SIGIR 2025
Padova
ITALY

Summary and Outlook

Yifei Yuan

ETH Zürich & University of Copenhagen

Outline

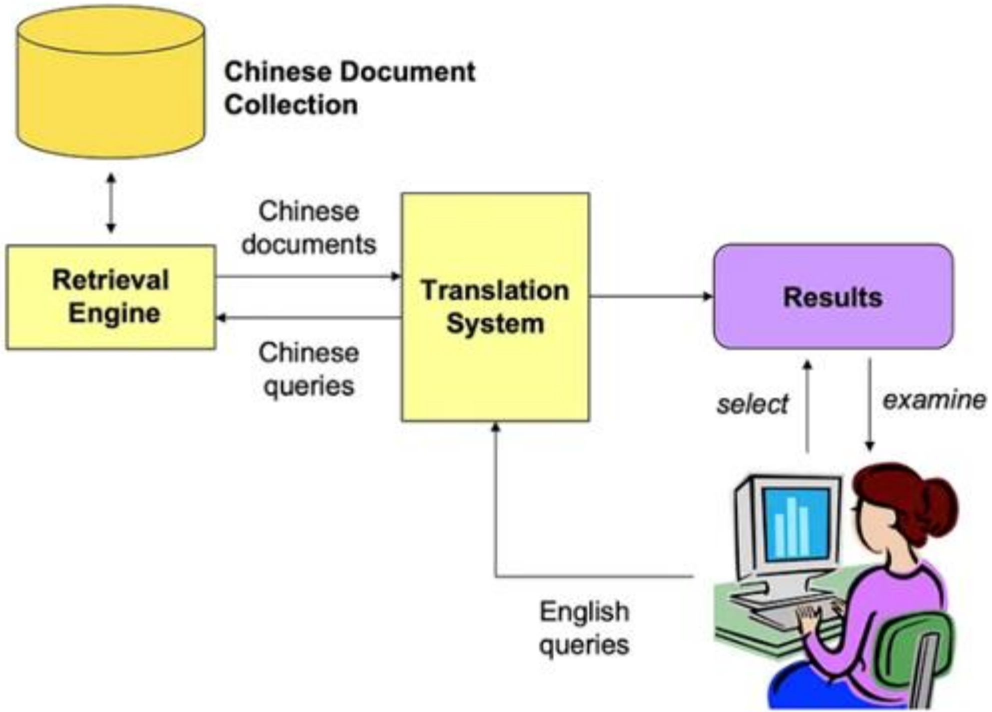
- **Definition & preliminaries - Yifei**
  - Query understanding
  - LLM-based conversational information seeking
- **LLM-based Query Enhancement - Yifei**
  - Resolving ambiguity in queries
  - Multimodal conversational query rewrite
- **LLM-based Proactive Query Management – Yang**
  - Unanswerable query mitigation
  - Uncertain query clarification
- **LLM-based Conversational Interaction – Mohammad**
  - Balancing user and system initiative
  - LLM-based user simulation
- **Conversational Query Understanding Evaluation– Zahra**
  - End-to-end evaluation
  - LLM-based relevance assessment

Open Challenges

- **Multilingual and Multimodal Extensions**
  - Multilingual and cross-cultural query understanding
  - Expanding query understanding beyond text
- **Real-time adaptation to evolving user needs**
  - Shift toward user-personalized dialogue agents
  - Increasing reliance on multi-turn reasoning in LLMs
  - Integration of retrieval-augmented generation for real-time knowledge access

Challenge: Multilingual and Multimodal Extensions

Translation, a common approach for handling multilingual information retrieval



http://www.ccs.neu.edu/home/jaa/IS4200.12S/Handouts/cross_language.pdf

Challenge: Multilingual and Multimodal Extensions

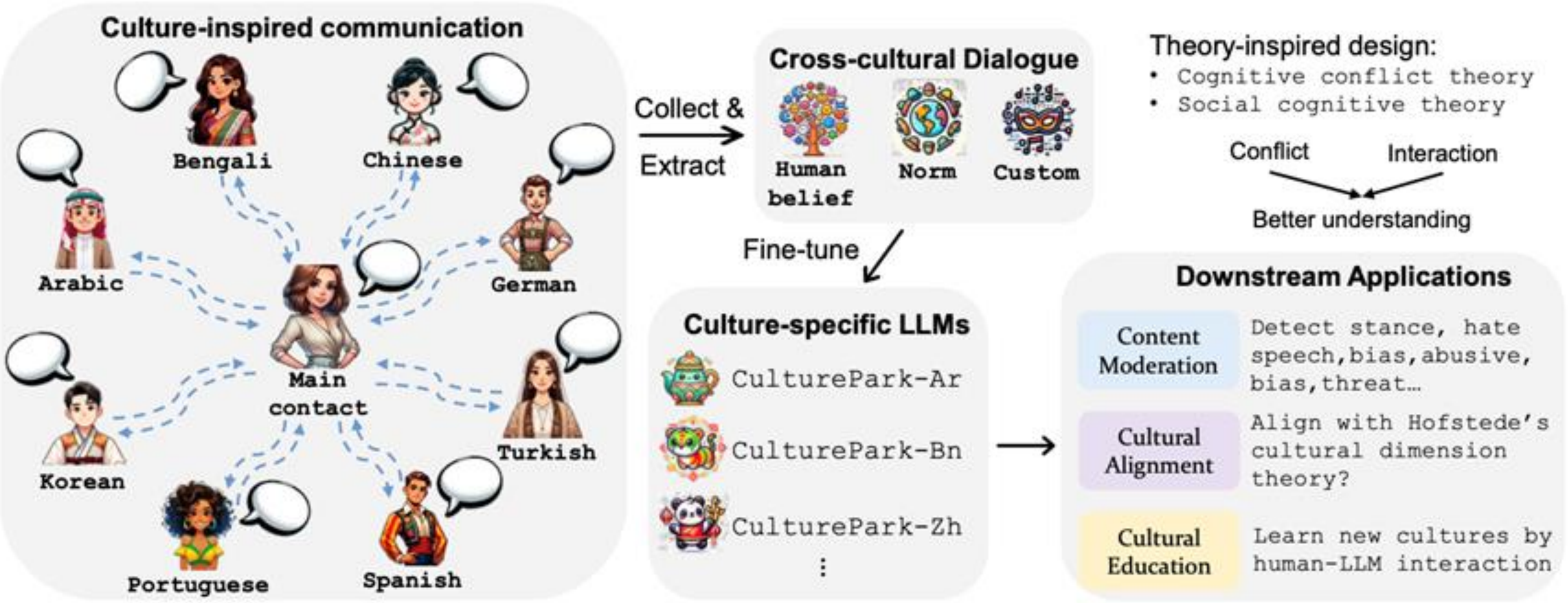
Cross-cultural query understanding requires:

- Multilingual semantic alignment
- Cultural context awareness
- Multimodal grounding
- Adaptivity to different user expectations

Even LLMs like GPT-4 can make cultural mistakes!

Hu et al, "Bridging Cultures in the Kitchen: A Framework and Benchmark for Cross-Cultural Recipe Retrieval" (EMNLP'24)
Sun et al, "CL-QR: Cross-Lingual Enhanced Query Reformulation for Multi-lingual Conversational AI Agents" (EMNLP'23)

Challenge: Multilingual and Multimodal Extensions



Li et al, "CulturePark: Boosting Cross-cultural Understanding in Large Language Models" (NeurIPS'24)

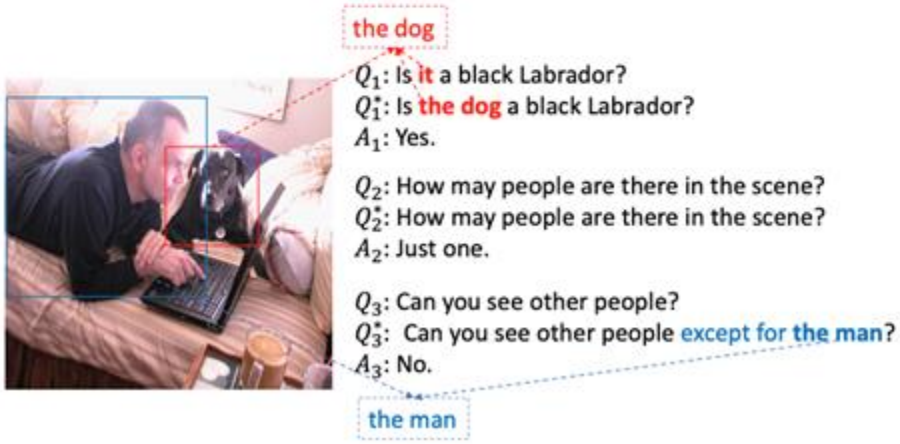
Challenge: Multilingual and Multimodal Extensions

Why Multimodality Matters

- Text-only input limits user expression
- Real-world queries often include contents alongside text
- Multimodal systems improve context awareness and intent resolution

Challenges

- Aligning different modalities in real-time
- Lack of high-quality multimodal training datasets
- Maintaining performance and interpretability across domains

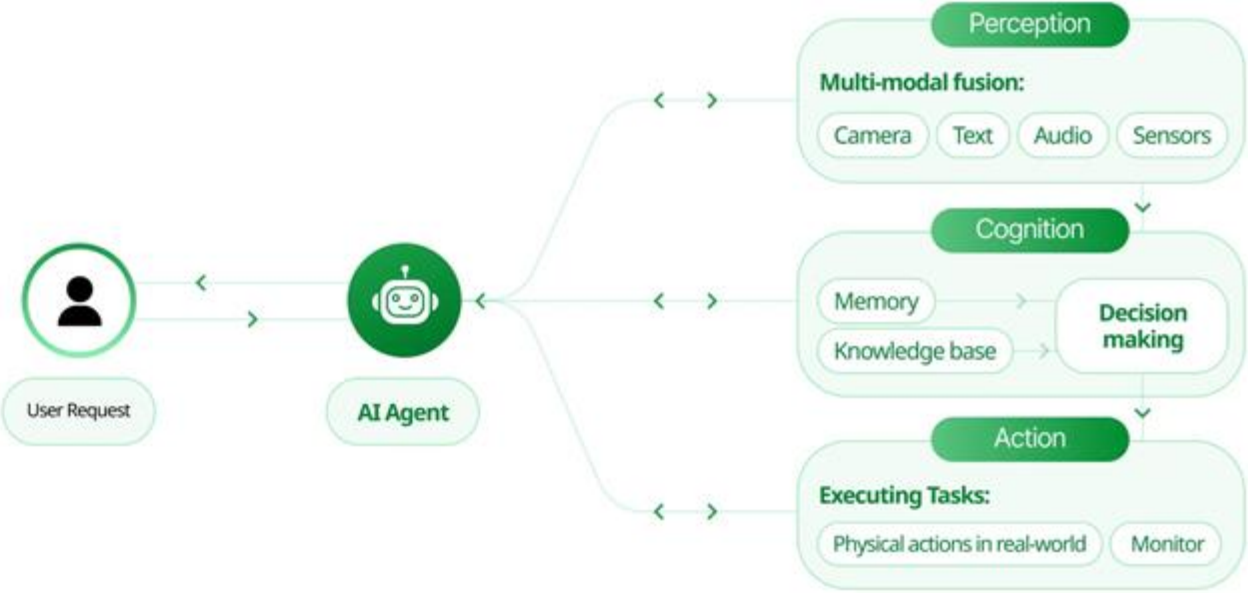


Yuan et al, "McQueen: A Benchmark for Multimodal Conversational Query Rewrite" (EMNLP'22)

Challenge: Real-time adaptation to evolving user needs

Shift toward user-personalized dialogue agents

Agentic AI in conversational search enables systems to proactively plan, reason, and act across multiple turns to deliver more context-aware, goal-oriented, and dynamic information-seeking experiences.



Zhang et al, "Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games" (ACL'25)

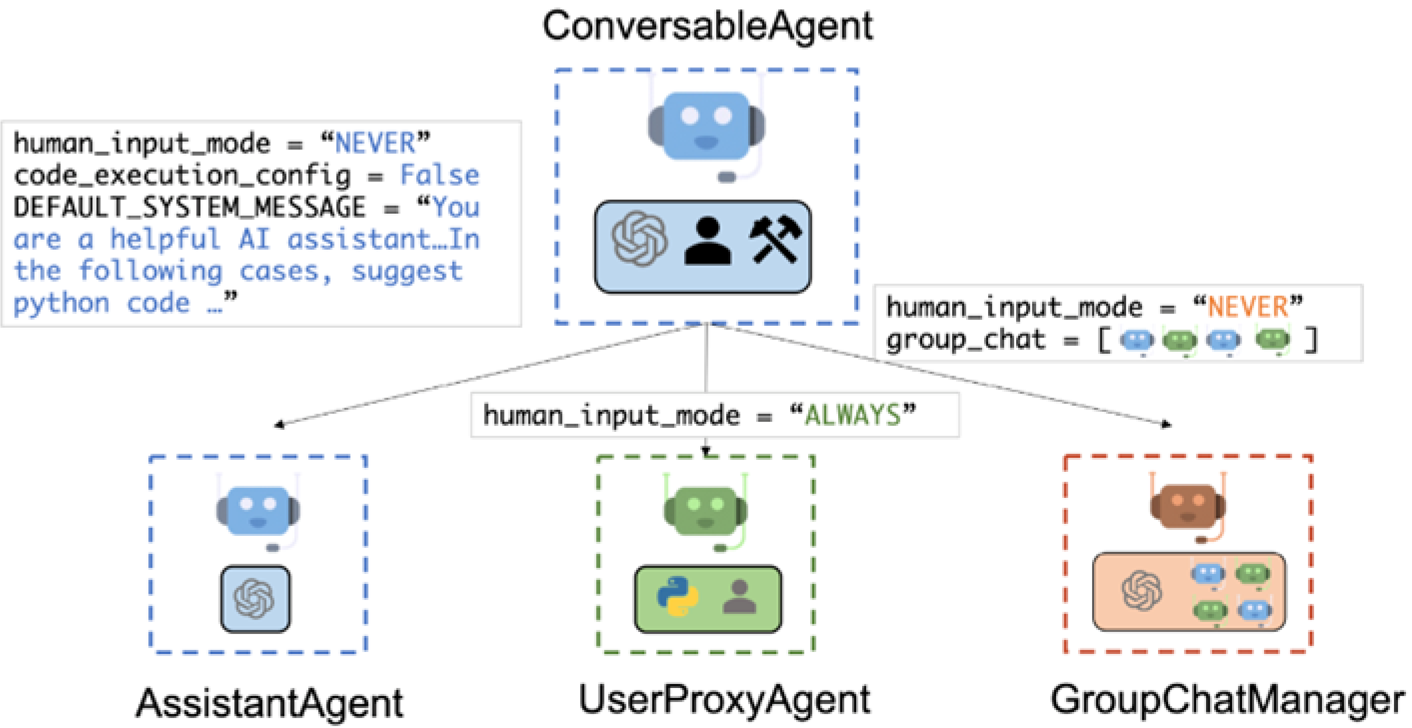
Challenge: Real-time adaptation to evolving user needs

From search to agentic search

Zhang et al, “From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents” (arXiv’25)

Challenge: Real-time adaptation to evolving user needs

AutoGen, a multi-agent framework for conversation interaction



Wu et al, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation" (COLM'24)

Challenge: Real-time adaptation to evolving user needs

Multi-turn Planning

- Increasing reliance on multi-turn reasoning in LLMs
- Current LLMs still face reasoning and planning challenges when generating clarifying questions.
- When faced with complex and hidden user needs, the agent faces the problem of redundancy, inconsistency, and early enumeration.

Reasoning issue (Inconsistency)

G: Is the entity a living being? *J: No.*
G: Is the entity an object or a thing? *J: Yes.*
G: Is the object typically found indoors? *J: Yes.*
G: Is the object used primarily for a specific purpose? *J: Yes.*
G: Is the object small and portable? *J: Maybe.*
G: Is the object electronic? *J: Yes.*
G: Is the object a device used for communication *J: No.*
G: Is the object a device used for **entertainment** or media *J: No.*
G: Is the object a device used for productivity or work *J: Yes.*
G: Is the object a laptop? *J: No.*
G: Is the object a **game console**? *J: No.*
...

Zhang et al, “Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games” (ACL’25)

Challenge: Real-time adaptation to evolving user needs

When tasked with generating clarifying questions to guess a hidden item, GPT-4 only achieves about a **26%** success rate, with the correct guess taking **nearly 18 turns on average**. 😟

	Things				Celebrities			
	#Turns (↓)	Success (↑)	#Yes	Score (↑)	#Turns (↓)	Success (↑)	#Yes	Score (↑)
GPT-4	**17.8±0.2**	**0.31±0.03**	5.9±0.1	**0.26±0.02**	**17.3±0.1**	**0.50±0.02**	6.8±0.2	**0.40±0.02**
GPT-3.5	18.2±0.2	0.28±0.02	6.3±0.1	0.23±0.02	18.8±0.2	0.27±0.03	7.4±0.2	0.21±0.03
Claude-2	18.4±0.3	0.21±0.03	5.0±0.1	0.18±0.03	17.6±0.2	0.31±0.02	5.6±0.1	0.26±0.02
Claude-1	18.8±0.1	0.16±0.02	4.2±0.1	0.13±0.02	17.7±0.2	0.29±0.03	5.3±0.2	0.25±0.02
Vicuna 13B	18.4±0.1	0.18±0.02	5.0±0.2	0.15±0.02	18.7±0.2	0.22±0.03	6.1±0.1	0.18±0.02
Vicuna 7B	19.5±0.2	0.09±0.02	5.7±0.2	0.07±0.02	19.6±0.3	0.06±0.02	5.9±0.2	0.05±0.02
Mistral 7B	18.9±0.1	0.13±0.02	3.8±0.5	0.11±0.02	18.2±0.1	0.22±0.04	4.3±0.1	0.20±0.03
V-FT 7B (All)	19.2±0.1	0.13±0.01	6.1±0.1	0.10±0.01	19.3±0.1	0.16±0.02	7.6±0.3	0.13±0.02
V-FT 7B (Suc.)	18.0±0.1	0.23±0.01	5.1±0.2	0.20±0.01	19.0±0.2	0.15±0.02	6.3±0.2	0.13±0.02
V-FT 13B (All)	18.6±0.2	0.21±0.03	6.1±0.2	0.17±0.02	18.8±0.2	0.22±0.01	6.2±0.2	0.18±0.01
V-FT 13B (Suc.)	18.0±0.2	0.25±0.02	4.5±0.1	0.21±0.03	18.4±0.3	0.23±0.04	5.9±0.2	0.19±0.03
V-RLGP 7B	17.8±0.1	0.26±0.02	4.7±0.1	0.22±0.01	18.8±0.1	0.16±0.01	5.9±0.1	0.14±0.00
V-RLGP 13B	17.9±0.1	0.27±0.02	4.5±0.1	0.23±0.01	18.5±0.2	0.26±0.03	6.1±0.1	0.21±0.02

Zhang et al, "Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games" (ACL'25)

Challenge: Real-time adaptation to evolving user needs

Possible ways for enhancing reasoning: Chain-of-Thought (CoT) Prompting, ReACT: Reasoning and Acting, RLHF, Self-reflection, …

Model	Turn-Level Reward		Outcome Reward	
	Tool Execution (0-0.2)	Search Answer (0-0.5)	XML Format (0-0.2)	Exact Match (0-1)
Qwen2.5-7B-Base	0.0559	0.0934	0.1562	0.0469
Qwen2.5-7B-Instruct	0.1626	0.2814	0.1982	0.1559
Qwen2.5-7B-Base + GRPO-OR	0	0	0.04	0
Qwen2.5-7B-Base + GRPO-MR	0.2	0.3724	0.1994	0.3346
Qwen2.5-7B-Base + MT-GRPO	0.2	**0.3926**	**0.1996**	**0.5010**

Variants	GPT-4o		Llama 3.1 70B		Llama 3.1 405B	
	1@1	1@3	1@1	1@3	1@1	1@3
Single-turn	17.0	27.6	23.2	27.3	27.8	32.9
+ CoT	$25.5_{+8.5}$	$29.0_{+1.4}$	$25.5_{+2.3}$	$28.9_{+1.6}$	$25.1_{-2.7}$	$31.8_{-1.1}$
+ Multi-turn	-	$23.1_{-4.5}$	-	$29.5_{+2.2}$	-	$35.4_{+2.5}$
+ Multi-turn CoT	-	$31.5_{+3.9}$	-	$31.5_{+4.2}$	-	$40.1_{+7.2}$

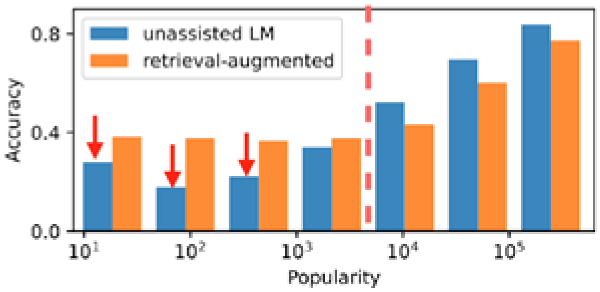
Zheng et al, "What Makes Large Language Models Reason in (Multi-Turn) Code Generation?" (ICLR'25)
Zeng et al, "Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Credit Assignment" (arXiv'25)

Challenge: Real-time adaptation to evolving user needs

Intergrating RAG for real-time knowledge access

RAG allows for instant knowledge update from an external knowledge base

What is Kathy Saltzman's occupation?



(Mallen et al., 2023)

GPT-3 davinci-003: 20%-30% accuracy

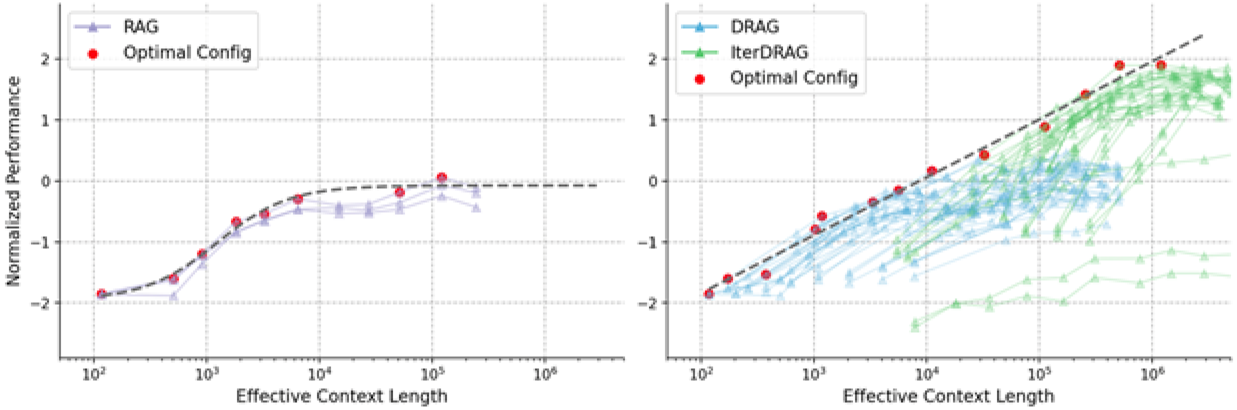
Challenge: Real-time adaptation to evolving user needs

Key issues:

What to retrieve?

When to retrieve?

How to retrieve?



In RAG, scaling becomes **multi-dimensional** due to the addition of a **retrieval system**.

Yue et al, "Inference Scaling for Long-Context Retrieval Augmented Generation" (ICLR'25)

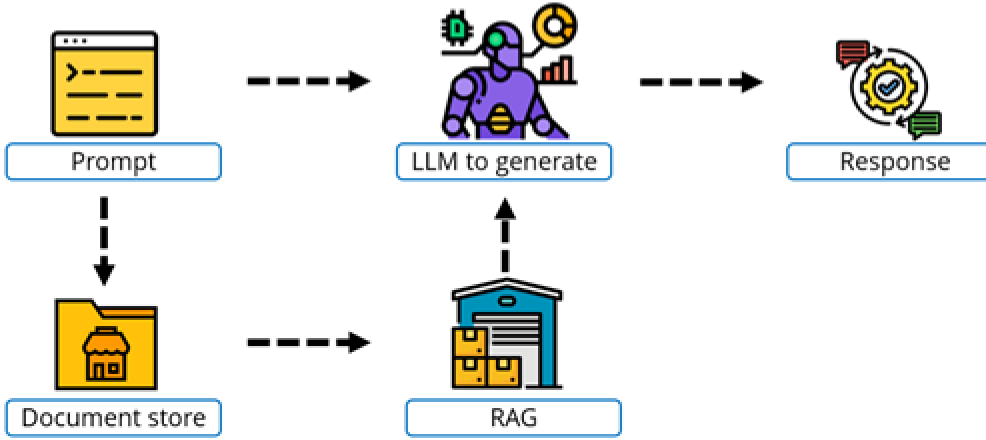
Future Direction

🧠Knowledge-Aware Query Interpretation

Challenge: Query understanding often ignores world knowledge or domain-specific constraints.

Direction: Inject structured knowledge (e.g., KBs, graphs, taxonomies) into LLMs to enable semantic grounding and facet-level disambiguation.

Research idea: Jointly learn facet extraction + semantic typing + query understanding using adapter layers or retrieval-enhanced decoding.



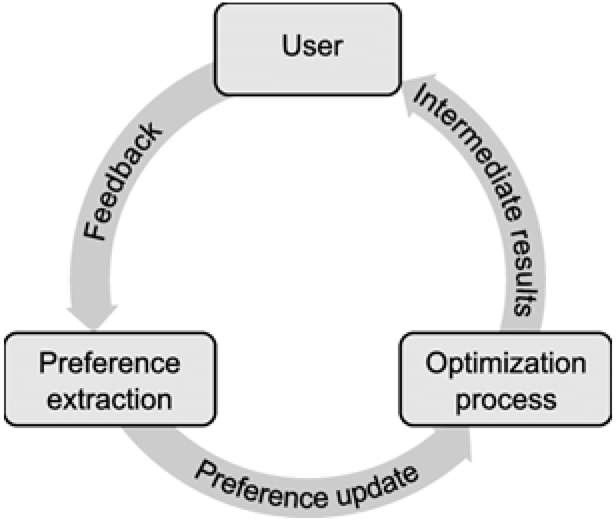
Future Direction

User-In-the-Loop Adaptive Query Understanding

Challenge: LLMs often hallucinate user intent

Direction: Use **relevance feedback**, **user corrections**, or **interaction signals** to continuously refine query interpretation during the session.

Research idea: Online LLM fine-tuning or reward shaping using bandit signals from user engagement.



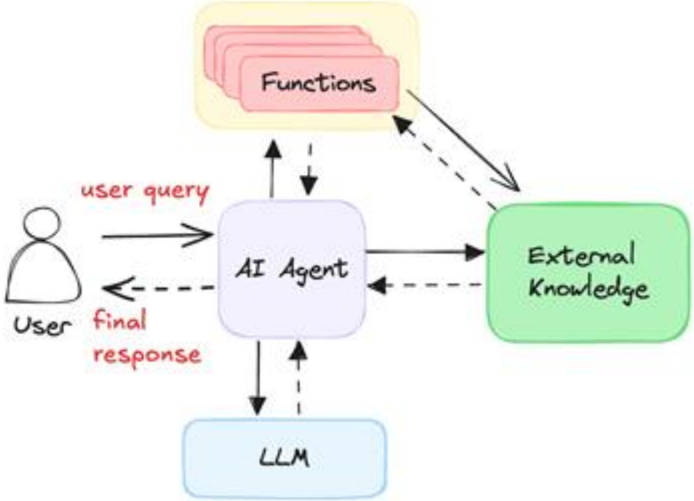
Future Direction

🔮 Agentic Query Understanding

LLMs-as-agents can "think" about whether their current interpretation is sufficient.

Meta-level Decision Making: When to Ask, When to Act

Agentic models can call APIs or retrieve from knowledge bases to clarify ambiguous queries.



Open Questions

- The best way to accurately understand and predict complex user needs through effective interaction remains largely underexplored.

- How can LLMs effectively understand user needs across multimodal and multilingual real-world scenarios?

- Evaluation metrics need to be designed for better capturing user satisfaction in conversational search.

Q & A

Thank you for joining us today!

All the materials at *https://sigirusertutorial.github.io/*