

The Future of Digital Libraries: Serving Humans and Artificial Intelligence Agents

ChengXiang (“Cheng”) Zhai

Siebel School of Computing and Data Science

School of Information Sciences (affiliate)

Department of Statistics (affiliate)

Carl R. Woese Institute for Genomic Biology (affiliate)

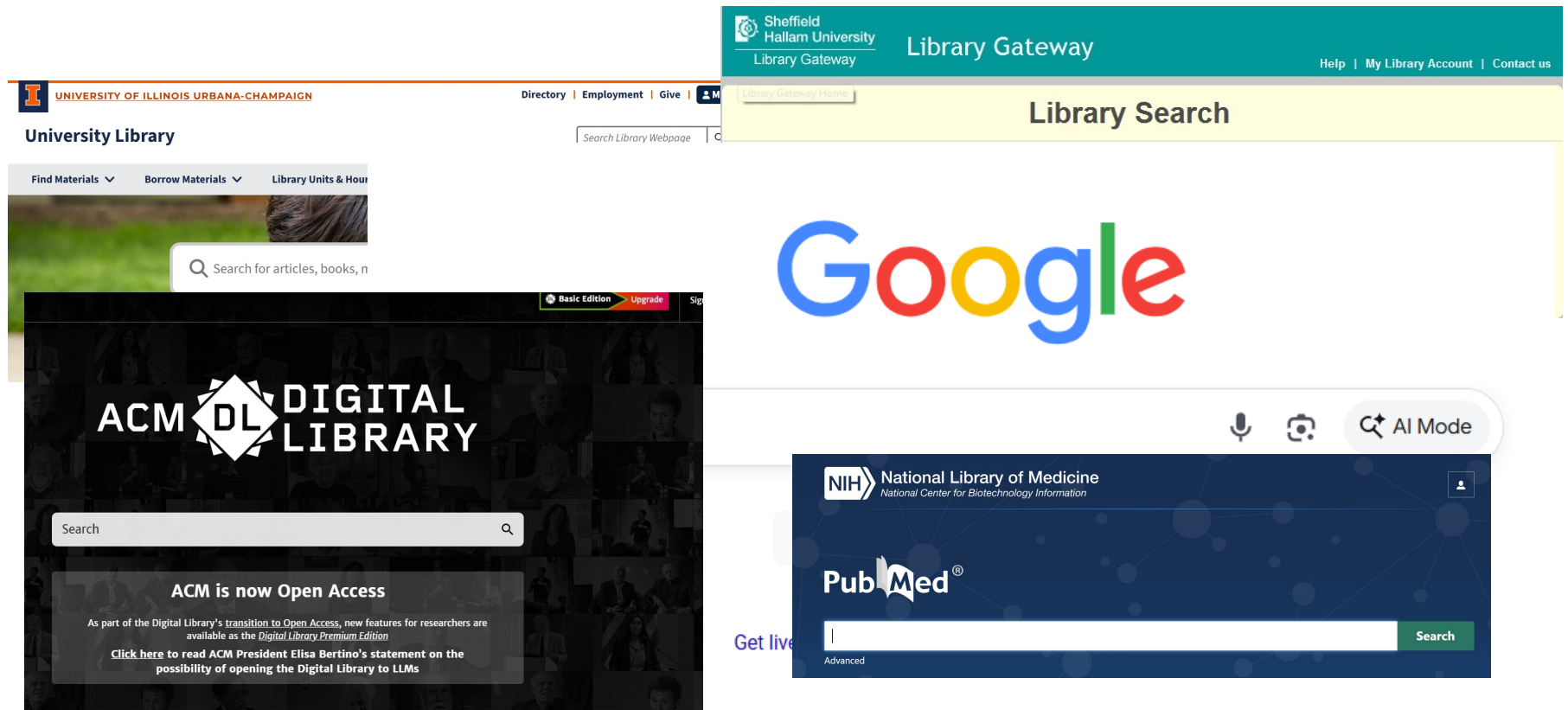
University of Illinois at Urbana-Champaign

czhai@illinois.edu

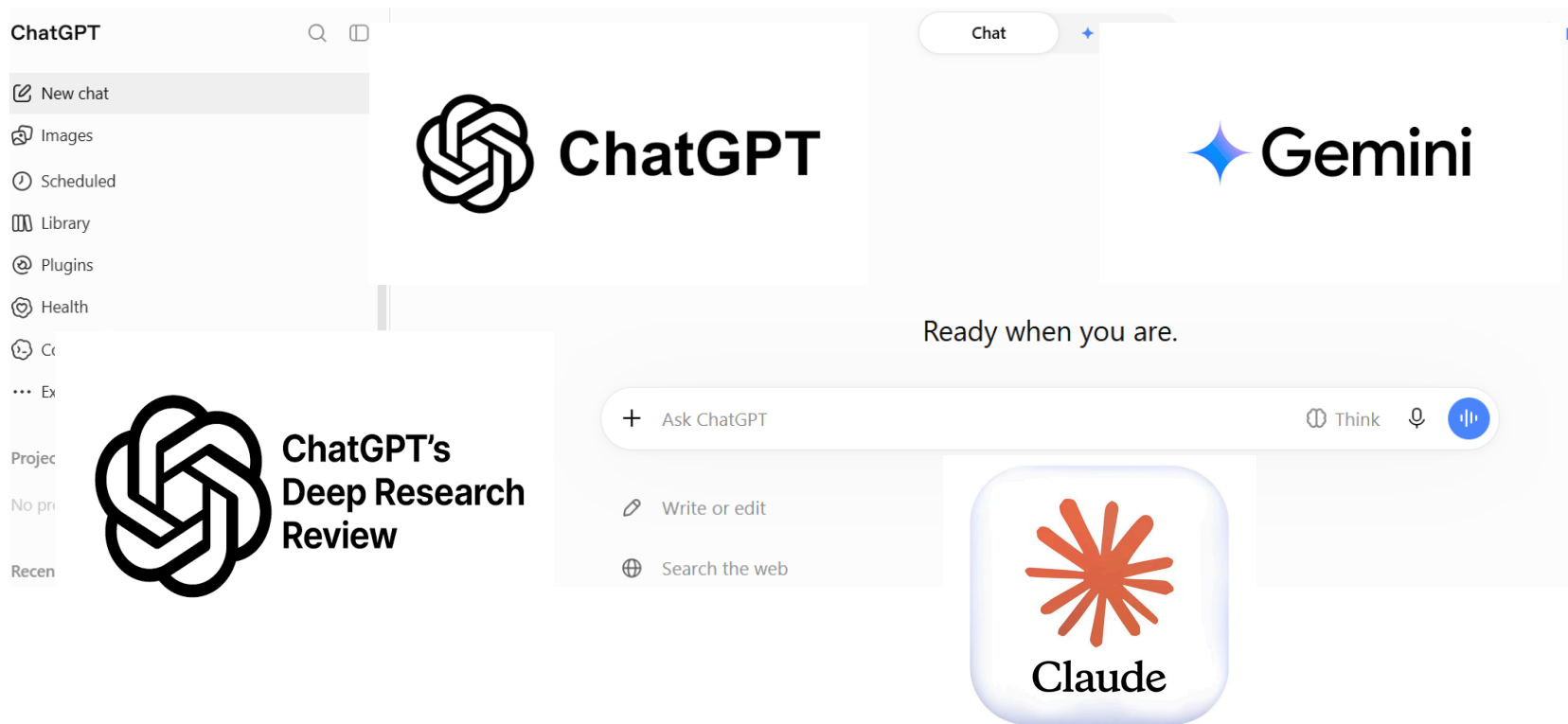
Once upon a time....



Then, Web was born ...



Now, Chatbots are available....



How is generative AI impacting digital libraries?

What's the future of digital libraries?

The Rest of the Talk

- Strengths and limitations of LLMs and agents 
- Impact of LLMs and agents on digital libraries
- Future digital library = Trustworthy Agentic Information Environment (TAIE)
- Summary

Traditional LM

vs.

Deep Neural Network LM

Directly memorize $P(\text{retrieval}|\text{information})$

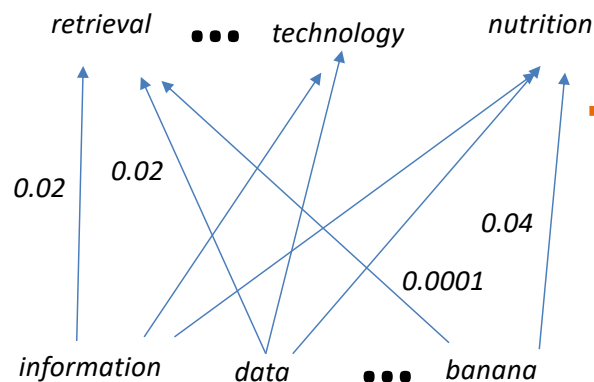
Interpretable, but redundant, poor generalization
(inefficient representation)

Construct $P(\text{retrieval}|\text{information})$ using
memorized generalizable latent associations

Non-interpretable, but compact & generalizable
(efficient representation)

Parameter size (bigram LM): $|V|^2$

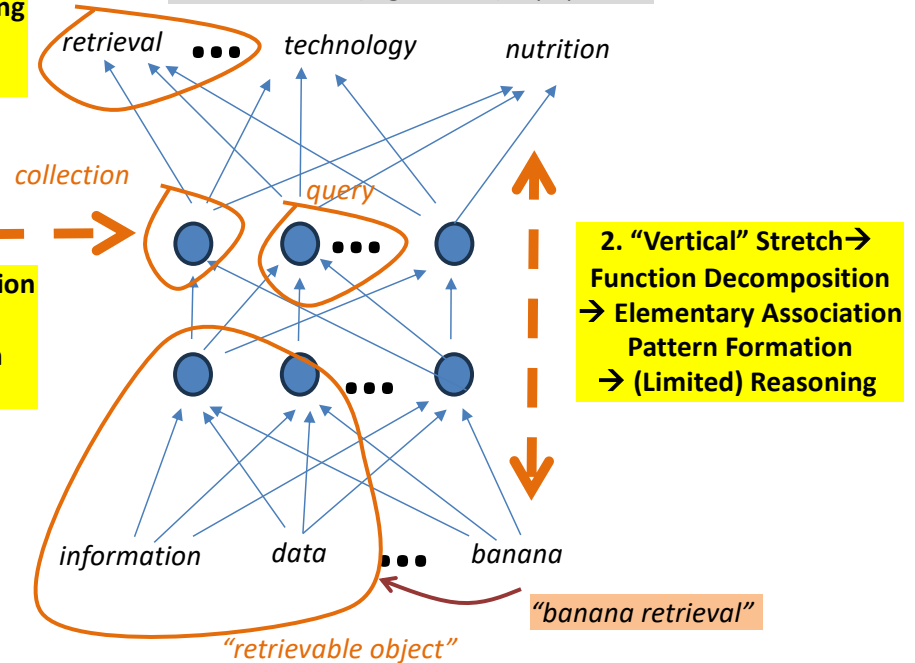
3. Self-attention & Positional encoding
→ Structure formation
→ Semantic composition



1. "Horizontal" compression
→ Clustering
→ Concept Formation
→ Generalization

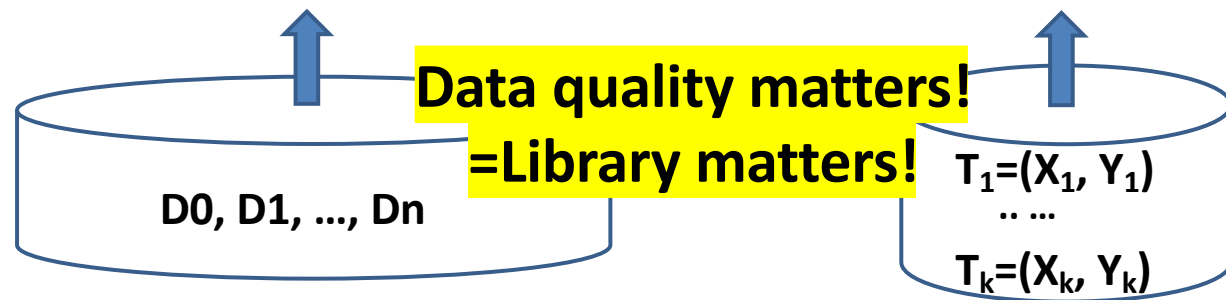
Compression → Learning → Knowledge Acquisition

Parameter size (bigram LM): $2|V|k + k^2$



LLMs: An Optimization View

$$\Phi^* = \operatorname{argmin}_{\Phi} \alpha_1 \operatorname{ReconstructError}(D_0, D_1, \dots, D_n | \Phi) + \alpha_2 \operatorname{TaskError}((X_1, Y_1), \dots (X_k, Y_k) | \Phi)$$

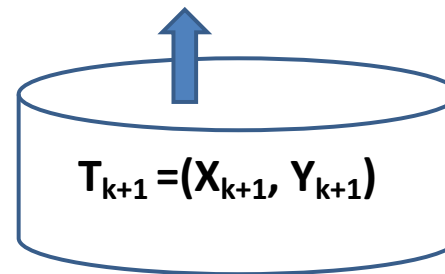


Fine-Tuning for a new task?

$$+ \alpha_3 \operatorname{TaskError}((X_{k+1}, Y_{k+1}) | \Phi)$$

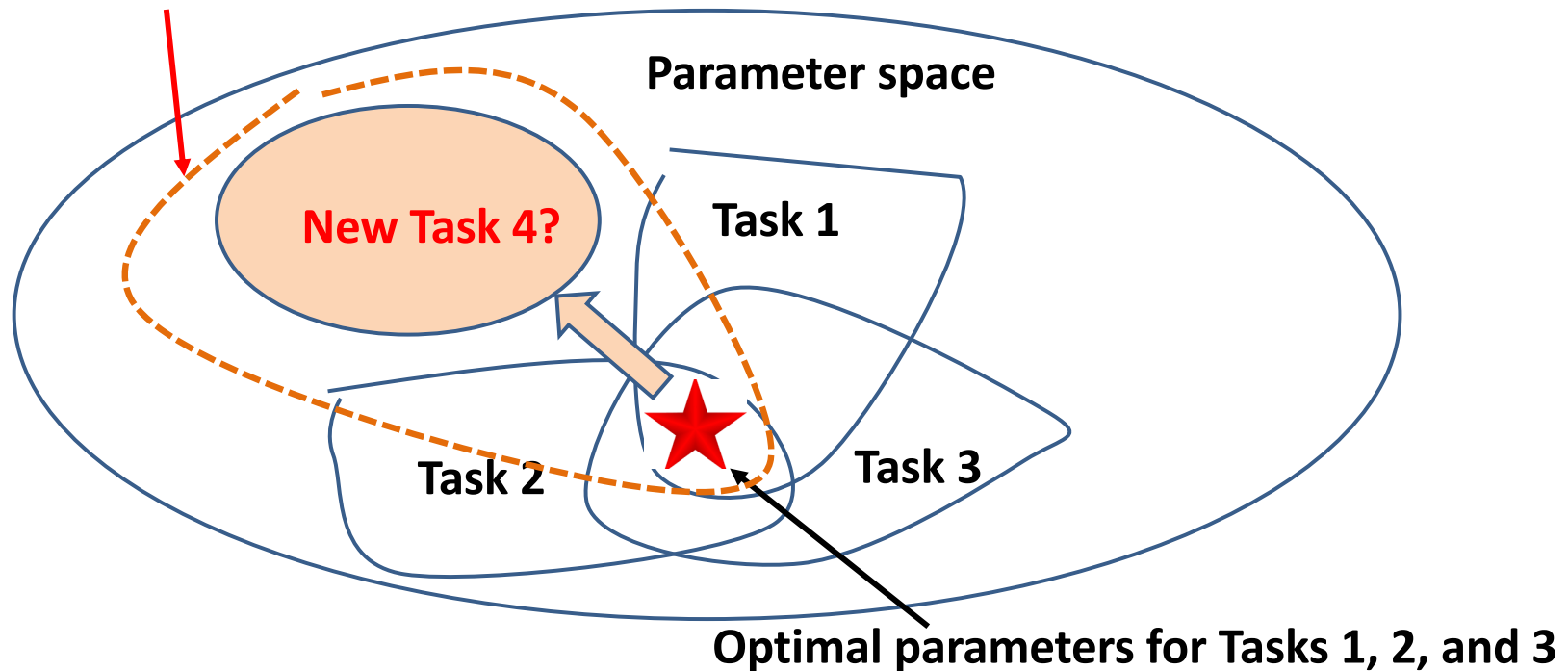
How to control and optimize α_3 ?

What if α_3 is too small or too large?



Inevitable forgetting!

Only possible if the parameter space is sufficiently large!



Limitations of Foundation Models

- Inevitable forgetting → Unpredictable behavior
- Limited context window
- Limited capacity of reasoning
- Cannot explain its behavior
- Cannot self-improve itself

ChatGPT can't do addition!

Show the detail steps of adding each pair of digits.

Great question — let's do this slowly, cleanly, and checkable, adding each pair of digits from right to left, just like grade-school column addition.

We are adding:

A = 10100100000001000000000000000001

B = 20202000000000001000001

Step 1: Align the numbers (right-justified)

10100100000001000000000000000001
+ 000000202020000000000001000001

Step 2: Add digit by digit from right to left

I'll show each column as:

(digit from A) + (digit from B) + carry → result digit, new carry

Starting from the rightmost digit:

1 + 1 + 0 = 2, carry 0	1 + 1 + 0 = 2, carry 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 1 + 0 = 1	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 0 + 0 = 0
0 + 0 + 0 = 0	0 + 1 + 0 = 1

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 0 + 0 = 0

1 + 0 + 0 = 1

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 0 + 0 = 0

0 + 2 + 0 = 2

0 + 0 + 0 = 0

0 + 2 + 0 = 2

0 + 0 + 0 = 0

1 + 2 + 0 = 3

0 + 0 + 0 = 0

1 + 0 + 0 = 1

0 + 0 + 0 = 0

1 + 0 + 0 = 1

= 0

= 3

= 0

= 1

= 0

= 1

• No carries occur at all, because no column exceeds 9.

Step 3: Read the result from top to bottom

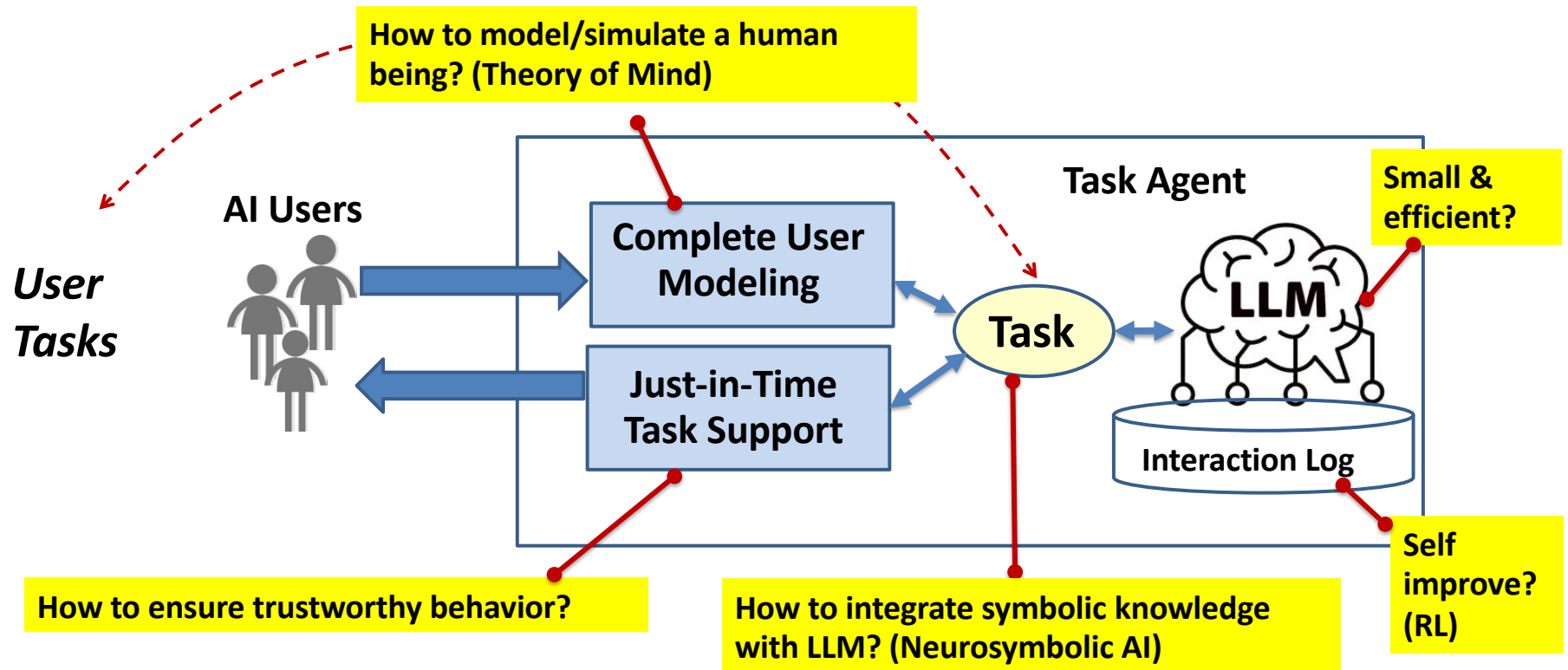
Putting all result digits together:

1010011002020100000001000002

Agent to Address Limitations of Foundation Models

- Inevitable forgetting: **Persistent memory + Regulated fine-tuning**
- Limited context window: **Working memory** to provide a larger buffer
- Limited capacity of reasoning: Reliable reasoning using **symbolic knowledge** (reasoning as **tool use in memory**)
- Cannot explain its behavior: **interpretability** and **explicit knowledge representation**
- Cannot self-improve itself: **reinforcement learning** from **episodic memory**

Five General Challenges in AI



Useful Resources

- User Simulation: <https://usersim.ai/>
- Trustworthy AI (recommender systems):
 - <https://fairness-tutorial.github.io/>
 - <https://arxiv.org/html/2205.13619v6>
- Neurosymbolic AI
 - <https://arxiv.org/abs/2501.05435>
 - <https://www.youtube.com/@neurosymbolic>
- Autonomous Agents: <https://arxiv.org/abs/2607.13104>
- Small LMs and light-weight agents:
<https://research.nvidia.com/labs/lpr/slm-agents/>

Agent = Foundation Model + Memory + Tools

- Adding memory and tools on top of a foundation model creates a powerful agent that can
 - interact with a task environment that includes users
 - maintain a current state representing its understanding of the current environment and a user's need
 - take an optimal action based on the current state to optimize the expected reward
 - learn continuously from its interactions with the environment to improve its reward function and behavior

How can we build an agent (foundation agent)?

- Agent memory management: working memory, episodic memory: <https://arxiv.org/abs/2512.13564>
- Tools learning: <https://arxiv.org/abs/2405.17935>
- Neurosymbolic architecture:
<https://www.sciencedirect.com/science/article/abs/pii/S1574013726000110>
- Harness to ensure safety: <https://picrew.github.io/LLM-Harness/>
- Continual learning: <https://arxiv.org/abs/2302.00487>

Outline

- Strengths and limitations of LLMs and agents
- Impact of LLMs and agents on digital libraries 
- Future digital library = Trustworthy Agentic Information Environment (TAIE)
- Summary

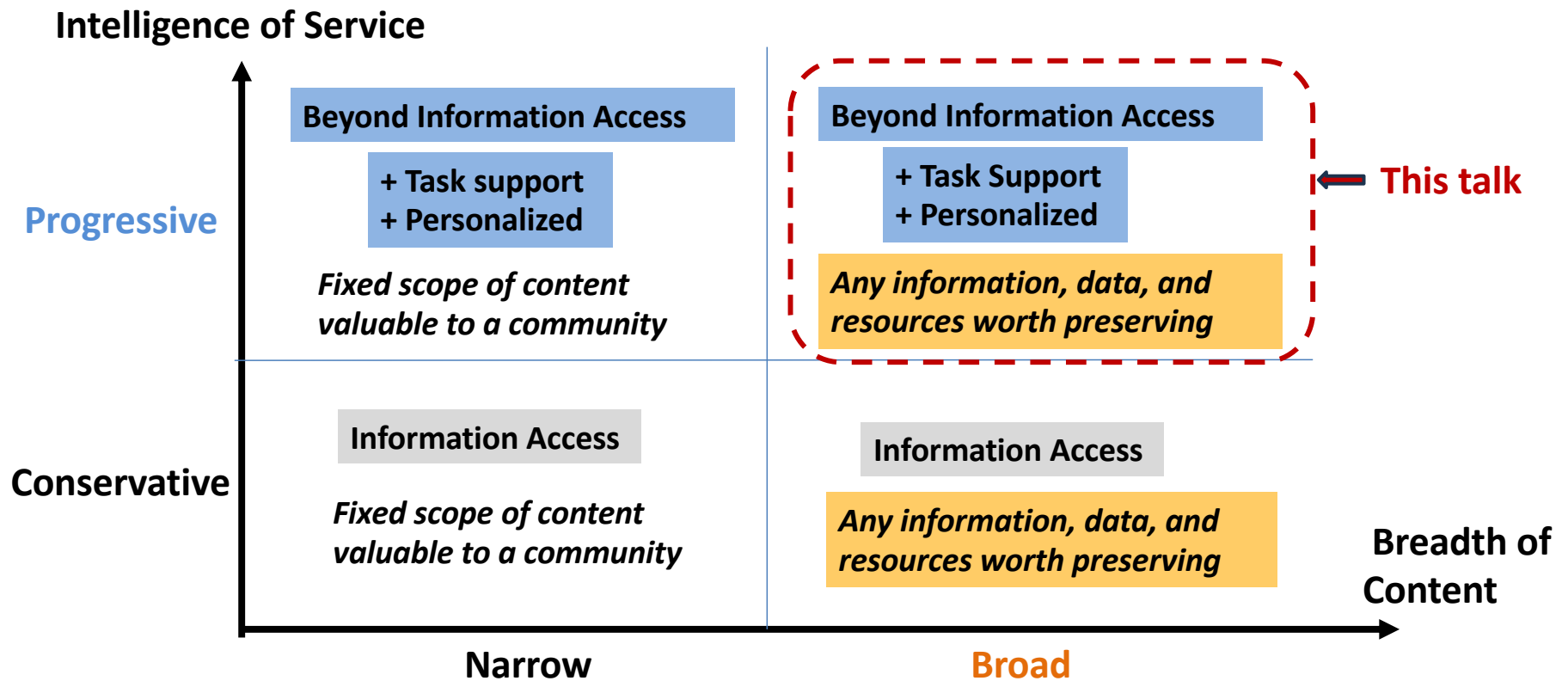
How has generative AI impacted digital libraries?

- **Interface displacement:** Users interact with ChatGPT rather than library search interfaces
 - Library collections may still supply information to AI systems, but the libraries themselves may lose control over how their content is discovered, interpreted, attributed, and trusted (no feedback from users).
- **Authority displacement:** users may trust synthesized answers without consulting the original sources or evaluating their provenance.
- **Value-capture displacement**
 - Digital libraries risk becoming invisible infrastructure in an AI-mediated information ecosystem.
 - AI platforms capture user relationships and attention while libraries bear the costs of collection development, curation, and preservation.

Questions

- Will digital libraries remain destinations for information seekers, or become invisible knowledge infrastructure for AI agents?
- How do we define a library in the generative AI era?
 - Different answers may lead to a different future of digital libraries

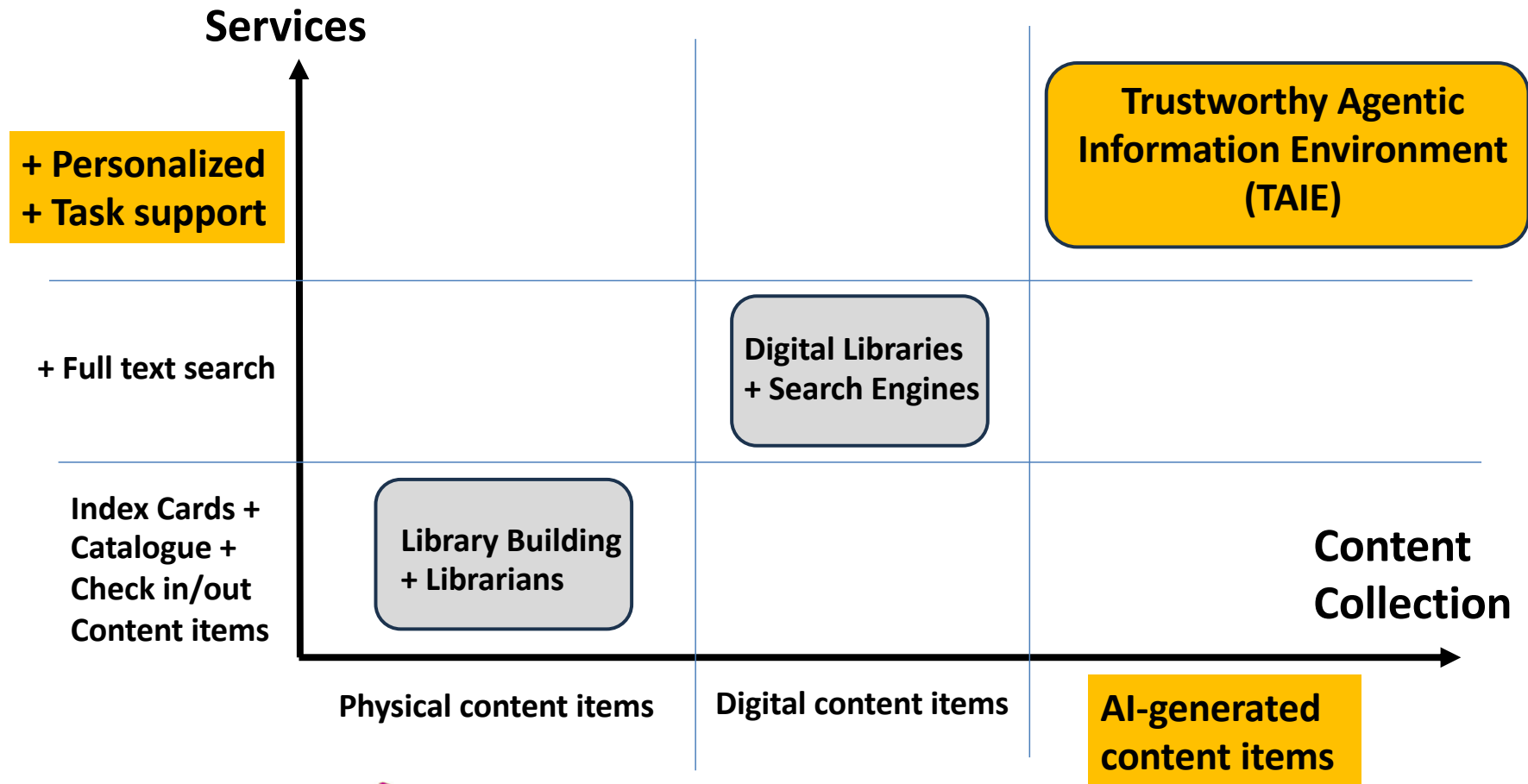
Definitions of Library: Narrow vs. Broad & Conservative vs. Progressive



Why a broad and progressive definition?

- **Progressive** → Higher system intelligence → **Higher utility** to users
- **Broad** → Higher coverage of content → **Higher utility** to users
- Core missions of a library require (preservation + quality assurance) can only be sustained by providing library services that users would *actually* use (**unused services** → inefficient use of resources → **unsustainable**)
- Users prefer direct task support (e.g., question answering) to minimize their effort → **Library-based AI agents are necessary**

Future Digital Library?



Outline

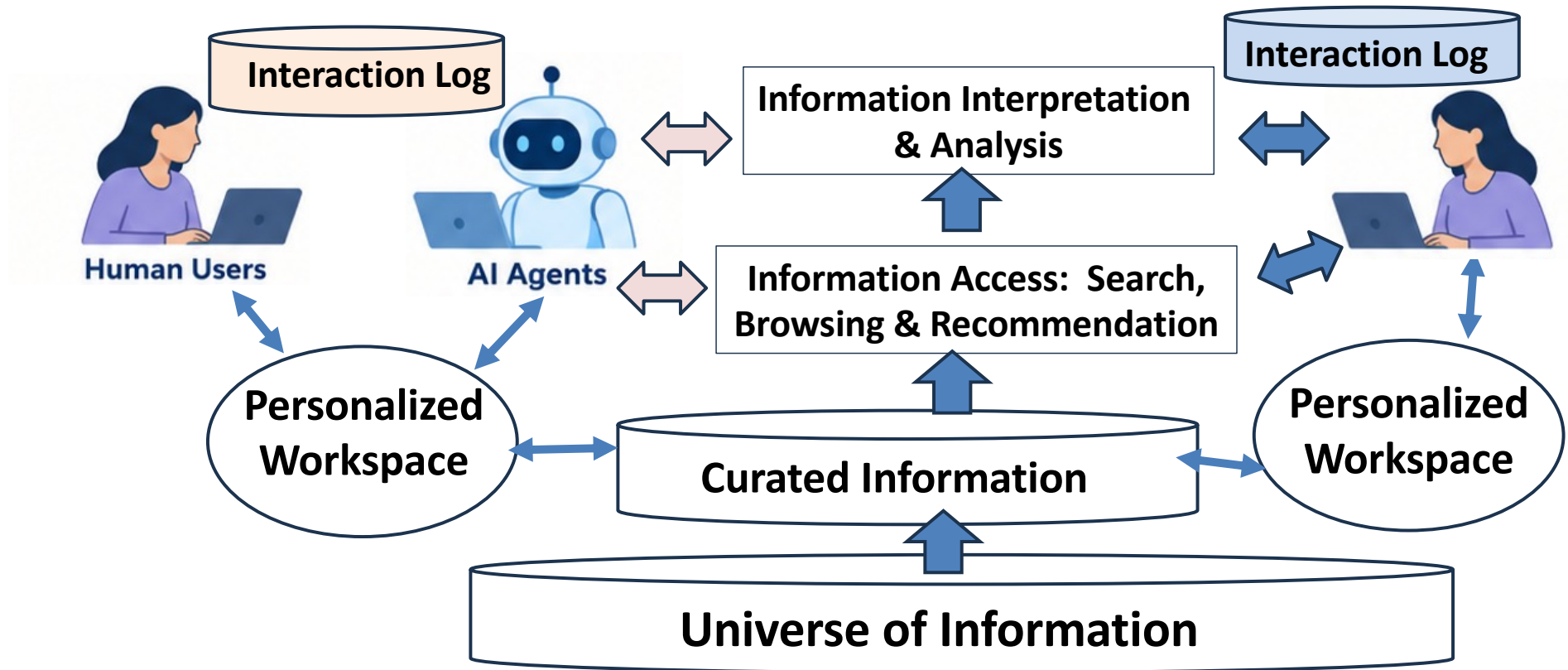
- Strengths and limitations of LLMs and agents
- Impact of LLMs and agents on digital libraries
- Future digital library = Trustworthy Agentic Information Environment (TAIE)
- Summary



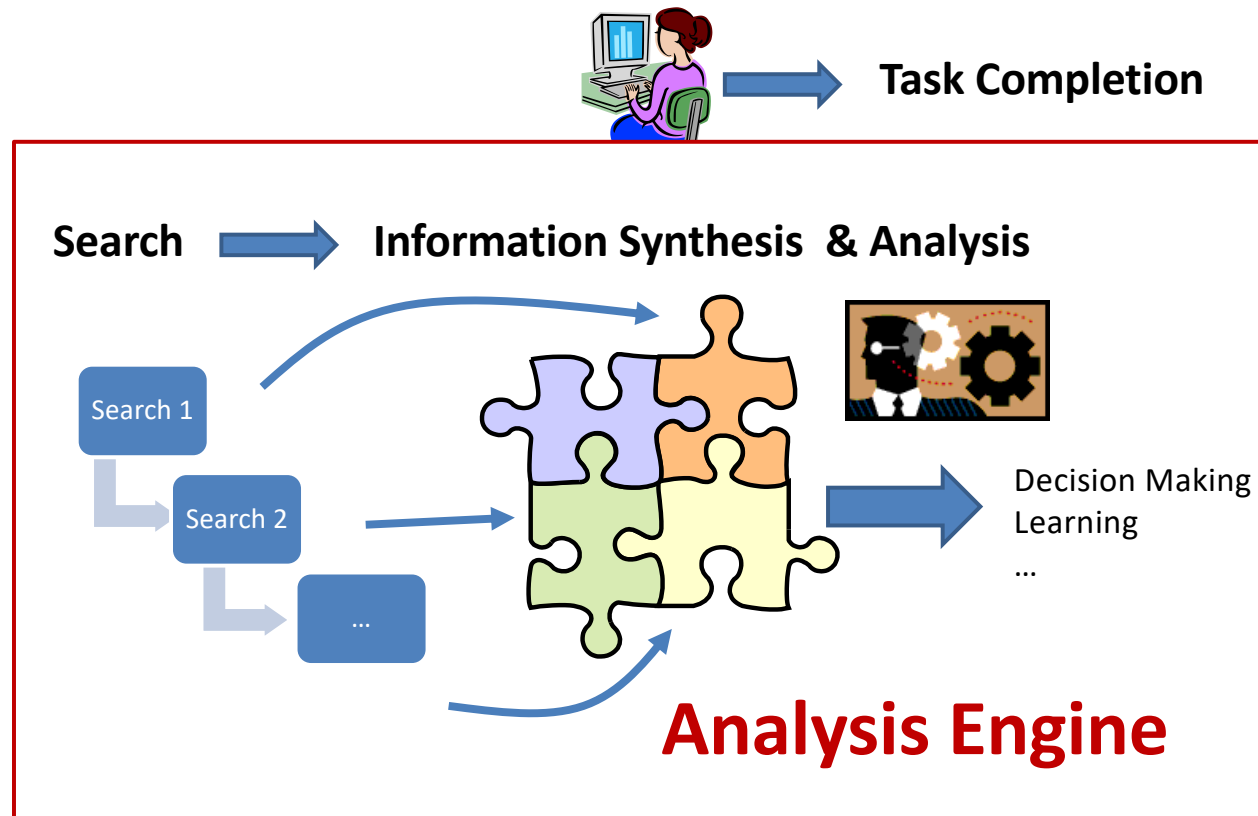
Trustworthy Agentic Information Environment (TAIE)

- **Trustworthiness:** Curated Collections + Transparent/Interpretable Services → **quality assurance**
- **Agentic:** Conversational Interface + Personalization + Self-Improvement → **regain user relationship, personalization**
- **Information:** Any information/data worth preserving
- **Environment:** Extensible set of tools/agents that can be combined to support a flexible workflow for information seeking, interpretation, analysis, and user tasks (multi-agent system)
→ **optimization of AI-user collaboration**

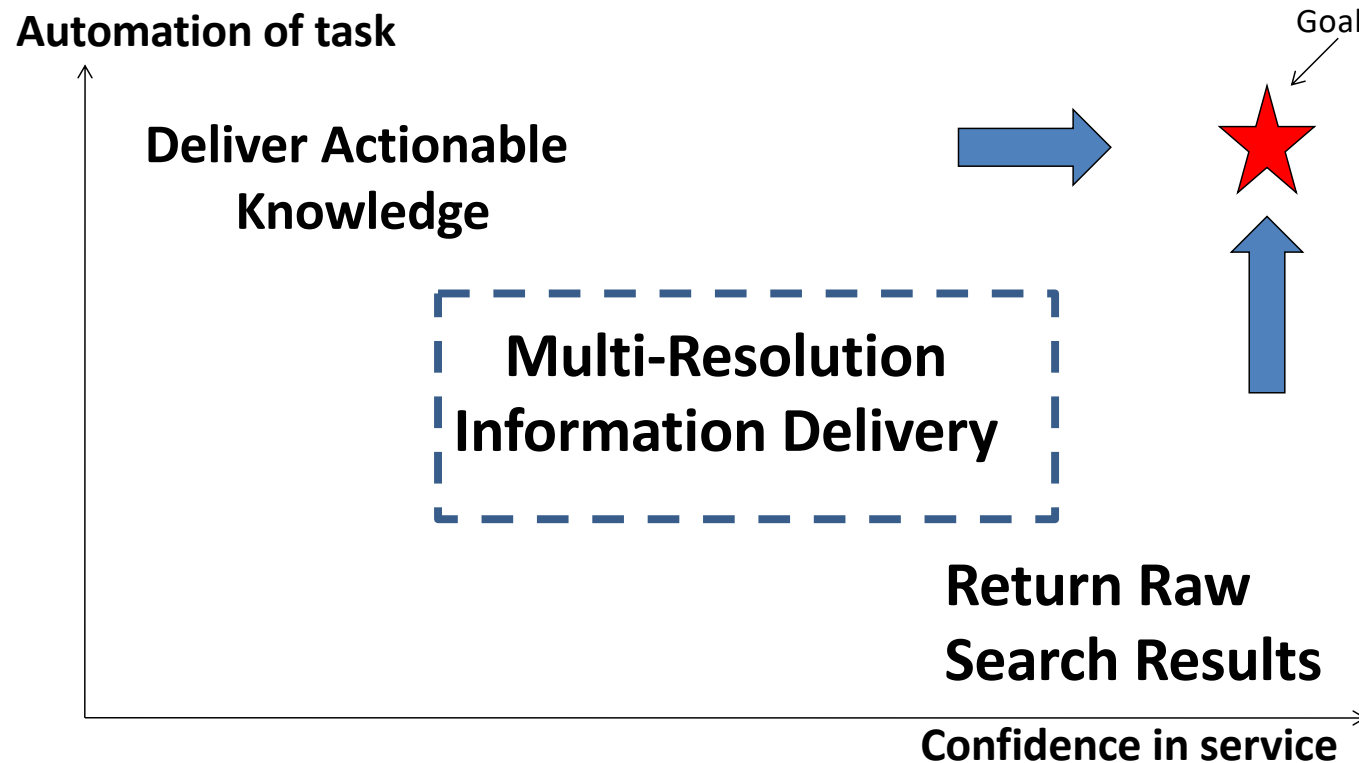
Architecture of TAIE: Serving (Multiple) Humans and AI Agents



Beyond Search: TAIE as a General Analysis Engine



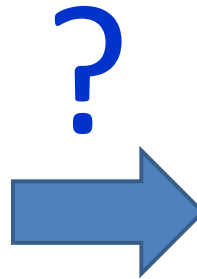
Challenge 1: Automation-Confidence Tradeoff



Automation-Confidence Tradeoff: Dining Analogy

Serve Raw-Food

Need further processing,
but flexible for making different dishes

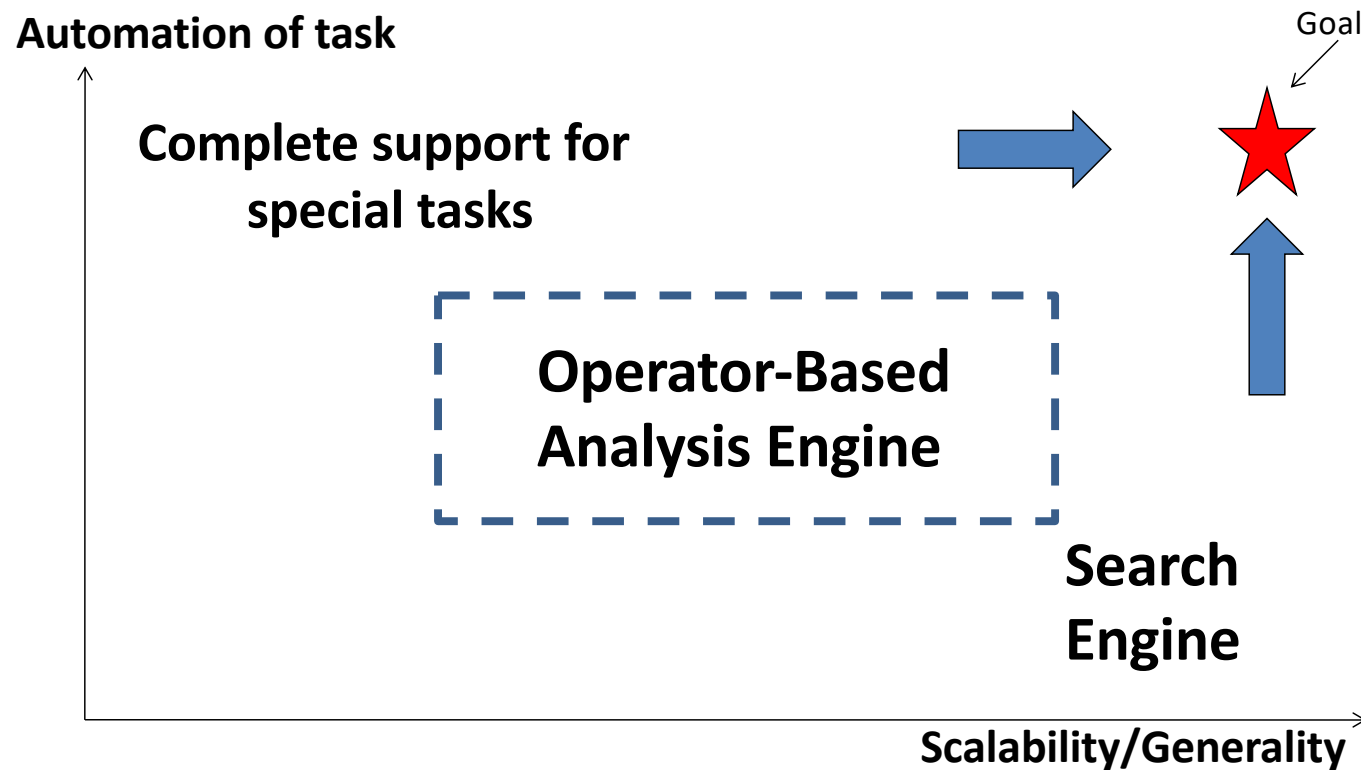


Serve Cooked Dishes

Directly useful for a task,
But would be worse if it's not the right dish



Challenge 2: Automation-Generality Tradeoff



Automation-Generality Tradeoff: Dining Analogy

What's the right paradigm? Need both paradigms?

Buffet Paradigm

Basic Components + Infinite Combination

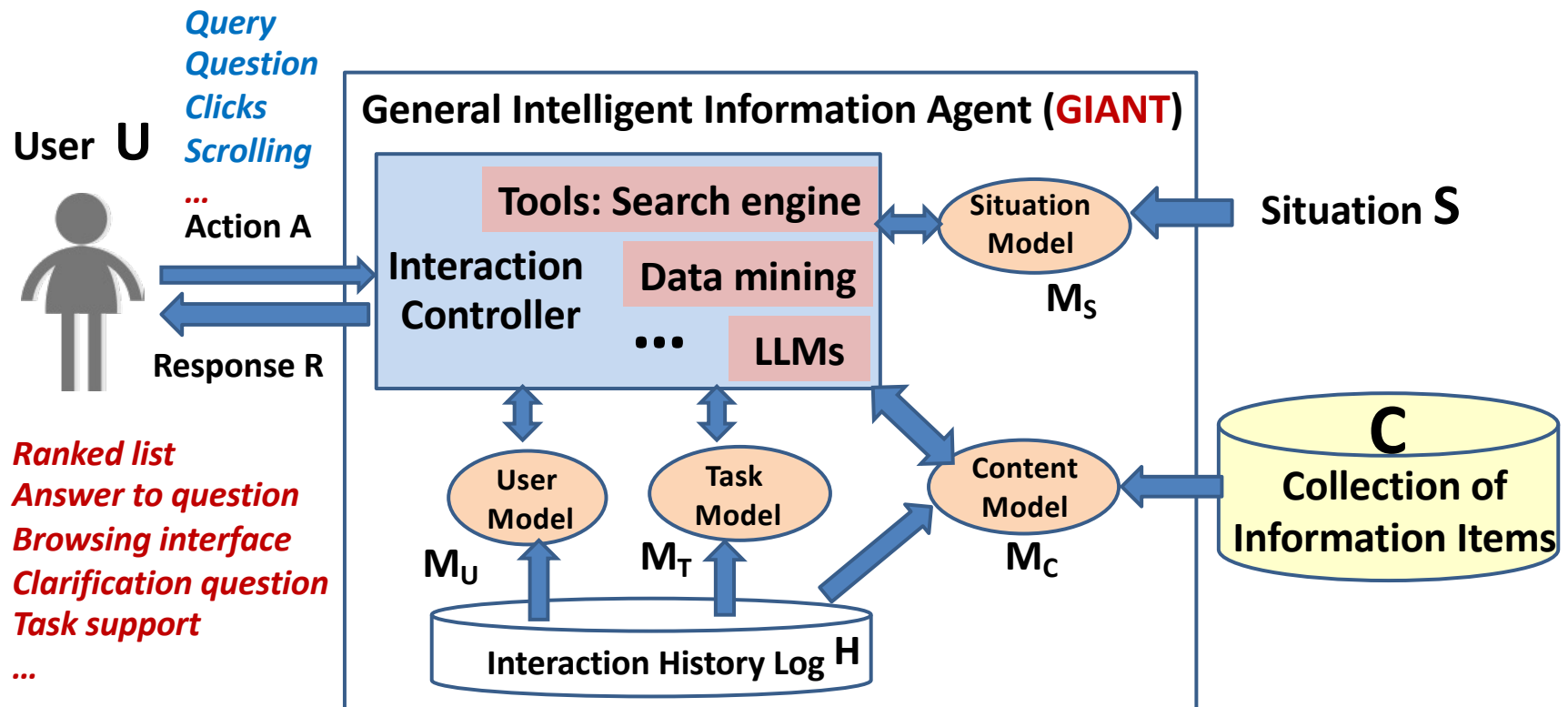


Food Court Paradigm

Finite Choices of Complete Packages



TAIE as a General Intelligent Information Agent (GIANT) [Zhai 2026]



Chengxiang Zhai. 2026. Towards a General Intelligent Information Agent: Framework, Models, and Evaluation. In Proceedings of ACM SIGIR 2026. <https://doi.org/10.1145/3805712.3808543>

A General IR Framework to Unify Search & Recommendation?

[Belkin & Croft CACM 1992]

Information
FILTERING
and Information
RETRIEVAL:
Two Sides of the
Same Coin?

We conclude that information retrieval and information filtering are indeed two sides of the same coin. They work together to help people get the information needed to perform their tasks. 

What is this “coin”?

What is an information retrieval (IR) system?

- IR systems = Search engines?
- IR systems = Search engines + Recommender Systems?
- IR systems = Search engines + Recommender Systems + QA systems?
- IR systems = Conversational assistants?
- IR systems = Intelligent information agents?

What is **the most general definition** of an IR system?

Why General Information Retrieval (GIR)?

- A **unified formulation** of an IR system **accelerates IR research and applications**
 - It provides a framework for developing **general models** and **algorithms** that are **applicable to all specific applications**
 - It enables the development of a **general IR system** that can be deployed in **many different application domains**
 - It enables a **unified evaluation methodology** for **accelerating research and development**

Three Key Questions

- **Framework:** How can we **formalize IR problems** with a unified framework for solving all the IR problems?

Probability Ranking Principle

→ GIR Task = Cooperative Game

- **Models:** How can we design **general models and algorithms** and implement a general system for solving the formulated IR problem?

Vector Space/Prob. Model, L2R
(Inverted Index, Lucene, Solr, ...)

→ GIR Algorithm = Sequential Interface Optimization
(LLMs, memory, agents, tools, GIR toolkits)

- **Evaluation:** How can we evaluate an IR system?



Cranfield

→ GIR Evaluation = Measures on Interactions with Simulated User

Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Found. Trends Inf. Retr.* 18, 1–2 (Jun 2024), 1–261.

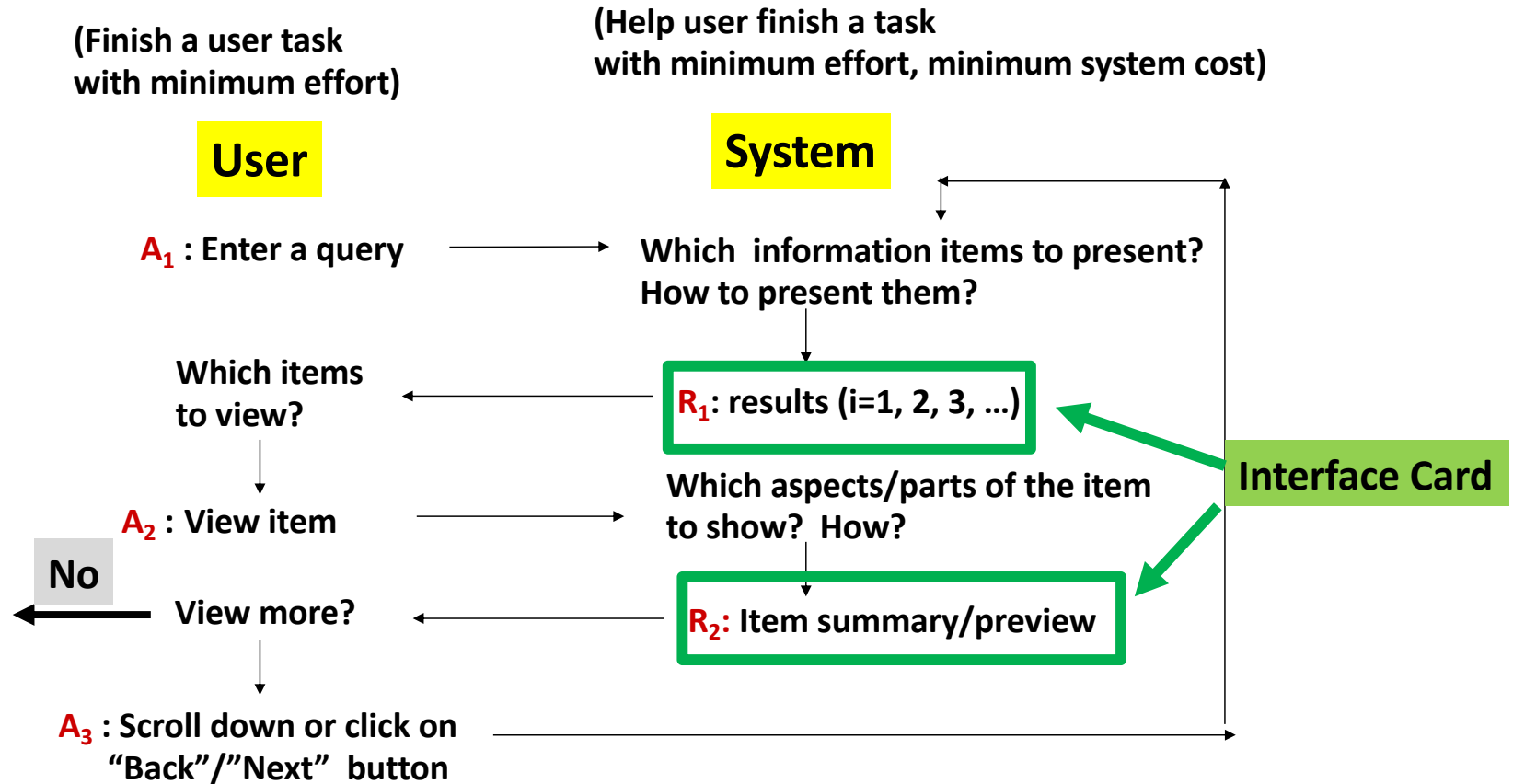
<https://usersim.ai/>



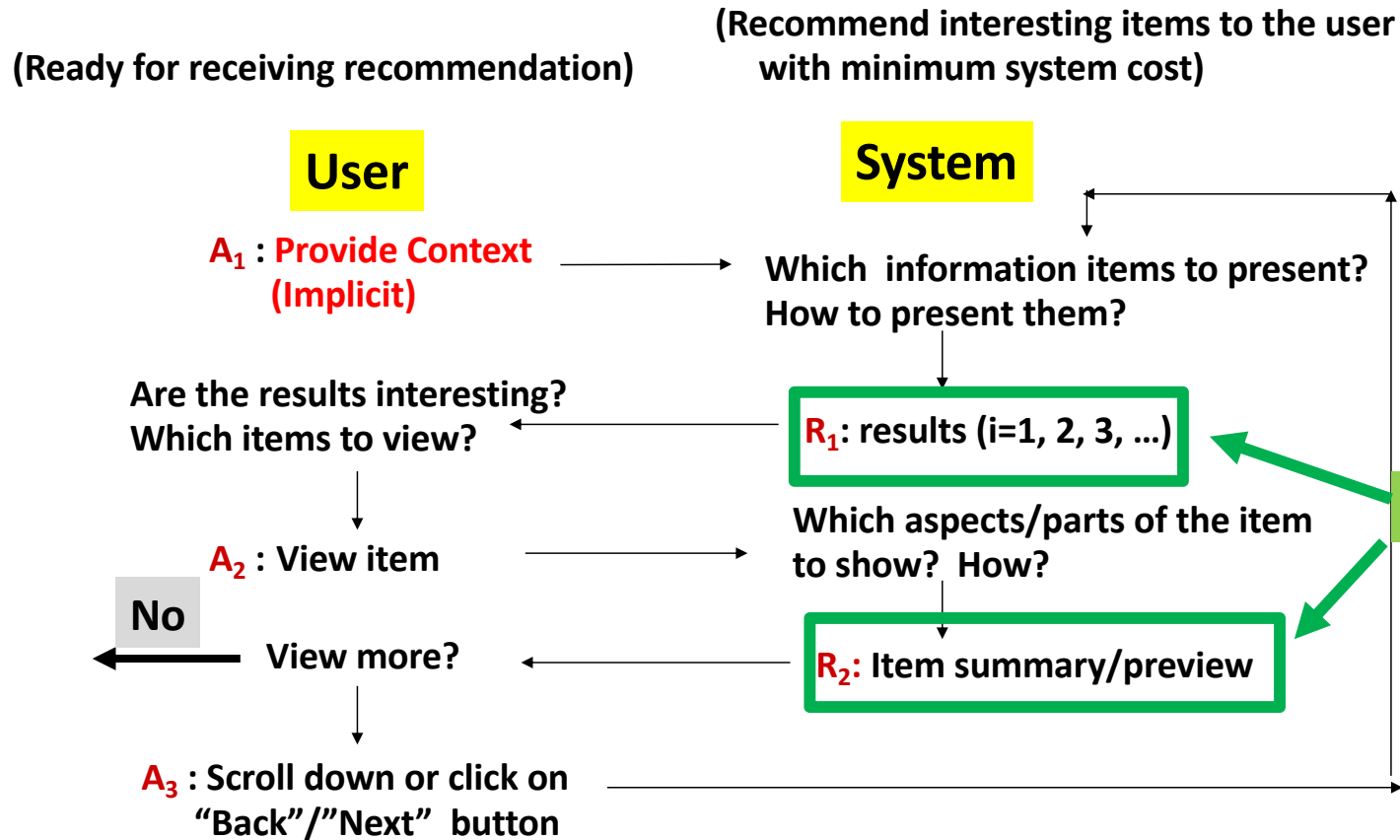
GIR = Cooperative Game-Playing

- **Players:** Player 1= IR System; Player 2= User
- **Rules of game:**
 - Players **take turns to make “moves”**
 - E.g., first move = “user entering the query” (in search) or “system recommending information” (in recommendation)
 - For each move of the user, the system makes a response move (shows an **interaction interface**)
 - Depending on the interaction interface, the user makes a subsequent move
 - User makes the last move (usually)
- **Objective:** help the user **complete a task** with **minimum effort & minimum operating cost** for the IR system

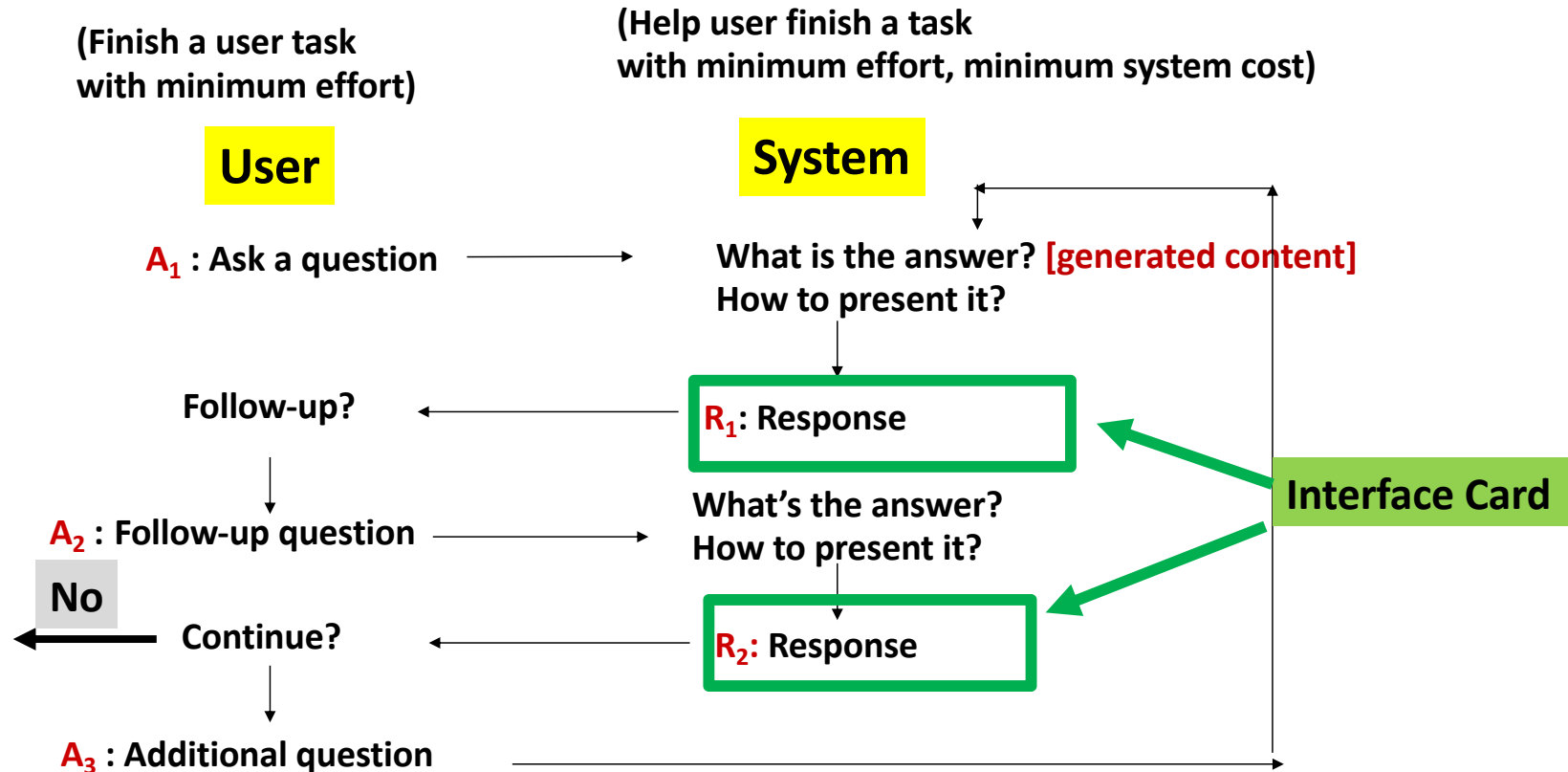
Search as cooperative game-playing



Recommendation as cooperative game-playing



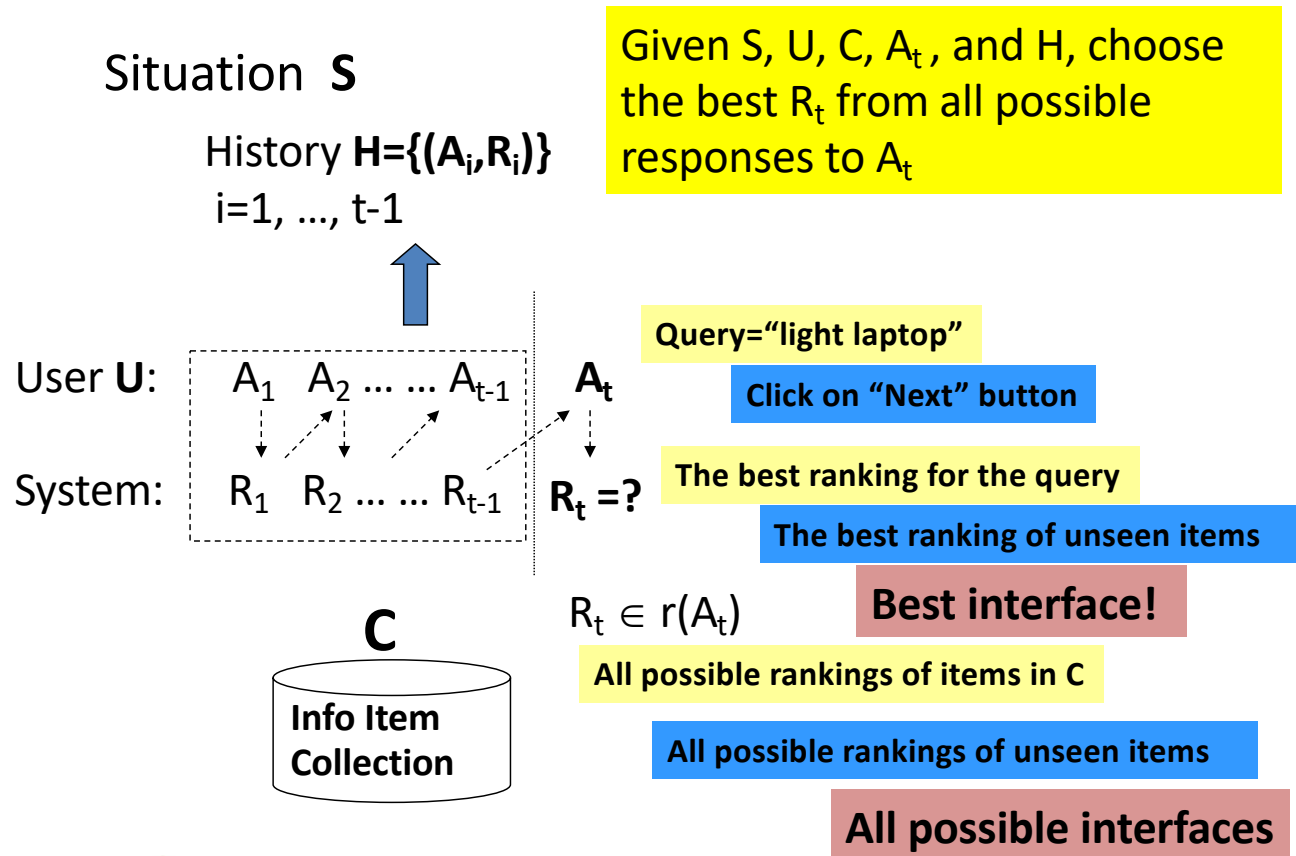
Question Answering as cooperative game-playing



4 Key Elements of the GIANT Framework (4 C's)

- **Collaboration:** Optimization of the collaboration (or **combined intelligence**, combined performance) of a user and a system
 - User knows well about what's useful, but doesn't know the whole information space
 - System "sees" the whole information space, but doesn't know which is most useful to the user
- **Communication:** Optimization of the **two-way communications** between a user and a system
 - Communication of the shared goal and plan
 - Explanation of both user actions and system responses
- **Cognition:** Optimization of cognition for user (**user learning**) and "cognition" for system (**machine learning**)
 - Modeling of knowledge state and helping users learn during search
 - Helping system learn knowledge about tasks and relevance from its users
- **Cost:** Optimization of system operation cost
 - Modeling operation cost and providing **cost-effective responses**

Optimal IR with GIANT = Sequential Interface Optimization

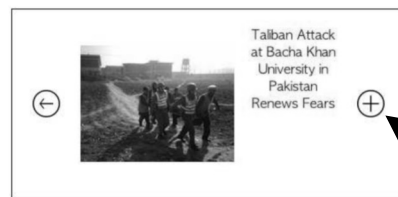


Example of interface optimization [Zhang & Zhai 15, 16]

Context c^t

After a user clicks on “Colleges & Universities”,
which interface card q^t to show?

If showing card q^t



If user action a^{t+1} = view content

surplus for a^{t+1} :

$$u(a^{t+1} | q^t, c^t) = \text{gain} - \text{cost}$$

If user action a^{t+1} = “see more”?

A different card



If user action a^{t+1} = “navigate”?

Zhang, Yinan, and Chengxiang Zhai. "Information retrieval as card playing: A formal model for optimizing interactive retrieval interface." In *SIGIR'15* pp. 685-694. 2015.

Zhang, Yinan, and Chengxiang Zhai. "A sequential decision formulation of the interface card model for interactive IR." *SIGIR'16*, pp. 85-94. 2016.



Expected surplus of an interface card: $E(u^t | q^t, c^t)$

$$E(u^t | q^t = \text{[card icon]}, c^t) \\ = p(a^t = \text{"view content"} | c^t, q^t) \times u(\text{[card icon]} | c^t, q^t)$$

$$u(\text{[card icon]} | c^t, q^t) = \text{Gain}(\text{[card icon]}) - \text{Cost}(\text{Viewing}) \\ \text{Gain}(\text{[card icon]}) = \text{Relevance}(\text{[card icon]})$$

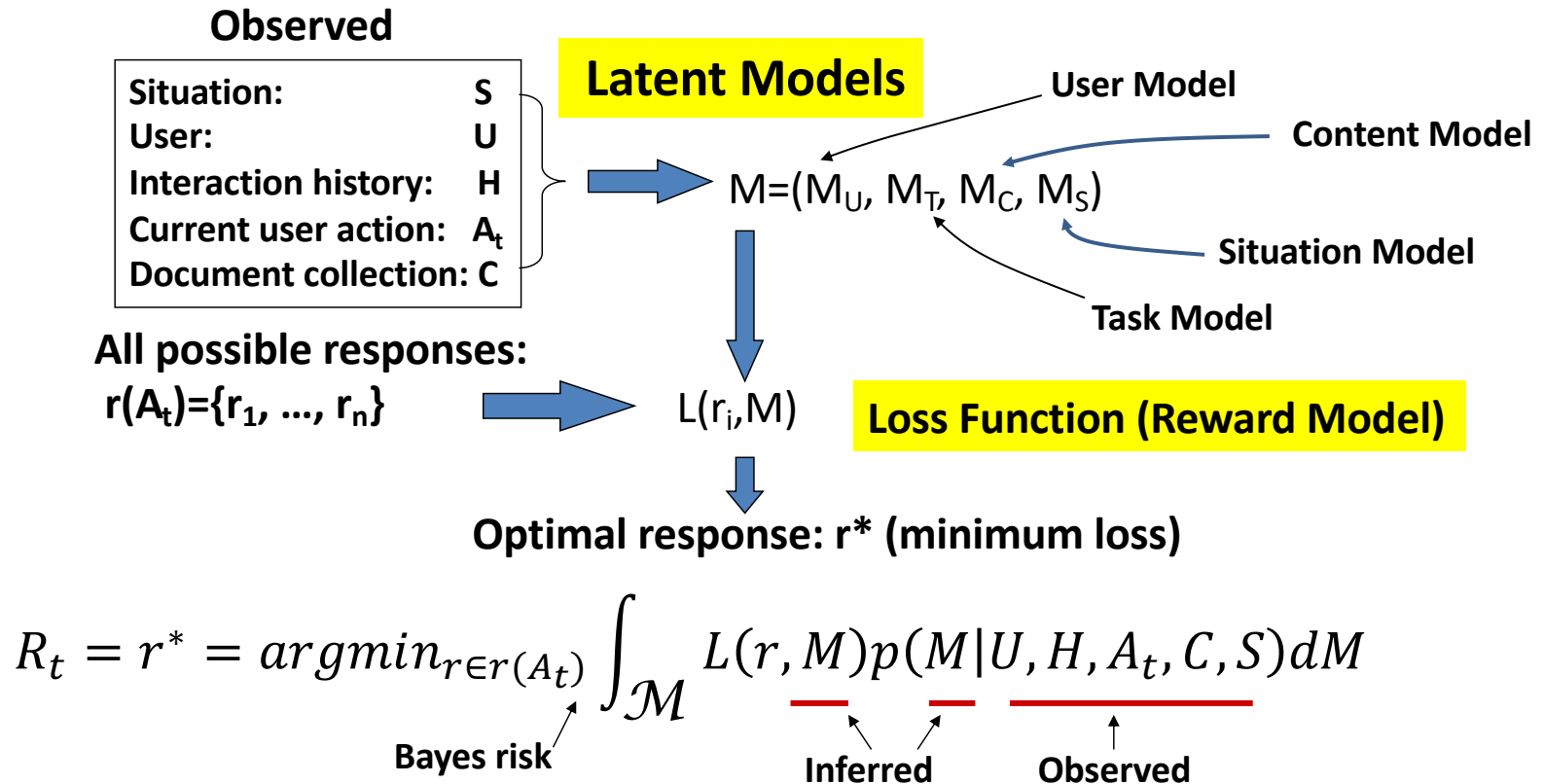
$$+ p(a^t = \text{"see more"} | c^t, q^t) \times u(\text{[card icon]} | c^t, q^t) + \dots$$

Depends on the next interface card q^{t+1}

Expected surplus of an interface card: $E(u^t | q^t, c^t)$

$$\begin{aligned}
 E(u^t | q^t = & \text{[Card Template: Oxford University, Coalition for Access, Affordability and Success, School Shootings and Armed Attacks]} , c^t) \\
 = & p(a^t = \text{"left-top tag"} | c^t, q^t) \times u(\text{[Card Template: Oxford University, Coalition for Access, Affordability and Success, School Shootings and Armed Attacks]} | c^t, q^t) \\
 & + p(a^t = \text{"right-top tag"} | c^t, q^t) \times u(\text{[Card Template: Oxford University, Coalition for Access, Affordability and Success, School Shootings and Armed Attacks]} | c^t, q^t) \\
 & + p(a^t = \text{"left-bottom tag"} | c^t, q^t) \times u(\text{[Card Template: Oxford University, Coalition for Access, Affordability and Success, School Shootings and Armed Attacks]} | c^t, q^t) \\
 & + \dots
 \end{aligned}$$

Bayesian Decision Theory for Optimal GIR [Zhai 16]



ChengXiang Zhai. Towards a game-theoretic framework for text data retrieval, IEEE Data Eng. Bull. 39(3): 51-62 (2016).

Simplification of Computation

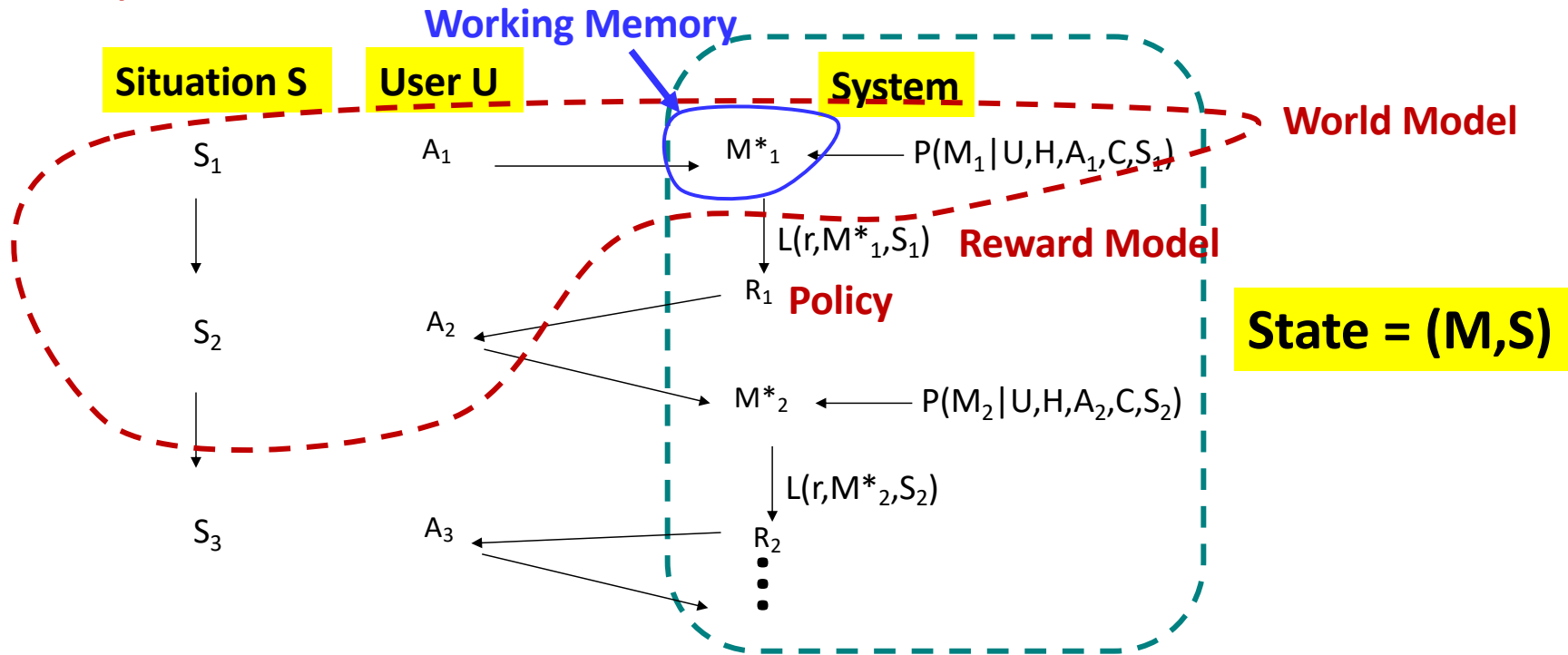
- Approximate the Bayes risk using posterior mode

$$\begin{aligned} R_t &= \operatorname{argmin}_{r \in r(A_t)} \int_{\mathcal{M}} L(r, M) p(M|U, H, A_t, C, S) dM \\ &\approx \operatorname{argmin}_{r \in r(A_t)} L(r, M^*) p(M^*|U, H, A_t, C, S) \\ &= \operatorname{argmin}_{r \in r(A_t)} L(r, M^*) \end{aligned}$$

where $M^* = \operatorname{argmax}_M p(M|U, H, A_t, C, S) \leftarrow$ **Posterior Mode**

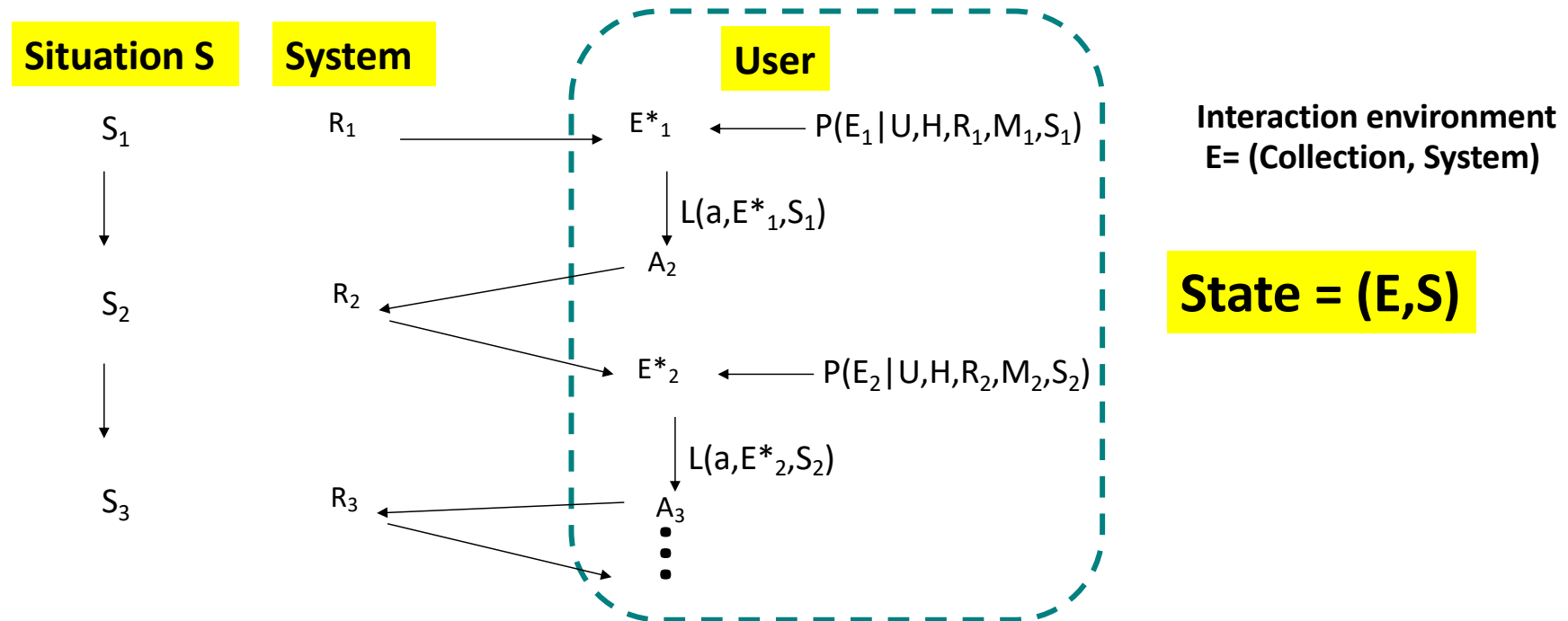
- Two-step procedure
 - Step 1: Compute an updated user model M^* based on all the currently available information
 - Step 2: Given M^* , choose an optimal response to minimize the loss function

Optimal Interactions of GIANT and Reinforcement Learning



A Partially Observable Markov Decision Process (POMDP) with (M, S) as State, naturally accommodating reinforcement learning

Duality of User & System Decision Making



User's decision process (behavior) can be modeled by a POMDP as well with (E,S) as State
Simulation of user agent for evaluating & optimizing interactive AI System

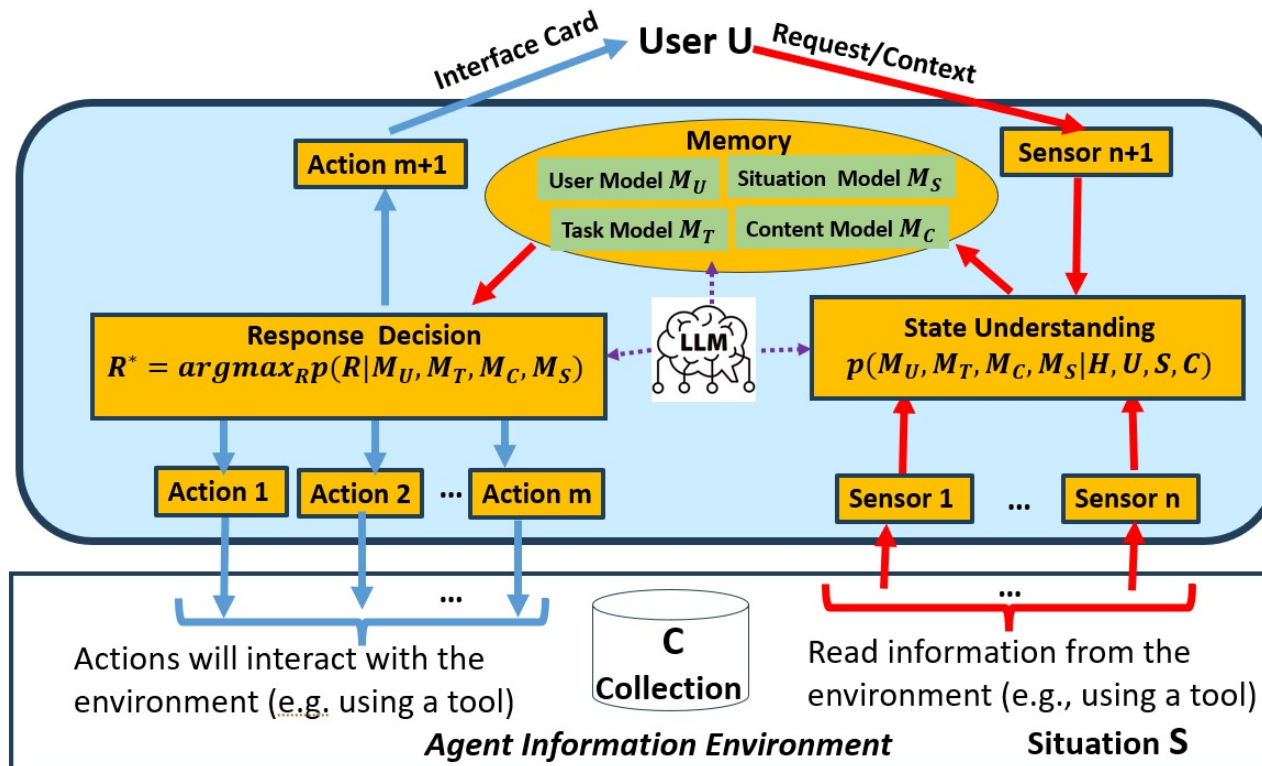
Instantiation of GIANT

- **Situation S:** can include time, location, and other environmental factors that are relevant to a task
- **Information Item Collection C:** naturally available in any application
- **User U:** can include any information we know about a user (or group)
- **User interaction history H:** naturally accumulated over time
- **User Actions and System Responses R(A):** all interfaces (moves of the game)
- **Loss Function L(R,M):** captures the objective of the game
- **Multi-dimensional Latent Model $M=(M_U, M_T, M_C, M_S)$:** can include everything that we can infer about a user M_U , user's task M_T , the content M_C , and situation M_S relevant to deciding how to respond to a user's action
- **Inference of Model: $P(M|U, H, A_t, C, S)$:** capture system's belief about the multidimensional latent model $M=(M_U, M_T, M_C, M_S)$.

Can go beyond search to support task

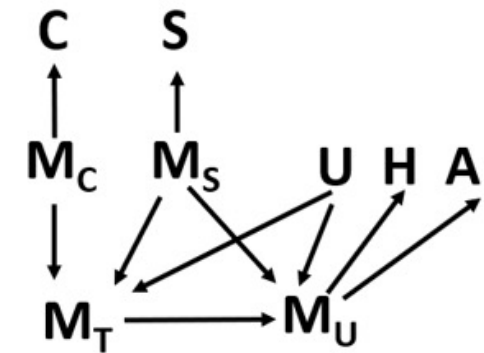
LLMs can be used for instantiation of all these

Implementation Details

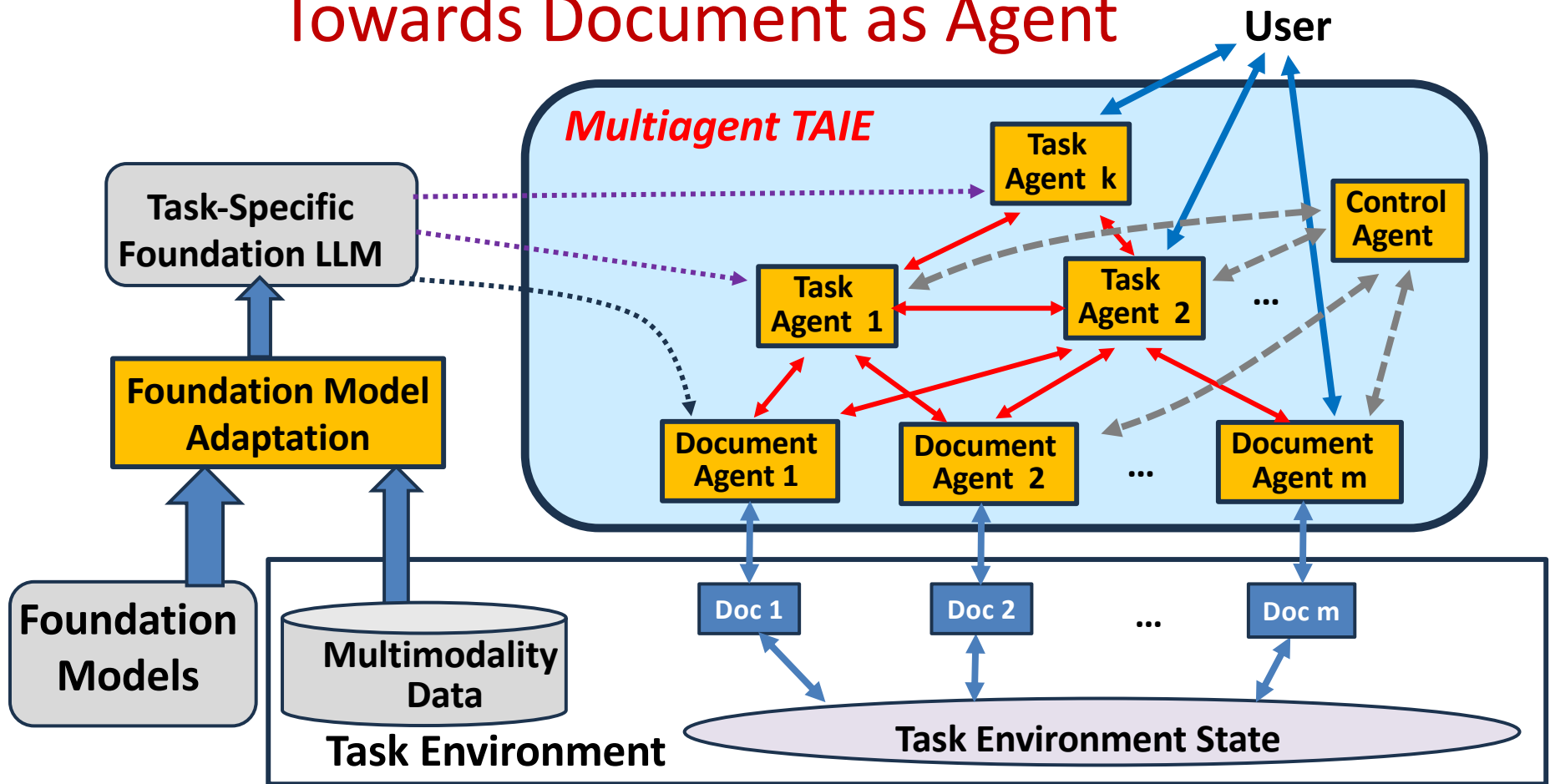


Model Estimation

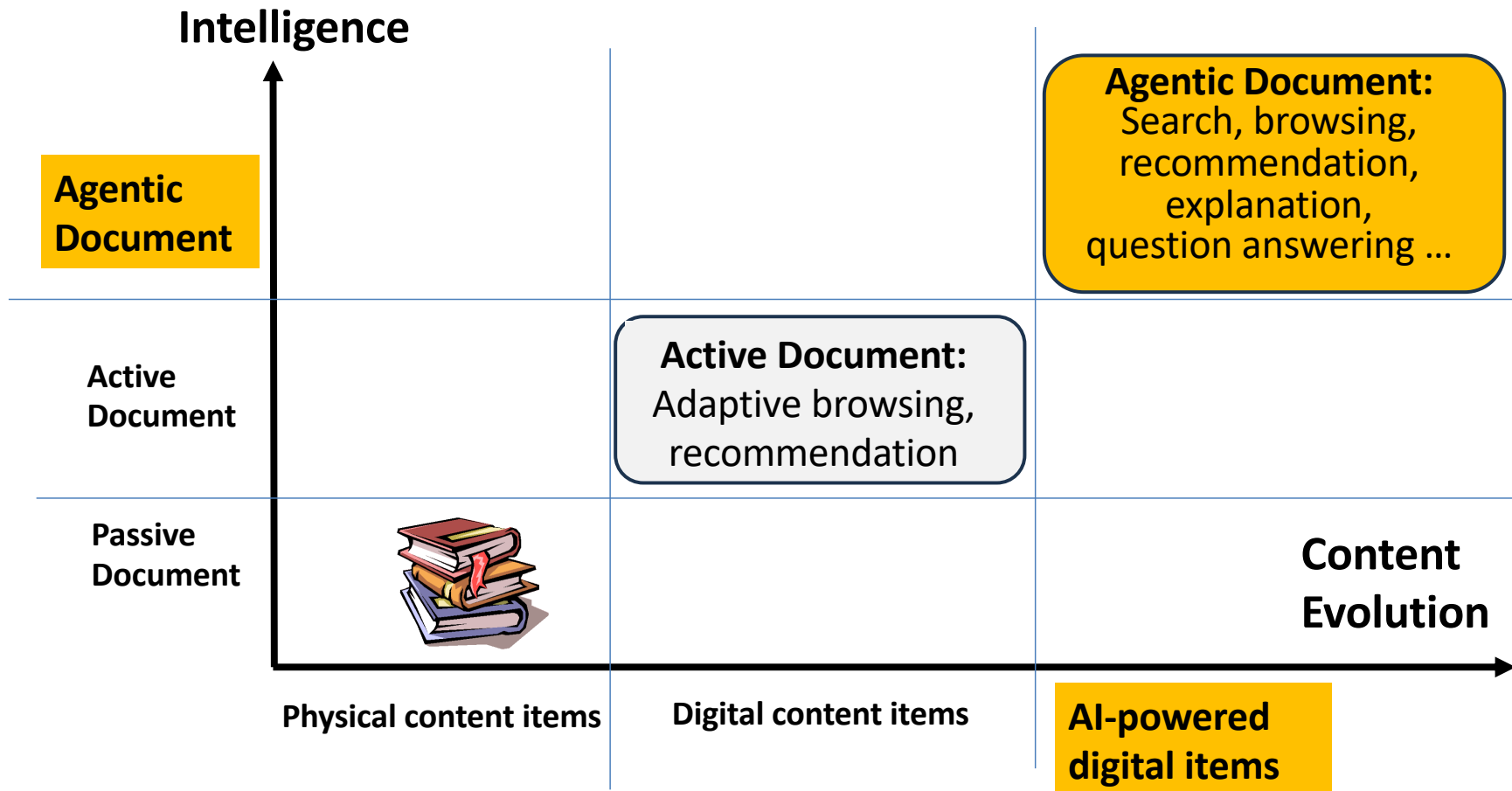
$M = (M_U, M_T, M_C, M_S)$



Towards Document as Agent



Document as an Agent



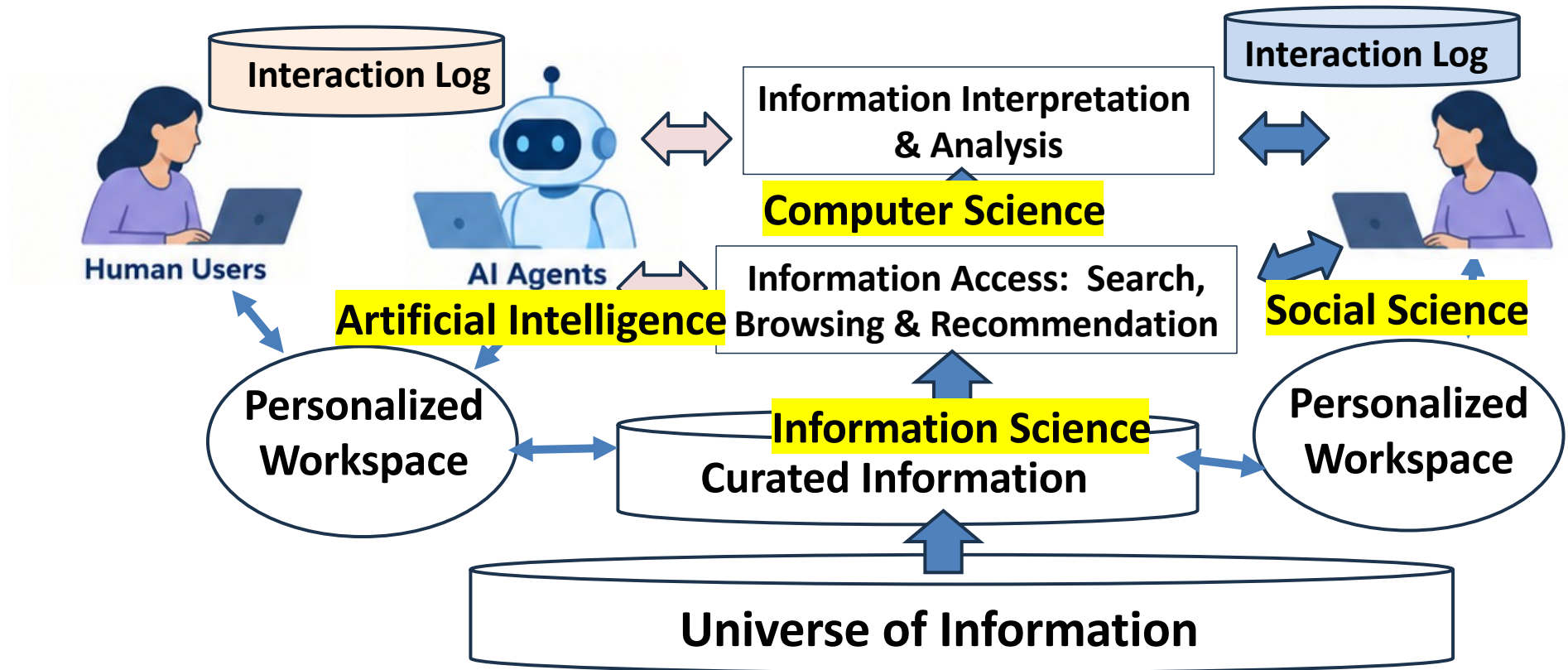
MIAS: A potential infrastructure for document as agent

- Demo: Navigable JCDL'26 Proceedings supported by MIAS (<https://www.iaaa.online/>)
- Techniques:
 - **Hypergraph of Text (HoT) as a general browsing structure** [Alvarez & Zhai 24]
 - HoT Browser enables navigation over any HoT
- **Browsing support = HoT construction**
- **A document can naturally become an agent**
- **Can be deployed to enhance any digital library**



Alvarez, Dean E., and ChengXiang Zhai. "Hypergraph of Text: a Mathematical Structure for Organizing and Analyzing Big Text Data." In *2024 IEEE International Conference on Big Data (BigData)*, pp. 8605-8607. IEEE, 2024.

Summary: Trustworthy Agentic Information Environment (TAIE)



Thank You!

Questions/Comments?