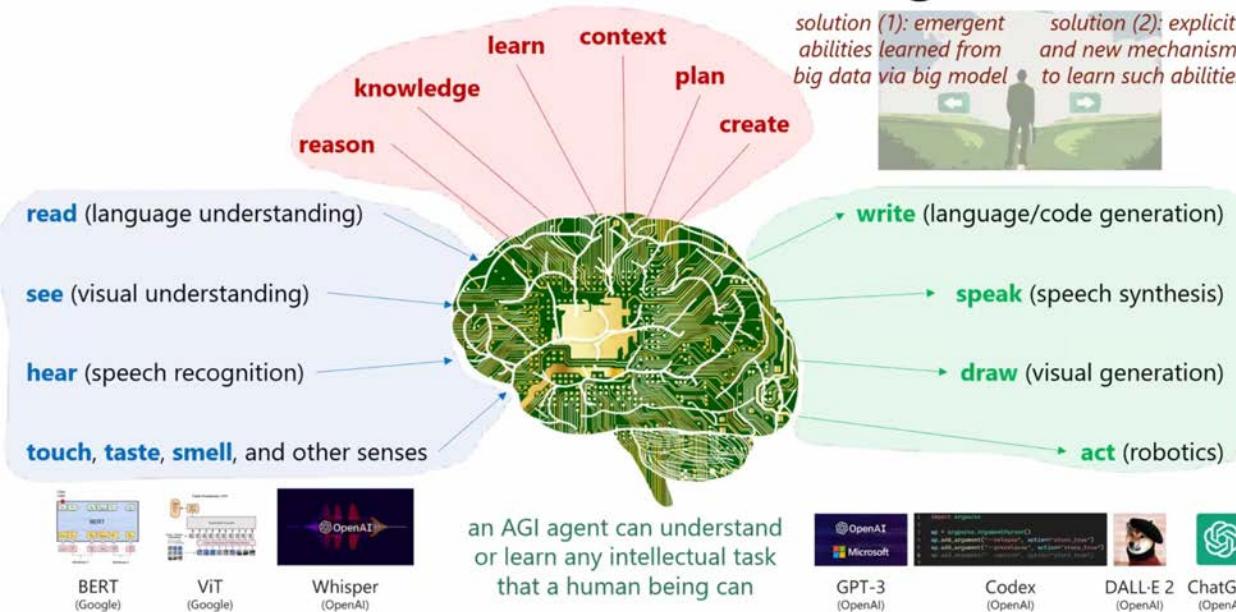


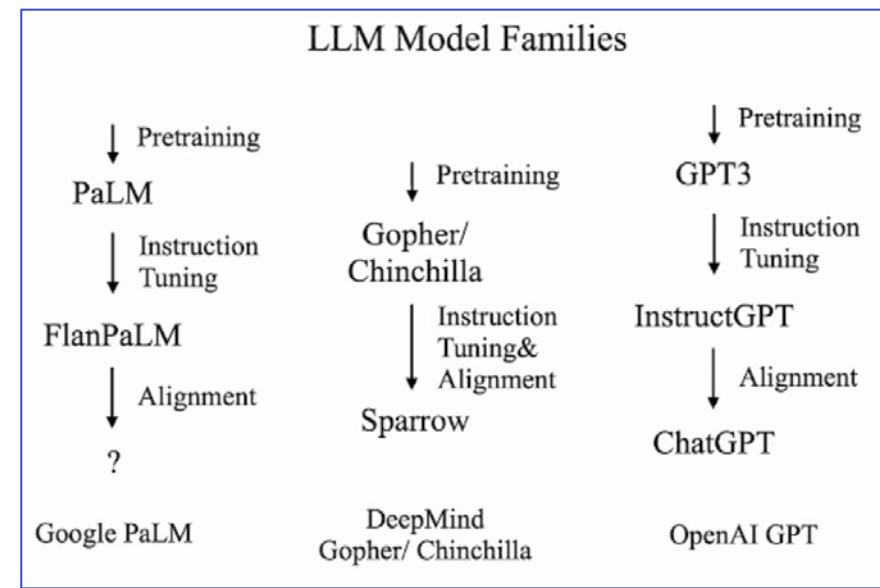
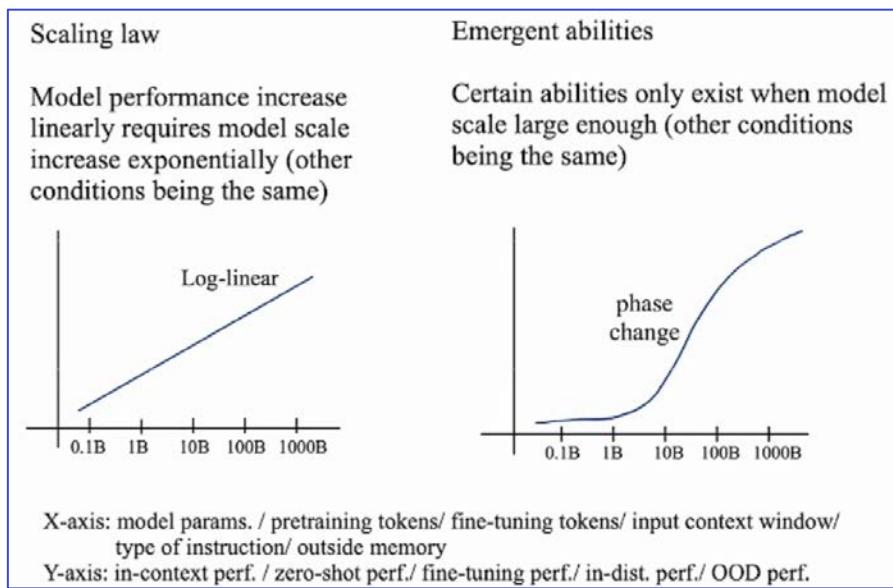
Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

Credit by Nan Duan, MSRA

A Survey of Large Language Models, 2023

The Emergent Abilities of LLM

- IT is only when the size of a model reaches a certain threshold that its capabilities such as **cross-domain transfer** and **complex reasoning** emerges.
- The evolution of large language models at Google, DeepMind, and OpenAI generally involves three stages: **pre-training**, **instruction tuning**, and **alignment**.

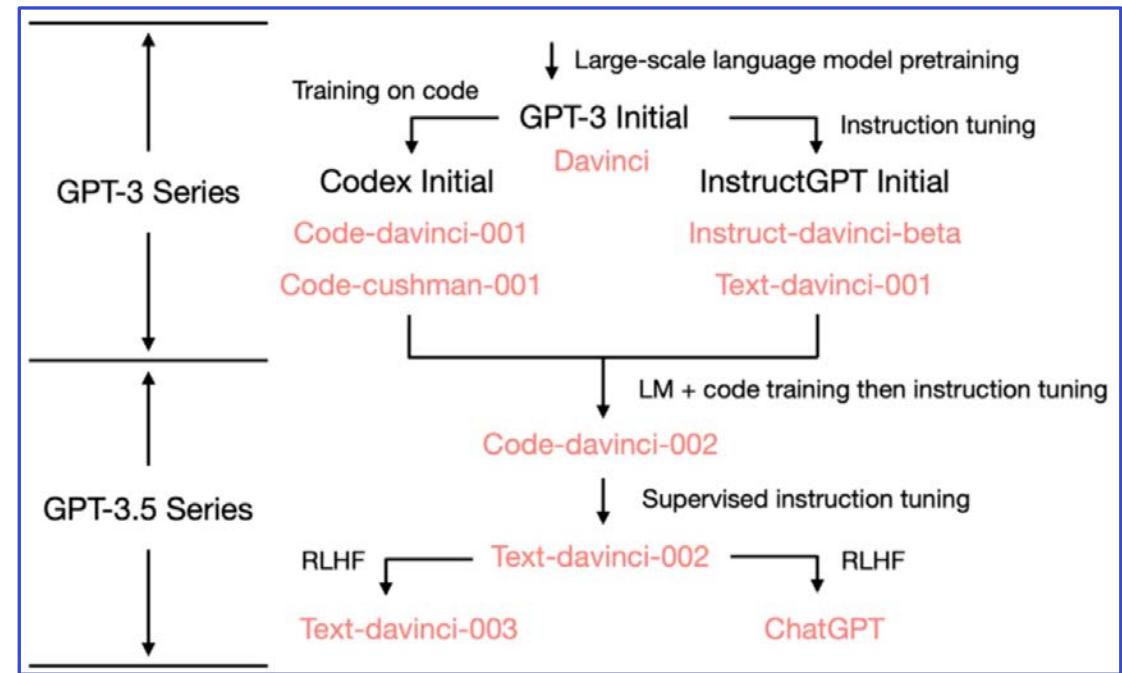


Take a Closer Look at GPT Series

The Timeline of GPT

Time	Corp	Paper/Model	Milestones
2017.6	Google	Attention is all you need	Transformer
2018.6	OpenAI	Improving Language Understanding by Generative Pre-Training	GPT-1
2019.2	OpenAI	Language Models are Unsupervised Multitask Learners	GPT-2
2020.5	OpenAI	Language Models are Few-Shot Learners	GPT-3
2022.2	OpenAI	Training language models to follow instructions with human	InstructGPT
2022.11.30	OpenAI	ChatGPT	/
2023.3.15	OpenAI	GPT-4	/

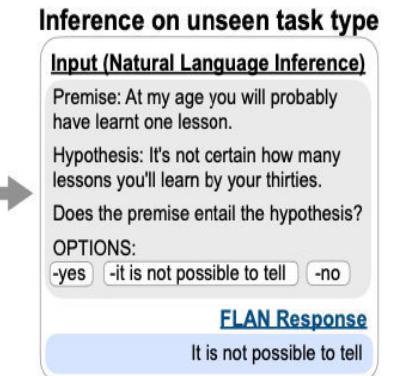
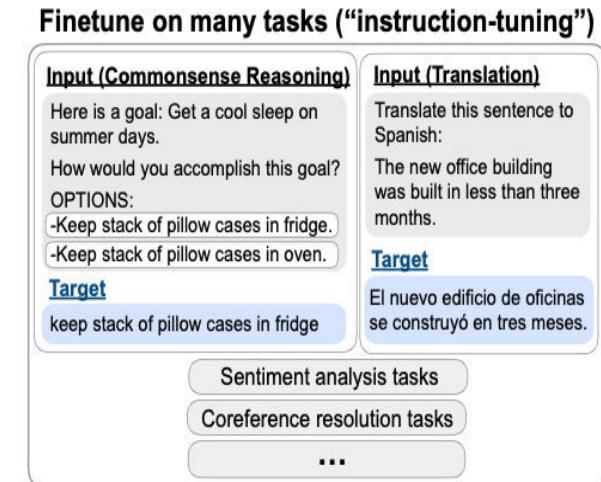
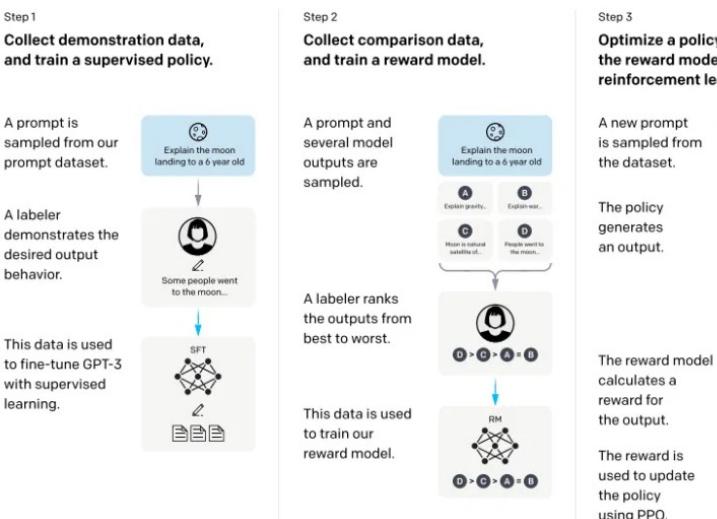
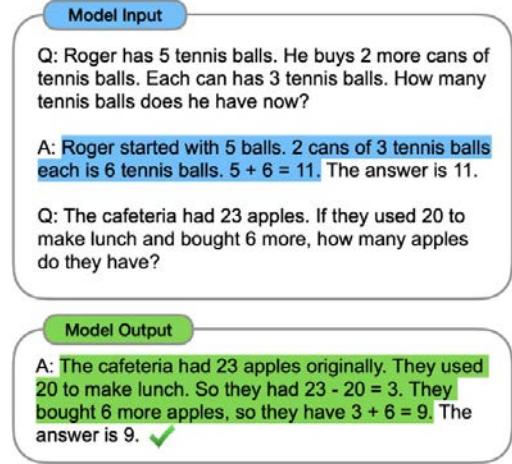
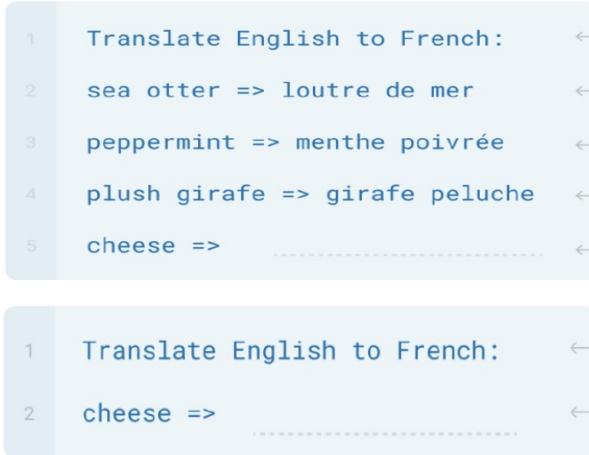
The Roadmap of GPT Tech Tree



Key Tech Development



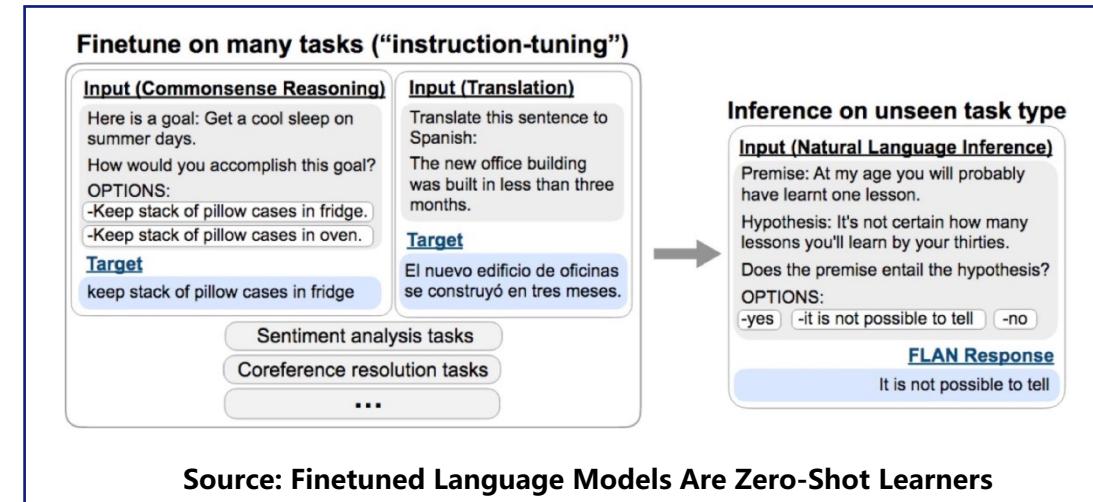
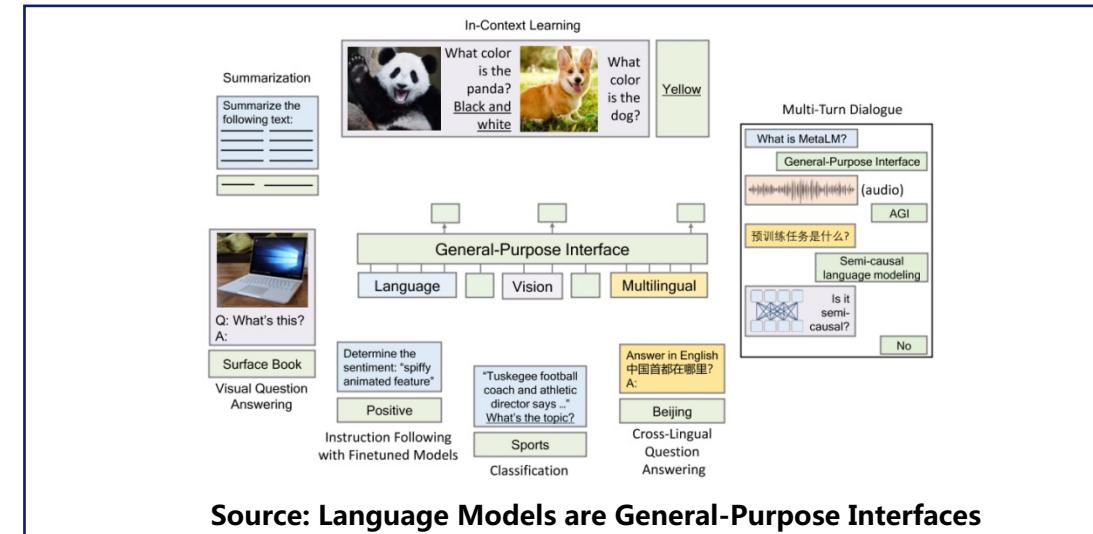
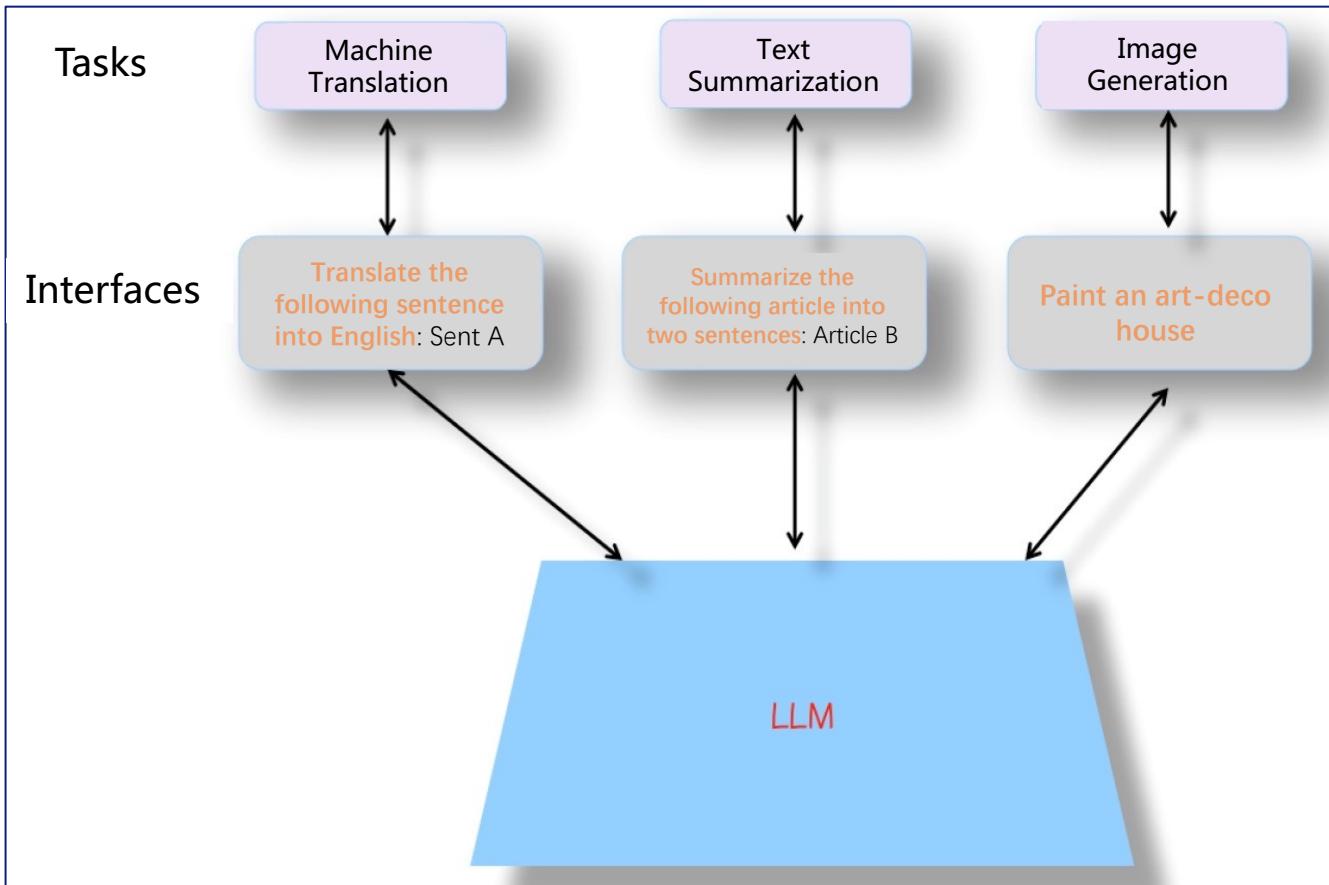
Take a Closer Look at Different Key Technologies



Instruction-tuning

LLM as a Basis for Providing Interfaces

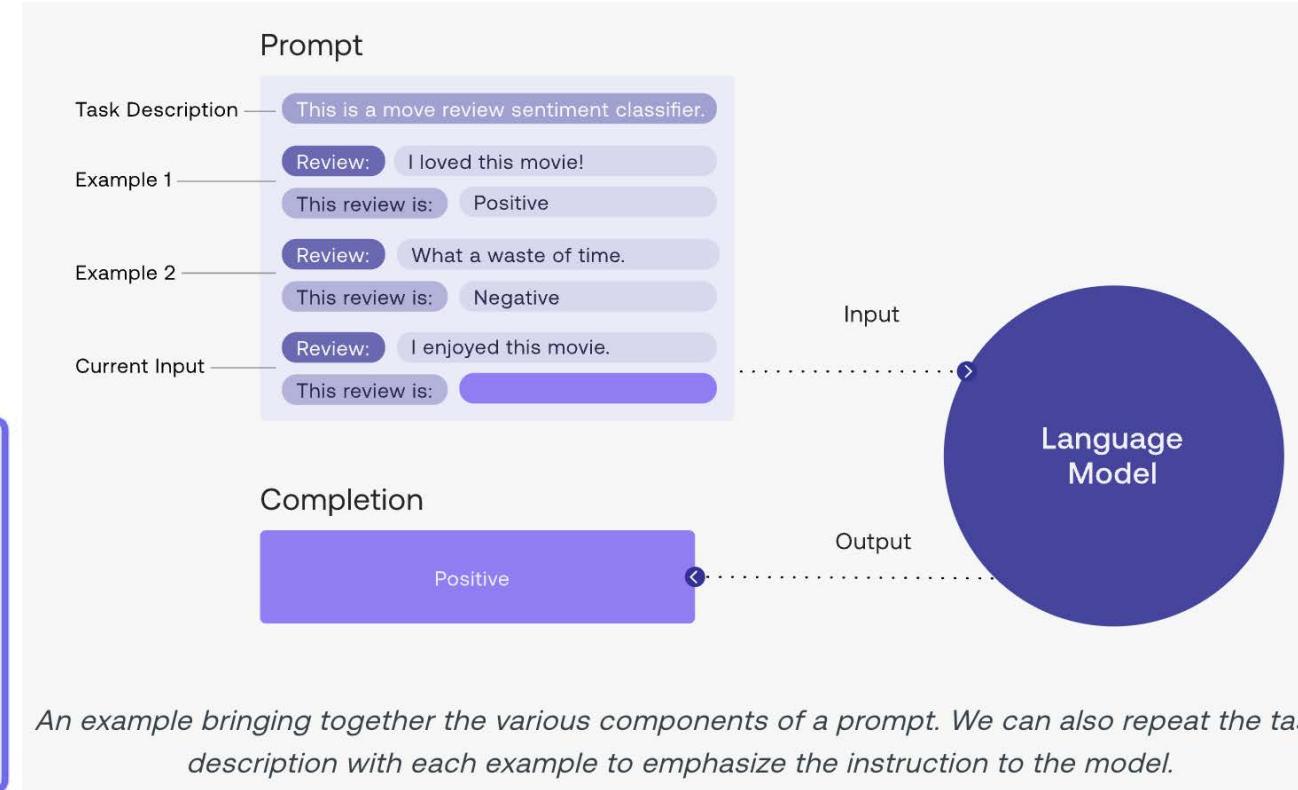
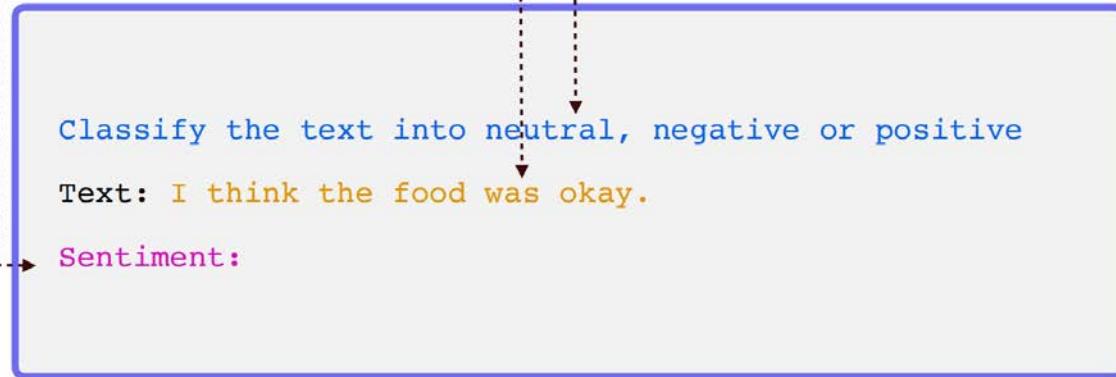
- Zero/few shot prompting
- In Context Learning
- Instruction



The Programming of LLM: Prompt Engineering

- A prompt is composed with the following components:

- Instructions
- Context
- Input data
- Output indicator



The Programming of LLM: Prompt Engineering

Standard Prompting

Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

T: The answer is :

LM

72

Few-shot Prompting

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
A: The answer is 5.

.....
Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

A: The answer is :

LM

72

Chain-of-Thought

Prompting

Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
C: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars.
A: The answer is 5.

.....
Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

LM

C: Natalia sold $48 / 2 = 24$ clips in May. Altogether, Natalia sold $48 + 24 = 72$ clips in April and May.
A: The answer is 72.

T

T

The Programming of LLM: Prompt Engineering

Zero-Shot

Chain-of-Thought

- Involves adding "Let's think step by step" to the original prompt

I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

11 apples

I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

Let's think step by step.

First, you started with 10 apples.

You gave away 2 apples to the neighbor and 2 to the repairman, so you had 6 apples left.

Then you bought 5 more apples, so now you had 11 apples.

Finally, you ate 1 apple, so you would remain with 10 apples.

Self-Consistency

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted $21 - 15 = 6$ trees. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.

...

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?
A: She bought 5 bagels for \$3 each. This means she spent 5

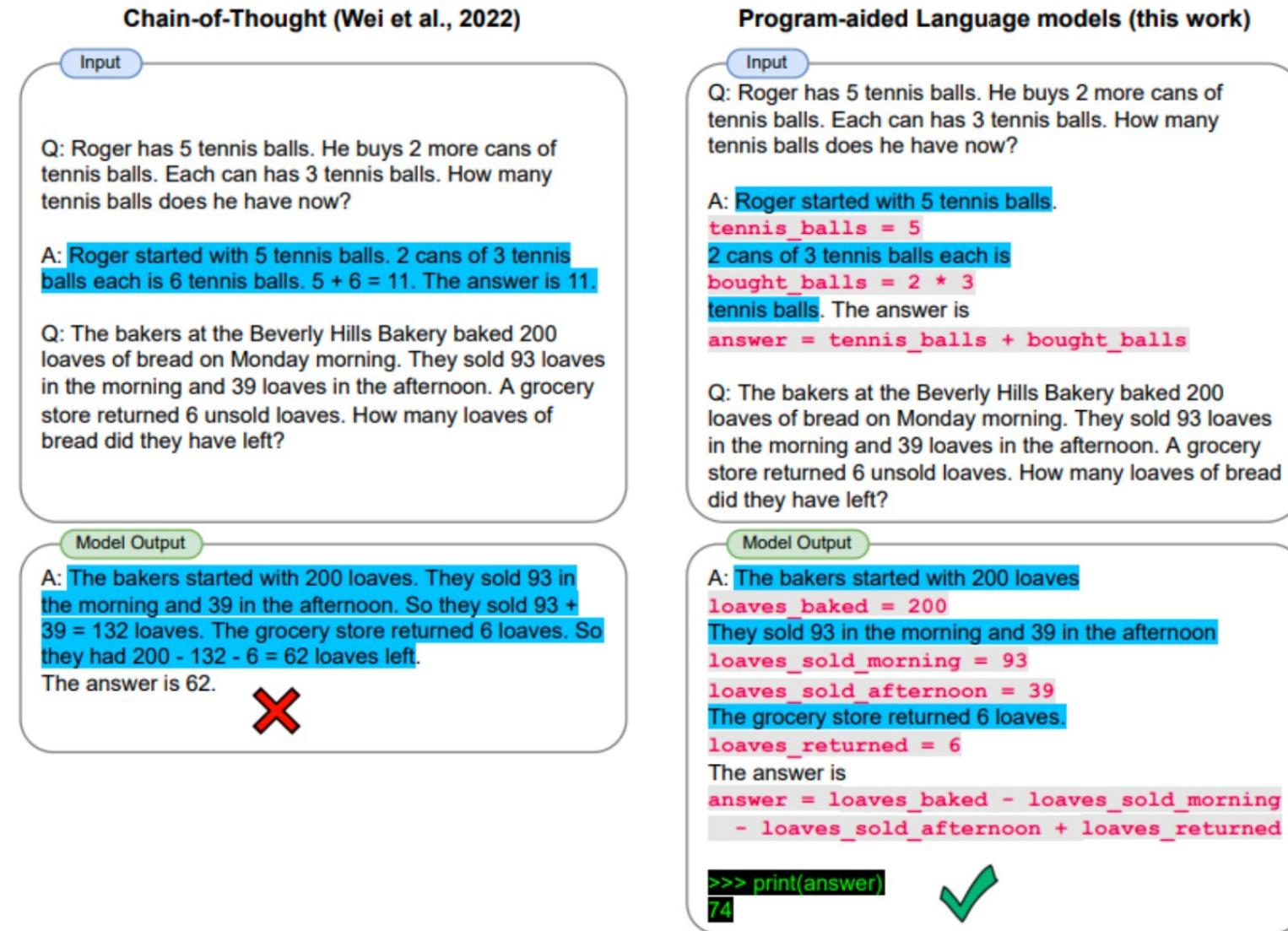
Q: When I was 6 my sister was half my age. Now I'm 70 how old is my sister?
A:

Output 1
When I was 6 my sister was half my age, so she was 3. Now I am 70, so she is $70 - 3 = 67$. The answer is 67.

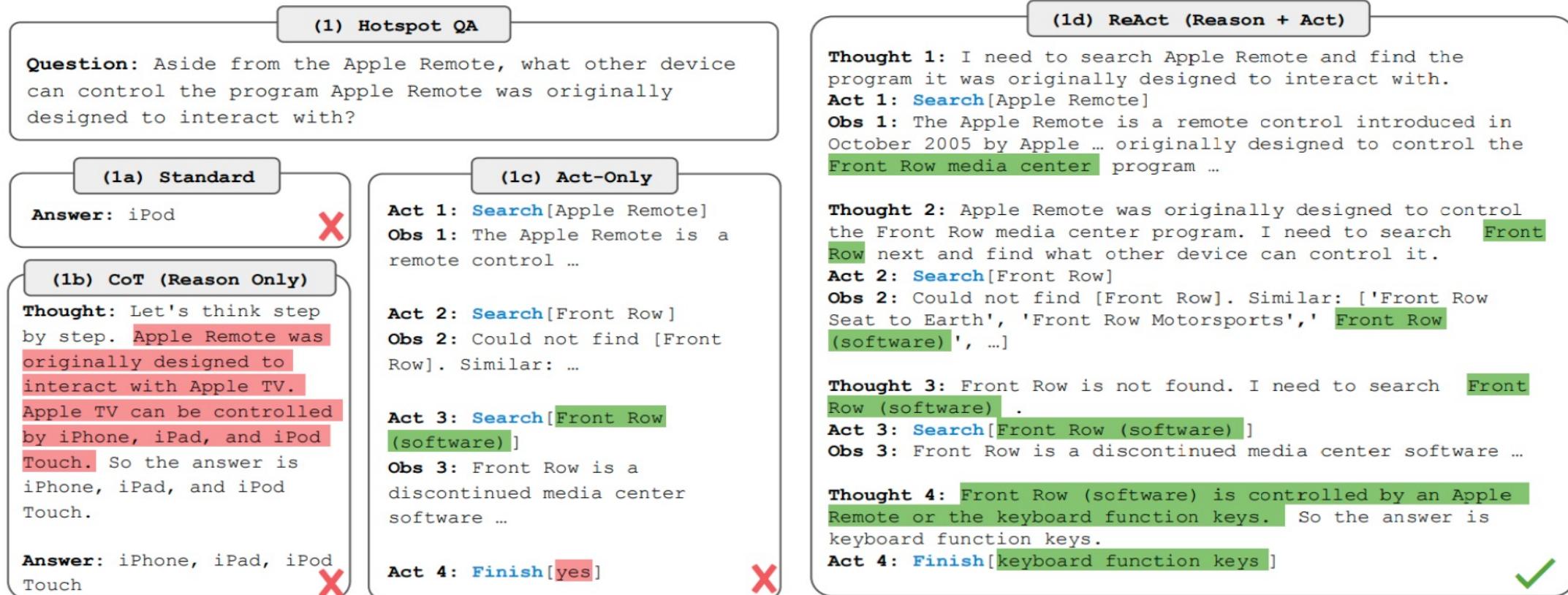
Output 2
When the narrator was 6, his sister was half his age, which is 3. Now that the narrator is 70, his sister would be $70 - 3 = 67$ years old. The answer is 67.

Output 3
When the narrator was 6, his sister was half his age, which is 3. Now that the narrator is 70, his sister would be $70 - 3 = 67$ years old. The answer is 67.

The Programming of LLM: Prompt Engineering

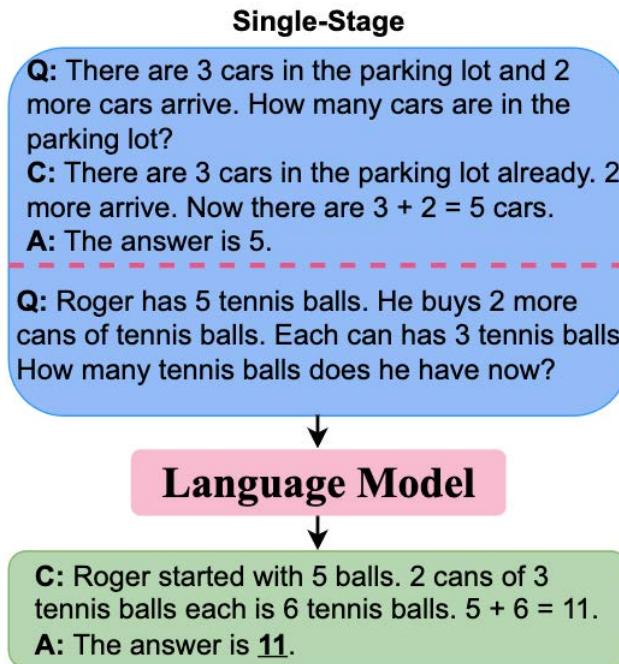


The Programming of LLM: Prompt Engineering



Source: [ReAct: Synergizing Reasoning and Acting in Language Models](#)

The Programming of LLM: Prompt Reasoning



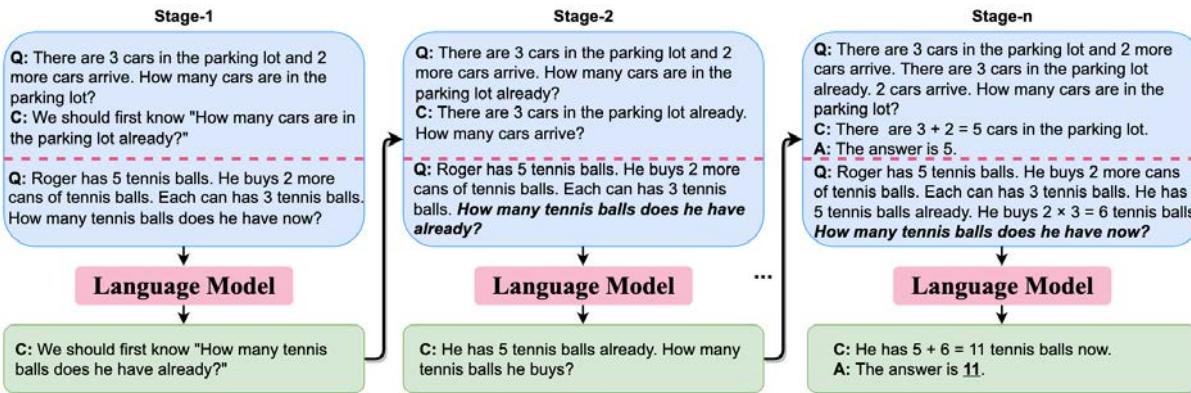
- Sensitivity of in-context learning
- complexity, diversity, explicity

Figure 2: Validation set performance. X-axis means reasoning steps and y-axis means accuracy. More reasoning steps in prompts overall achieve higher accuracy when prompts are in-distribution (left), noisily labeled (middle), and out of distribution (right).

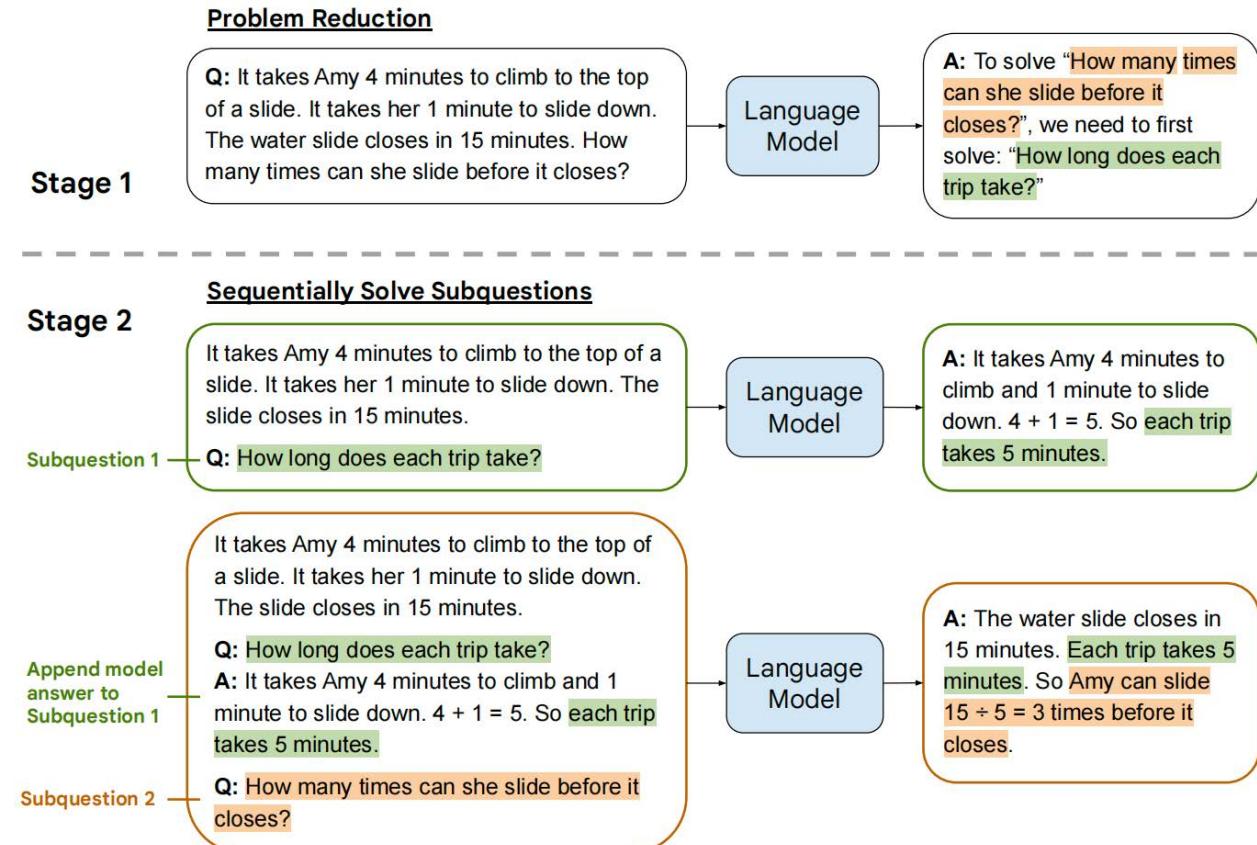
A simple and direct improvement on CoT methods. Existing work has shown that in-context learning is very sensitive to the order and quality of the examples contained in it. A very small change may cause a large decline in the effect of the model. To further optimize CoT, an intuitive method is to make its examples more complex, diverse or discriminative.

The Programming of LLM: Prompt Reasoning

Multi-stage

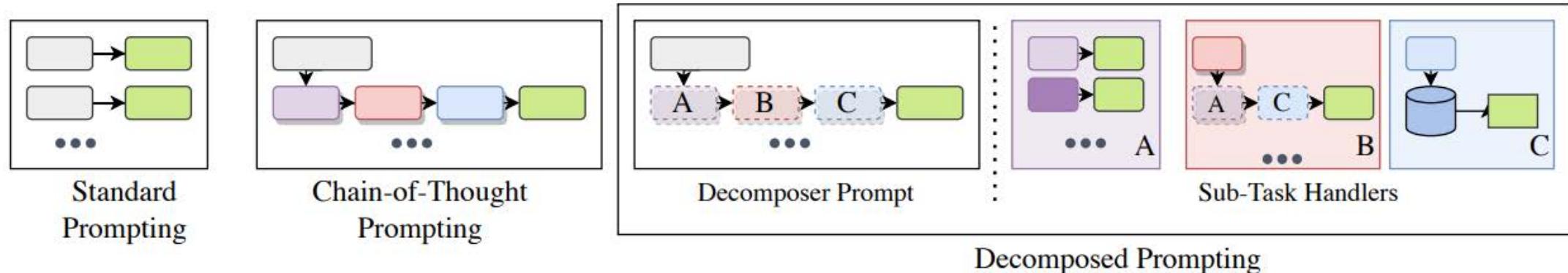
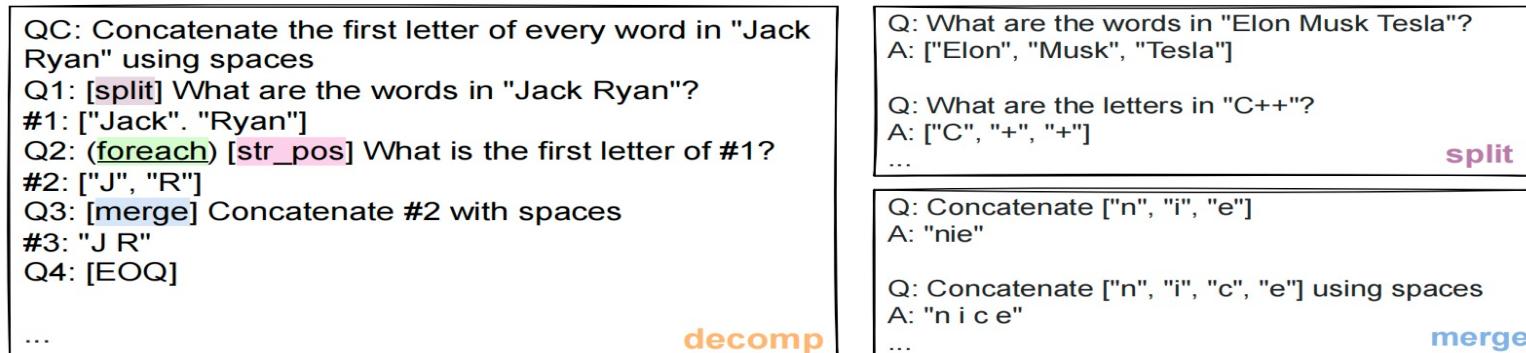


Just like human-like reasoning, we always need to iterative reason given a complex problem. Inspired by it, a multi-stage method reduces a problem into sub-questions and sequentially solve them.



The Programming of LLM: Prompt Reasoning

Decomposed Prompting



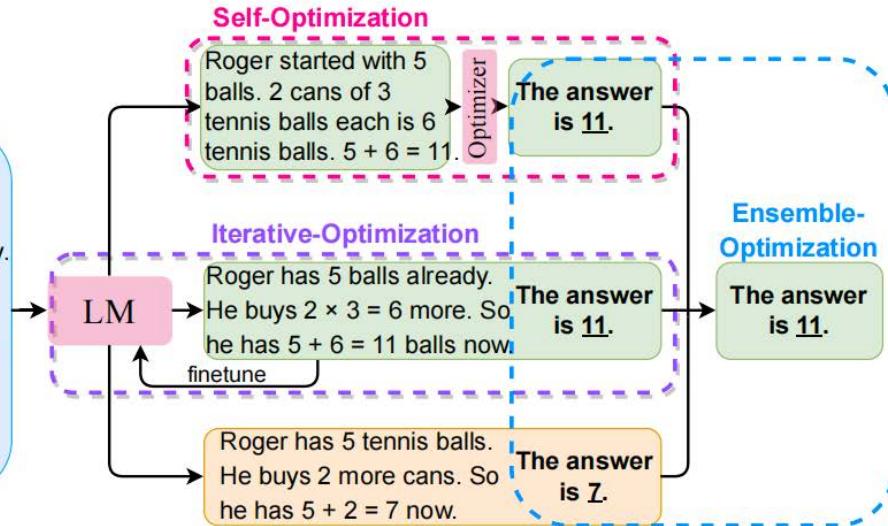
For different types of sub-questions, design specific prompts so that each type can be solved more precisely

The Programming of LLM: Prompt Reasoning

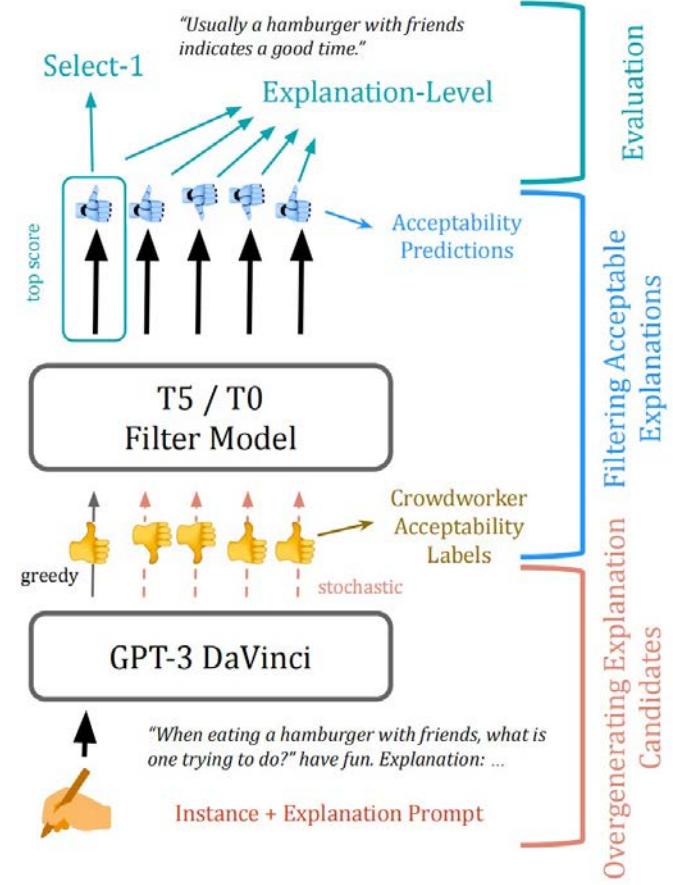
Q: There are 3 cars in the parking lot and 2 more cars arrive. How many cars are in the parking lot?
C: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars.
A: The answer is 5.

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?

Self-optimization

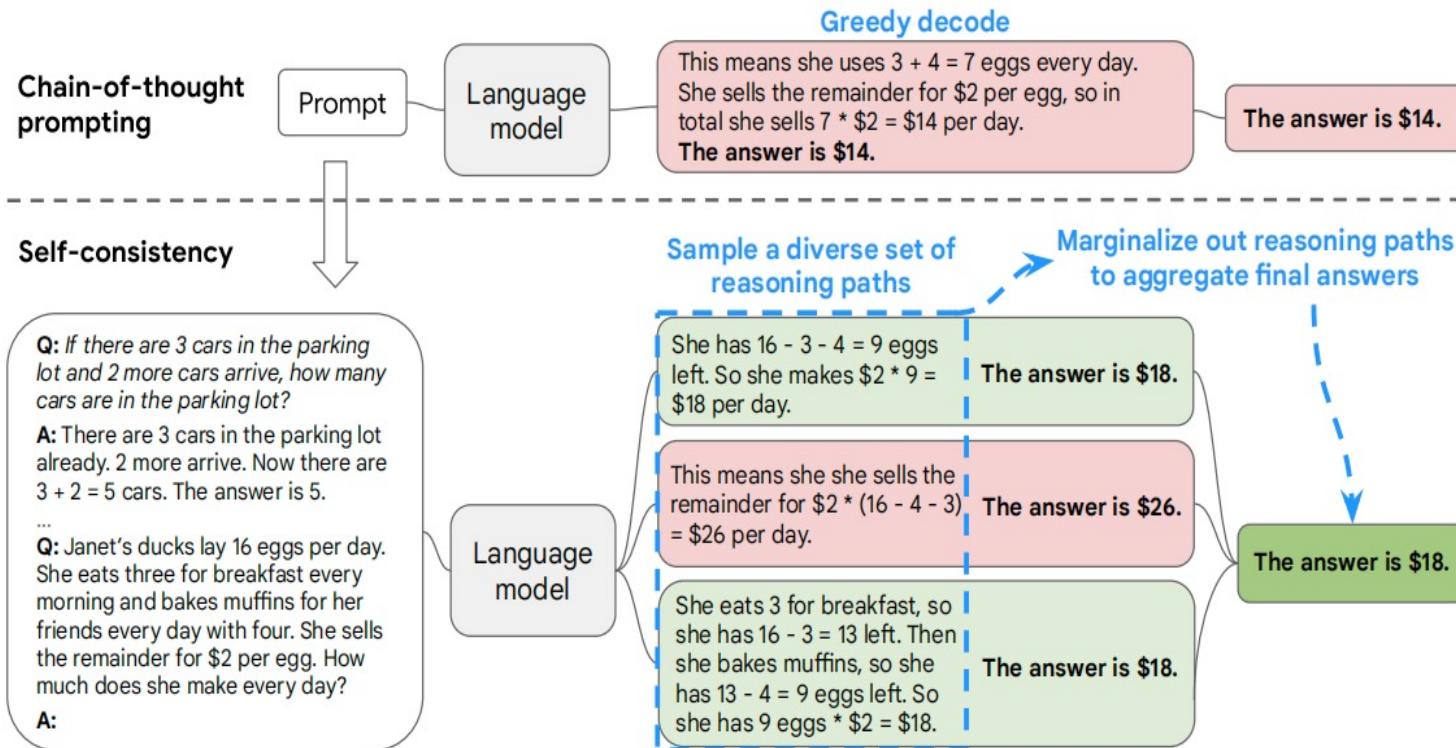


Both single-stage and multi-stage methods focus on the input optimization during prompting, we can also optimize the output of LLM. Since reasoning is the key of CoT prompting, self-optimization involves refinement or filter modules to optimize the reasoning paths during post-processing.



The Programming of LLM: Prompt Reasoning

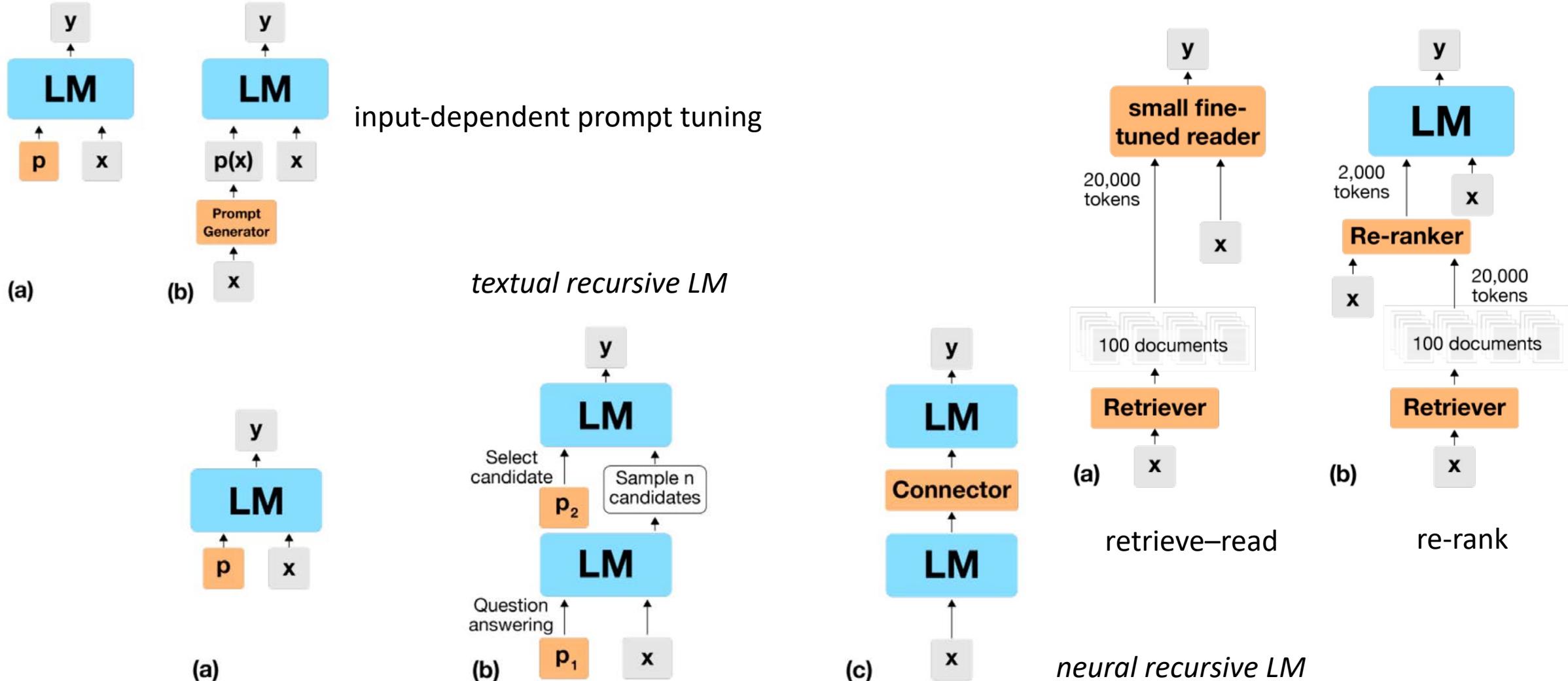
Ensemble-Optimization



All roads lead to Rome, reasoning problems often have more than one path to the correct answer, and there are often multiple solutions to a problem, thus resulting in an ensemble optimization method, represented by Google's self-consistency. It uses sampling methods commonly used in the generative decoding process, such as temperature, top-k, etc., to allow the language model to generate multiple paths, and summarize the answers of all paths to generate the final answer

Self-Consistency Improves Chain of Thought Reasoning in Language Models 2022

The Programming of LLM: Design Patterns of Prompts



The Ecosystem of LLM

ChatGPT Plugins



Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



FiscalNote

Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.



Instacart

Order from your favorite local grocery stores.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



Klarna Shopping

Search and compare prices from thousands of online shops.



Milo Family AI

Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?



OpenTable

Provides restaurant recommendations, with a direct link to book.



Shop

Search for millions of products from the world's greatest brands.



Speak

Learn how to say anything in another language with Speak, your AI-powered language tutor.



Wolfram

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.

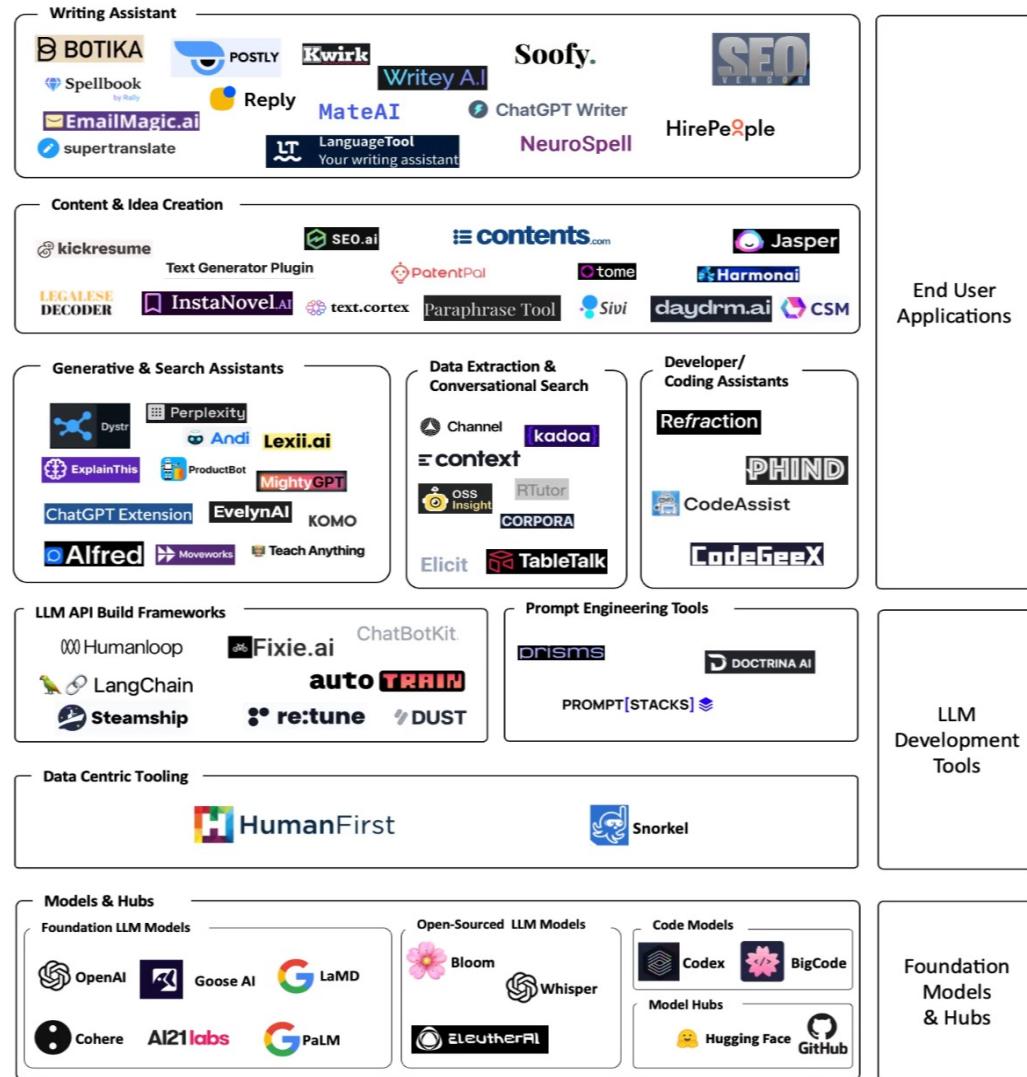


Zapier

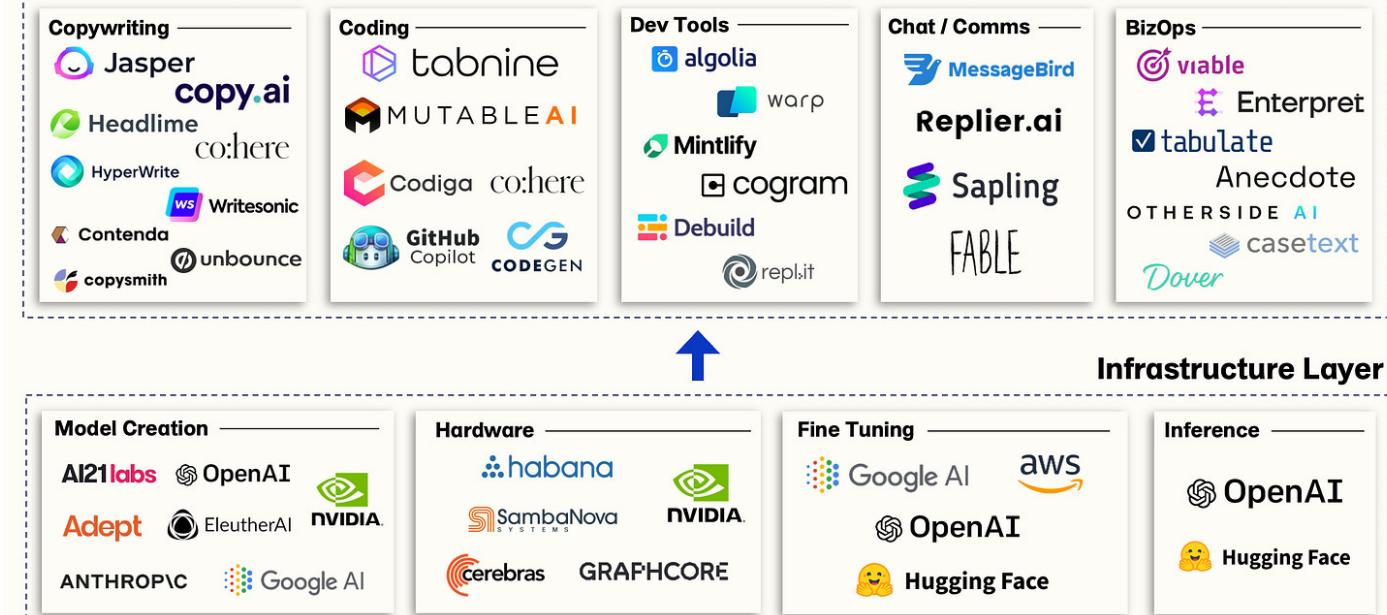
Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

The Ecosystem of LLM

Foundation Large Language Model Stack



Large Language Models

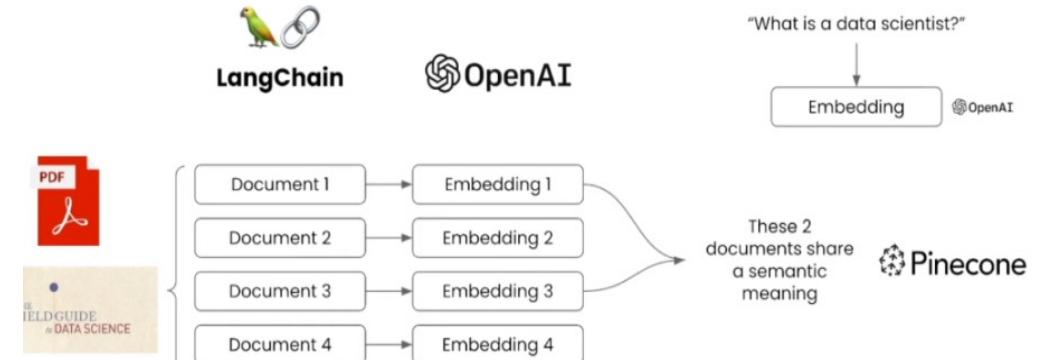


The Ecosystem of LLM: Open-source Players

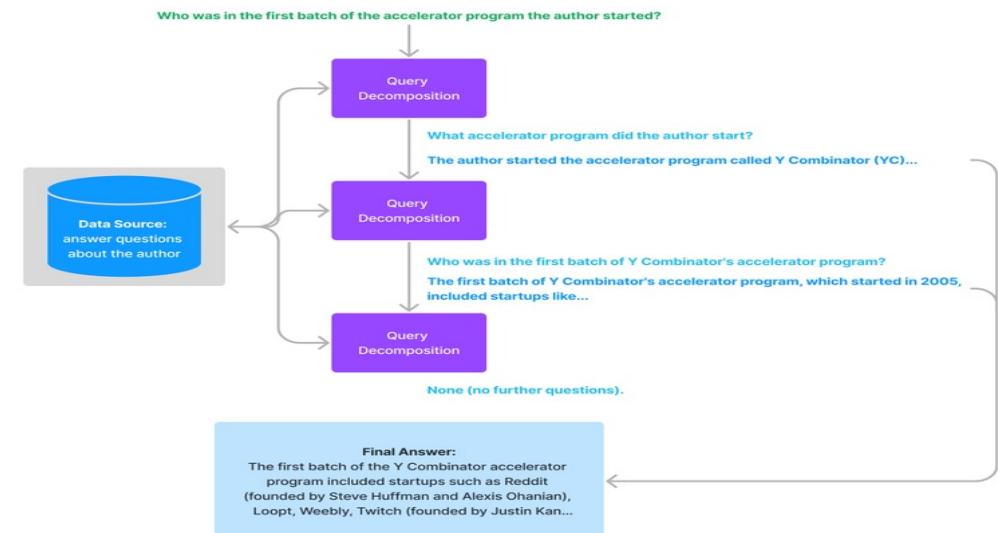
Chat language models tracker

Model	HuggingFace	License	Training Data
Foundation Models (Recent)			
Cerebras-GPT	13B	Apache 2.0	
LLaMa			
LLaMa	7-65B	Restricted	CCNet, C4, GitHub, Wikipedia, Books, ArXiv
Chat Models			
OpenChatKit	NeoX-20B	Apache 2.0	LAION OIG (Apache 2.0) Alpaca, CodeAlpaca, Guanaco, GPT4All (OIG, GP3 samples and more)
ChatRWKV	RWKV-4-Raven-v7-Eng	Apache 2.0	OpenAI GPT3 davinci-003 (52K samples)
Alpaca (LLaMa)	Weights no released	Restricted	OpenAI GPT4 (>20k samples)
Alpaca 13B	13B (tweet)	Restricted	Alpaca, GPT4All, CodeAlpaca, OIG
Alpaca w GPT4 data	13B	Restricted	Alpaca (GPT3), ChatGPT, Quora, StackOverflow and MedQuAD
Baize (LLaMa)	7-30B	Restricted	ShareGPT/ChatGPT (70K samples)
Vicuna (LLaMa)	13B (LLaMa delta), 13B model	Restricted	ChatGPT, LAION OIG, Alpaca (GPT3) and more
Koala (LLaMa)	Weights no released	Restricted	Alpaca (GPT3)
BLOOM-Alpaca	7B (tweet)	RAIL	Flan + Alpaca/GPT4All/ShareGPT/Vicuna
Flan-Alpaca	220M-11B	Apache 2.0	Alpaca (GPT3)
Dolly (GPT-J-Alpaca)	6B	Apache 2.0	GPT4All (OIG, GP3 samples and more)
GPT4All (LLaMa)	7B	Restricted	GPT4All (OIG, GP3 samples and more)
GPT4All (GPT-J)	Coming soon	Open	InstructionWild
ColossalChat (LLaMa)		Restricted	

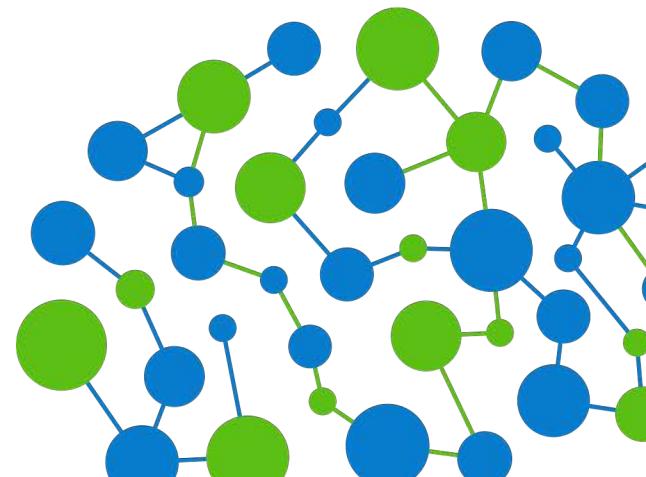
<https://www.chatorg.ai/blog/chat-language-models-tracker>



Welcome to LlamaIndex 🐾 (GPT Index)!

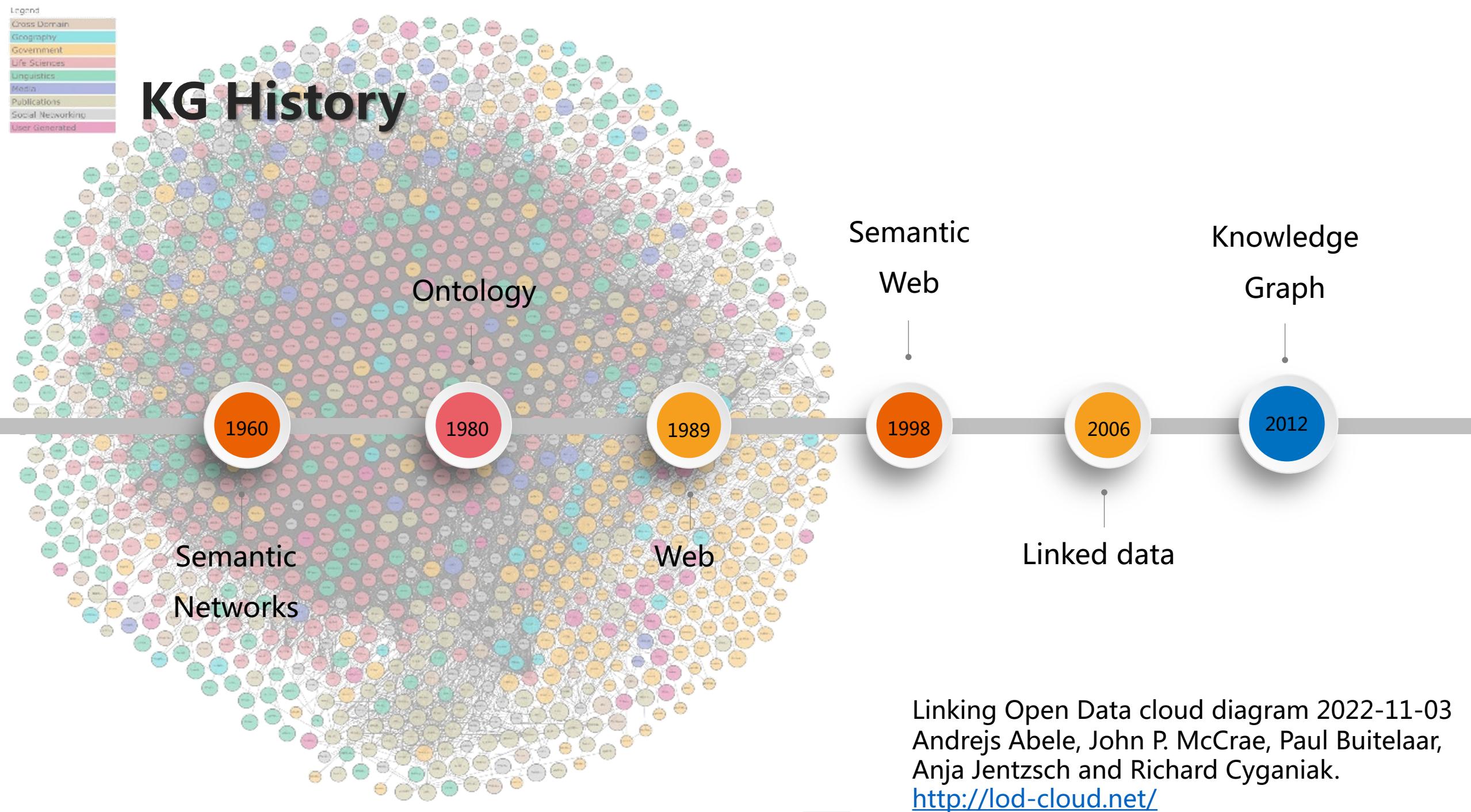


- The Rise of LLM
- **The Development of KG**
- The Emerging Paradigm
- Outlook



Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated

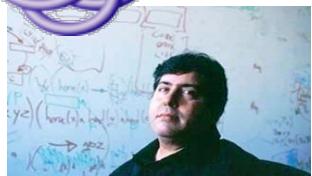
KG History



What is Knowledge Graph (KG) – Popular KBs and Characteristics



Cyc



WordNet



guitarist $\subset \{\text{player}, \text{musician}\}$
 $\subset \text{artist}$

algebraist

$\subset \text{mathematician}$
 $\subset \text{scientist}$

$\forall x: \text{human}(x) \Rightarrow$
 $(\exists y: \text{mother}(x,y) \wedge$
 $\exists z: \text{father}(x,z))$

$\forall x,u,w: (\text{mother}(x,u) \wedge$
 $\text{mother}(x,w)$
 $\Rightarrow u=w)$

1985

1990

2000

2005

2010

2015

2023

Key Features

In the early stage, KG is High-quality, manually-built, and for human consumption; in the middle age, KG is constructed by algorithms and used to enhance the understanding capability of machines; nowadays KG is evolving towards multi-modality and subsymbolic representations

By Human
For Human

Wikipedia



4.5 Mio. English articles
20 Mio. contributors

WolframAlpha™ computational knowledge engine

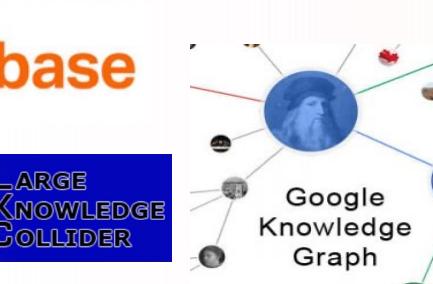
By Algorithm
For Machine

DBpedia

yAGO
select knowledge

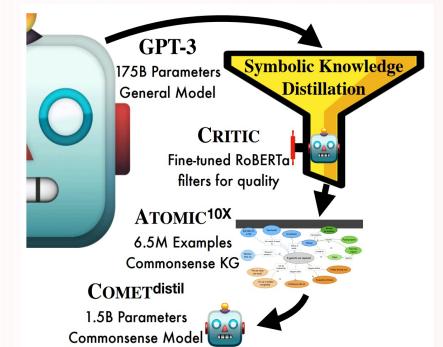
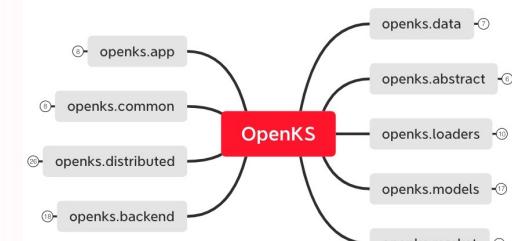
Freebase

LARKC
LARGE KNOWLEDGE COLLIDER



OpenKG.CN
链上的开放知识图谱

Human Machine
Collaboration



Knowledge Graph

Knowledge Graph (KG) is an explicit representation of human knowledge, which is stored in the form of graph and used for reasoning and computing.

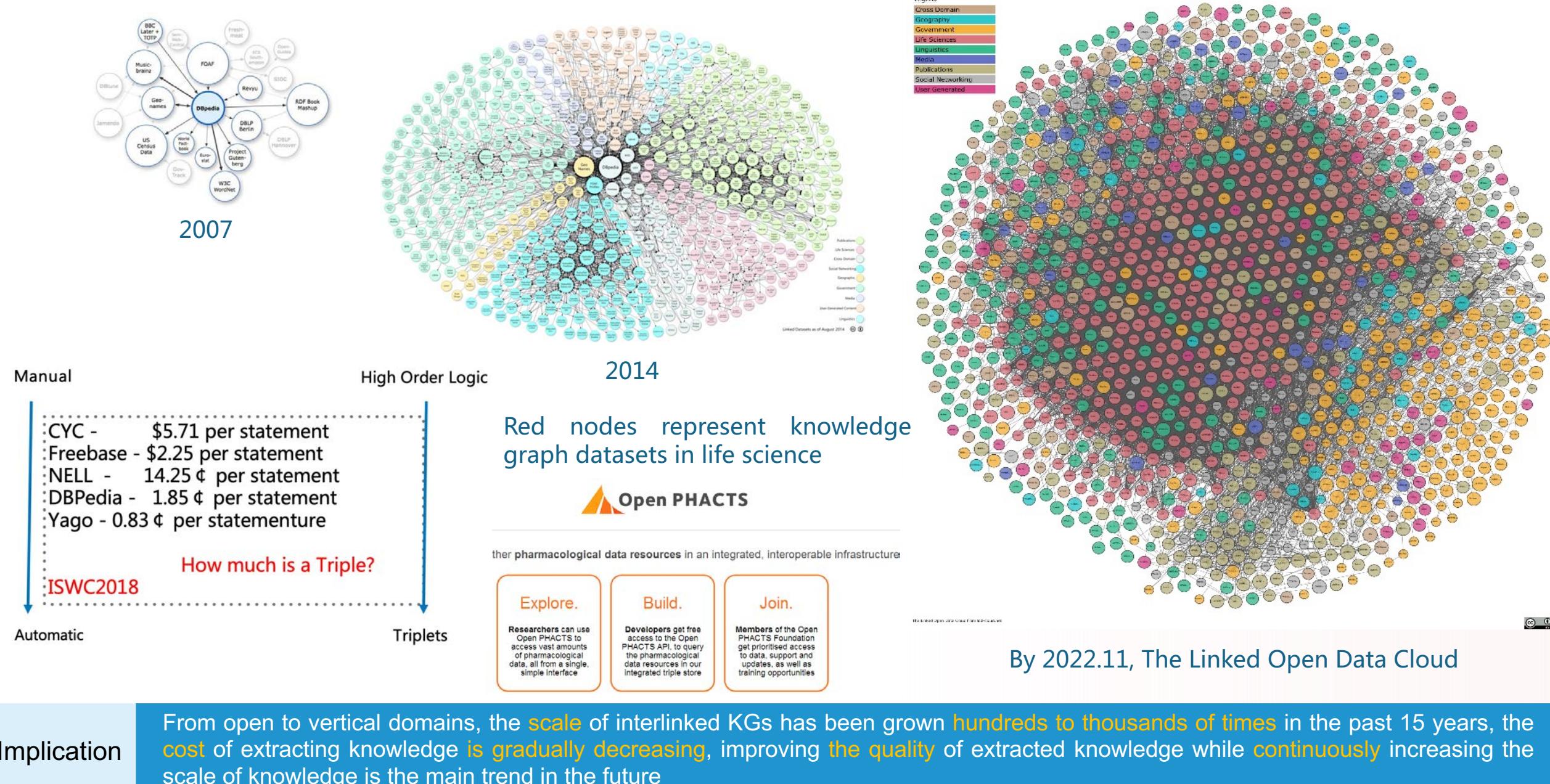


- General domain oriented
- Commonsense knowledge
- Structured encyclopedia knowledge
- Emphasize the breadth of knowledge
- For general users

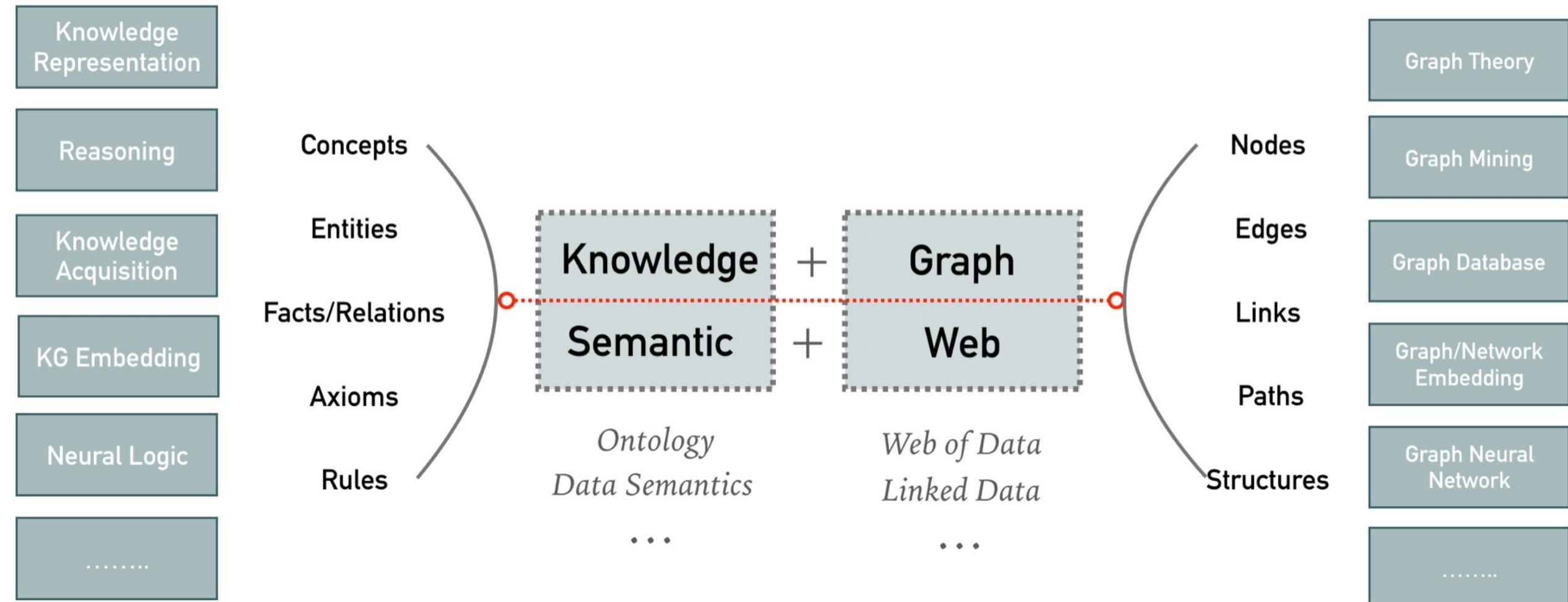


- Industrial domain oriented
- Industrial data
- Semantic industrial knowledge base
- Emphasize the depth of knowledge
- For industry users

What is Knowledge Graph (KG) – Rapid Growth and Lower Cost



Knowledge Graph is more expressive than *pure Graph* but less complex than *formal logic*.



What is Knowledge Graph (KG) – Perspective and Implication

KG as a World Model

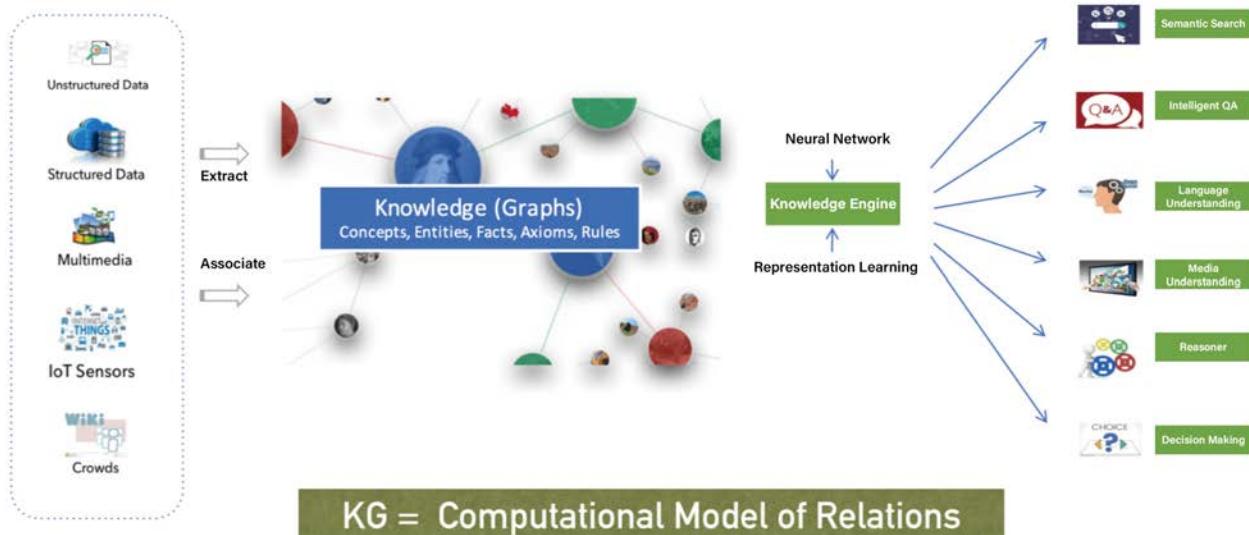


The Good Old Fashioned AI

The Semantic Web & Linked Knowledge

The Knowledge Graph

Graph Structure as Knowledge Base



KG = Computational Model of Relations

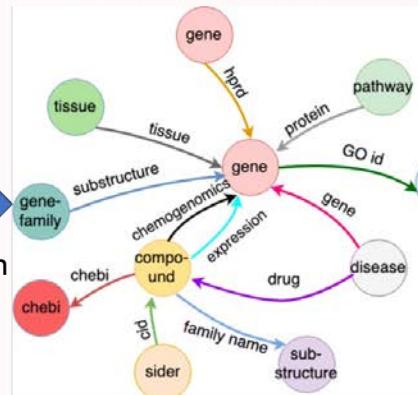
Knowledge Graph originates from how machines represent knowledge, use the graph structure to describe the relationship between things, developed in the rise of Web technologies, and landed in application fields such as search engine, intelligent QA, and recommender systems.

Text as Knowledge Base



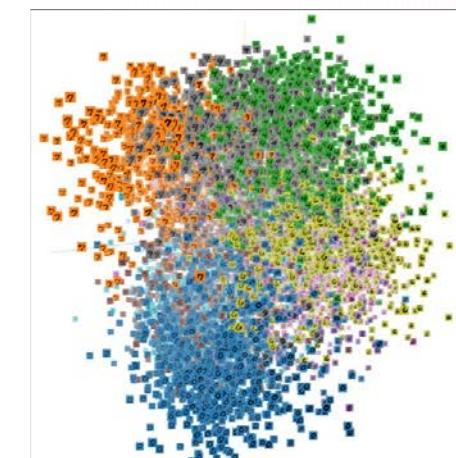
Text2Knowledge

Language Pretrain
BERT, GPT.....

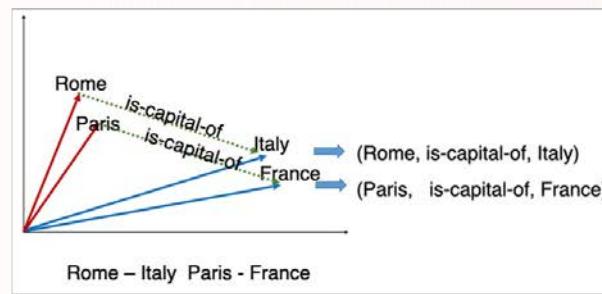


Embeddings as Knowledge Base

Embeddings : Distributed Vector Representation

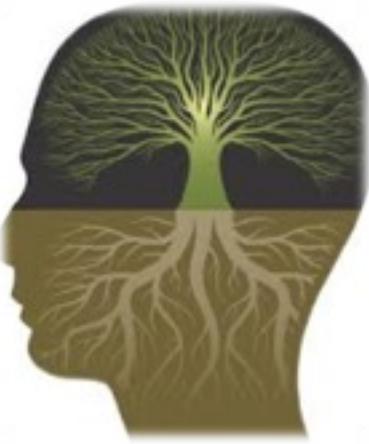


- Text : Learn a vector of each word in a sentence
- KG: Learn a vector for each entity or property
- Image/Video : Learn a vector for each visual object



Implication

Smart AI vs. Knowledgeable AI

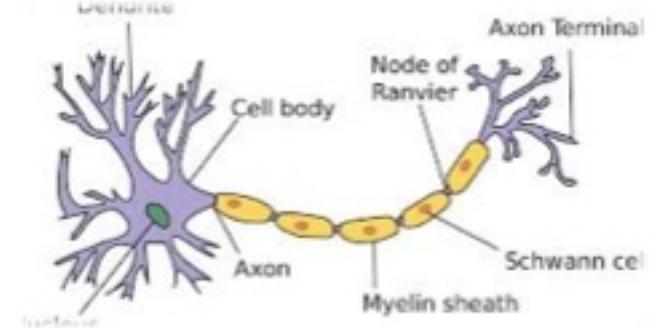


Smart
AI

perception
recognition
judgment



Deep Learning



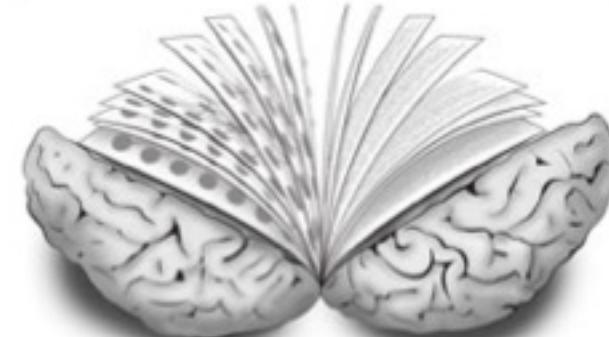
Human brain can conduct reasoning and understanding
based on acquired knowledge

Knowledgeable
AI

thinking
language
reasoning

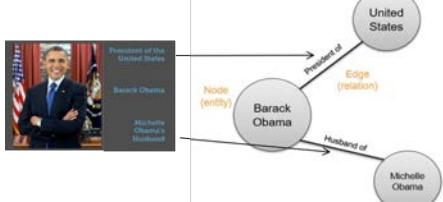


Knowledge Graph



SOTA and Trend of KG – Knowledge Representation and Reasoning

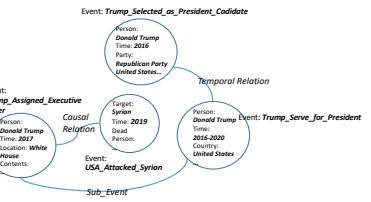
Entity Centric KG



(Barack Obama, Spouse, Michelle Obama)

Triples in form of (s, p, o)

Event Centric KG



Event Frames

Event Relations

Causal relation, Temporal relation, Co-reference relation, Sub-class relation...

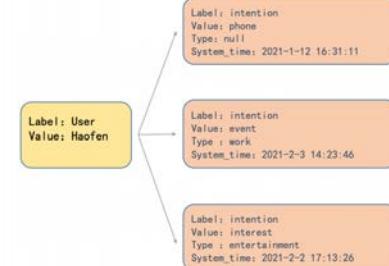
Prompt KG

```

1 Translate English to French:
sea otter => loutre de mer
2
3 peppermint => menthe poivrée
4 plush girafe => girafe peluche
5 cheese =>
    
```

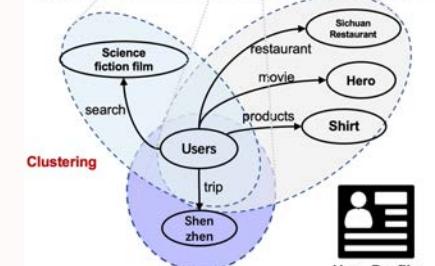
LLM as knowledge base and use prompts to elicit the corresponding facts

Personal KG



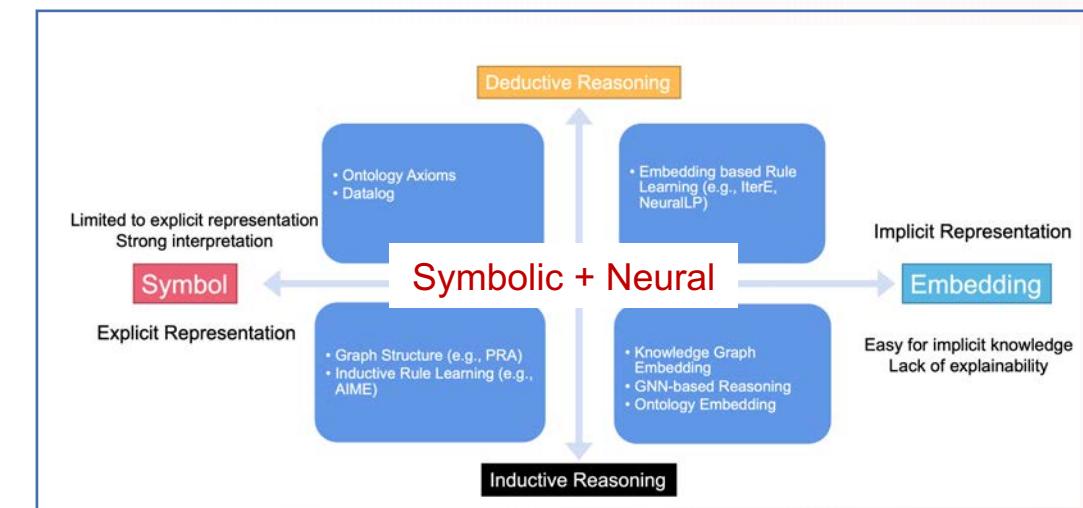
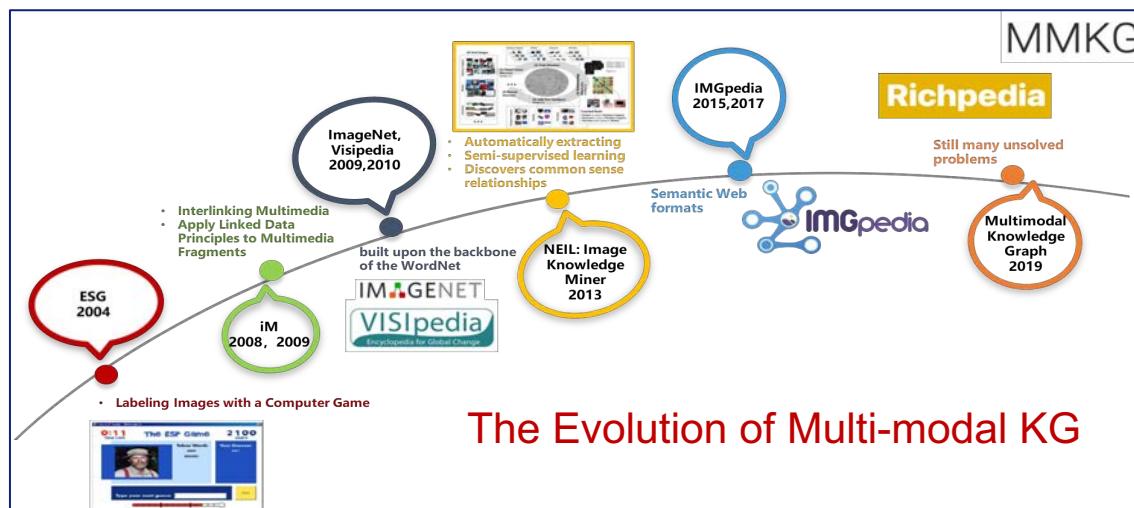
Graph composing events, interests, behaviors of an individual under the protection of privacy and security

Spatiotemporal KG



Expand knowledge representation level in time series and space dimensions

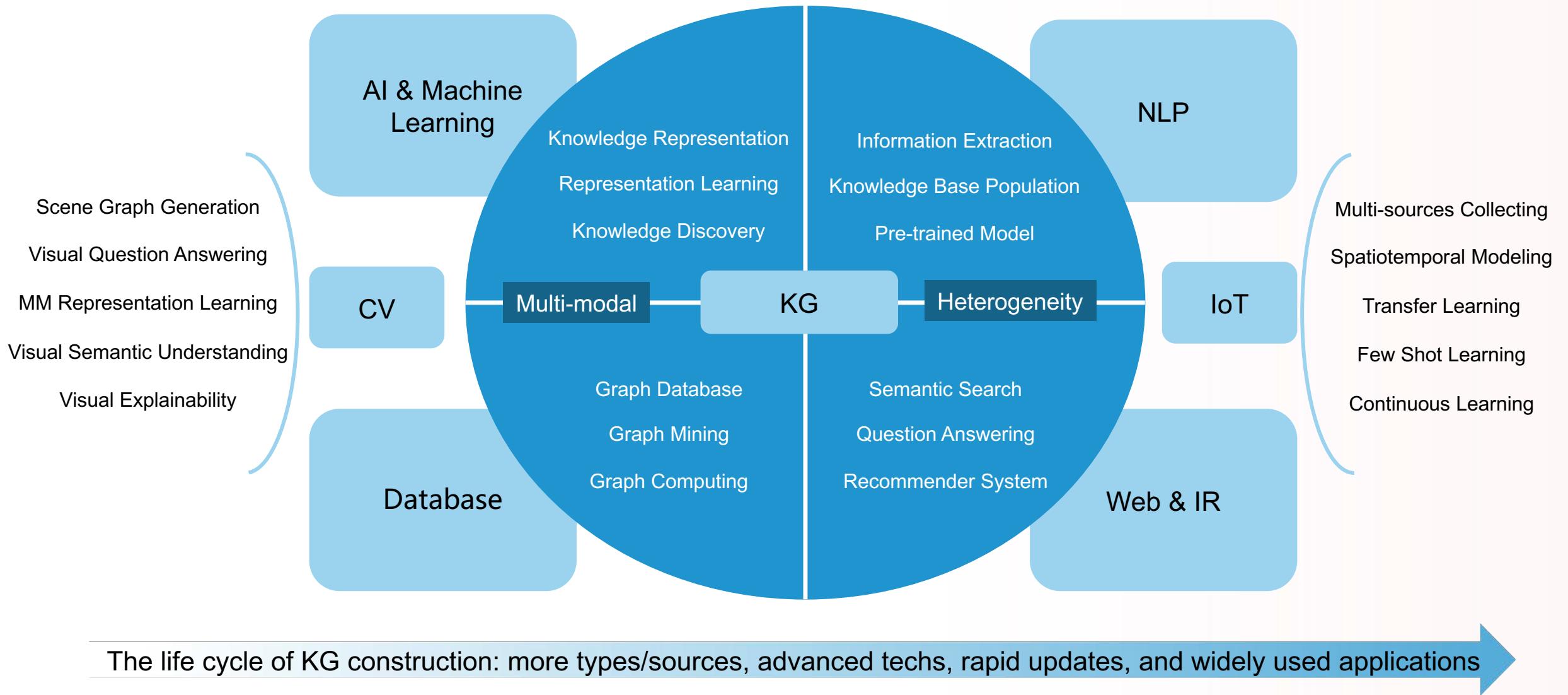
Knowledge types: simple -> complex, static -> dynamic, community -> personal, plain -> spatiotemporal



Challenges

Traditional symbolic knowledge representation methods are difficult to accurately represent complex knowledge such as **dynamics**, **processes**, and **cross-modalities**. At the same time, how to **combine symbolic reasoning methods based on knowledge graphs and neural reasoning methods** is extremely challenging.

SOTA and Trend of KG – Interdisciplinary



Challenges

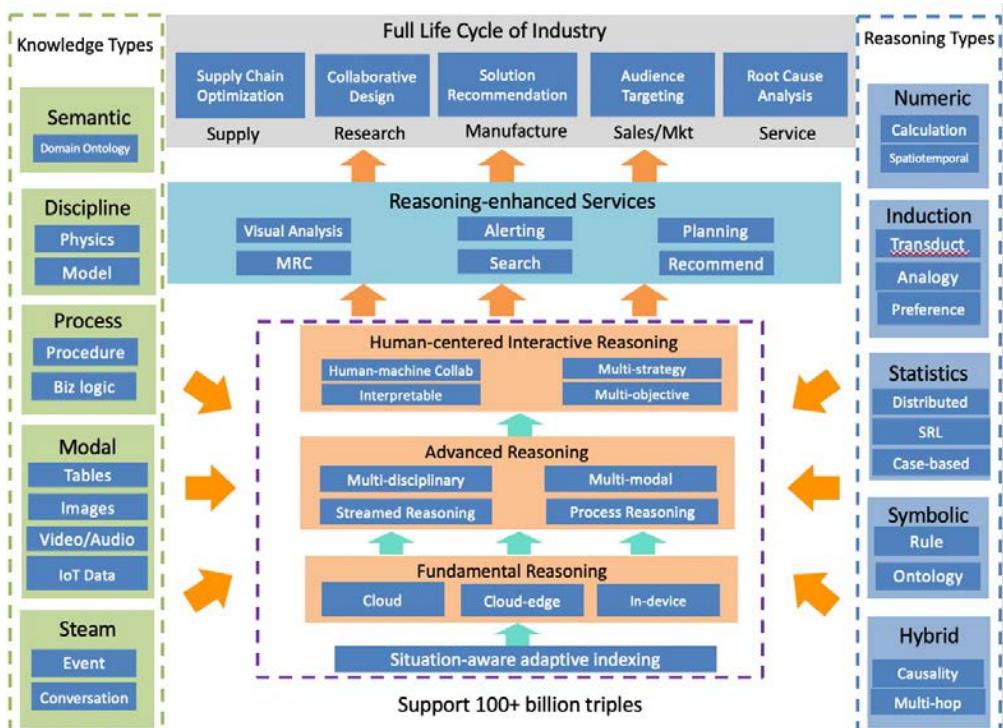
The multi-scale, multi-modal, and multi-disciplinary characteristics of data have put forward new requirements for knowledge representation, collection, extraction, storage, computing, and application. Among them, it is necessary to overcome few shots, explainability, and domain adaptation issues. At the same time, how to realize knowledge update at a low cost is also extremely

SOTA and Trend of KG – System Engineering View

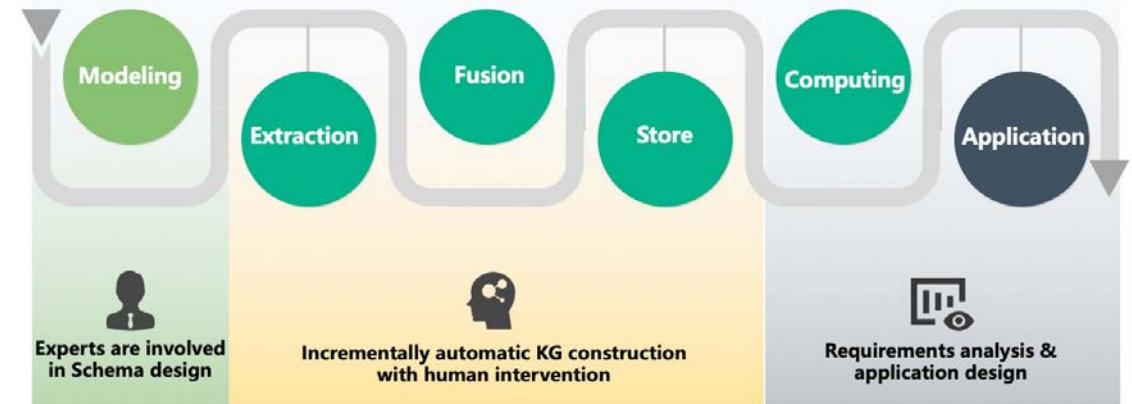


AI should focus on small data and **data centric AI**. Especially in the manufacturing industry, we must rely on **domain knowledge**

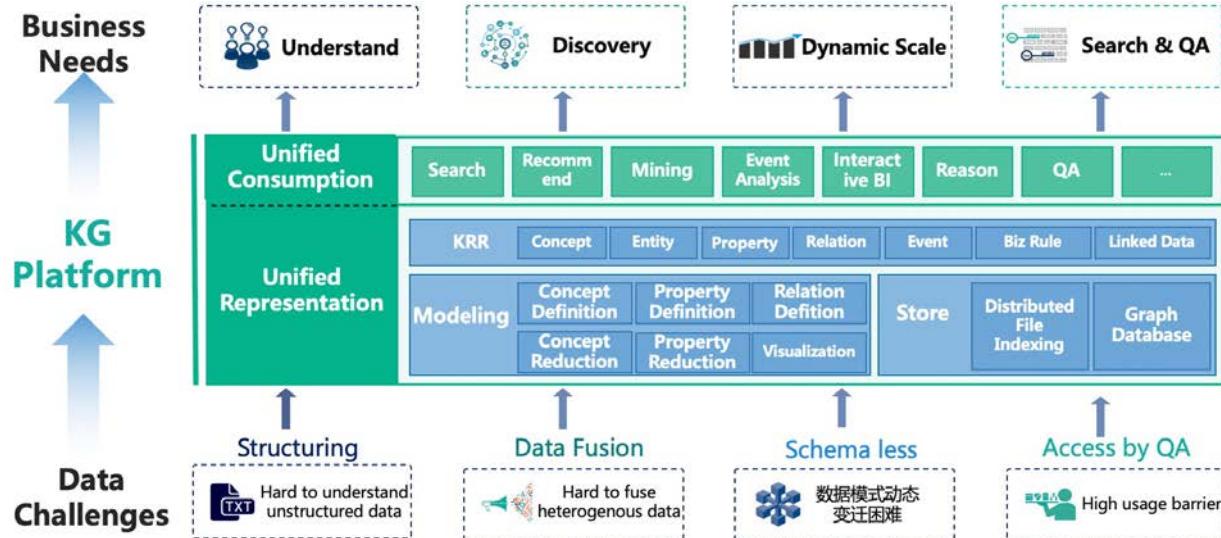
Andrew Ng



Knowledge Graph **System Architecture** in Industry



Different computing manners, "offline - near real time - real-time", depend on the type of knowledge



New Paradigm of Technology with Knowledge as the Core

Challenges: Data characteristics and knowledge differences in different fields lead to low knowledge coverage, intensive labor input, shallow usage. In applications, poor computing efficiency, difficult & weak sustainable operation and long time cost.

Trends of the Interdisciplinary Development of KG – Applications

Search

The screenshot shows two search results pages from a search engine.

Search results for "tim berners lee":

- Google search bar: tim berners lee
- Results: About 4,420,000 results (0.46 seconds). Includes links to w3.org, Tim Berners-Lee - World Wide Web Consortium, Wikipedia, and a knowledge card for Tim Berners-Lee.
- knowledge card:** Tim Berners-Lee, Engineer. Includes a photo, a grid of smaller photos, and a summary: "Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA FBCS, also known as TimBL, is an English engineer and computer scientist best known as the inventor of the World Wide Web. He is a Professorial Fellow of Computer Science at the University of Oxford and a professor at the Massachusetts Institute of Technology. Wikipedia".
- search results:** Includes a sidebar for "Famous computer scientists" featuring Alan Turing, Larry Page, Vint Cerf, Ada Lovelace, Robert E. Kahn, Grace Hopper, and Steve Wozniak.
- recommendation:** A sidebar for "Famous computer scientists" featuring Alan Turing, Larry Page, Vint Cerf, Ada Lovelace, Robert E. Kahn, Grace Hopper, and Steve Wozniak.

Search results for "how old is yao ming's wife":

- Google search bar: how old is yao ming's wife
- Results: About 12,600,000 results (0.55 seconds). Includes a summary for Ye Li: "Ye Li / Age 38 years November 20, 1981".
- knowledge card:** Ye Li, 38 years. Includes a photo of a woman playing basketball.
- People also search for:** Yao Ming (39 years), Yao Qinqlei (10 years), Fang Fengdi (72 years).

Question Answering

Machine Reading Comprehension

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.

Q: Where was the milk before the den?

A : Hallway

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.

Q: What color is Brian?

A: White

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.

Q: Where is the apple?

A: Bedroom

Multi-modal QA

At the base of a muddy ditch is the first primrose of my spring - glowing in the grey, a little spot of hope, brave, beautiful and perfect.



Hi Chris, wow well spotted with the beautiful flower, I love walking alongside the river where there is a bluebell way

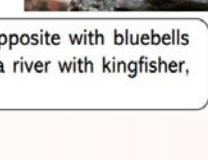


Love every photo. 😊😊 Especially the weeping willow.



Thanks. It's nice to enjoy the wildlife nature and walk all the way to Winchester's great scenery.

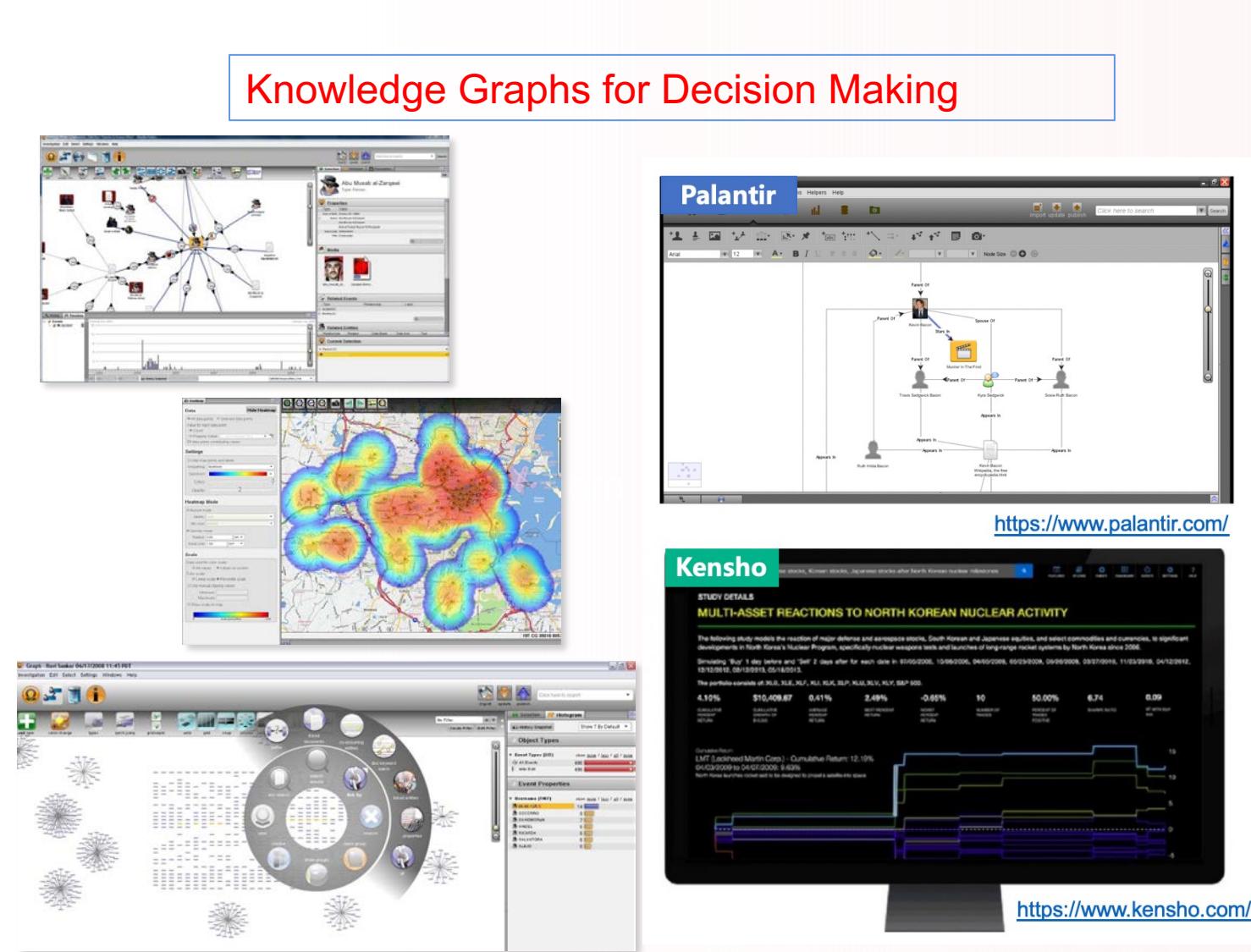
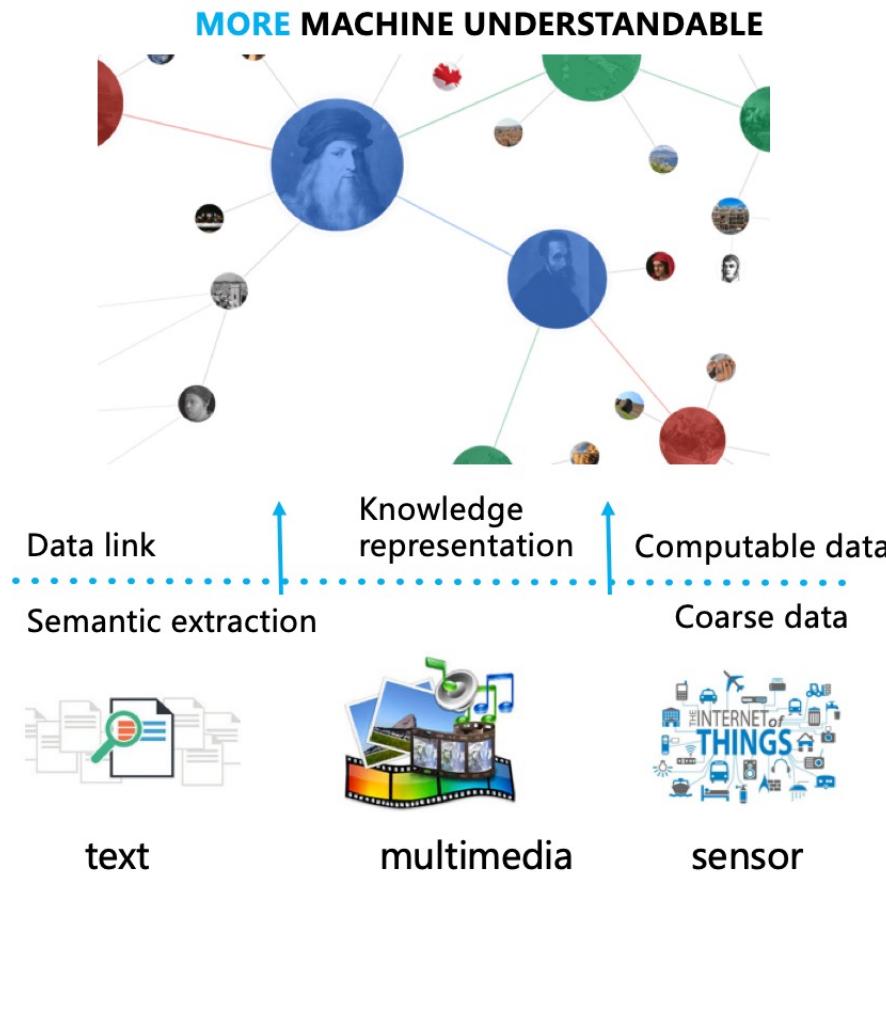
I live in Scotland. We have woods opposite with bluebells but not as thick as yours, but have a river with kingfisher, Heron and dipper.



Challenges

To build a multi-source and multi-modal knowledge graph, not only quality but also coverage should be considered. In the process of model training, the alignment of heterogeneous and multimodal knowledge is the difficulty of knowledge fusion and learning

Trends of the Interdisciplinary Development of KG – Applications



In each specific field, the explainability of the model and predictions are the most important to realize the application value. How to balance the advantages and disadvantages of **symbolic models** and **neural networks**, and learn from each other is a hot topic in academia and industry.

Life cycle of Knowledge Graph

- Top-down method
- Bottom-up method

- Schema graph fusion
- Data graph fusion

- Graph computing
- Ontology reasoning
- Rule-based reasoning

**Reasoning
is important !**

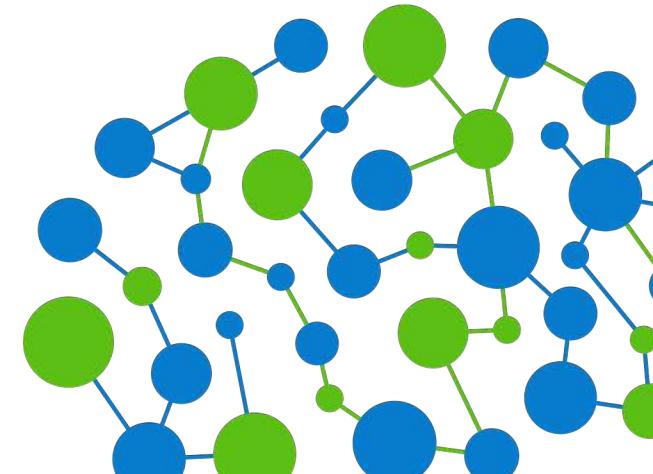


- Linked data: graph mapping
- Structured data: D2R
- Semi-structured data: wrapper
- Text: information extraction

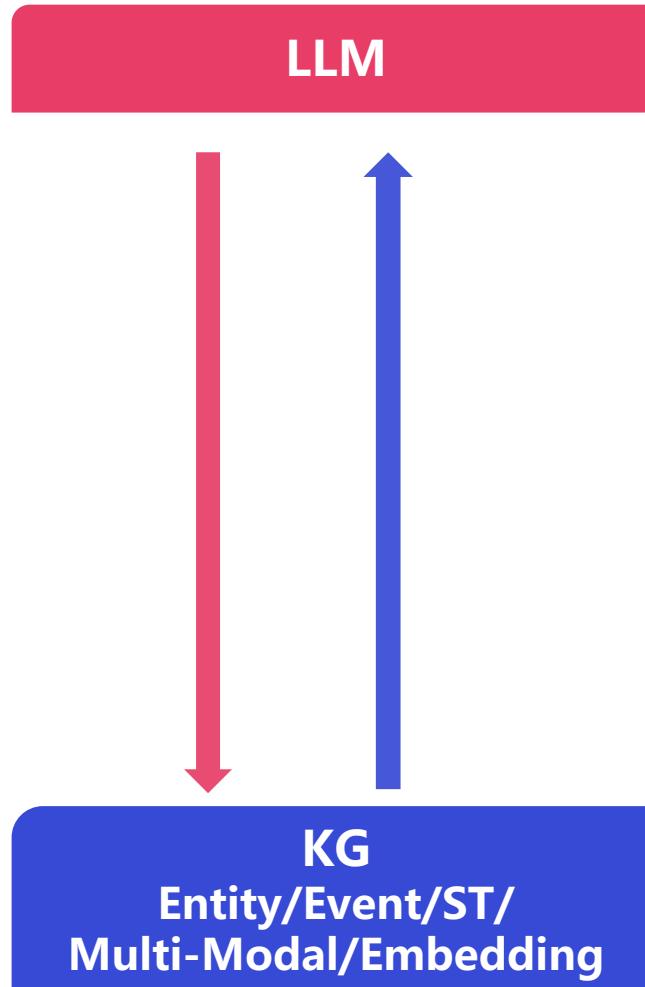
- Triples
- Event information
- Temporal information
- Multi-modal

- Semantic search
- Question answering
- Recommendation
- Assistant decision

- The Rise of LLM
- The Development of KG
- **The Emerging Paradigm**
- Outlook



When KG meets LLM: Two Sides of a Coin



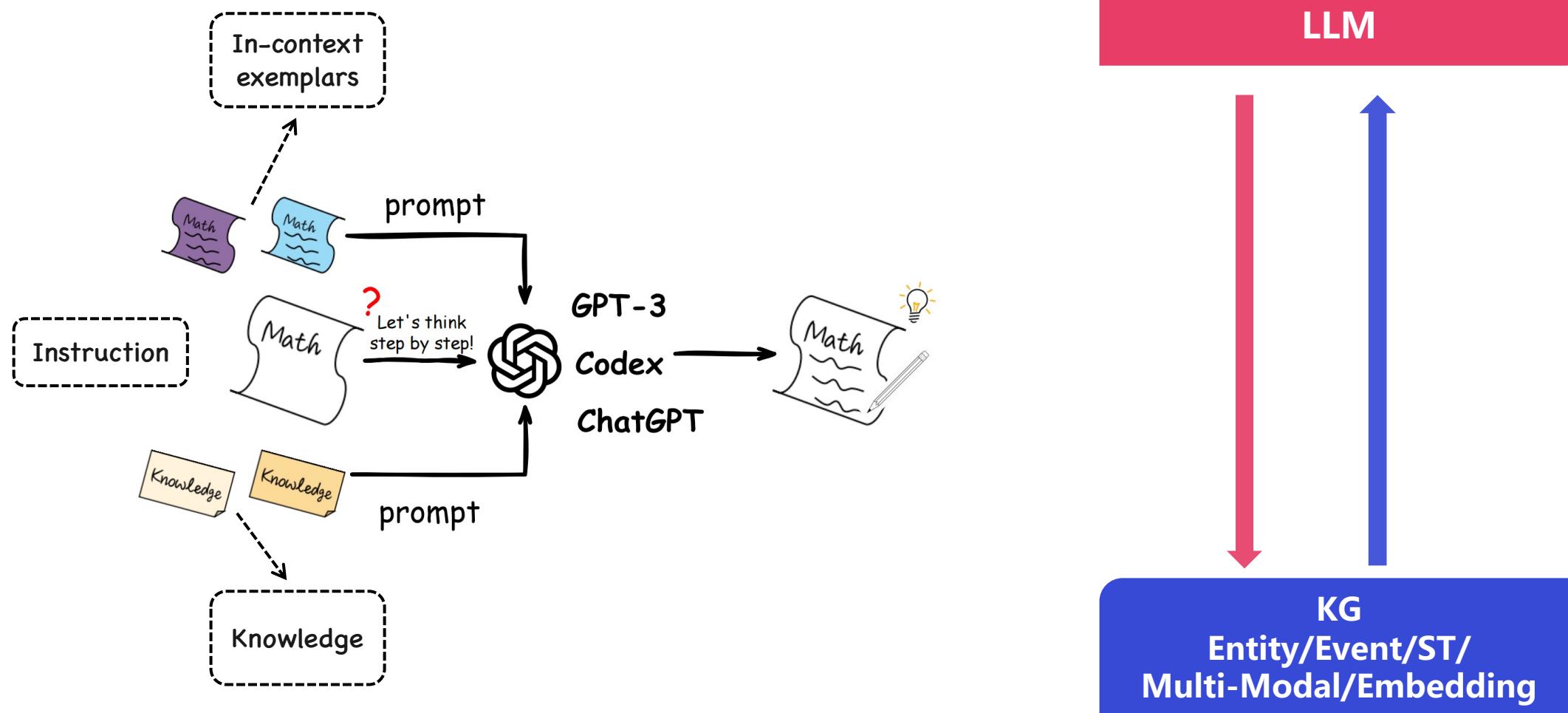
Boost Rapid KBP: Knowledge Extraction/Fusion

- Leverage the **understanding** and **generation** capabilities of LLM for automatic **extraction** and **fusion** of knowledge

Knowledge enhanced PTM/Prompt Engineering/Complex Knowledge Reasoning/Provenance/Integrating Dynamic Knowledge

- **Knowledge guided** automatic prompt engineering
- The emergent ability of LLM especially CoT , combined with **KG reasoning** to solve complex tasks
- Knowledge in form of triples, instructions, rules or codes can be injected into LM for data augmentation and bring more **trustworthy** and **interpretability**
- Linking generated responses of LLM to KG for **references**, **provenance** or **fact checking**

When KG meets LLM: Knowledge-guided Prompting



When KG meets LLM: Knowledge-guided Prompting

Implicit Knowledge

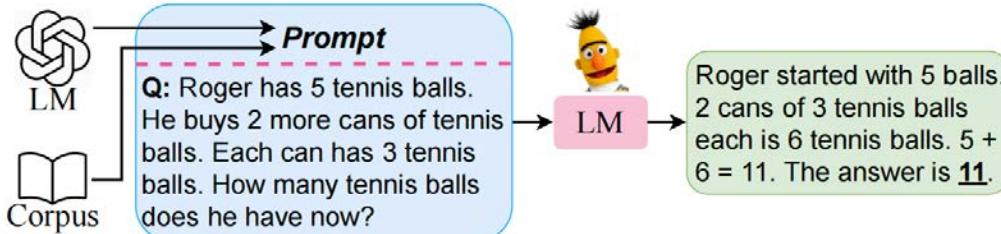
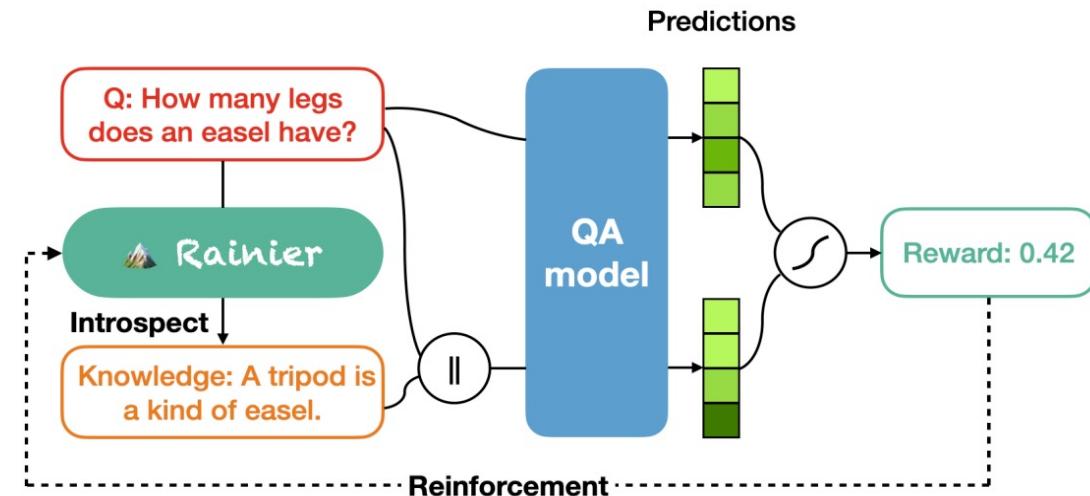


Figure 6: Knowledge Enhanced Reasoning (§3.2). Prompts are generated by LMs (**Implicit Knowledge**) or retrieved from external resources (**Explicit Knowledge**).

- LLM contains massive implicit knowledge
- Knowledge distillation



Few-shot prompting over GPT for downstream tasks like QA, and introspect knowledge via reinforcement learning

When KG meets LLM: Knowledge-guided Prompting

Explicit Knowledge

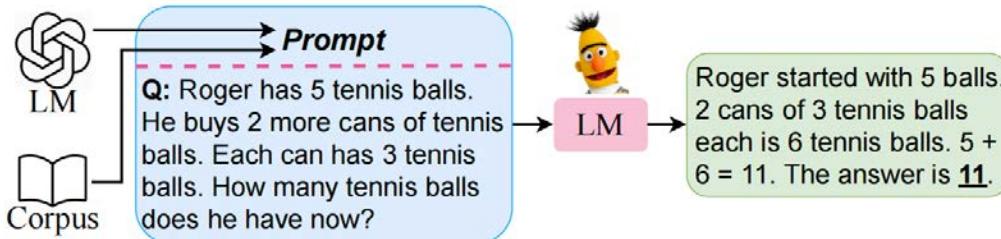
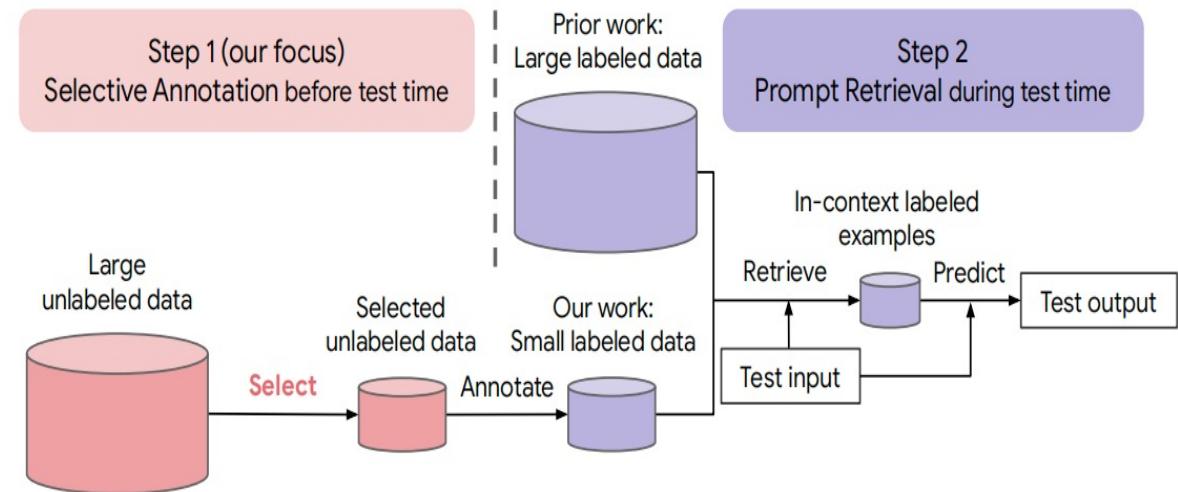


Figure 6: Knowledge Enhanced Reasoning (§3.2). Prompts are generated by LMs (**Implicit Knowledge**) or retrieved from external resources (**Explicit Knowledge**).

- Despite exhibiting strong generative abilities, LLM tends to hallucinate and generate inconsistent knowledge
- Retrieved from external explicit knowledge



Develop a selective annotation framework to avoid reliance on retrieval of large-scale annotated corpora

When KG meets LLM: Reasoning Capability Comparison

LLM Reasoning

- **Code pre-training:** enhance LLM reasoning during training
- **Prompt Engineering:** eliciting LLM reasoning during inference

KG Reasoning

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

LLM Reasoning

- zero-shot prompting
- Few-shot prompting
- CoT prompting
- Instruction



KG Reasoning

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

When KG meets LLM: Two Kinds of Tasks

端到端的感知类任务、创造性任务、探索性任务
End-to-end perception tasks, creative tasks, exploratory tasks

强调数据可靠、结果确定、计算精准的任务
Tasks emphasizing data reliability, outcome determinism, and computational precision

自然人机交互界面
Natural Human-machine Interface

自然人机交互界面
Natural Human-machine Interface

提示工程
Prompt Engineering

微调
Fine-tuning

提示工程
Prompt Engineering

微调
Fine-tuning

非确定性计算系统
Non-deterministic Computing

非确定性计算系统
Non-deterministic Computing

多模态大模型
Multimodal Foundation Models

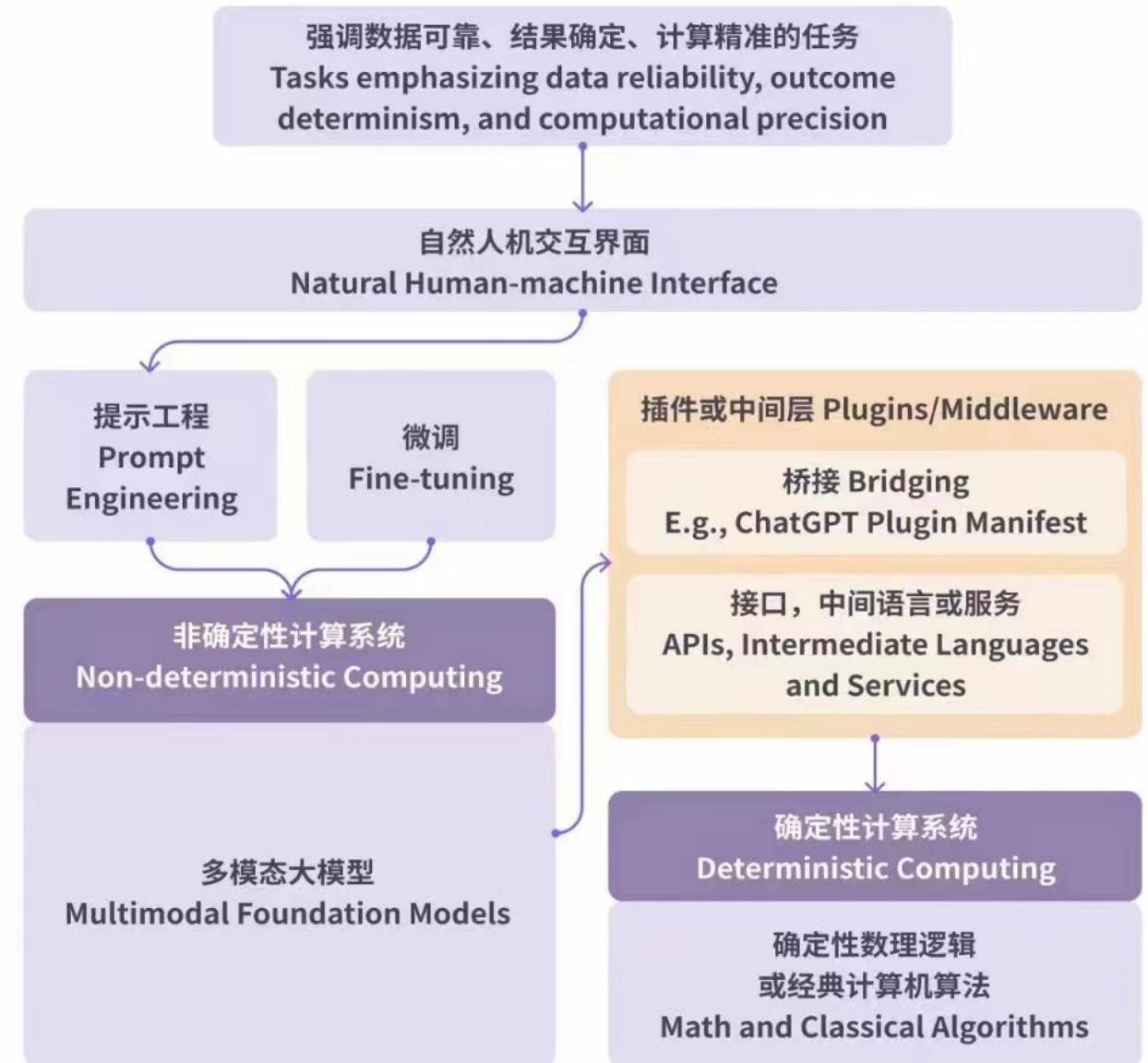
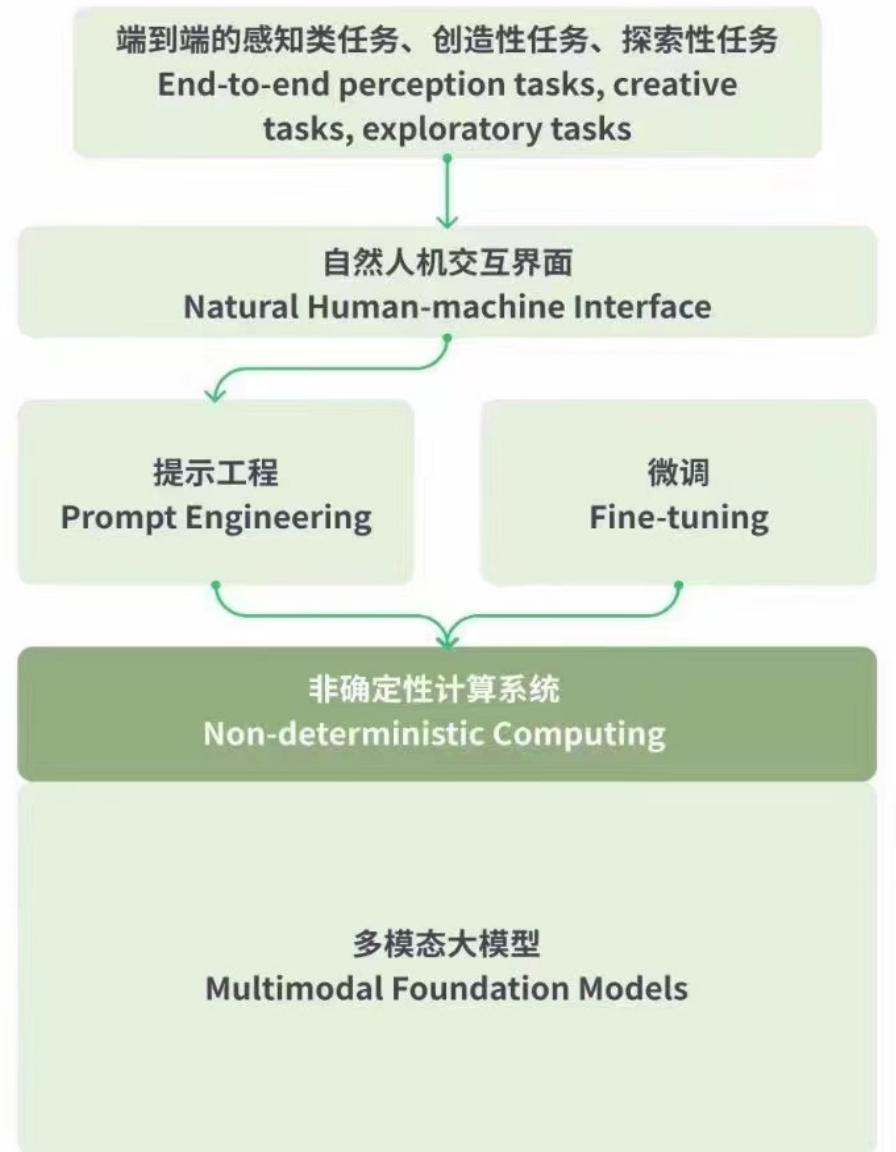
多模态大模型
Multimodal Foundation Models

插件或中间层 Plugins/Middleware
桥接 Bridging
E.g., ChatGPT Plugin Manifest

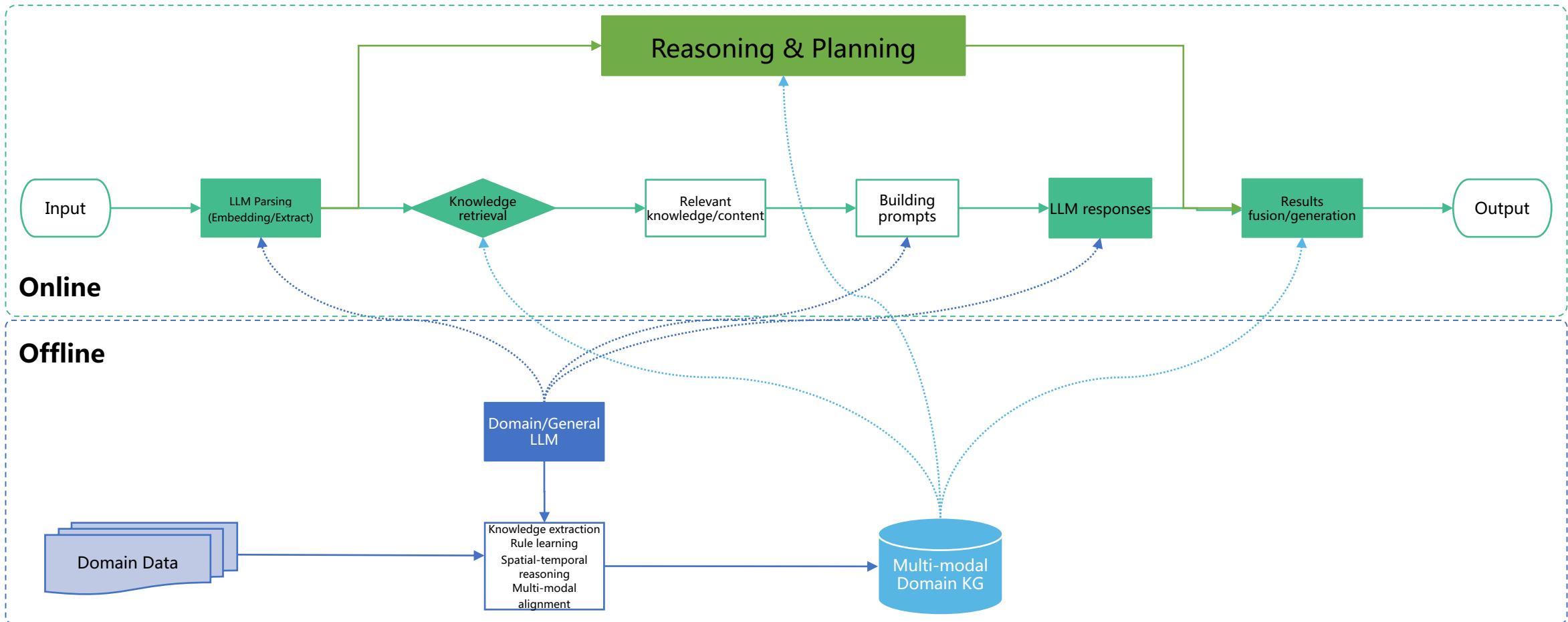
接口，中间语言或服务
APIs, Intermediate Languages and Services

确定性计算系统
Deterministic Computing

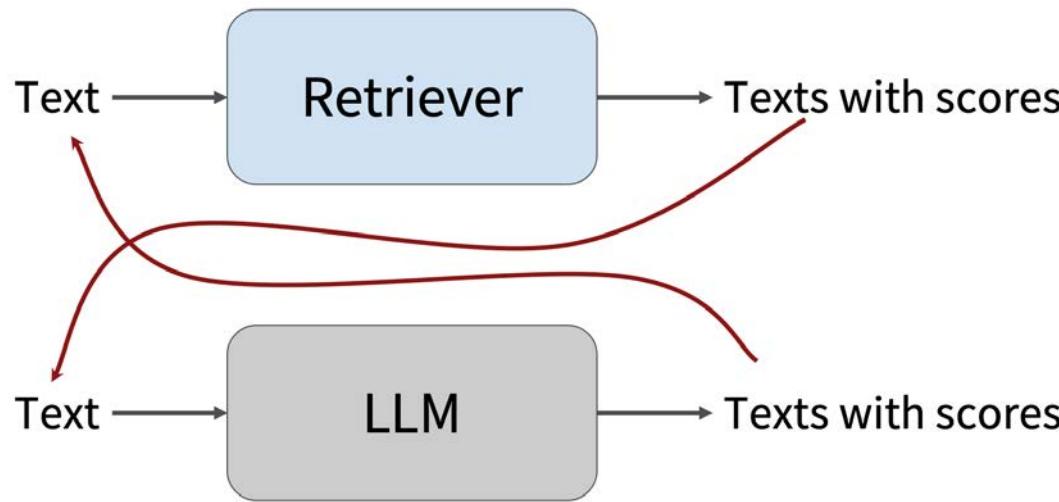
确定性数理逻辑
或经典计算机算法
Math and Classical Algorithms



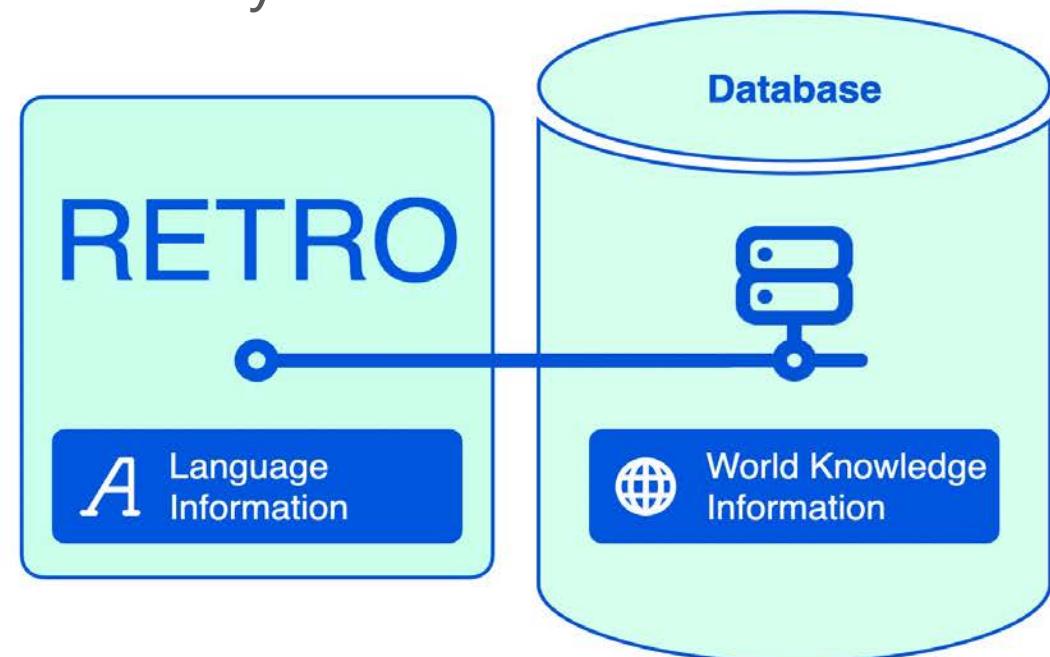
When KG meets LLM: Knowledge Augmentation



When KG meets LLM: Retrieval & Knowledge Augmentation



- Efficiency
- Updateability
- Provenance
- Effectiveness
- Synthesis



DeepMind's RETRO (Retrieval-Enhanced TRansfOrmer)

When KG meets LLM: Knowledge Augmentation

👤 Which 2013 Los Angeles Galaxy player transferred in from the team with 12 international titles ?

Consolidate evidence from external knowledge

WIKIPEDIA Transfers In

Wiki Loves FOLKLORE

Photograph your local culture, help Wikipedia and win!

2013 Los Angeles Galaxy season

Juninho (footballer, born January 1989)

Juninho grew up in the city of São Paulo and played for the São Paulo youth team, winning the U-17 Paulista Championship side in 2000 made one appearance for the São Paulo senior side in 2007. He was sent out on loan to Major League Soccer team Los Angeles Galaxy in 2010 along with fellow Brazilians from the club Al Cazumba and Leonardo.^[1] He made his debut for the team on 27 March 2010, in Galaxy's opening game of the 2010 MLS season against New England Revolution,^[2] and scored his first goal for the Galaxy in a 2-0 win over AC St. Louis.

Transfers

No.	Pos.	Player	Transferred from	Fee/No
1	GK	Carlo Cudicini	Tottenham Hotspur	Free
	MF	Juninho	São Paulo	Undisclosed

Honours

São Paulo FC is one of the most successful clubs in Brazil, having won a total of 30 domestic honours, in addition to their 12 international successes. It is the Brazilian club with the most international titles.

Major competitions

Worldwide

Cup	Titles	Years
Intercontinental Cup	2	1992, 1993
FIFA Club World Cup	1	2005

Continental

Competitions	Titles	Years
Copa Libertadores	3	1992, 1993, 2005

AI Agent (LLM-Augmenter + LLM)

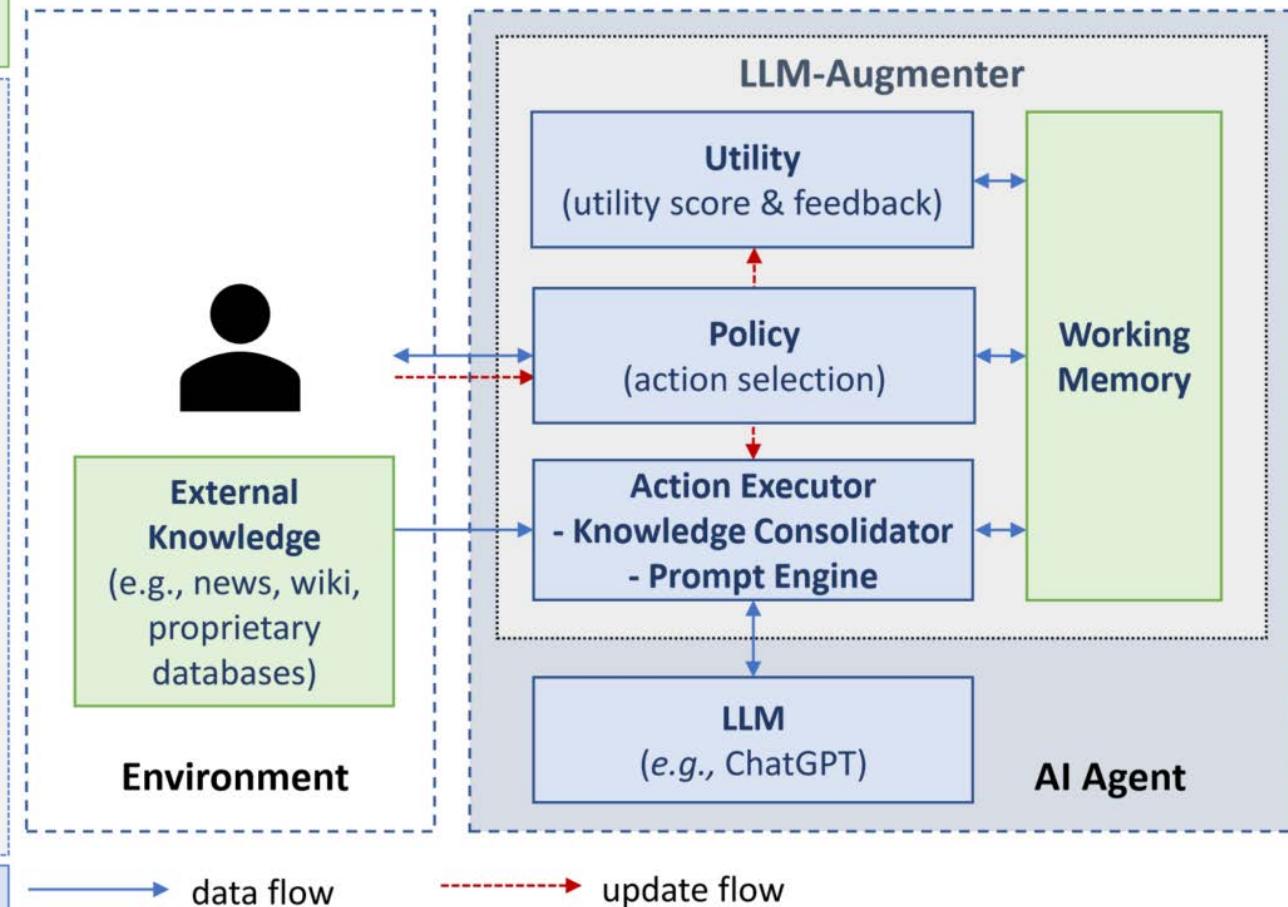
Revise response via automatic feedback

Candidate response:
Jaime Penedo is transferred in from C.S.D. Municipal, a team with 12 international titles.

Feedback:
The player Jaime Penedo is transferred in from C.S.D. Municipal, but there is no information about the number of international titles of this team.

Revised candidate response:
Juninho is transferred in from São Paulo, a team with 12 international titles.

🤖 Juninho is transferred in from São Paulo, a team with 12 international titles.

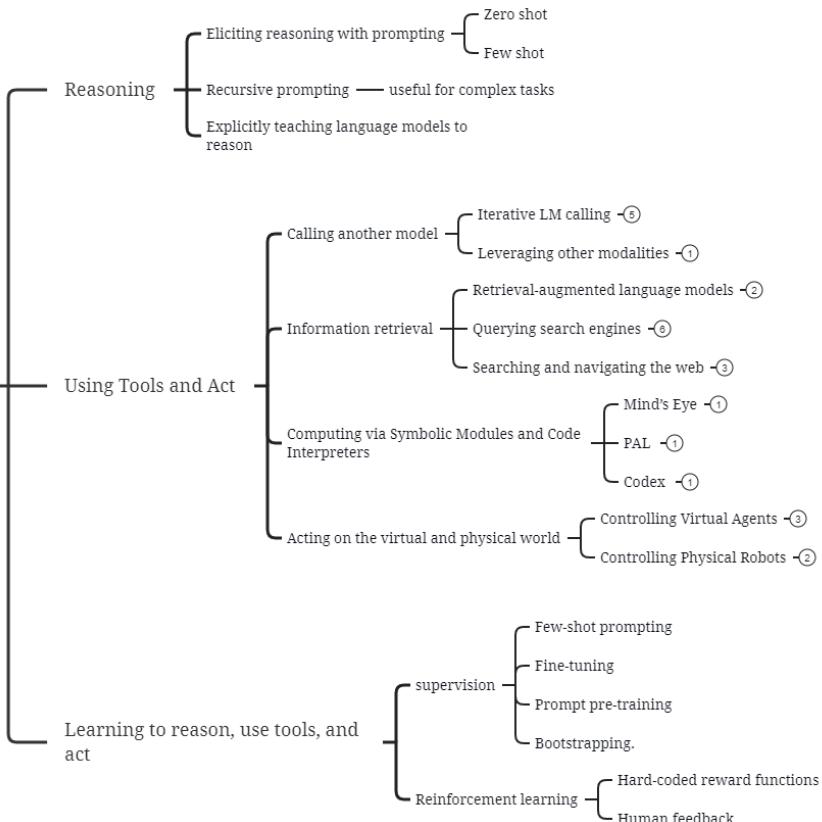


Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback 2023

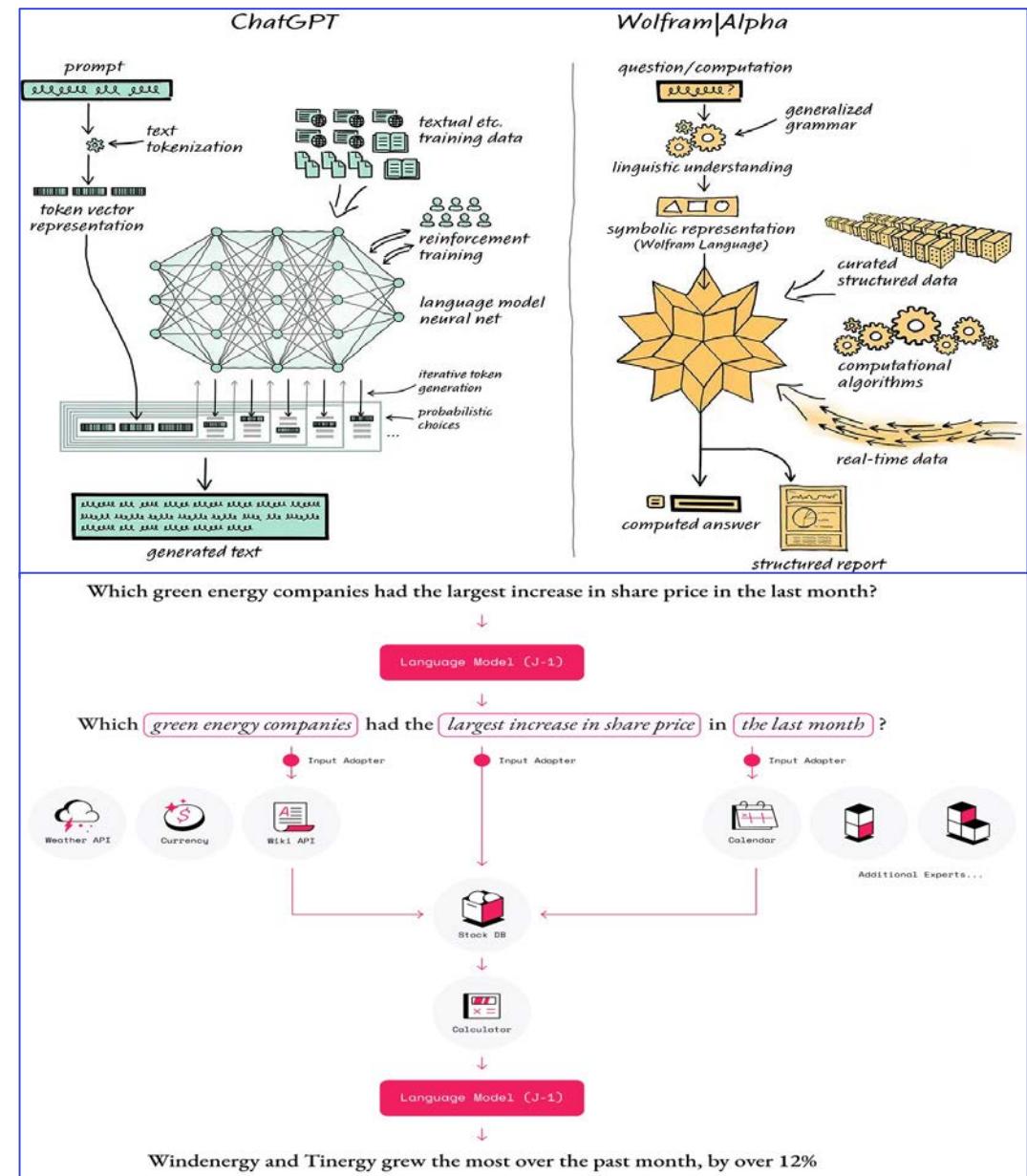
Augmented Language Models

Important Note: This week, the WL will be op

ToolFormer



Source: Augmented Language Models: a Survey (Yann Lecun et al.)



When KG meets LLM: Interoperability & Orchestration

KG uses ontology to realize the representation of domain data, knowledge and interaction, and completes the automation of the whole process from real-time data access, knowledge update to user interaction.

Interoperate & Orchestrate

Integrate the following steps into a whole end-to-end process for hyper-automation

Data Access

Real-time, dynamic, or temporal-spatial data access

Knowledge Update

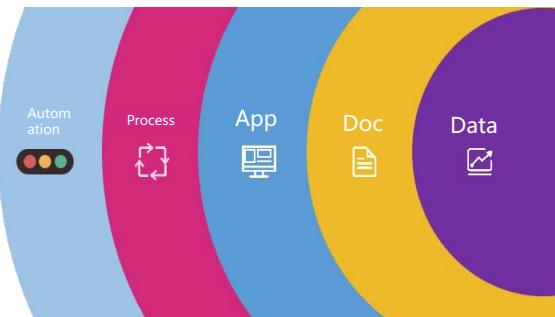
Update knowledge via extraction by LLM

Reasoning

Combine reasoned results from KG with those LLM responses

Act

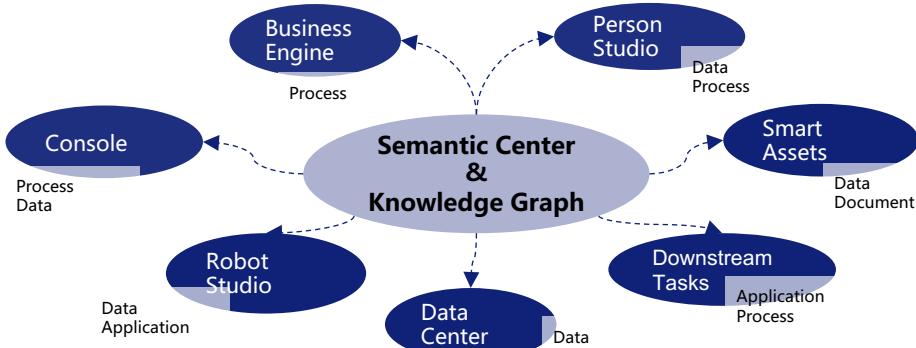
- User: chat, speak, recommend
- System: API invoke, send commands



- Task mining
- Consistency checking
- Process discovery
- Modeling via simulation
- Case prediction
- Suggestion via history data



When KG meets LLM: Interoperability & Orchestration



RPA + Tool use

New Task (Physics QA) Answer this high-school physics question
Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.

Tool Library:

- Code operations
- CoT-style reasoning
- Search
- Arithmetic
- String operations

A → **TASK LIBRARY** → **B** → **TOOL LIBRARY** → **Google**

Google Response:

Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.
Q1: [search] What is the formula for the horizontal component of the tension force?
#1: The formula for the horizontal component of the tension force is $T\cos\theta$. The horizontal component (F_x) can be calculated as $F_x = T\cos\theta$ where θ is the angle which the force makes with the horizontal in radians.

LLM Interaction:

Input: ... Q1: [search] ...
#1: ... can be calculated as $F_x = T\cos\theta$ where θ is ...
Q2: [generate code] Use the formula $F_x = T\cos\theta$ to solve: Hank ...
#2: T = 72.0, theta = 35.0
radians= math.pi*theta/180
Fx = T*math.cos(radians)

OpenAI Codex:

Input: ...Q1: [search] ...#1: ...
Q2: [generate code] Use the formula $F_x = T\cos\theta$ to solve: Hank ...
#2: ... Fx = T*math.cos(radians)
Q3: [code execute] Execute the python code and get the value of "Fx"
#3: 58.9789
Q4: [EOQ] Ans: 58.9789

Python:

python

Code Snippets:

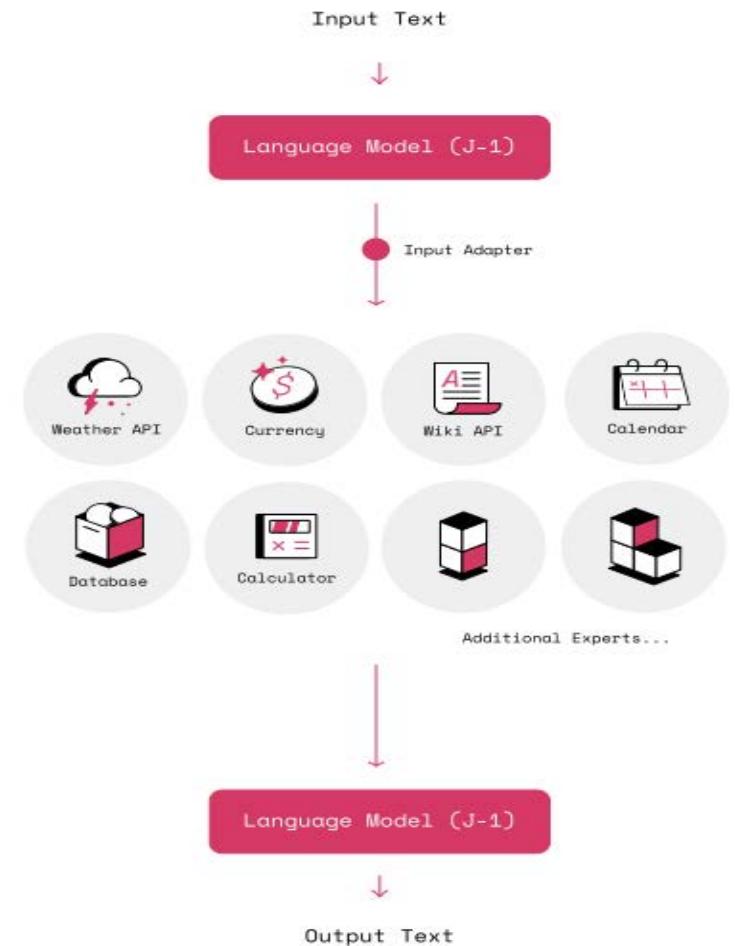
```
viola_bought = 167, nancy_took = 137
ans = viola_bought - nancy_took
Q2: [code execute] Execute snippet #2: 30 Q3: [EOQ] Ans: No
```

Arithmetic Problems:

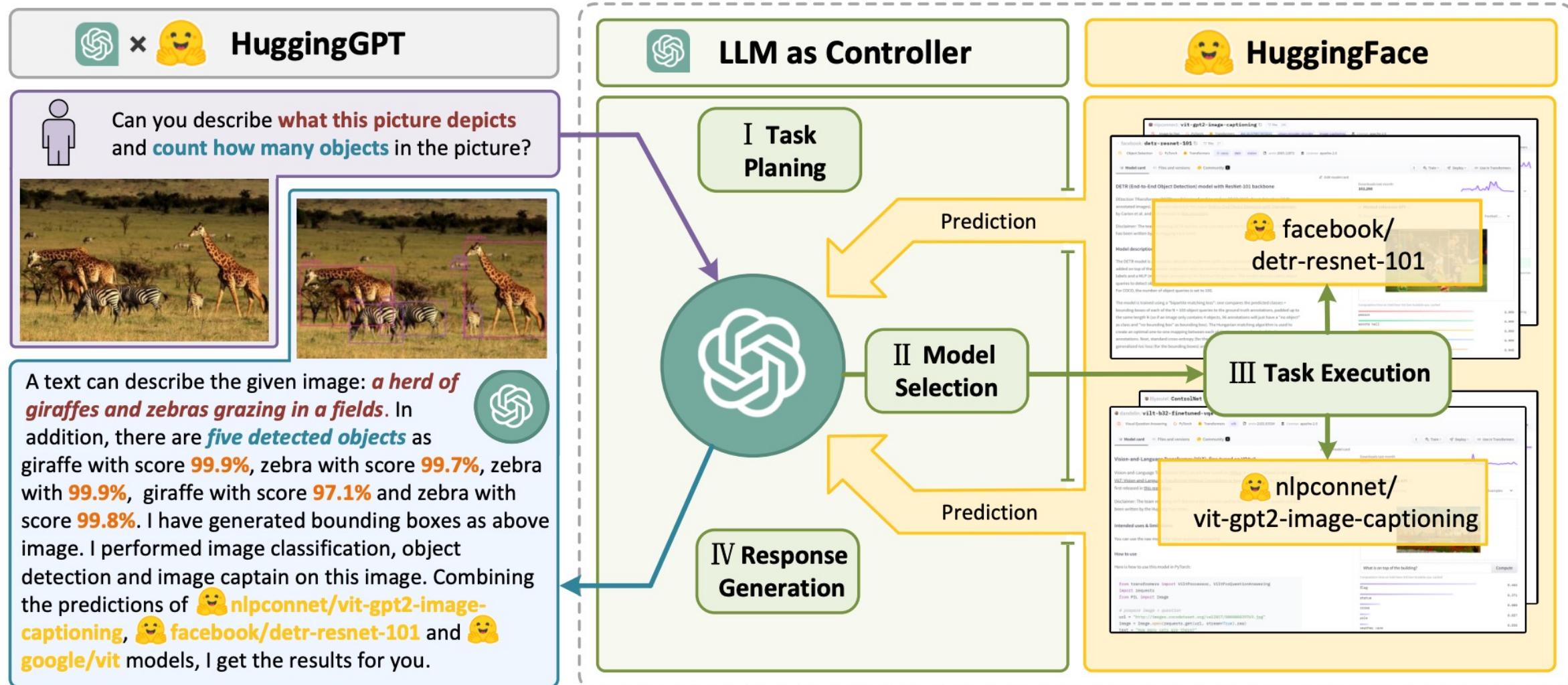
Solve these arithmetic problems using python code
Input: Viola had 167 breads. Nancy took 137 from him. How many does Viola have now?
Q1: [generate code] Write down arithmetic as python code
#1: viola_bought = 167, nancy_took = 137
ans = viola_bought - nancy_took
Q2: [code execute] Execute snippet #2: 30 Q3: [EOQ] Ans: No

Does the sentence contain an anachronism? Yes/No.
Input: President George H. W. Bush called his generals at the outset of the Gulf War.
Q1: [search] When was President George H. W. Bush, president?
#1: George H. W. Bush's tenure started on January 20, 1989, and ended on January 20, 1993.
Q2: [search] When was the Gulf War fought? #2: The Gulf War was a 1990–1991
Q3: [subquestion] Could these entities have co-existed? #3: Yes. Their time periods intersect.
Q4: [generate output] Is this an anachronism? #4: No Q5: [EOQ] Ans: No

Adaptive Routing

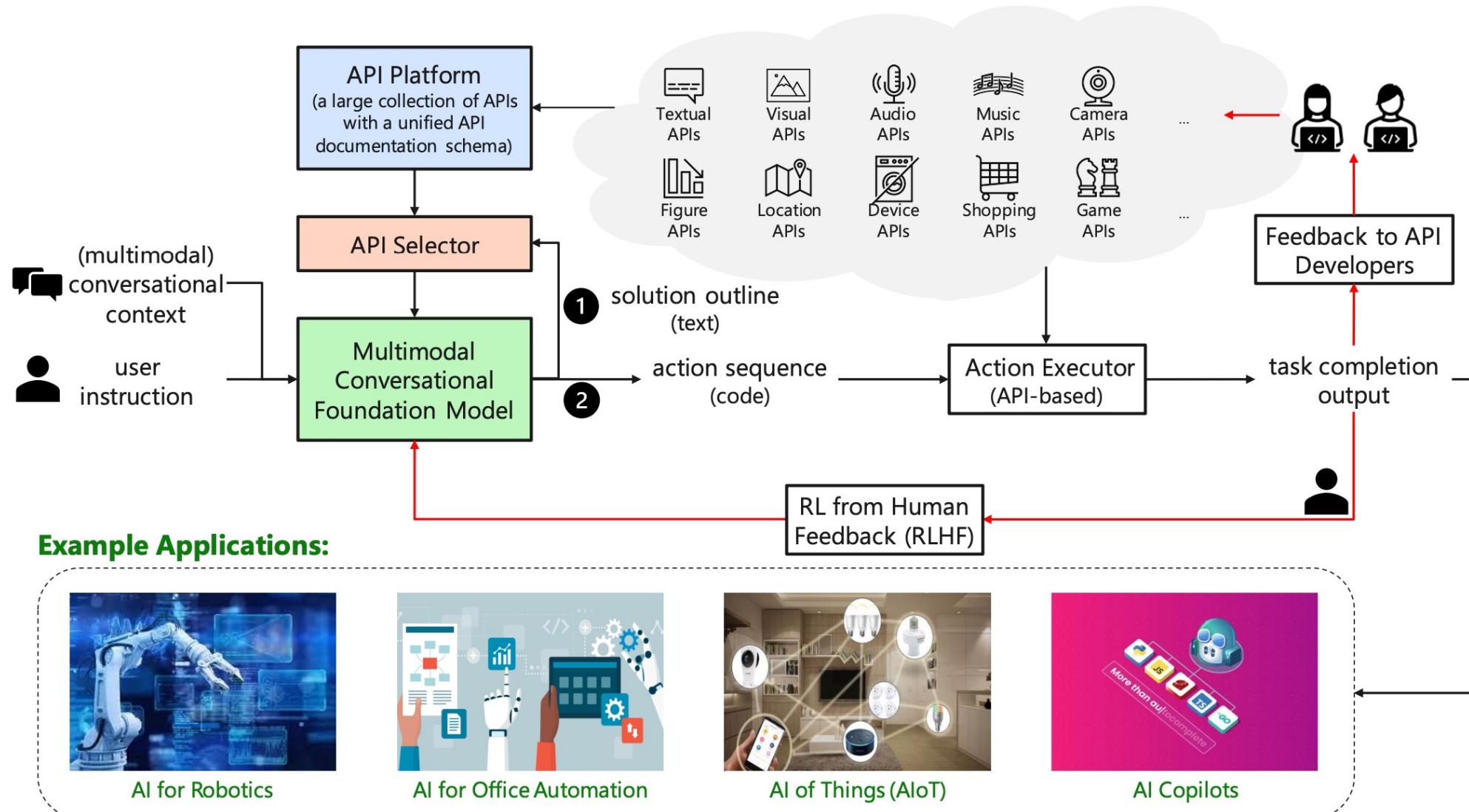


When KG meets LLM: Interoperability & Orchestration

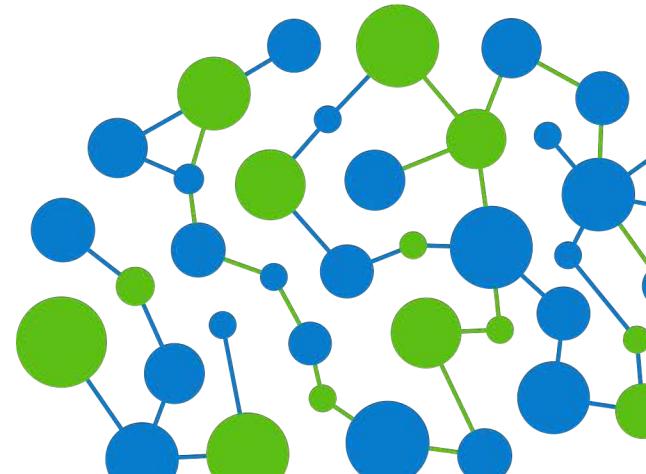


HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace 2023

When KG meets LLM: Interoperability & Orchestration

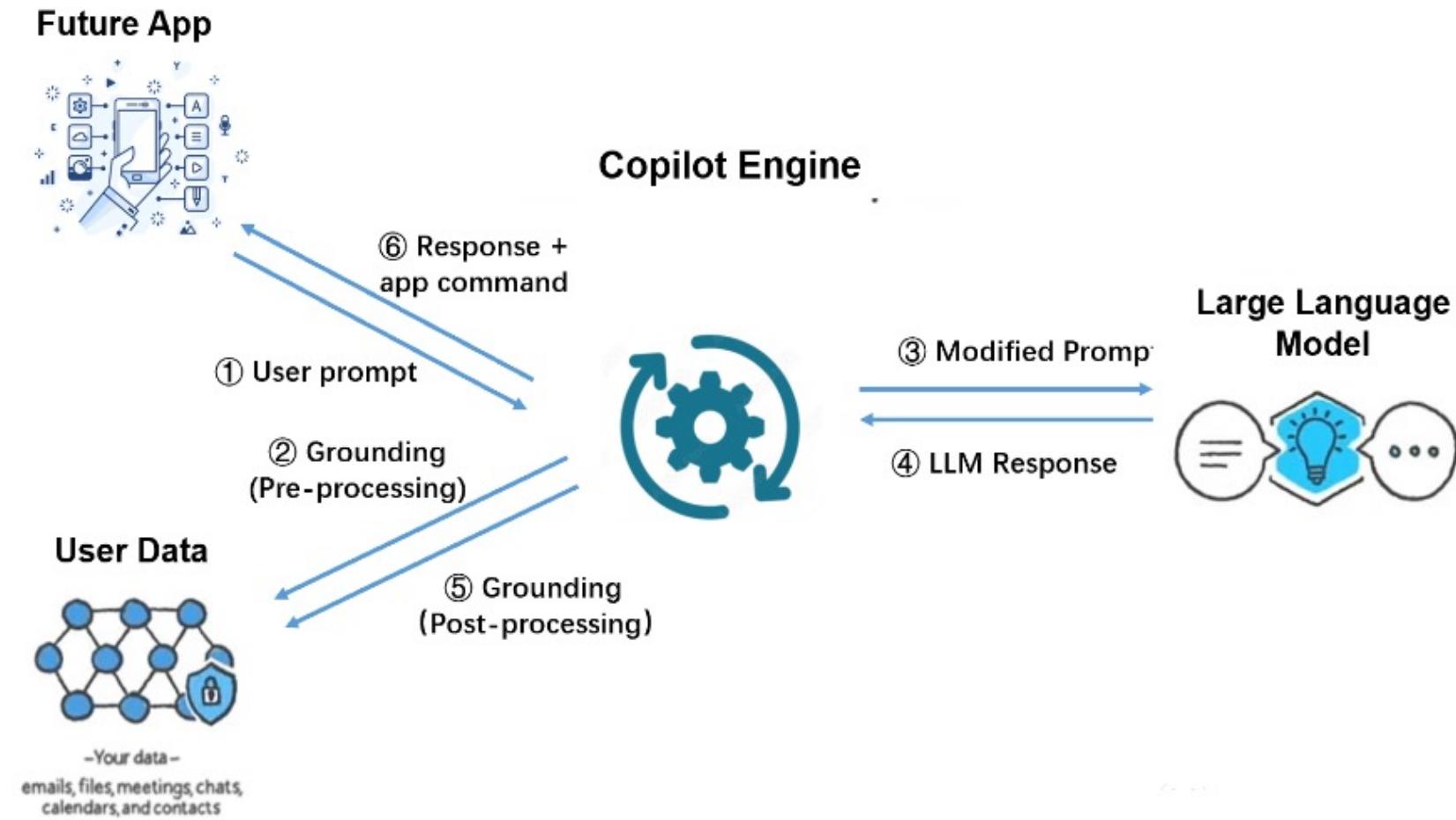


- The Rise of LLM
- The Development of KG
- The Emerging Paradigm
- **Outlook**



Augmented Language Models: Copilot Engine

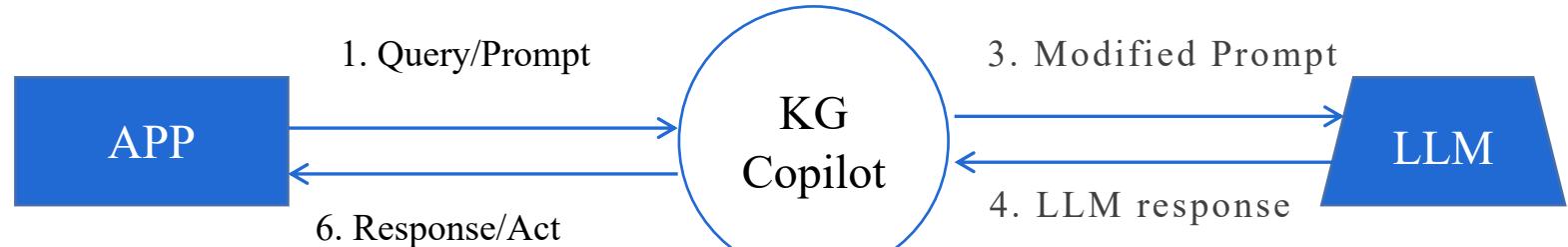
- ❖ Grounding (pre-processing): use Microsoft Graph to extract relevant user data, and treat them as a part of prompts
- ❖ Modified prompt: combine user prompt with data, and optimize the prompt for richer contexts to be fed into LLM
- ❖ Grounding (post-processing): perform post-processing for results from LLM with further validation and referencing
- ❖ Response + App command: convert results into commands for applications like Word, Excel, or PPT. For example, Office supports VBA and the output of a command can be any code.



Augmented Language Models: KG Copilot Engine

- ❖ KG for unified knowledge modeling

- ❖ Multi-modal data
- ❖ Multi-source data



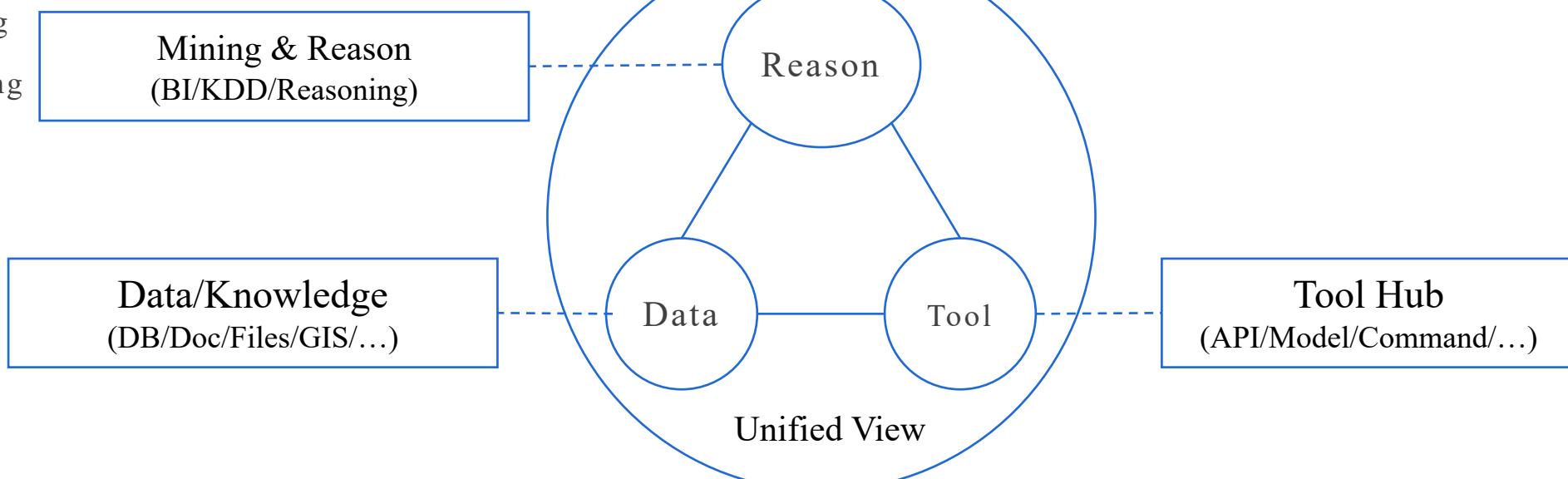
- ❖ KG for unified programming as a semantic gateway

- ❖ LLM + KG + Tool

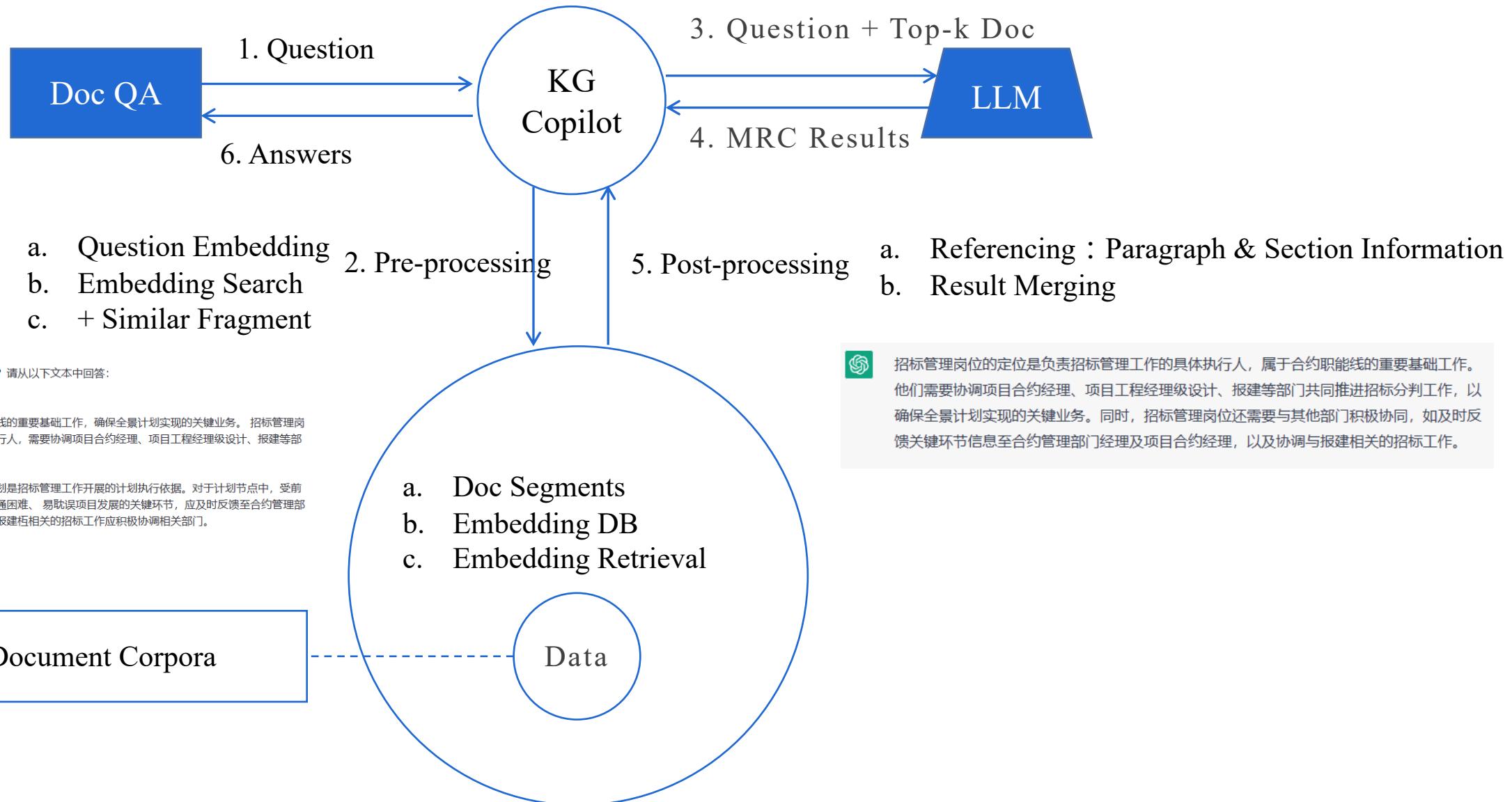
- a. Prompt build
 - b. refinement
2. Pre-processing
 5. Post-processing
- a. Fact checking
 - b. Data query
 - c. Tool invoke
 - d. Reasoning
 - e. Result fusion

- ❖ KG for automatic prompt building and result checking

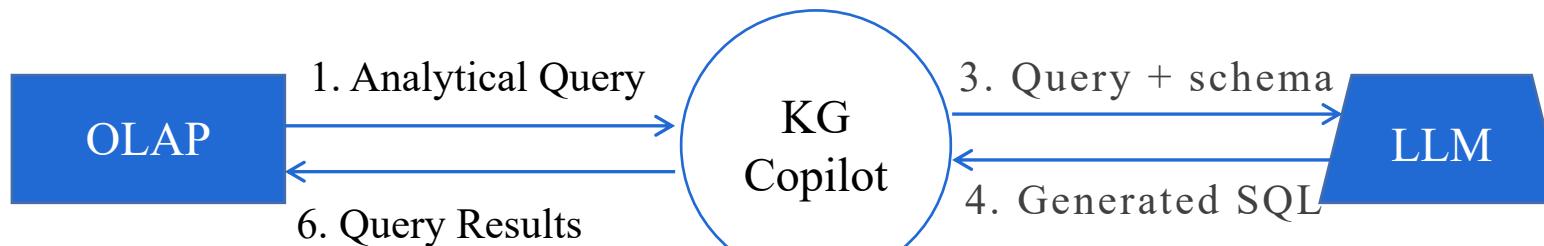
- ❖ Pre-processing grounding
- ❖ Post-processing grounding



KG Copilot Engine Example: ChatDocument



KG Copilot Engine Example: ChatTable



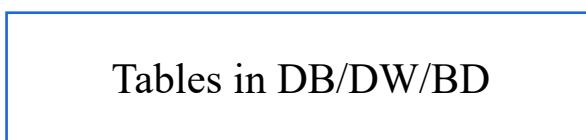
- a. Lookup Tables
- a. SQL Querying
- b. Schema2Prompt
- 2. Pre-processing
- 5. Post-processing

GU
数据库表公民信息表 citizen, 表格的列为“唯一标识id”, “姓名name”和“身份证号ID_number”
写一个SQL查询: 查找出所有“身份证号相同但姓名不同”的记录

要查询所有身份证号相同但姓名不同的公民信息记录, 可以使用子查询来选择具有多个不同名称的唯一身份证号码, 然后在唯一身份证号码上将子查询与原始表进行联接:

```
sql
Copy code

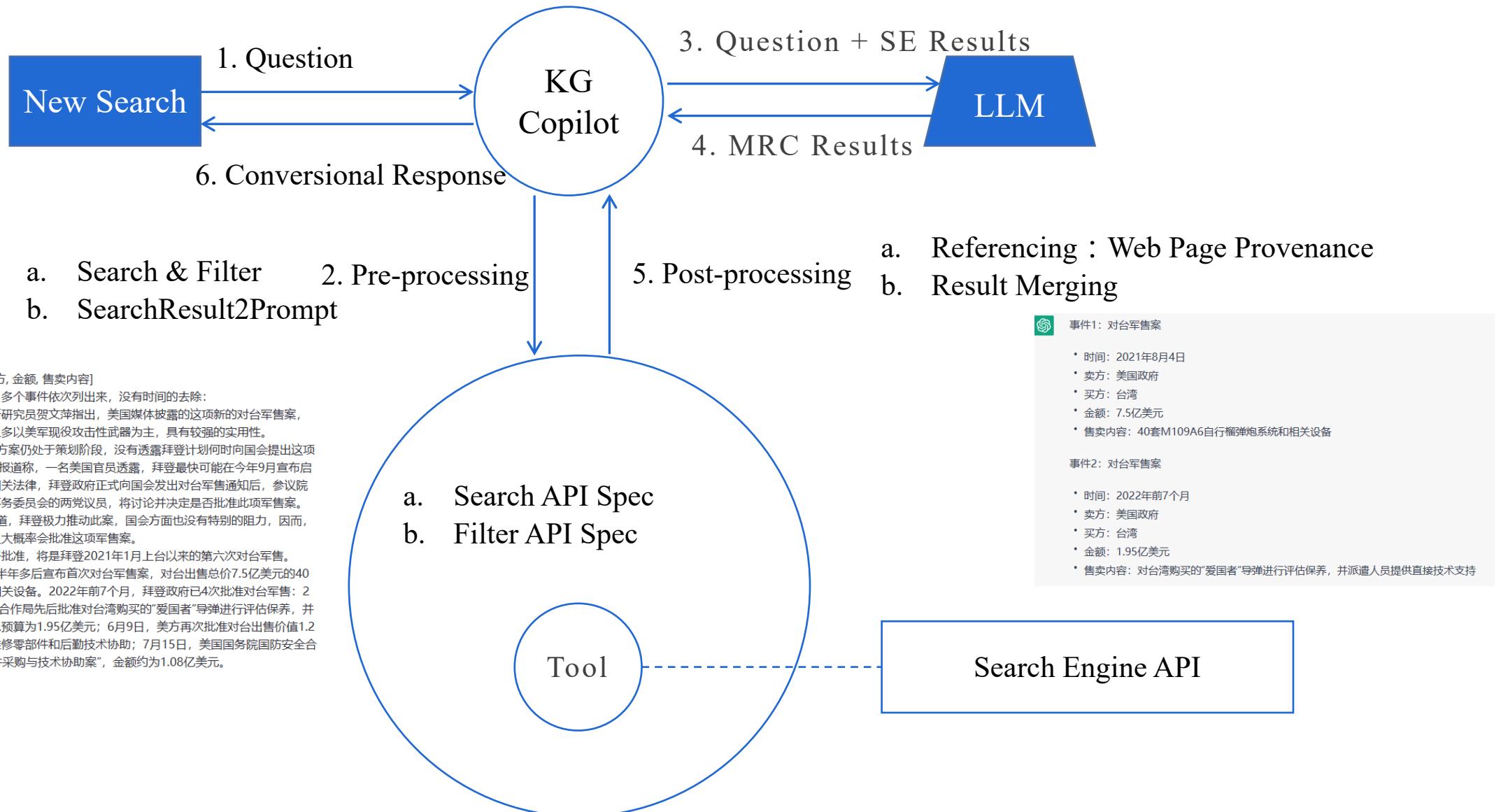
SELECT c.id, c.name, c.ID_number
FROM citizen c
INNER JOIN (
    SELECT ID_number
    FROM citizen
    GROUP BY ID_number
    HAVING COUNT(DISTINCT name) > 1
) c2 ON c.ID_number = c2.ID_number
ORDER BY c.ID_number, c.name
```



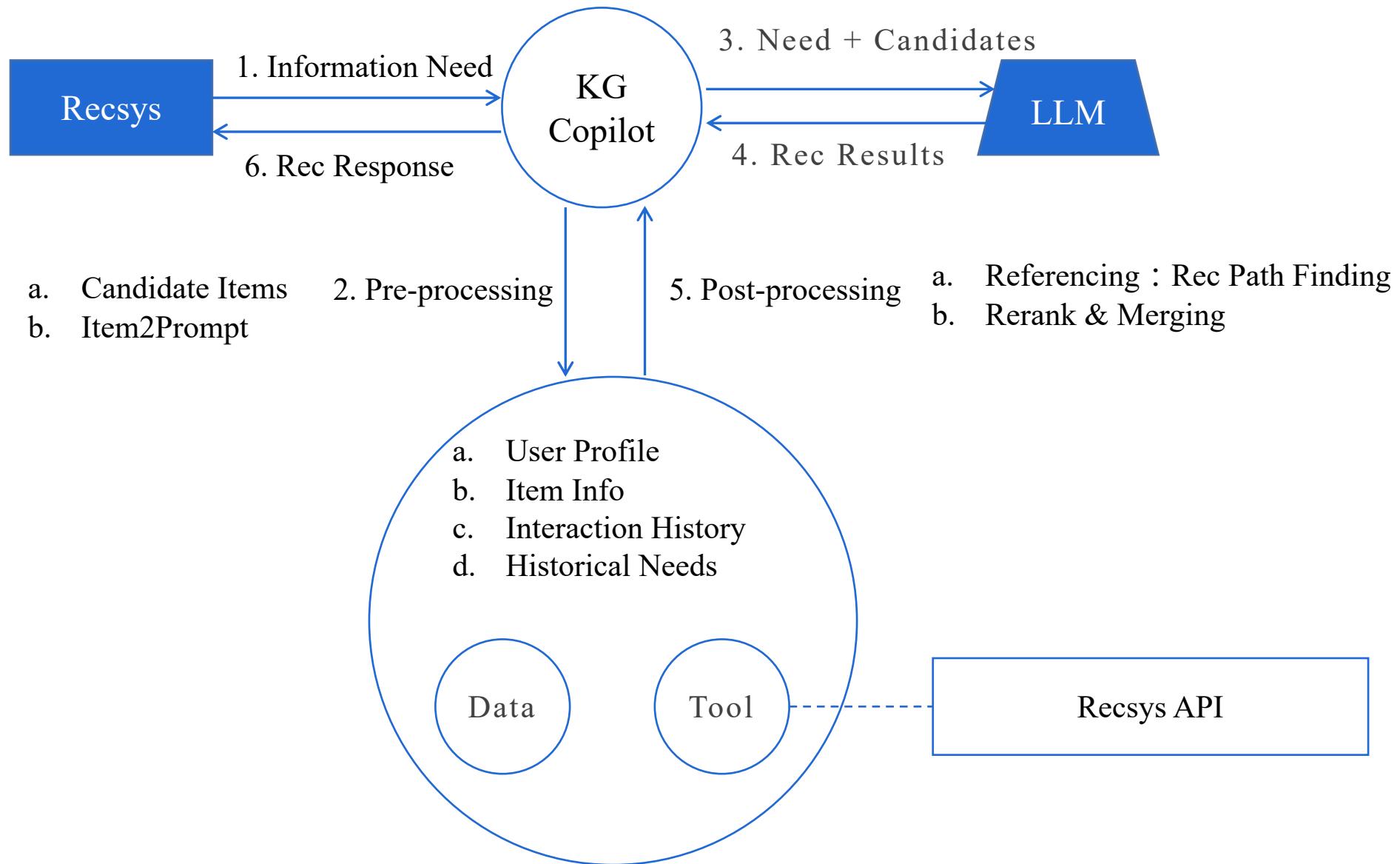
- a. Table schema Description

Data

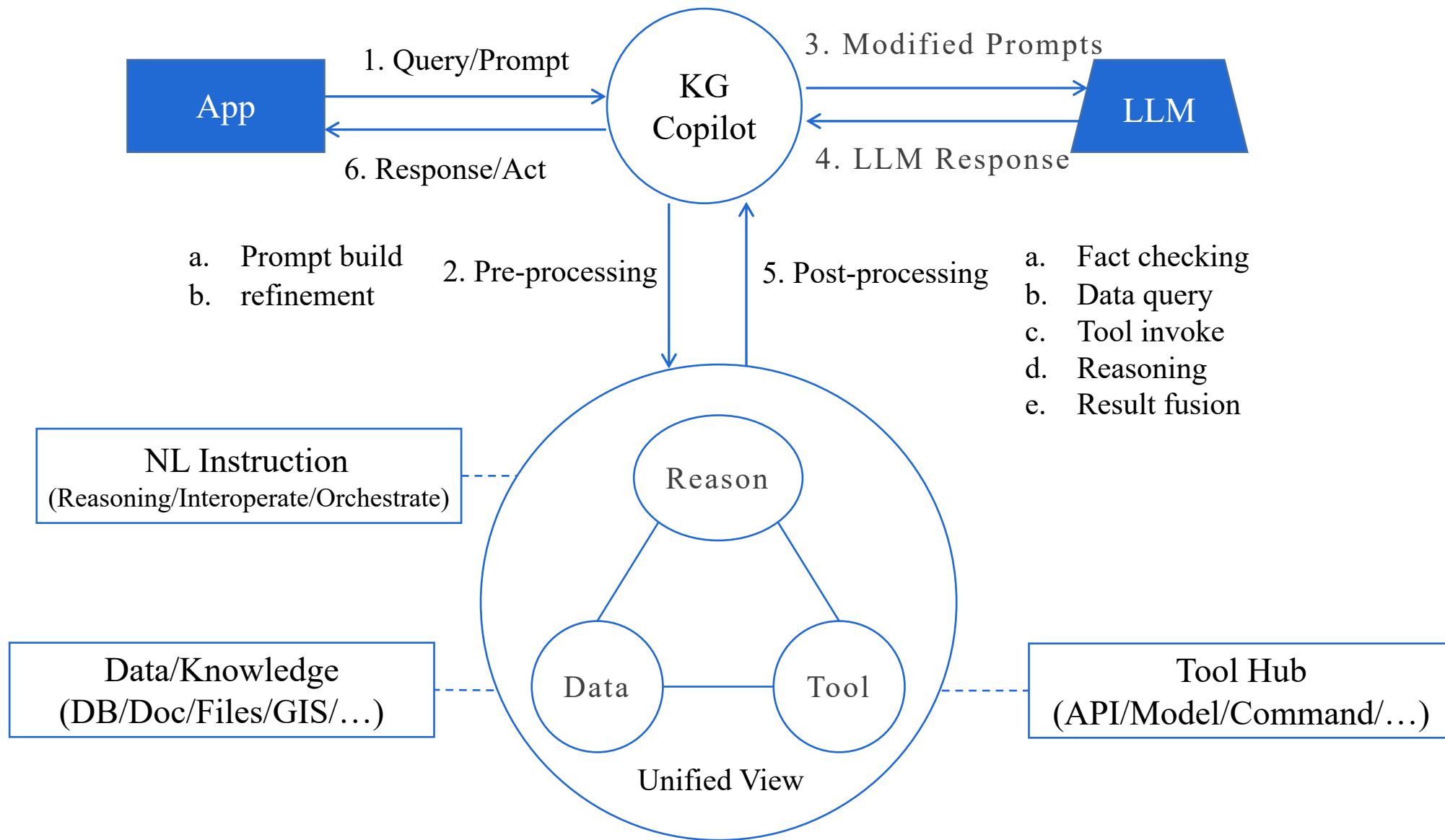
KG Copilot Engine Example: ChatWeb



KG Copilot Engine Example: ChatRec



KG Copilot Engine Example: Chat & Analyze All Data



Thank you



 OpenKG.CN
开放的中文知识图谱

<http://www.openkg.cn/>