Reasoning with Structures for Large Language Models

JIAWEI HAN
SIEBEL SCHOOL OF COMPUTING AND DATA SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
JUNE 29, 2026

Outline

- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

Using LLMs: Prompt Design and Engineering

- Prompt: An instruction to an LLM
  - A prompt is the textual input provided by users to guide the model's output. This could range from simple questions to detailed descriptions or specific tasks.
- Prompts generally consist of instructions, questions, input data, and examples
  - A prompt must contain either instructions or questions, with other elements being optional
- Good prompts follow two basic principles: **clarity** and **specificity**
  - **Clarity**: Use unambiguous language that avoids jargon and overly complex vocabulary, making your point sufficiently clear to the LLM
  - **Specificity**: Tell your model as much as it needs to know to answer the question
- Prompt engineering tricks
  - Do say "do," don't say "don't" (since "do" is by nature more specific than "don't")
    - Ex. "Each title should be 2-5 words long" vs. "don't make the title too long"
  - Use few-shot prompting (i.e., give a few examples)
  - Use "leading words": Guide the model step-by-step for problem-solving

Prompting Strategy

- **In-Context Learning** (ICL): Allow LLMs to adapt to new tasks by *presenting them with a minimal set of examples within the prompt*
  - Use the model's extensive pre-training on diverse text data, enabling it to generate responses tailored to the specifics of the task at hand
  - Zero-shot (no example) vs. one-shot vs. few shots (given a few examples)
- **Multi-step Reasoning**
  - Strategies such as Chain-of-Thought and Self-Consistency prompting guide the model through multi-step reasoning, allowing it to explore various pathways to arrive at an answer
  - Improve the output accuracy and enhance transparency and insight into the reasoning process
  - Representing the evolving interaction between users and LLMs, showcasing how adaptability and improved task execution can be achieved without further training

Zero-Shot Prompting: Cloze Patterns

- Zero-shot: Language models as knowledge bases
  - Even without any training, knowledge can be extracted from PLMs through cloze patterns
- PLMs can serve as knowledge bases
  - Pros: require no schema engineering, and support an open set of queries
  - Cons: retrieved answers are not guaranteed to be accurate
- Could be used for unsupervised open-domain QA systems

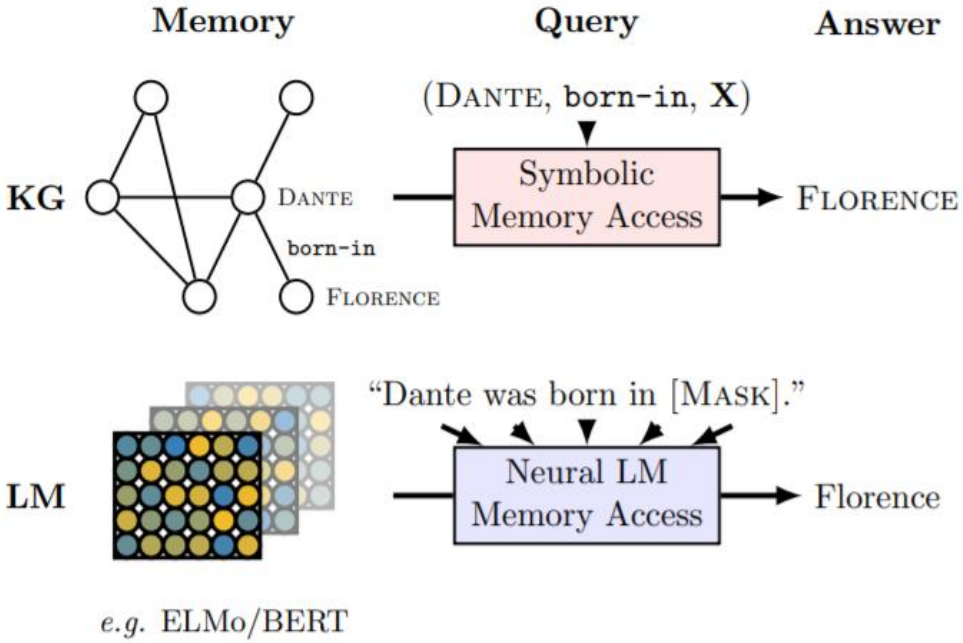


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

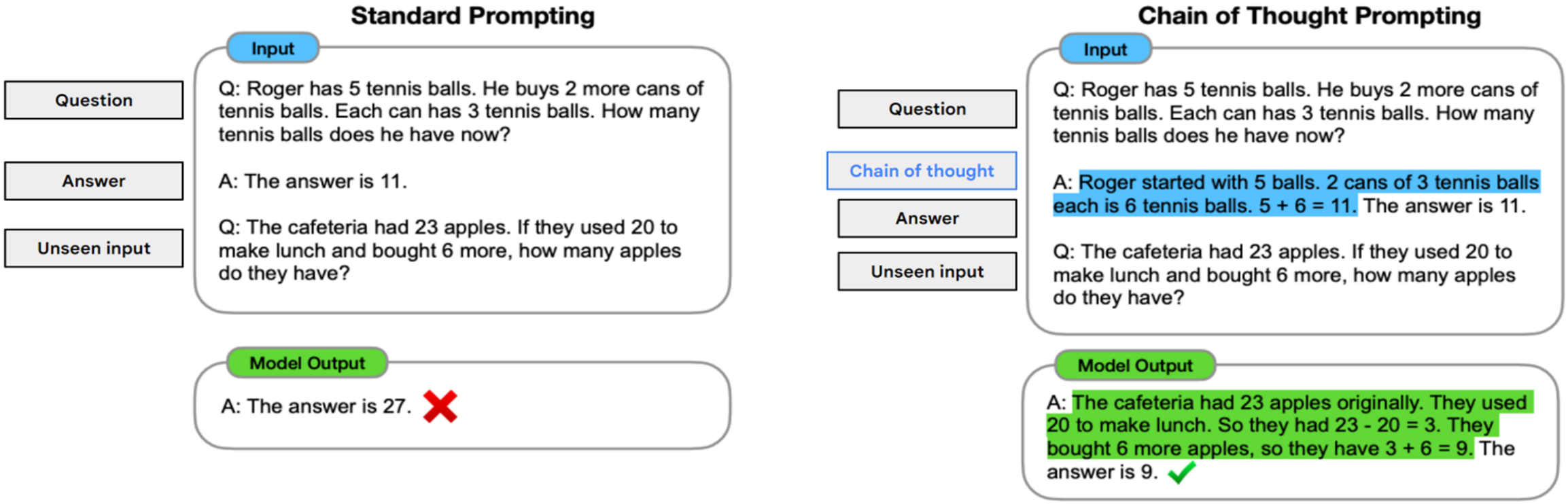
Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? EMNLP.

Prompt Engineering: From Prompt to Chain-of-Thought

- Prompt engineering is a rapidly evolving discipline that shapes the interactions and outputs of LLMs
  - The essence of prompt engineering lies in crafting the optimal prompt to achieve a specific goal with a generative model
  - Prompt engineering transcends the mere construction of prompts; it requires a blend of domain knowledge, understanding of the AI model, and a methodical approach to tailor prompts for different contexts
- Some typical prompt engineering approaches
  - Chain of Thought (CoT)
  - Tree of Thought (ToT)
  - Self-Consistency
  - Reflection
  - Expert Prompting
  - Chains
  - Rails
  - Automatic Prompt Engineering (APE)

Chain-of-Thought : Prompting Multi-step Reasoning

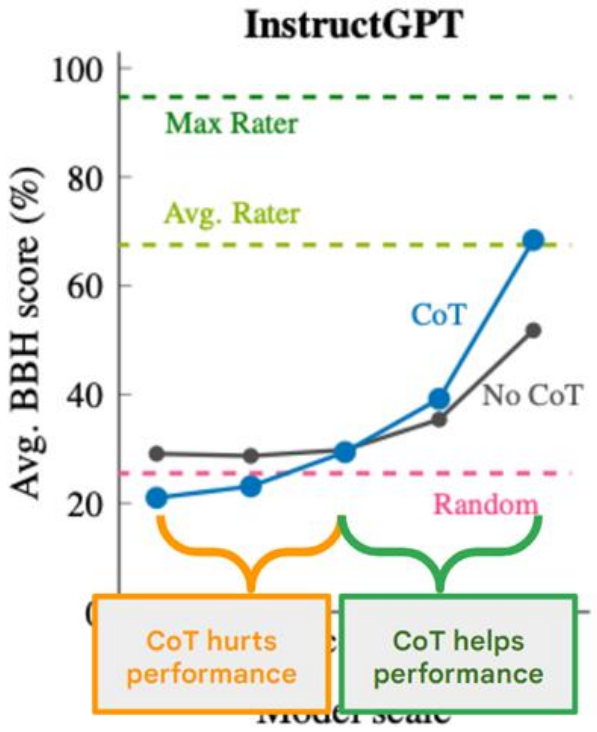
- Chain-of-Thought (CoT): Simulates human-like reasoning processes by delineating complex tasks into a sequence of logical steps towards a final resolution
  - A structured mechanism for problem-solving: Break down elaborate problems into manageable, intermediate thoughts that sequentially lead to a conclusive answer



Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS.

Chain-of-Thought Requires Scaling

- Only Larger LMs (e.g., GPT 3.5) can be improved with CoT prompting



Prompt

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?
A: Let's think step by step. Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5+6=11. The answer is 11.

Q: What is half of (3 + 7) plus one?
A:

text-ada-001

The answer is the result of adding 1 more ball (3 + 7) plus 1. ✗

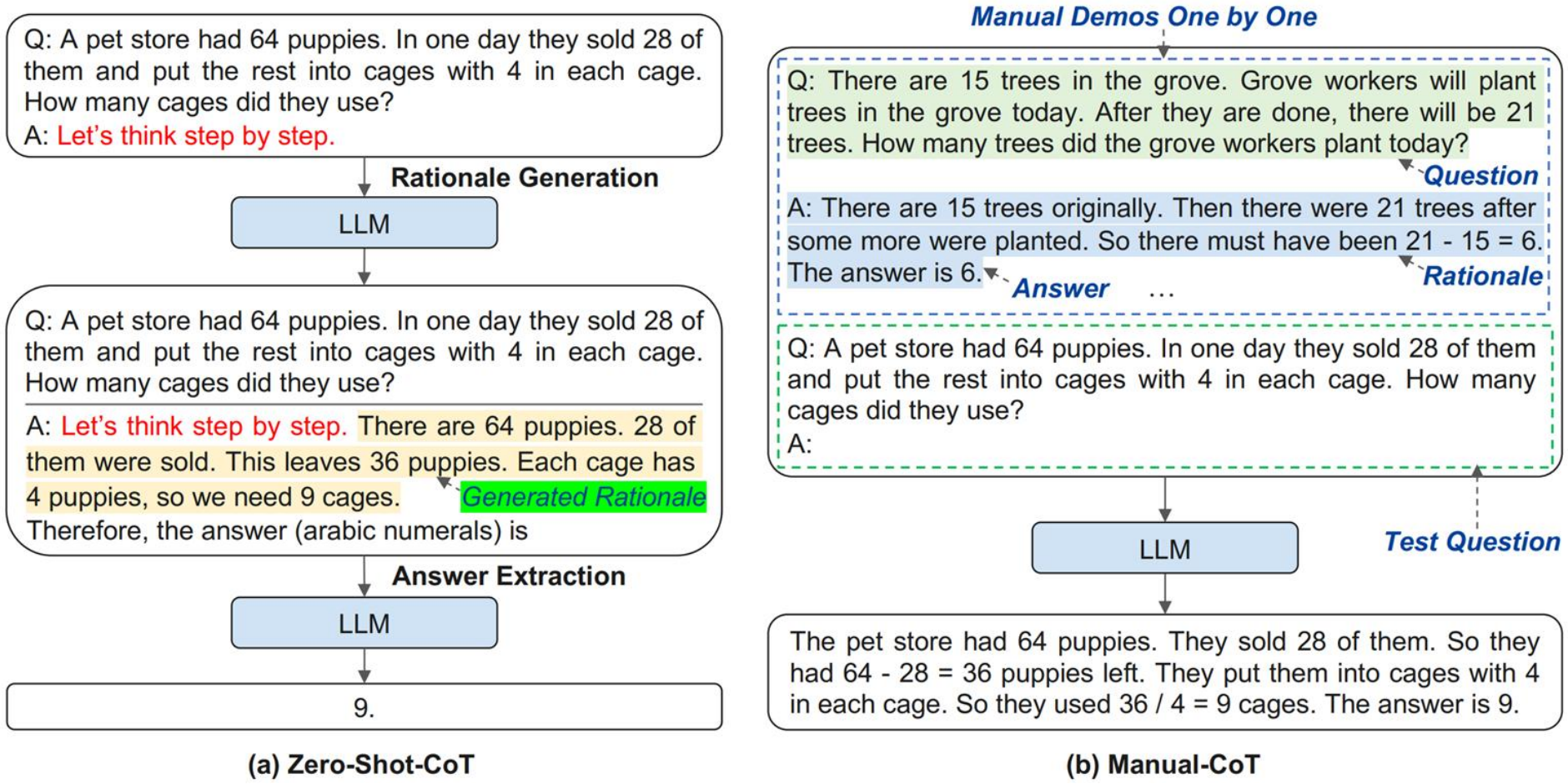
text-davinci-002

Let's think step by step. 3+7=10. 10/2=5. 5+1=6. The answer is 6. ✓

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., Wei, J. (2023). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ACL Findings.

Variants of CoT

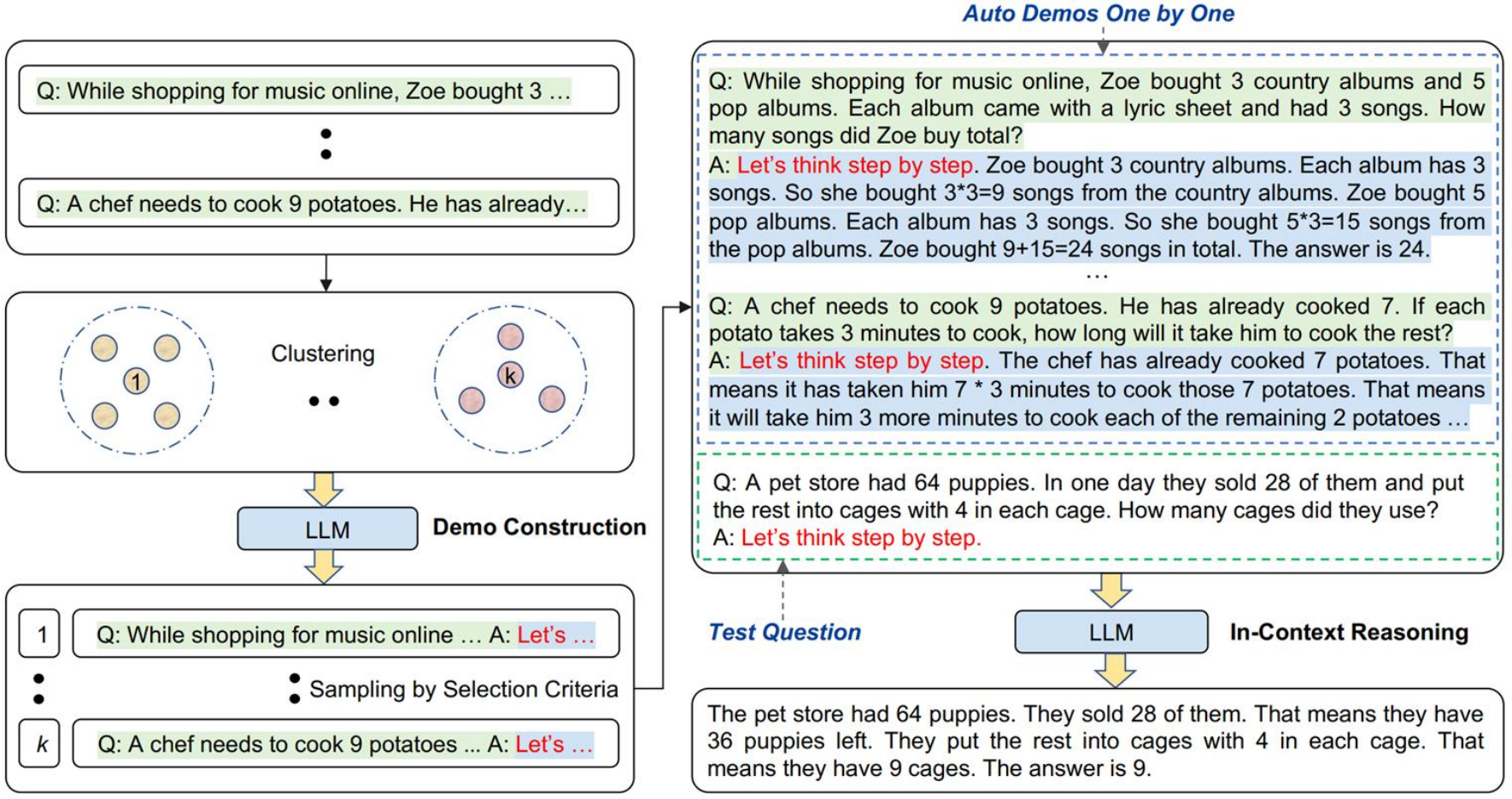
- Zero-Shot-CoT: using the “Let’s think step by step” prompt
- Manual-CoT: using manually designed demonstrations one by one



Omar Shaikh, et al. (2022), "On Second Thought, Let’s Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning" arXiv.2212.08061

Automatic Chain-of-Thought

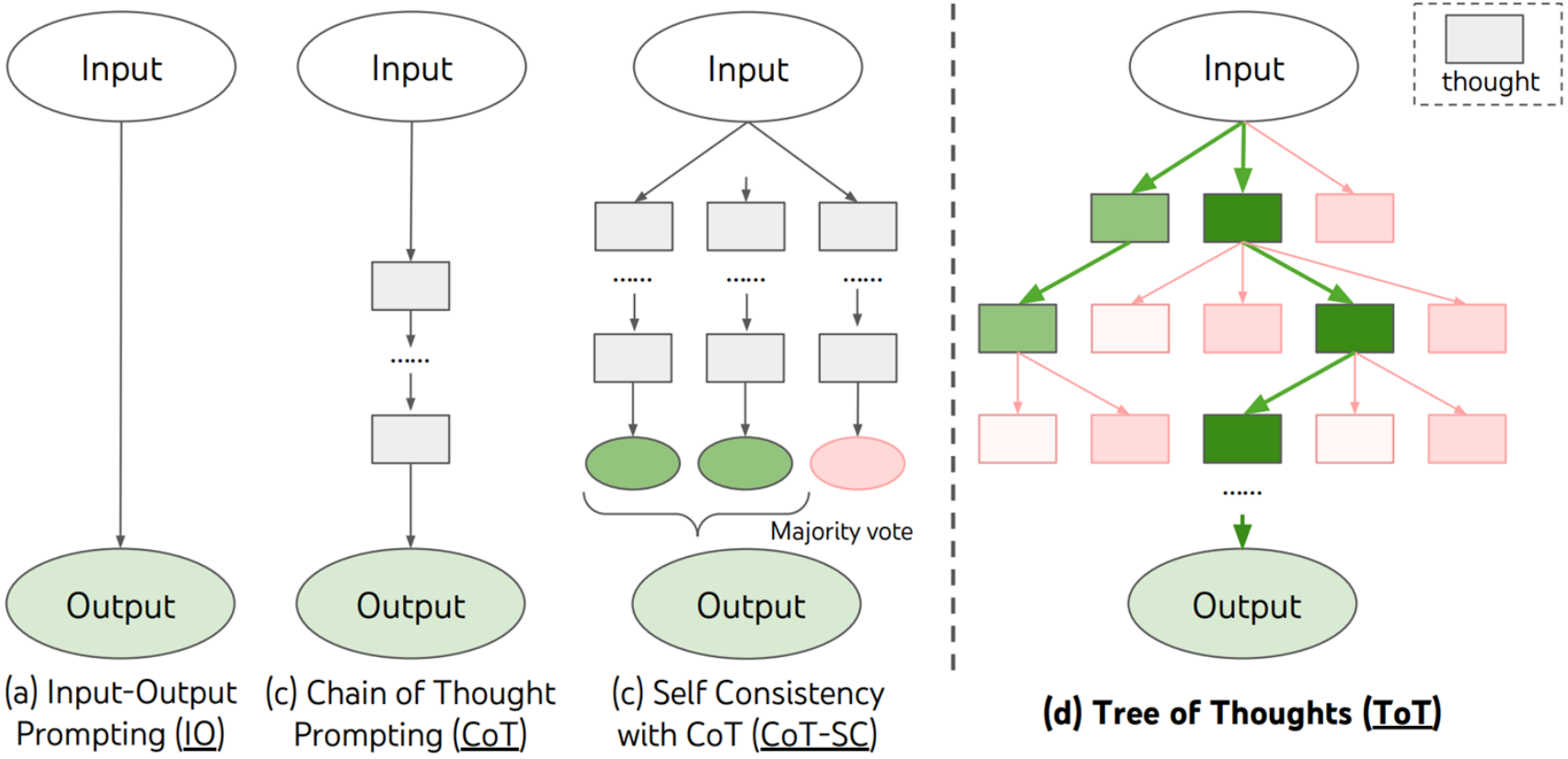
- Stage 1: Partition questions of a given dataset into a few clusters
- Stage 2: Select a representative question from each cluster and generate its reasoning chain



Zhang, Z., Zhang, A., Li, M., Smola, A. (2023). Automatic Chain of Thought Prompting in LLMs. ICLR.

CoT, Self-Consistency with CoT, and Tree-of-Thought

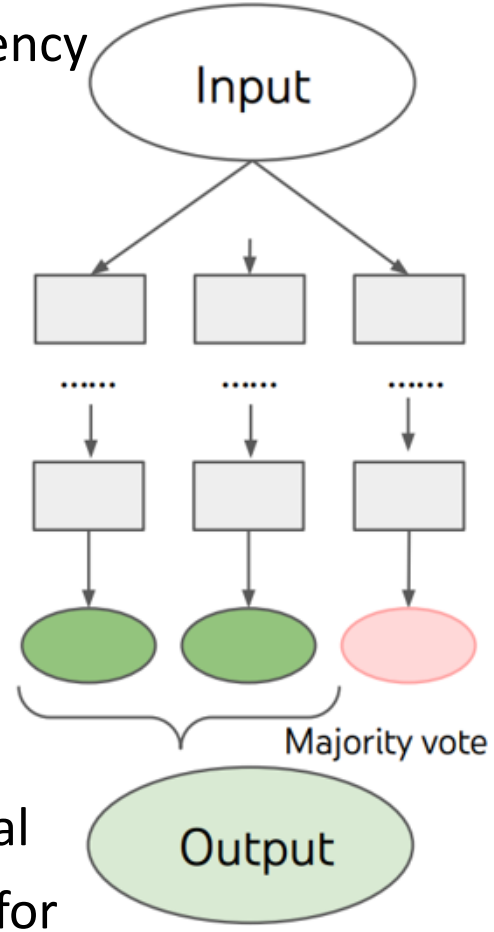
- Consider multiple different reasoning paths through self-evaluating voting
- Look ahead or backtracking when necessary to make global choices



Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS.

Self-Consistency: Ensemble of Multiple Responses

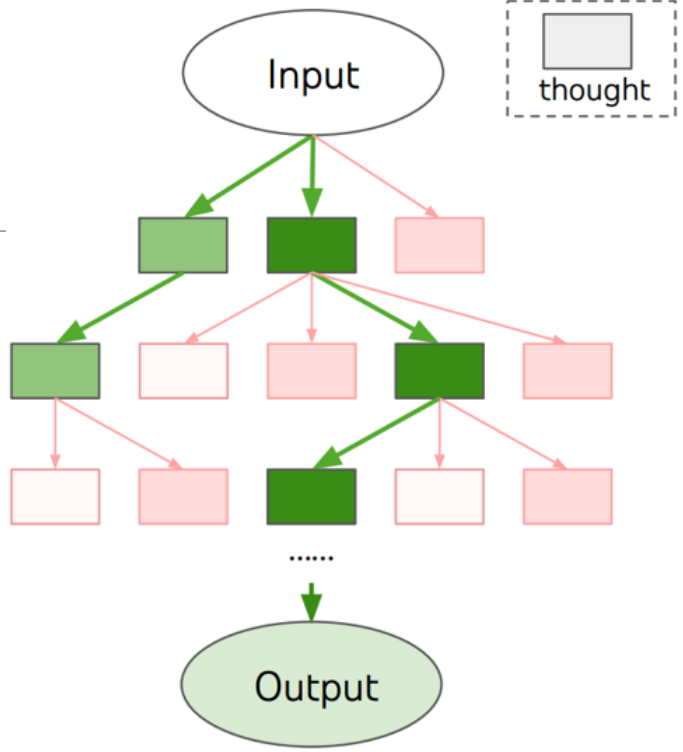
- LLM is prompted to **generate multiple responses to the same query**: The consistency among these responses serves as an indicator of their accuracy and reliability
  - If an LLM generates multiple, similar responses to the same prompt, it is more likely that the response is accurate
- Measuring the consistency of responses
  - Analyzing the overlap in the content of the responses
  - Comparing the semantic similarity of responses
  - Employing more sophisticated techniques like BERT-scores or n-gram overlaps
- Self-Consistency has significant applications where **the veracity of information is critical**
  - Ex. Fact-checking: Ensure the accuracy of info. provided by AI models is essential
- This technique enhance the trustworthiness of LLMs, making them more reliable for tasks that require high levels of factual accuracy



Self-consistency of CoT

P. Manakul, A. Liusie, and M. J. F. Gales, "SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models," 2023

Tree-of-Thought (ToT) Prompting



- Tree of Thought (ToT) prompting: Consider various alternative solutions or thought processes before converging on the most plausible one
  - Branch out into multiple "thought trees", each representing a different line of reasoning
  - Allow the LLM to explore various possibilities and hypotheses, much like human cognitive processes—multiple scenarios considered before determining the most likely one
- LLM generates different branches of thought, each is assessed for its validity/relevance to the query
  - This process involves real-time analysis and comparison of the branches, leading to a selection of the most coherent and logical outcome
- Used in complex problem-solving scenarios where a single line of reasoning might not suffice
  - It allows LLMs to mimic a more human-like problem-solving approach, considering a range of possibilities before arriving at a conclusion
  - It enhances the model's ability to handle ambiguity, complexity, and nuanced tasks, making it a valuable tool in advanced AI applications

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS.

Tree-of-Thought: An Example

- A complex question that needs ToT

Adapted from "GitHub - dave1010/tree-of-thought-prompting: Using Tree-of-Thought Prompting to boost ChatGPT's reasoning"

Bob is in the living room.
He walks to the kitchen, carrying a cup.
He puts a ball in the cup and carries the cup to the bedroom.
He turns the cup upside down, then walks to the garden.
He puts the cup down in the garden, then walks to the garage.
Where is the ball?

ChatGPT 3.5: The ball is in the garden. ❌
ChatGPT 4: The ball is in the bedroom. [followed by an explanation] ✅

Prompt given by the author:

Imagine three different experts are answering this question.
All experts will write down 1 step of their thinking,
then share it with the group.
Then all experts will go on to the next step, etc.
If any expert realizes they're wrong at any point then they leave.
The question is...

Then ChaptPRT3.5 will use ToT to do the reasoning:

Expert 1: The ball is in the living room.
Expert 2: The ball is in the kitchen.
Expert 3: The ball is in the bedroom.
Expert 1: Bob carries the cup to the bedroom, so the ball must be in the cup.
Expert 2: Oh, I see my mistake. Yes, the ball is in the cup.
Expert 3: Agreed, the ball is in the cup in the bedroom.
Expert 1: Next, Bob turns the cup upside down in the bedroom.
Expert 2: After that, Bob walks to the garden and puts the cup down.
Expert 3: Therefore, the ball must have fallen out of the cup when Bob turned it upside down in the bedroom. So, the ball is in the bedroom, not in the cup anymore.

Expert 1: Oh, you're right. I made a mistake. The ball is in the bedroom, not in the cup.
Expert 2: Agreed, the ball is in the bedroom.
Expert 3: Bob then walks to the garage, so the ball remains in the bedroom. It is not in the garden or the garage.
Expert 1: Absolutely, the ball is still in the bedroom.
Expert 2: Yes, the ball hasn't moved from the bedroom.
Expert 3: Therefore, the ball is in the bedroom, not in the garden or the garage.
All three experts agree that the ball is in the bedroom.

ChatGPT 3.5 ToT: The ball is in the bedroom. ✅

Reflection: Assess and Potentially Revise their Own Outputs

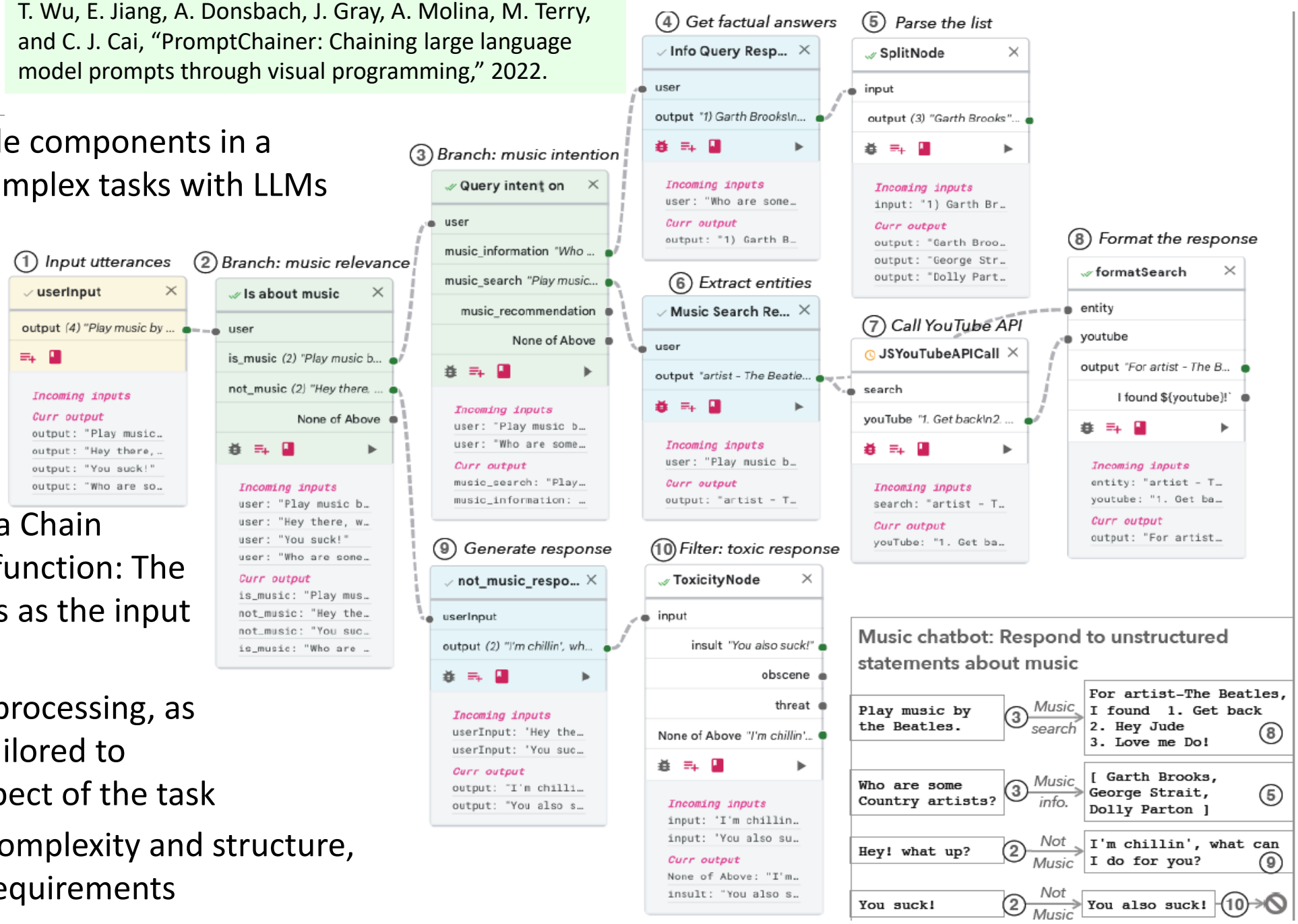
- Reflection: *Prompting LLMs to **assess and potentially revise their own outputs** based on reasoning about the correctness and coherence of their responses*
  - LLMs can engage in a form of **self-evaluation**
    - After generating an initial response, the model is prompted to reflect on its own output, considering factors like factual accuracy, logical consistency, and relevance
    - This introspective process can lead to the generation of revised or improved responses
- A key aspect of Reflection is the LLM's capacity for self-editing
  - By evaluating its initial response, the model can identify potential errors or areas of improvement
  - This ***iterative process of generation, reflection, and revision*** enables the LLM to refine its output, enhancing the overall quality and reliability of its responses

N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," 2023.

Chains

T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai, "PromptChainer: Chaining large language model prompts through visual programming," 2022.

- **Chains**: Linking multiple components in a sequence to handle complex tasks with LLMs
  - Idea: Construct a workflow where different stages or components are sequentially arranged
  - Each component in a Chain performs a specific function: The output of one serves as the input for the next
  - Allows for complex processing, as each stage can be tailored to handle a specific aspect of the task
  - Chains can vary in complexity and structure, depending on the requirements



Rails: Following Predefined Rules or Templates

- Rails: Guide and control the output of LLMs through predefined rules or templates
  - Ensure the model's responses adhere to certain standards or criteria, enhancing the relevance, safety, and accuracy of the output
- Rails involves setting up a framework or a set of guidelines that the LLM must follow while generating responses
  - These guidelines are typically defined using a modeling language or templates known as Canonical Forms, which standardize the way natural language sentences are structured and delivered.
- Rails can be designed for various purposes, depending on the specific needs of the application
  - Topical Rails: Ensure that the LLM sticks to a particular topic or domain
  - Fact-Checking Rails: Aimed at minimizing the generation of false or misleading information
  - Jail-breaking Rails: Prevent the LLM from generating responses that attempt to bypass its own operational constraints or guidelines

Automatic Prompt Engineering

- Automatic Prompt Engineering (APE): Automating the process of prompt creation for LLMs
- APE seeks to streamline and optimize the prompt design process, leveraging the capabilities of LLMs themselves to generate and evaluate prompts
- APE involves using LLMs in a self-referential manner where the model is employed to generate, score, and refine prompts. This recursive use of LLMs enables the creation of high-quality prompts that are more likely to elicit the desired response or outcome.
- The methodology of APE can be broken down into several key steps:
  - Prompt Generation: The LLM generates a range of potential prompts based on a given task or objective
  - Prompt Scoring: Each generated prompt is then evaluated for its effectiveness, often using criteria like clarity, specificity, and likelihood of eliciting the desired response
  - Refinement and Iteration: Based on these evaluations, prompts can be refined and iterated upon, further enhancing their quality and effectiveness

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” 2023.

Outline

- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

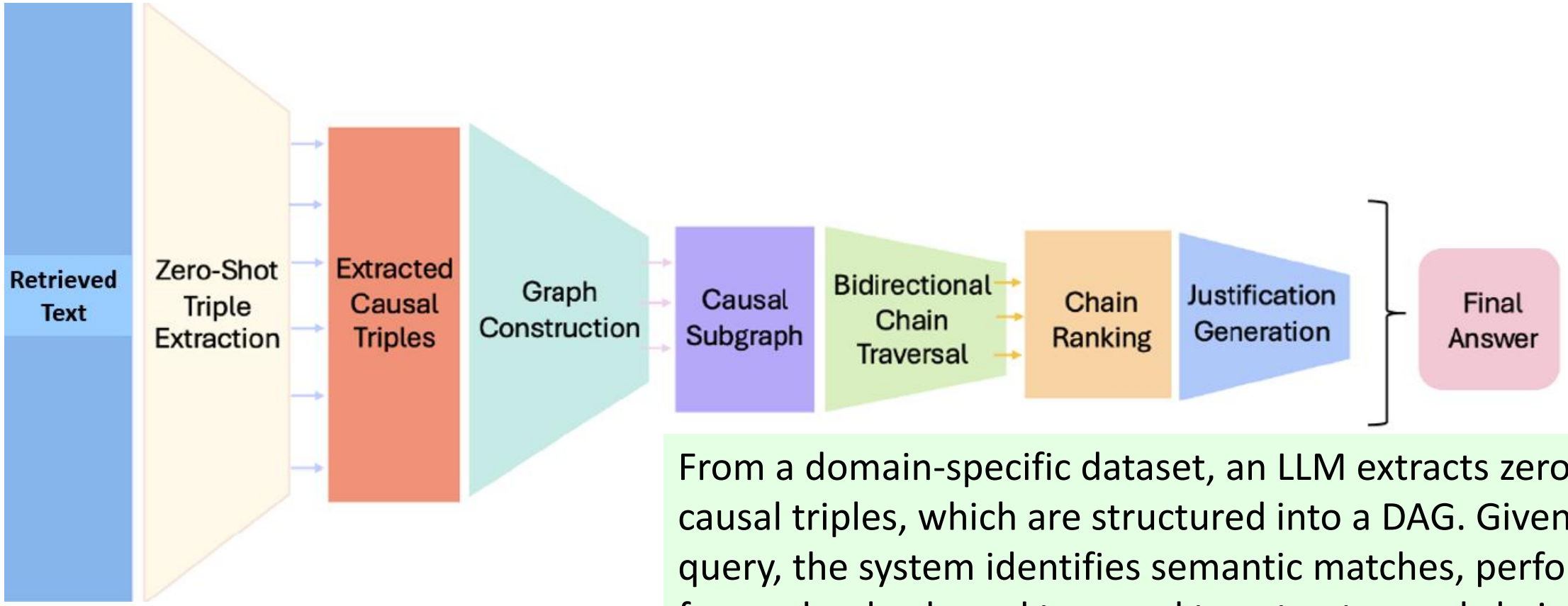
Why SARG: Structure-Augmented Reasoning Generation?

- **Standard RAG** treats evidence as *flat context*, lacking the structure required to model true causal dependencies
- **Classical causal inference** focuses on statistical associations and are not well equipped to extract narratives from unstructured, implicit text, across documents
- **Reasoning Generation** integrates *zero-shot triple extraction* and *theme-aware graph chaining* into the retrieved text, enabling structured multi-hop inference
- Given a domain specific corpus, it *constructs a DAG* of ⟨cause, relation, effect⟩ triples and uses *forward/backward chaining* to guide structured answer generation
- Experiments on two real-world domains: **Bitcoin price** (BP) & **Gaucher disease** (GD)
  - SARG outperforms standard RAG and zero-shot LLMs on *chain similarity, information density, lexical diversity, LLM-as-a-Judge*, and *human evaluations*
- Explicitly modeling causal structure enables LLMs to *generate more accurate and interpretable responses*, especially in specialized domains where flat retrieval fails

Jash Parekh, P. Jiang, J Han, "Structured Multi-Hop Augmented Reasoning Generation“, arXiv:2506.08364

The SARG Framework

- SARG: Structuring and Construction of Reasoning Graphs for LLM Reasoning Generation



From a domain-specific dataset, an LLM extracts zero-shot causal triples, which are structured into a DAG. Given a query, the system identifies semantic matches, performs forward or backward traversal to extract causal chains and generates a justification-based answer using an LLM

Jash Parekh, Pengcheng Jiang, Jiawei Han, "SARG: Structure Augmented Reasoning Generation", arXiv:2506.08364

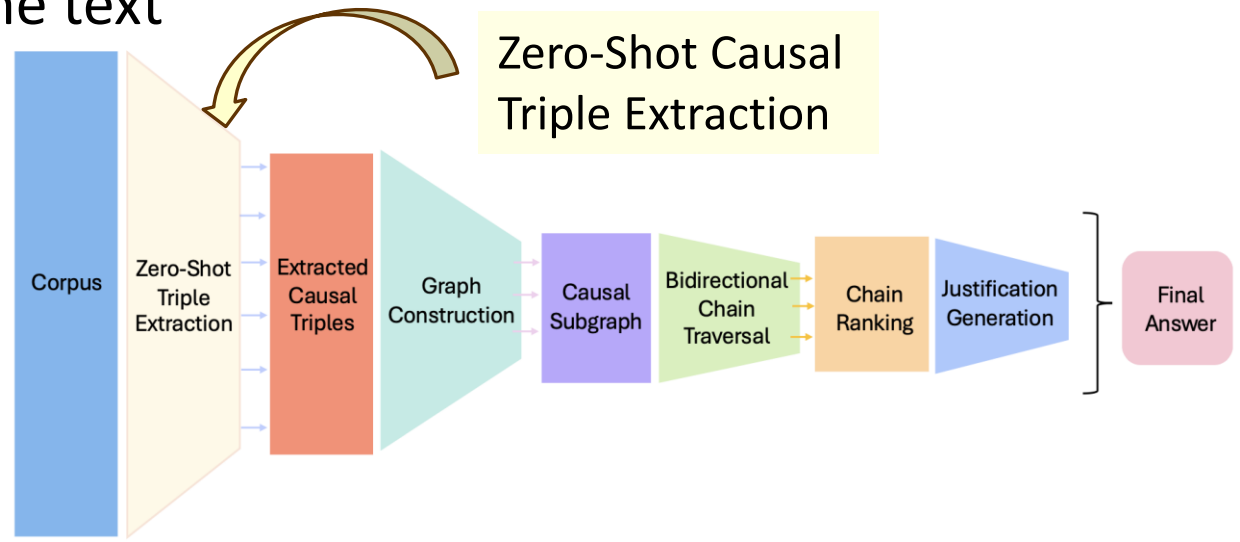
Zero-Shot Extraction of Causal Triples

- SARG uses GPT-4o to extract both explicit and implicit ⟨cause, relation, effect⟩ triples from unstructured text, without requiring labeled training data
  - **Cause** refers to an entity, event, or action that triggers an outcome, even if the causal connection is not explicitly stated
  - **Relation** is a *causal verb or phrase* (e.g., caused, led to, resulted in, triggered, influenced), or an *inferred connection* that is understood contextually
  - **Effect** represents the resulting entity, event, or action, regardless of whether the causal relationship is directly stated in the text

$$\text{Traditional RAG:} \quad q, \mathcal{D} \rightarrow \text{LLM} \rightarrow r$$

$$\text{SARG:} \quad q, \mathcal{D} \rightarrow \mathcal{G} \rightarrow \text{paths} \rightarrow \text{LLM} \rightarrow r$$

- G represents the constructed knowledge graph
- *paths* capture multi-hop inference trajectories



SARG Methodology: Graph-Based Multi-Hop Reasoning

- Entity Extraction and Graph Construction
  - For each document $di \in$ D, we extract a set of knowledge triples Ti using zero-shot prompting: $T_i = \mathrm{LLM}_{\mathrm{extract}}(d_i, P_{\mathrm{extract}})$
    - Each triple $t = \langle c, r, e \rangle \in Ti$ represents a directed relationship where c and e are entities and r is the relation type
    - The complete set of extracted triples constructs a directed knowledge graph G
- Entity Clustering: Cluster semantically similar entities w. SentenceBERT embeddings
  - Reduce graph fragmentation & improve reasoning chain connectivity
- Semantic Node Matching: To identify relevant starting points for graph traversal, perform semantic matching between query terms and graph nodes
- Graph-Based Multi-Hop Traversal
  - Direction classification: {forward, backward, bi-directional}
  - Path discovery (by depth-first traversal)

SARG: Chain Ranking and LLM-Guided Answer Generation

- **Chain-Ranking and Selection:** All reasoning chains are scored using semantic similarity between the original query and the aggregated evidence
  - This semantic scoring ensures that selected chains are both on-topic and coherent in supporting the user's question
  - Top-k chains are selected for final generation
- **LLM-Based Answer Generation:** synthesize selected reasoning chains into a coherent response
  - *Evidence Compilation*: For each selected chain c, compile supporting evidence by retrieving original text snippets that led to each triple, creating structured evidence packages E(c) with source traceability
  - *Prompt Construction:* Pgen = Pinst $\oplus$ Pquery $\oplus$ Pchains $\oplus$ Pevidence
    - where $\oplus$ denotes concatenation, combining task instructions, the original query, serialized reasoning chains, and compiled evidence.
- Response Generation. The final response is generated as: r = LLMgen (Pgen)
  - The model is instructed to maintain logical coherence with provided reasoning chains while producing natural, human-readable text

LLM-Powered Output Generation with Justification

- Case study comparing SARG w. expert-annotated triples for answering a biomedical question
- SARG successfully reconstructs a multi-hop pathway, which the human-annotated KG fails to recover

SARG's key advantages:

- **Plug-and-Play Compatibility**: integration into any existing RAG pipeline
- **Interpretability:** graph and chain selection provides clear reasoning visibility
- **Domain Adaptability:** zero-shot triple extraction; no domain-specific training

Question: How do susceptibility genes ultimately impact the timely treatment of diabetic ketoacidosis?

Full text

Fulminant type 1 diabetes (FT1D) is a novel type of type 1 diabetes that is caused by extremely rapid destruction of the pancreatic β-cells. Early diagnosis or prediction of FT1D is critical for the prevention or timely treatment of diabetic ketoacidosis, which can be life-threatening. Understanding its triggers or promoting factors plays an important role in the prevention and treatment of FT1D. In this review, we summarised the various triggering factors of FT1D, including susceptibility genes, immunological factors (cellular and humoural immunity), immune checkpoint inhibitor therapies, drug reactions with eosinophilia and systemic symptoms or drug-induced hypersensitivity syndrome, pregnancy, viral infections, and vaccine inoculation. This review provides the basis for future research into the pathogenetic mechanisms that regulate FT1D development and progression to further improve the prognosis and clinical management of patients with FT1D.

Human Annotated Chains

<immune checkpoint inhibitor therapies, FT1D>, <susceptibility genes, FT1D>, <susceptibility genes, diabetic ketoacidosis>, <FT1D, diabetic ketoacidosis>, <pancreatic β-cells, diabetic ketoacidosis>, <viral infections, pancreatic β-cells>, <immunological factors, FT1D>, <pregnancy, FT1D>, <drug reactions with eosinophilia and systemic symptoms, FT1D>

SARG Extracted Chains

<extremely rapid destruction of the pancreatic β-cells, causes, FT1D>, <FT1D, leads to, diabetic ketoacidosis>, <early diagnosis or prediction of FT1D, is critical for, prevention of diabetic ketoacidosis>, <**FT1D, necessitates, diagnosis or prediction**>, <**early diagnosis or prediction of FT1D, is critical for, timely treatment of diabetic ketoacidosis**>, <understanding its triggers or promoting factors, plays an important role in, prevention of FT1D>, <understanding its triggers or promoting factors, plays an important role in, treatment of FT1D>, <**susceptibility genes, triggers, FT1D**>, <cellular and humoural immunity, triggers, FT1D>, <immune checkpoint inhibitor therapies, triggers, FT1D>, ..., <pregnancy, triggers, FT1D>, <viral infections, triggers, FT1D>, <vaccine inoculation, triggers, FT1D>, ..., <pathogenetic mechanisms that regulate FT1D development and progression, inform, future research into FT1D>, <pathogenetic mechanisms that regulate FT1D development and progression, underlie, improved prognosis and clinical management of patients with FT1D>

The causal context provides the following relevant causal paths:
1. Susceptibility genes directly cause FT1D.
2. FT1D then causes diabetic ketoacidosis.

Additionally:
- Susceptibility genes directly cause diabetic ketoacidosis.

However, the context does not provide any causal link between susceptibility genes or diabetic ketoacidosis and the timely treatment of diabetic ketoacidosis. Therefore, based on the provided causal context, "**The causal context is insufficient to answer this question.**"

To determine the impact of susceptibility genes on the timely treatment of diabetic ketoacidosis, we can trace the causal chain as follows:

1. Susceptibility genes --triggers--> FT1D
2. FT1D --necessitates--> diagnosis or prediction
3. Early diagnosis or prediction of FT1D --is critical for--> timely treatment of diabetic ketoacidosis

From these relationships, we can deduce that susceptibility genes, by triggering FT1D, indirectly necessitate the need for diagnosis or prediction of FT1D. Early diagnosis or prediction of FT1D, in turn, is critical for the timely treatment of diabetic ketoacidosis. Thus, susceptibility genes ultimately impact the timely treatment of diabetic ketoacidosis by being an initial trigger within this causal chain.

Performance Comparison: SARG vs. RAG vs. Zero-Shot

Data Statistics: BP (Bitcoin Price) and GD (Gaucher Disease)

Dataset	Documents	Entities	Triples
BP	20	317	172
GD	20	573	419

- **BERTScore** (Chain similarity): semantic similarity using contextual embeddings from RoBERTa-large
- **Conciseness** (Info. density): Ratio of content words to total words, scaled by the inverse log of response length
- FactCC (Factual Consistency): Ensure that generated answers remain grounded in the retrieved corpus

Automatic Evaluation: SARG vs. RAG vs. Zero-Shot

			Qwen			LLaMA			DeepSeek		
Metric	**GraphRAG**	**HippoRAG**	**SARG**	**RAG**	**Zero**	**SARG**	**RAG**	**Zero**	**SARG**	**RAG**	**Zero**
Bitcoin Price Fluctuations (BP)											
BERTScore	0.84	0.88	**0.90**	0.83	0.74	0.88	0.82	0.77	0.86	0.83	0.80
FactCC	0.96	0.91	**0.99**	0.85	0.35	0.98	0.86	0.49	0.98	0.86	0.46
Conciseness	0.07	**0.14**	0.09	0.05	0.07	0.08	0.04	0.02	0.09	0.06	0.07
Gaucher Disease (GD)											
BERTScore	0.82	0.85	**0.89**	0.80	0.71	0.85	0.80	0.72	0.86	0.82	0.79
FactCC	0.73	**0.85**	**0.85**	0.76	0.63	0.81	0.72	0.65	0.8	0.75	0.70
Conciseness	0.07	**0.15**	0.09	0.06	0.07	0.08	0.05	0.06	0.09	0.06	0.06

Evaluation of Summarization Quality by LLM and Human

- **LLM-as-a-Judge: A** blinded LLM-as-a-Judge evaluation
  - Using a panel of four LLMs: GPT-4, GPT-4o, LLaMA 3.1-8B-Instruct, and Mistral-7B-Instruct
  - Each judge model was prompted with a fixed template and shown anonymized answers from all systems. Models were asked to select the best response based on accuracy, interpretability, and conciseness
  - Use majority voting to determine the preferred answer
- **Human Evaluation**: Aggregate the preferences of three independent reviewers for each form to assess performance

LLM-as-a-Judge: Bitcoin Price Fluctuations (BP)

Method	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
GraphRAG	5.56	0.00	0.00
HippoRAG	0.00	11.11	11.11
Qwen-SARG	**72.22**	5.56	5.56
DeepSeek-SARG	5.56	**50.00**	5.56
LLaMA-SARG	5.56	11.11	**50.00**
Qwen-RAG	0.00	11.11	5.56
LLaMA-RAG	0.00	0.00	5.56
Qwen Zero-Shot	5.56	5.56	5.56
DeepSeek Zero-Shot	0.00	5.56	5.56
LLaMA Zero-Shot	5.56	0.00	5.56

Method	Accuracy (%)
Qwen-SARG	**83.0**
HippoRAG	81.0
GraphRAG	80.0
Qwen-RAG	67.0
Qwen Zero-Shot	28.0

Accuracy on a HotPotQA hard-100 subset

Human evaluation results showing preferred responses across BP and GD datasets. #s indicate votes received out of total questions evaluated

Method	BP	GD
Qwen-SARG	8/18	12/25
Qwen-RAG)	2/18	1/25
Qwen Zero-Shot	0/18	1/25
GraphRAG	3/18	4/25
HippoRAG	5/18	7/25

Case Study I: Bitcoin Price Fluctuation

Query 1

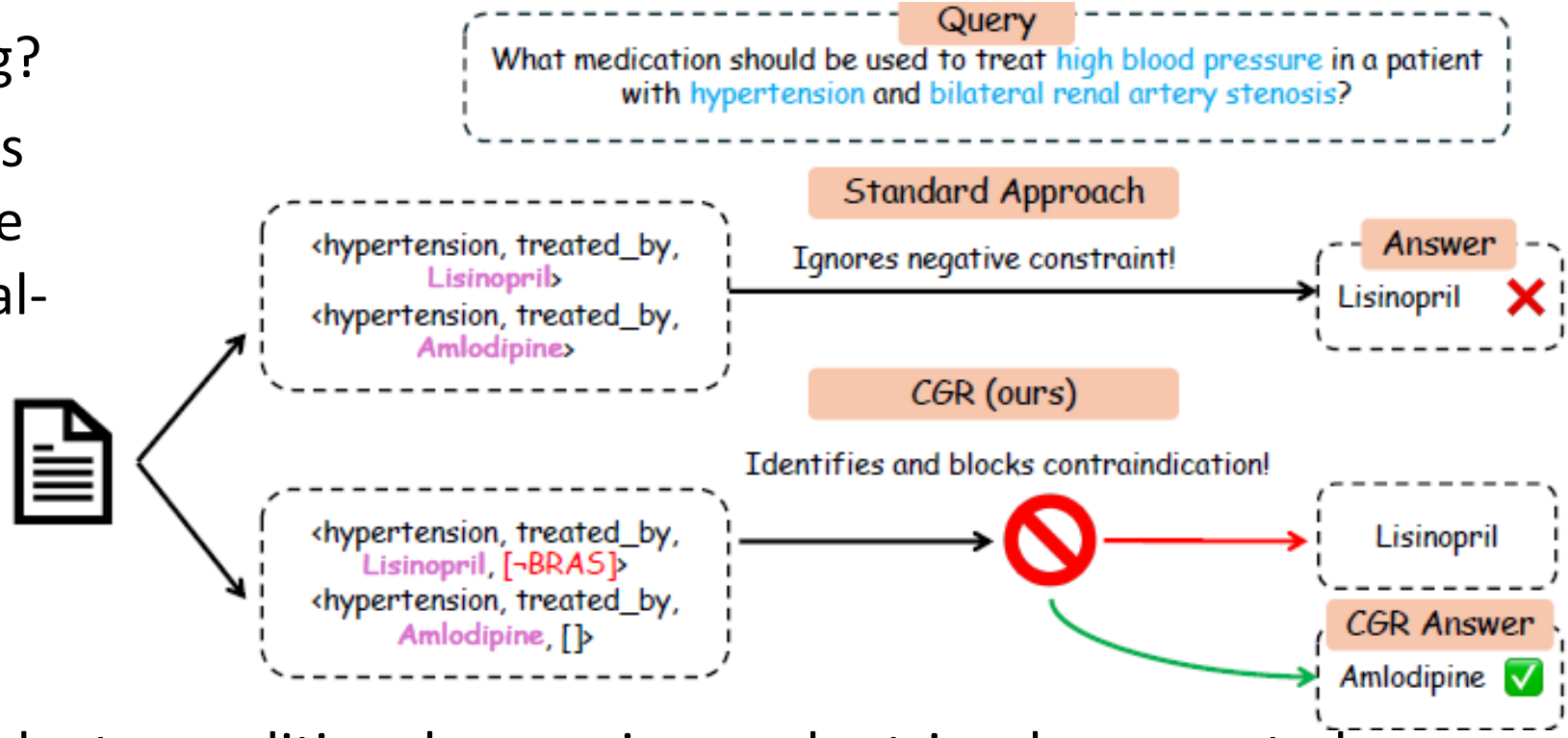
Query	How did Andrew Yang's advocacy for blockchain technology influence Bitcoin's adoption?
CC-RAG	Andrew Yang's advocacy for blockchain technology brought increased public and political awareness to cryptocurrencies, such as Bitcoin. By discussing the potential benefits of blockchain, such as transparency and efficiency, Yang helped legitimize the technology in the eyes of the public and policymakers. This likely contributed to broader acceptance and interest in Bitcoin, although direct causation to adoption rates is harder to measure.
GPT-4o with RAG	The context provided does not mention Andrew Yang or his advocacy for blockchain technology. Therefore, there's no information on how Andrew Yang's advocacy might have influenced Bitcoin's adoption.
Zero-Shot GPT-4o	N/A

Query	How did global economic uncertainty contribute to Bitcoin's price rise?
CC-RAG	Global economic uncertainty can lead to increased demand for Bitcoin as investors look for alternative assets that are perceived as safe havens. This heightened demand for Bitcoin drives up its price.
GPT-4o with RAG	Several contributing factors likely include: 1. **Speculative Investment**: The overall sentiment captured in the passage indicates that part of Bitcoin's price surge can be attributed to increased speculative investment, driven by optimism about a favorable regulatory environment and inflows of new investors, as detailed by the remarks from Chris Akhavan and Joel Kruger. 2. **Mainstream Adoption**: Another key influence is the increased mainstream adoption of Bitcoin, highlighted by the SEC's approval of new exchange-traded funds directly investing in cryptocurrency. This move allowed major global asset managers like BlackRock Inc. to participate, signaling institutional confidence and involvement in the market. 3. **Market Momentum**: Bitcoin's historic performance and recognition, compared to assets like gold and influenced by significant market movements and high-profile endorsements, have contributed to its rising status as a major asset class. Overall, while the context does not provide a direct explanation of how global economic uncertainty played a role, the speculative nature of investments and the moves towards mainstream adoption in a possibly volatile economic environment could be seen as indirect contributors to Bitcoin's price rise.
Zero-Shot GPT-4o	N/A

Query 2

CGR: Condition-Gated Reasoning for Context-Dependent Biomedical Question Answering

- Why Condition-Gated Reasoning?
  - Existing biomedical QA models often treat medical knowledge as static and universal, but real-world clinical decisions are **inherently conditional**—they depend on factors like comorbidities, allergies, and contraindications

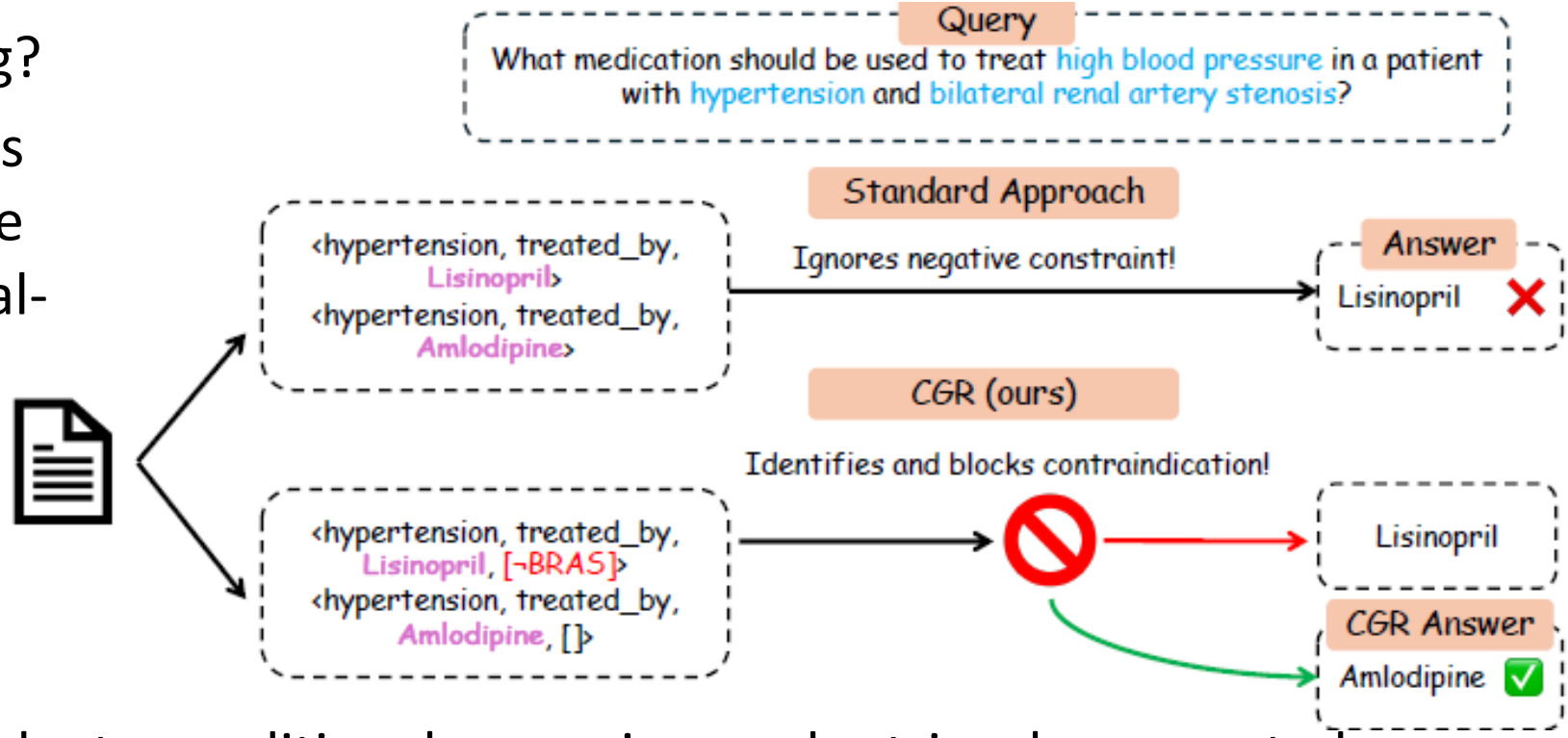


- Many benchmarks do not evaluate conditional reasoning, and retrieval-augmented or graph-based methods lack mechanisms to ensure retrieved knowledge is relevant to a given patient's context

Jash Rajesh Parekh, Wonbin Kweon, Joey Chan, Rezarta Islamaj, Robert Leaman, Pengcheng Jiang, Chih-Hsuan Wei, Zhizheng Wang, Zhiyong Lu, Jiawei Han, "Condition-Gated Reasoning for Context-Dependent Biomedical Question Answering", KDD 2026

CGR: Condition-Gated Reasoning for Context-Dependent Biomedical Question Answering

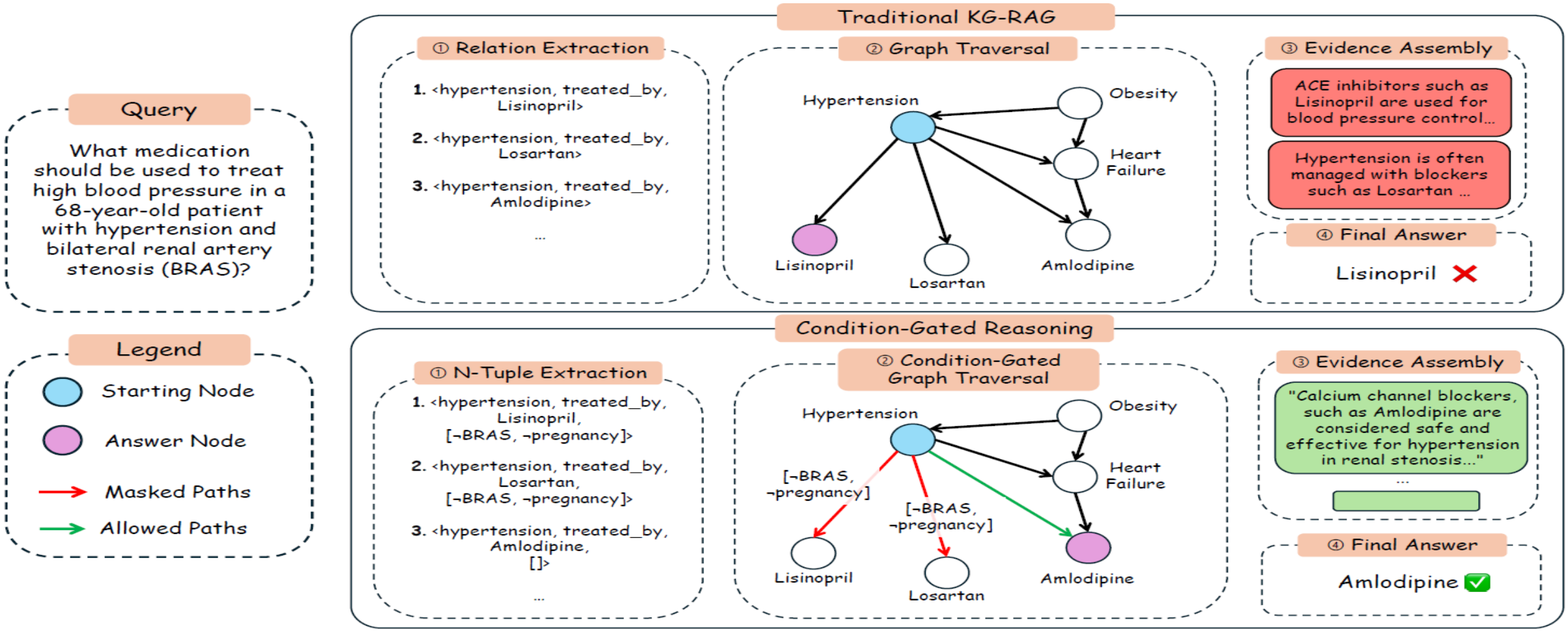
- Why Condition-Gated Reasoning?
  - Existing biomedical QA models often treat medical knowledge as static and universal, but real-world clinical decisions are **inherently conditional**—they depend on factors like comorbidities, allergies, and contraindications



- Many benchmarks do not evaluate conditional reasoning, and retrieval-augmented or graph-based methods lack mechanisms to ensure retrieved knowledge is relevant to a given patient's context

Jash Rajesh Parekh, Wonbin Kweon, Joey Chan, Rezarta Islamaj, Robert Leaman, Pengcheng Jiang, Chih-Hsuan Wei, Zhizheng Wang, Zhiyong Lu, Jiawei Han, "Condition-Gated Reasoning for Context-Dependent Biomedical Question Answering", arXiv:2602.17911

Overview of Condition-Gated Reasoning (CGR) compared to traditional KG-RAG



Given a clinical query about a patient with hypertension and bilateral renal artery stenosis (BRAS), traditional KG-RAG extracts standard relation triples and traverses all paths indiscriminately, retrieving evidence for contraindicated treatments (e.g., Lisinopril, an ACE inhibitor). CGR extends triples to n-tuples that include patient-specific conditions as gating constraints (e.g., ¬BRAS, ¬pregnancy), masking contraindicated paths during graph traversal and assembling only condition-appropriate evidence, yielding the correct answer (Amlodipine).

Types of Conditional Multi-hop Reasoning Required in CondMedQA

Reasoning Category	%	Example(s)
Comorbidity & Organ-Based Contraindications	57	**Doc 1:** Treatment for *major depressive disorder* includes *antidepressants* such as **bupropion**, sertraline, and *fluoxetine*. ***Bupropion*** acts as a norepinephrine-dopamine reuptake inhibitor. **Doc 2:** ***Bupropion*** lowers the seizure threshold and is contraindicated in patients with *bulimia nervosa* or other *eating disorders* due to elevated seizure risk. **Q:** Which antidepressant for MDD is contraindicated in patients with bulimia? **A:** **Bupropion** (w/o condition: Sertraline)
Diagnostic Modality Selection	23	**Doc 1:** For suspected *appendicitis*, diagnostic imaging options include CT scan, **ultrasound**, and *MRI*. *CT scan* has the highest sensitivity in adult populations. **Doc 2:** In *pediatric patients*, ***ultrasound*** is the preferred first-line imaging modality to avoid *ionizing radiation exposure*, with CT reserved for inconclusive cases. **Q:** What imaging test is preferred for suspected appendicitis in children? **A:** **Ultrasound** (w/o condition: CT Scan)
Special Population Safety	16	**Doc 1:** First-line antibiotics for *Lyme disease* include doxycycline, **amoxicillin**, and *cefuroxime*. ***Doxycycline*** is preferred in adults due to co-coverage of *Anaplasma* co-infection. **Doc 2:** ***Doxycycline*** is contraindicated during *pregnancy* due to risks to fetal *bone and teeth development*. **Amoxicillin** is considered safe across all trimesters. **Q:** What antibiotic is recommended for Lyme disease in a pregnant patient? **A:** **Amoxicillin** (w/o condition: Doxycycline)
Drug-Drug Interactions & Pharmacogenomics	4	**Doc 1:** Standard therapy for *tuberculosis* includes rifampin, *isoniazid*, *pyrazinamide*, and *ethambutol*. ***Rifabutin*** is an alternative rifamycin with a similar mechanism. **Doc 2:** Rifampin is a potent inducer of CYP3A4 and is contraindicated with *HIV protease inhibitors*. ***Rifabutin*** has fewer CYP3A4 interactions and is preferred in patients on *antiretroviral therapy*. **Q:** Which drug replaces rifampin in TB treatment for HIV patients on protease inhibitors? **A:** **Rifabutin** (w/o condition: Rifampin)

Our Method: Condition-Gated Reasoning

- **Condition-Gated Reasoning (CGR) Framework**
  - **Constructs condition-aware knowledge graphs**: Nodes and edges encode medical facts with explicit condition tags (e.g., “safe for patient with diabetes”).
  - **Gates reasoning paths**: Selectively activates or prunes reasoning paths based on the query’s patient conditions
  - Ensures only condition-appropriate knowledge is used in inference

Each edge $e = \langle u, r, v, C \rangle$

collect all unique conditions $C_G = \bigcup_{e \in E} C_e$

eval : $C_G \times q \rightarrow \{\texttt{True}, \texttt{False}, \texttt{Null}\}^{|C_G|}$

gating function:

$$\mathcal{G}(e, \mathcal{L}) = \prod_{c \in C_e} \mathbb{1}[\mathcal{L}(c) \neq \texttt{false}]$$

$$\text{score}(q, p) = \sum_{k \in K} \cos\left(\phi(p_{\text{end}}), \phi(k)\right),$$

- $\langle$hypertension, treated_by, lisinopril, [¬BRAS, ¬pregnancy]$\rangle$
  — **blocked**, since $\mathcal{L}(\neg\text{BRAS}) = \texttt{false}$
- $\langle$hypertension, treated_by, losartan, [¬BRAS]$\rangle$
  — **blocked**, since $\mathcal{L}(\neg\text{BRAS}) = \texttt{false}$
- $\langle$hypertension, treated_by, amlodipine, []$\rangle$
  — **traversed**, no conditions violated

$$\mathcal{E}_q = \{(V_p, E_p, S_p, C_p)\}_{p \in \mathcal{P}_q^N},$$

$$A = \text{LLM}\Big(\underbrace{q \oplus \mathcal{P}_q^N \oplus \mathcal{E}_q \oplus \texttt{instructions}}_{\text{prompt}} \Big),$$

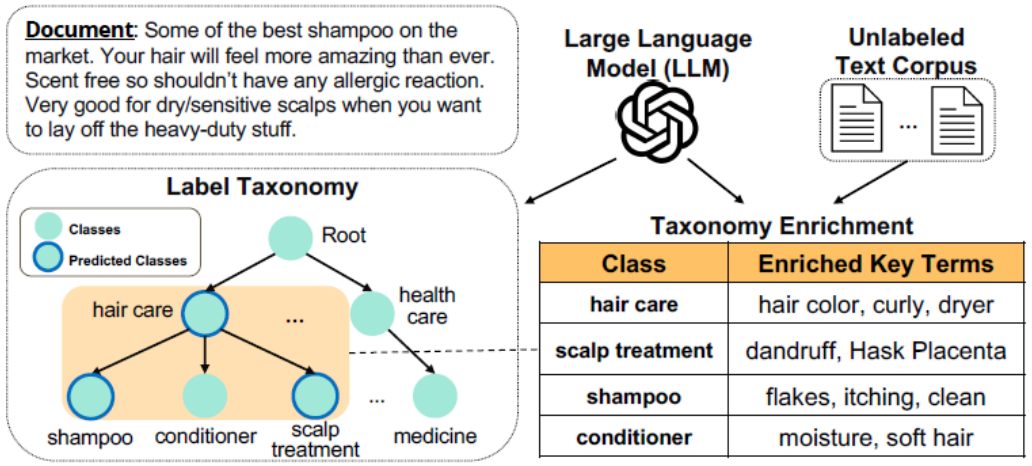
Performance across biomedical multi-hop QA Datasets and our Custom Benchmark

Method	CondMedQA		MedHopQA		MedHopQA (Cond)		BioASQ-B	
	EM	F1	EM	F1	EM	F1	EM	F1
GPT-5.2								
Zero-Shot	22.00	41.29	20.50	47.87	28.27	53.52	13.00	40.21
RAG	44.00	52.36	51.25	62.61	45.71	55.04	31.80	54.80
CoT	40.00	52.81	49.50	57.55	50.00	59.38	25.60	47.81
HippoRAG2	46.00	67.56	45.00	60.81	48.57	72.82	22.60	43.57
HyperGraphRAG	57.00	64.48	37.50	41.90	54.29	65.24	28.60	54.44
StructRAG	62.00	71.22	71.75	79.61	65.71	78.57	31.80	55.07
PathRAG	53.00	66.54	30.00	34.36	42.86	54.07	24.55	51.04
MedRAG	57.00	70.52	75.75	84.42	74.29	**89.31**	35.40	55.51
MKRAG	60.00	71.14	51.25	65.56	62.86	79.33	28.80	55.35
CGR (ours)	**82.00**	**86.67**	**86.75**	**89.91**	**80.00**	85.44	**40.60**	**57.99**
Qwen2.5-14B-Instruct								
Zero-Shot	34.00	39.04	43.20	51.14	34.29	43.09	15.80	33.44
RAG	44.00	52.01	45.50	55.54	42.86	50.19	31.80	**47.81**
CoT	49.00	53.50	42.50	52.44	37.14	46.24	20.80	34.88
CGR (ours)	**55.00**	**64.47**	**61.30**	**67.40**	**51.43**	**54.11**	**34.60**	44.08
LLaMA-3.1-8B-Instruct								
Zero-Shot	30.00	36.17	34.75	43.75	37.14	48.78	15.60	35.07
RAG	47.00	52.36	46.75	55.17	45.71	50.09	33.40	**53.66**
CoT	50.00	**58.67**	40.50	47.82	37.14	48.16	18.40	32.01
CGR (ours)	**52.00**	54.17	**61.50**	**67.36**	**48.57**	**50.48**	**35.80**	47.21

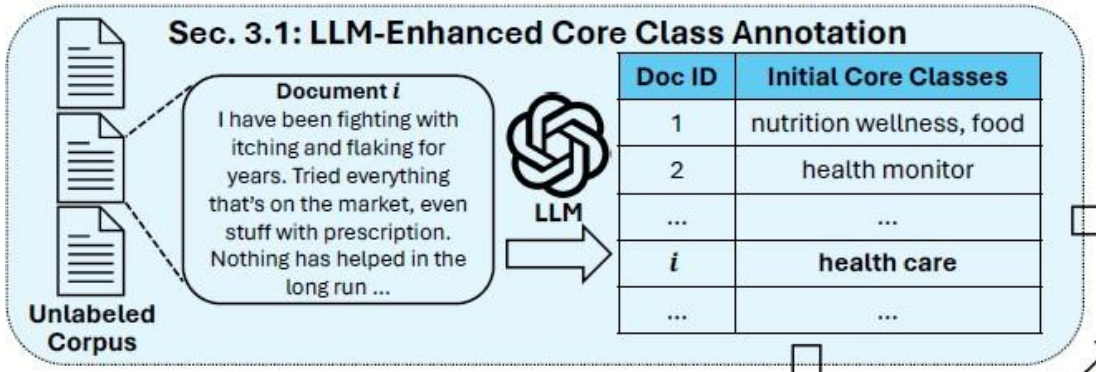
Outline

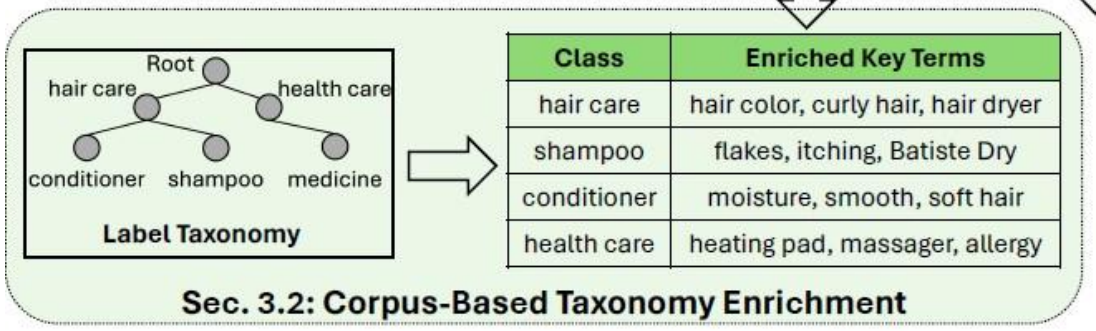
- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring 
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

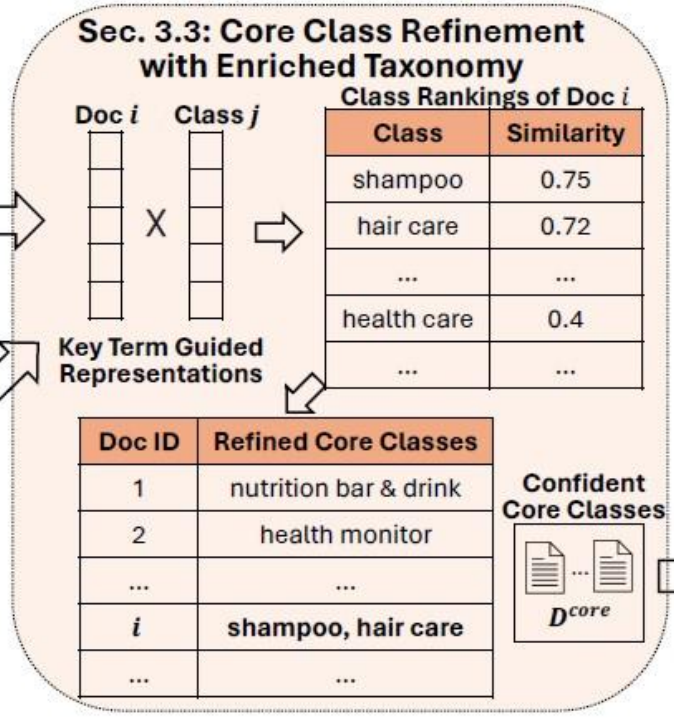
TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision

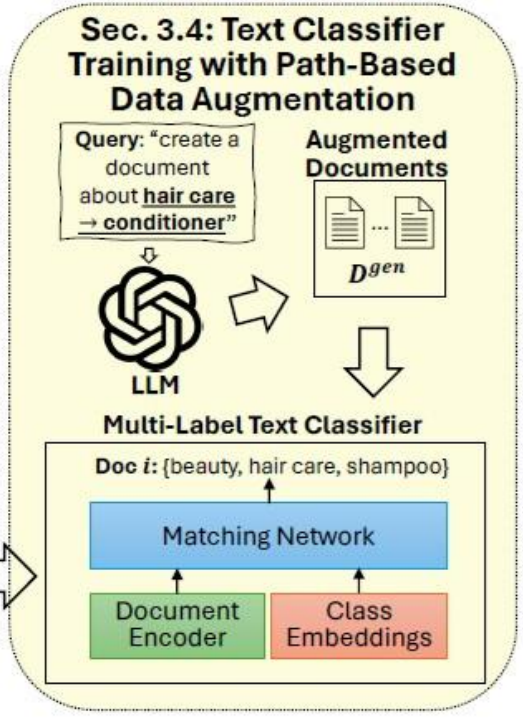


- Task: *Classifying documents into 10^2 to 10^3 classes, without human annotation*?
- Automatically enrich the label taxonomy with class-indicative topical terms mined from the corpus to facilitate classifier training
- Use LLMs for both data annotation and creation tailored for the hierarchical label space









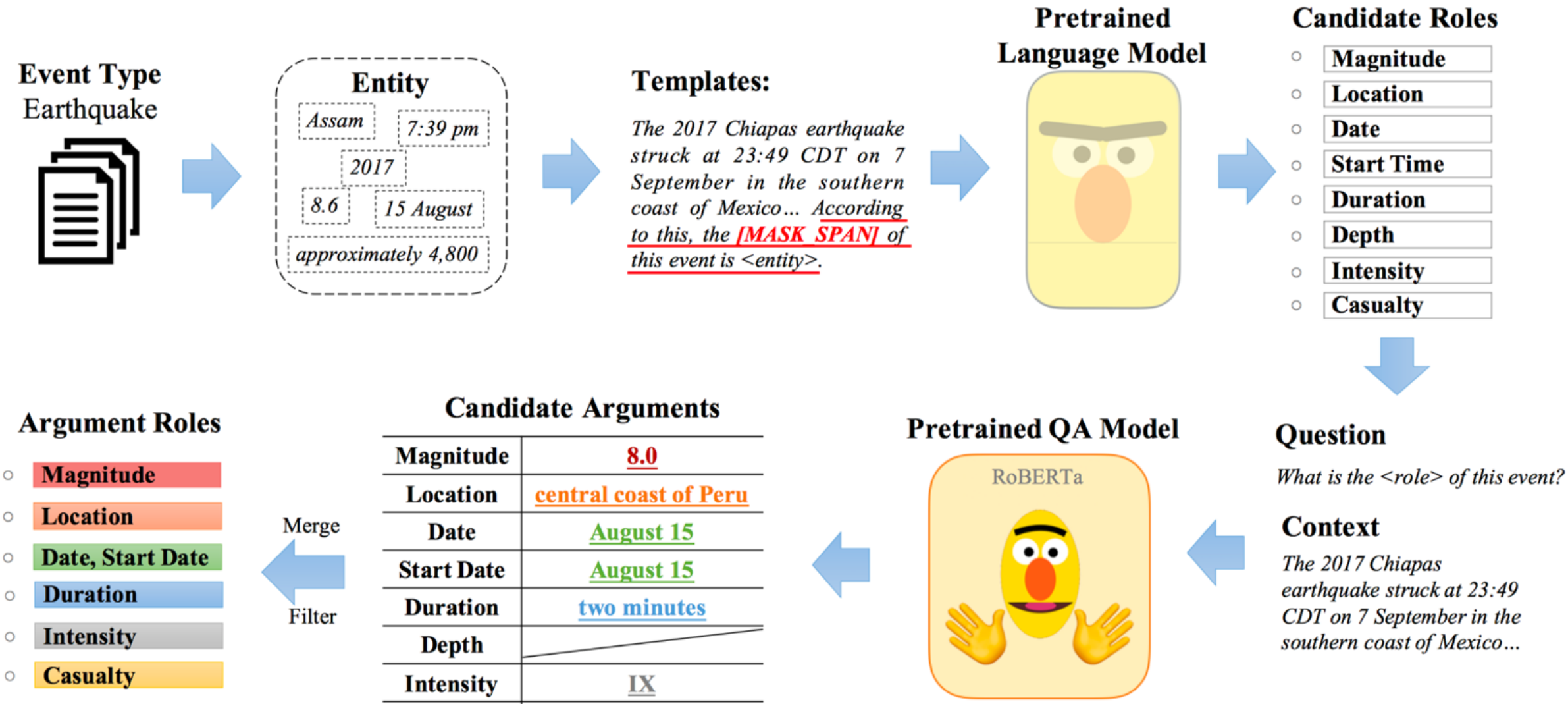
TELEClass: Performance Study and Cost for Text Classification

Supervision Type	Methods	Amazon-531				DBPedia-298			
		Example-F1	P@1	P@3	MRR	Example-F1	P@1	P@3	MRR
Zero-Shot	Hier-0Shot-TC†	0.4742	0.7144	0.4610	—	0.6765	0.7871	0.6765	—
	GPT-3.5-turbo	0.5164	0.6807	0.4752	—	0.4816	0.5328	0.4547	—
Weakly-Supervised	Hier-doc2vec†	0.3157	0.5805	0.3115	—	0.1443	0.2635	0.1443	—
	WeSHClass†	0.2458	0.5773	0.2517	—	0.3047	0.5359	0.3048	—
	TaxoClass-NoST†	0.5431	0.7918	0.5414	0.5911	0.7712	0.8621	0.7712	0.8221
	TaxoClass†	0.5934	0.8120	0.5894	0.6332	0.8156	0.8942	0.8156	0.8762
	TELEClass	**0.6330**	**0.8439**	**0.6269**	**0.6664**	**0.8684**	**0.9293**	**0.8684**	**0.8880**
Fully-Supervised		0.8843	0.9524	0.8758	0.9085	0.9786	0.9945	0.9786	0.9826

Methods	Amazon-531					DBPedia-298				
	Example-F1	P@1	P@3	Est. Cost	Est. Time	Example-F1	P@1	P@3	Est. Cost	Est. Time
GPT-3.5-turbo	0.5164	0.6807	0.4752	$60	240 mins	0.4816	0.5328	0.4547	$80	400 mins
GPT-3.5-turbo (level)	0.6621	0.8574	0.6444	$20	800 mins	0.6649	0.8301	0.6488	$60	1,000 mins
GPT-4‡	0.6994	0.8220	0.6890	$800	400 mins	0.6054	0.6520	0.5920	$2,500	1,000 mins
TELEClass	0.6330	0.8439	0.6269	<$1	3 mins	0.8684	0.9293	0.8684	<$1	7 mins

Yunyi Zhang, et al., "TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision", WWW'25

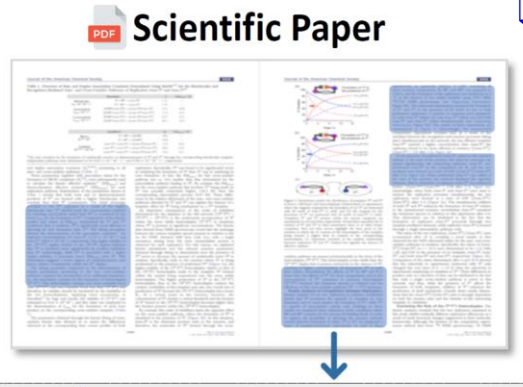
RolePred: Argument Role Prediction [EMNLP'22]



Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji and Jiawei Han "Open-Vocabulary Argument Role Prediction for Event Extraction", EMNLP'22

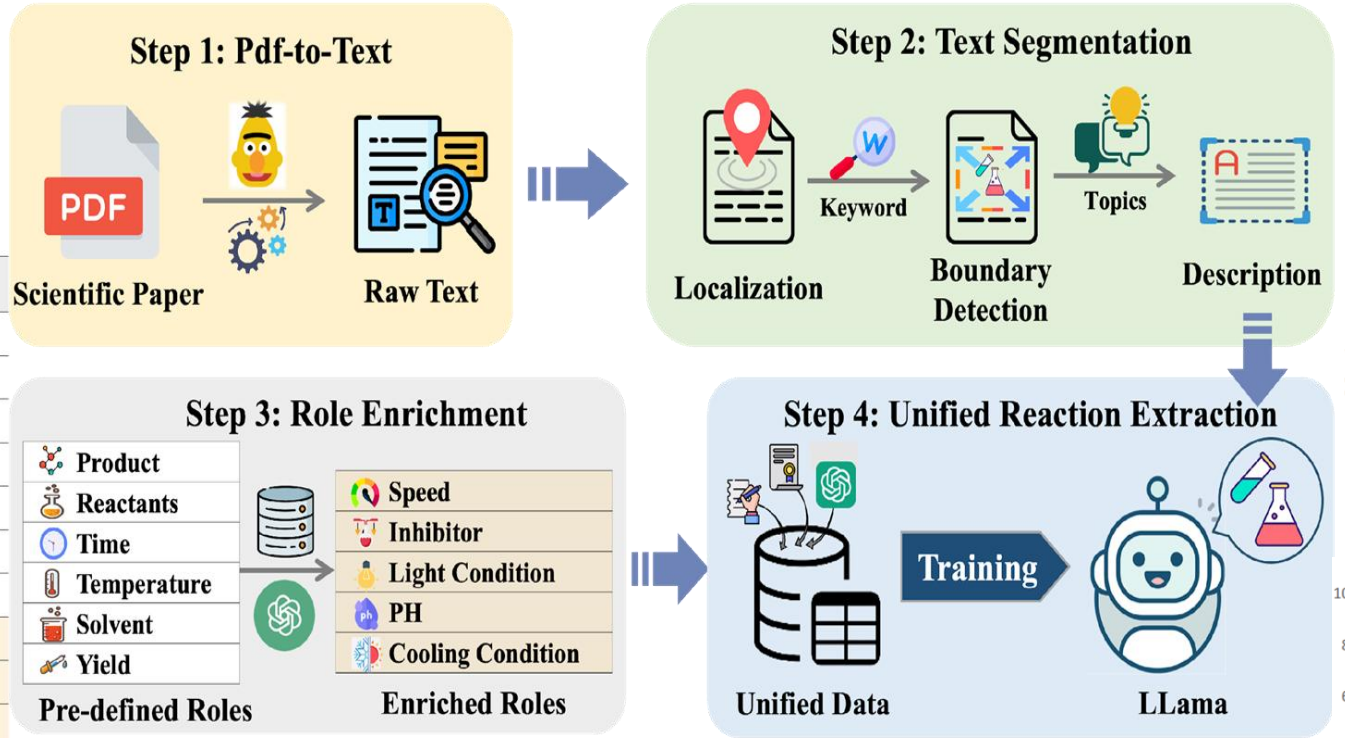
Reaction Miner: Chemical Reaction Info Extraction from Text Data

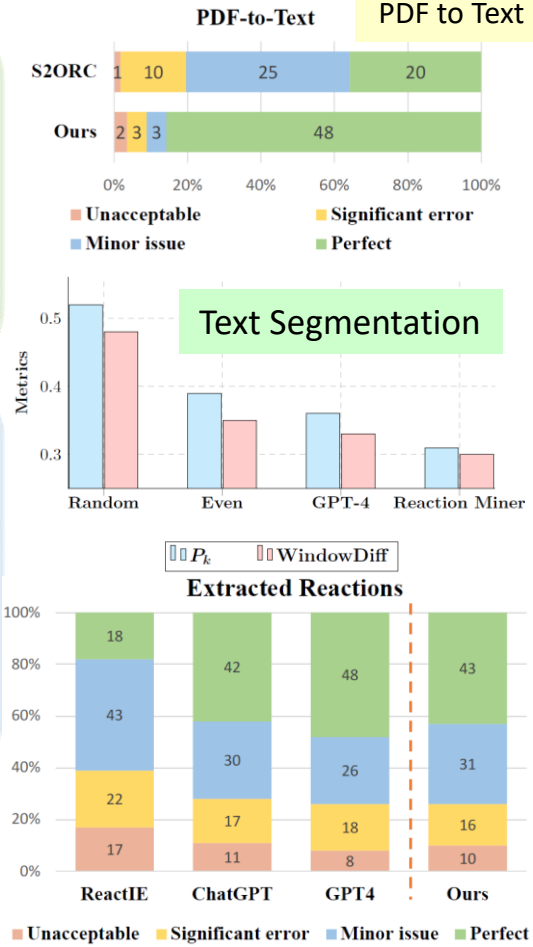
- Automatic extraction of chemical reaction information (e.g., reactants, catalyst, temperature, etc.)
  - Role enrichment: Use RolePredict to extract new reaction roles and generate synthetic data corresponding to each role based on GPT-4 for training



Chemical Reaction

Product	2-chlorophenol
Reactants	phenol, oxalyl chloride
Time	3 hours
Temperature	10 °C
Solvent	NaOH solution
Yield	78% (2-chlorophenol)
Workup Reagent	6 M HCl
Speed	500 rpm
Inhibitor	BHT
Light Condition	dark
PH	13.2
Cooling Condition	ice bath





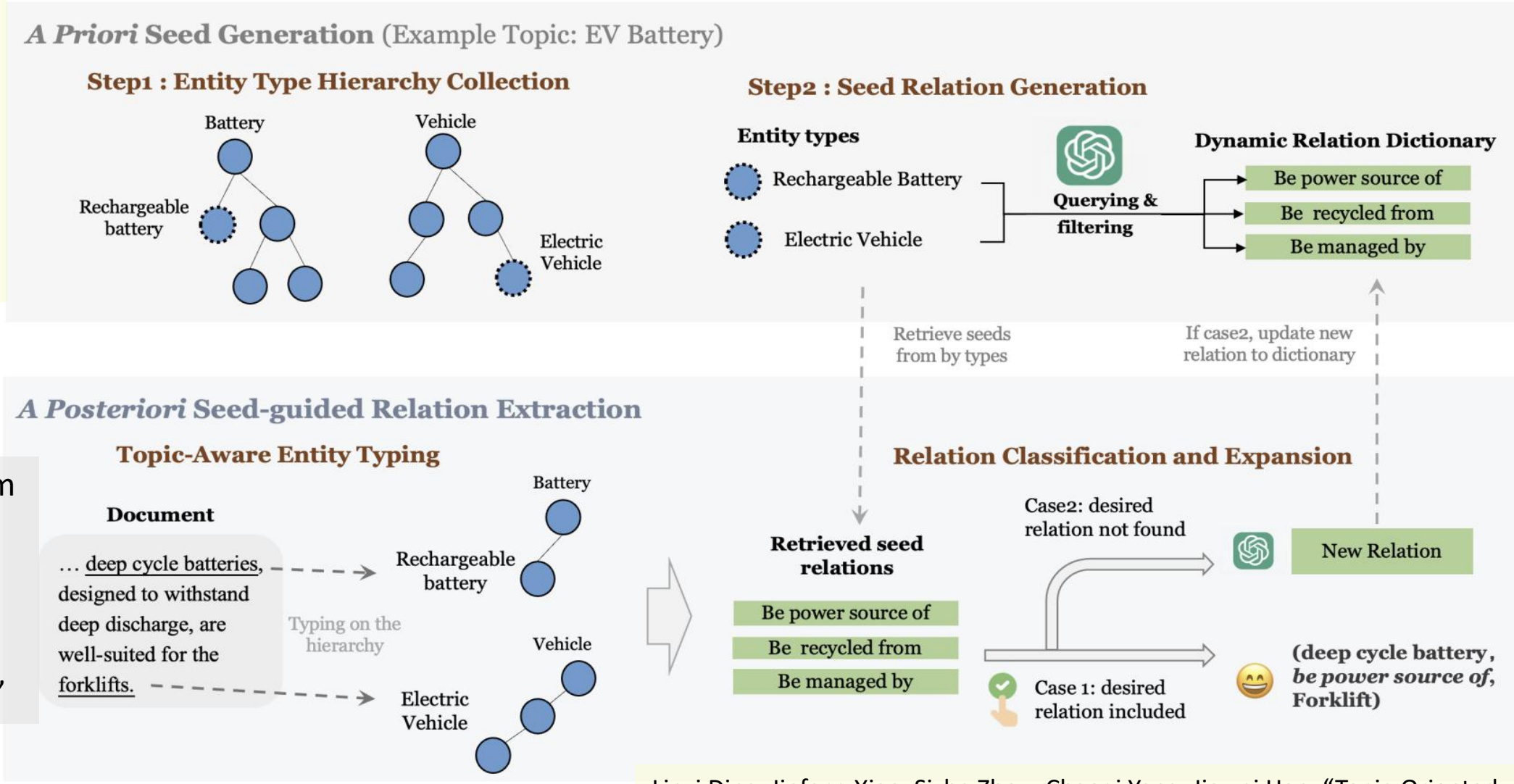
Performance study: ReactionMiner outperforms GPT4 on extraction quality

Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, et al., "Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data" EMNLP'2023

Open Relation Extraction: A Priori Seed Generation

PriORE: Open Relation Extraction with a Priori seed generation.

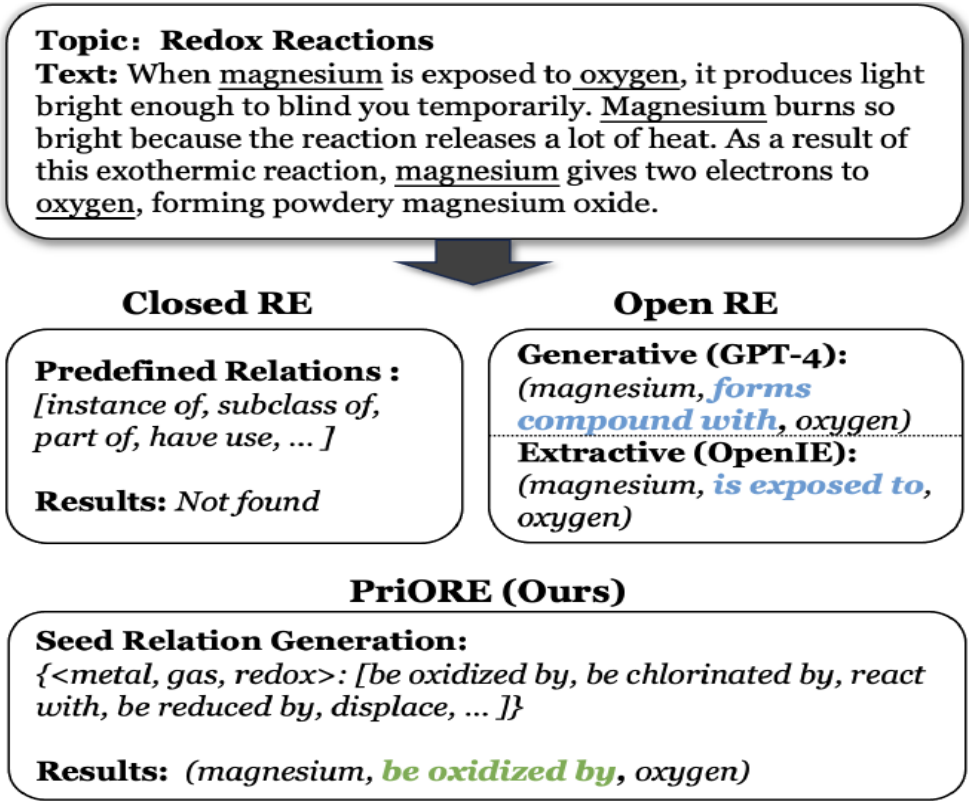
Use LLMs to generate seed relation candidates and select the best based on the contexts



Relation extraction from the text: "… ***deep cycle batteries designed to withstand deep discharge, are well-suited for the forklifts***."

Linyi Ding, Jinfeng Xiao, Sizhe Zhou, Chaoqi Yang, Jiawei Han, "Topic-Oriented Open Relation Extraction with A Priori Seed Generation", in EMNLP'24

Relation Extraction: From Close/Open RE to PriORE



- Close-Domain RE
  - Rely on the predefined relations
- Open-domain RE
  - Extractive methods（OpenIE）
    - misled by topic-irrelevant relations
    - implicit expressions
  - Generative methods （GPT-4）
    - overly general relations
    - overly specific relations
- Question: **How to align RE to user-specified topics?**

PriORE: Reduce randomness; Enhance the robustness of long-tail scenarios.
Two Steps: (1) Context-agnostic, seed Relations Generation (*A Priori*); (2) Context-based relation classification

Multiple Measures to be considered for Rel Extraction

Dimension	Extracted Relation	Explanation
Factualness	gives to	Wrong given text and topic
Relevance	is exposed to	Correct given text but not directly relevant to topic
Informativeness	form compound with	Correct but conceptually too general given text and topic
Coverage	form powdery magnesium oxide	Covers too few instances
Uniformity	{be oxidized by, give electrons to, . . . }	Multiple correct expressions extracted for the same relation

PrioRE: Experiment Results

Evaluation on coarse-grained topics

Method	DocRED (general)				FewREL DA (bio-medical)				
	Info	Fac	Cov	Acc	Info	TRel	Fac	Cov	Acc
Llama-3	0.91	**0.98**	0.92	0.82	0.57	0.97	0.88	0.98	0.49
Topic Llama-3	0.91	**0.98**	0.92	0.82	0.59	0.95	0.87	0.92	0.48
PriORE+Llama-3	0.90	0.92	0.97	0.81	0.79	0.98	0.85	0.97	0.70
GPT-4	0.89	0.96	0.94	0.78	0.61	0.94	0.90	0.91	0.49
Topic GPT-4	0.89	0.96	0.94	0.78	0.63	0.97	**0.90**	0.87	0.51
PriORE+GPT4 w/o type	**0.94**	0.93	0.86	0.80	0.66	0.97	0.79	0.83	0.52
PriORE+GPT4 w/o expan	0.86	0.88	**0.99**	0.74	0.74	**0.99**	0.75	**0.98**	0.64
PriORE+GPT4	0.90	0.94	0.98	**0.84**	**0.82**	0.98	0.88	0.94	**0.73**

- Datasets
  - General： DocRED
  - Domain： FewRel Domain Adaption （Biomedical）
  - Theme： Electric vehicle (EV) batteries & Redox reactions

Evaluation metrics: Info: informativemess; TRel: topic relevance; Fac: factualness; Cov: coverage; Acc: overall accuracy

Evaluation on fine-grained topics

Method	EV Battery					Redox Reaction				
	Info	TRel	Fac	Cov	Acc	Info	TRel	Fac	Cov	Acc
Llama-3	0.49	0.89	0.82	0.84	0.37	0.47	0.84	0.85	0.86	0.38
Topic Llama-3	0.54	0.89	0.79	0.79	0.40	0.53	0.88	0.81	0.83	0.43
PriORE+Llama-3	0.82	0.92	0.81	0.90	0.69	0.77	0.87	0.83	0.95	0.67
GPT-4	0.52	0.89	0.84	0.81	0.40	0.48	0.85	0.87	0.80	0.37
Topic GPT-4	0.55	0.90	0.80	0.77	0.39	0.51	**0.93**	**0.89**	0.75	0.39
PriORE+GPT4 w/o type	0.63	0.88	0.75	0.80	0.45	0.59	0.86	0.71	0.77	0.36
PriORE+GPT4 w/o expan	0.68	**0.93**	0.74	**0.94**	0.61	0.65	0.90	0.68	**0.96**	0.57
PriORE+GPT4	**0.83**	0.91	**0.85**	0.90	**0.70**	**0.80**	0.92	0.88	0.92	**0.69**

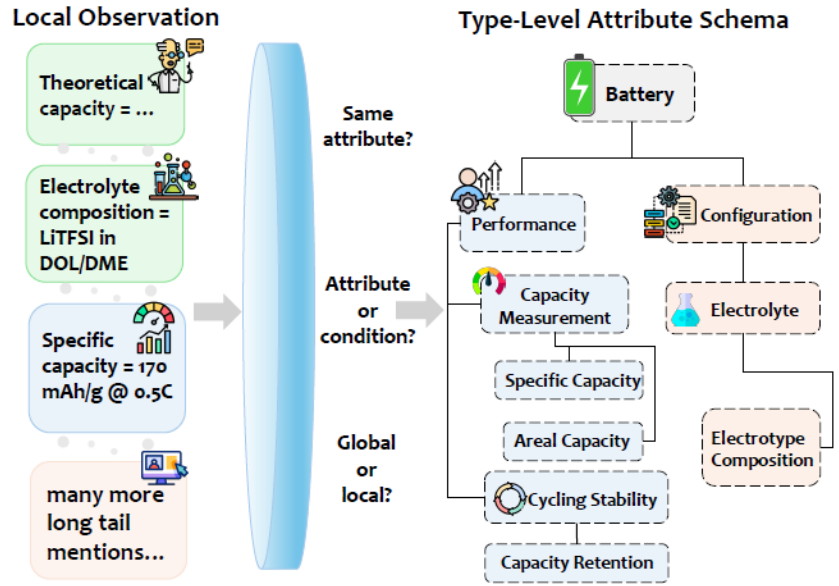
Method	Doc.	Few.	EVB.	Redox.
Llama3	0.41	0.25	0.27	0.20
Topic Llama3	0.41	0.22	0.25	0.24
PriORE+Llama3	**0.56**	**0.45**	0.46	0.41
GPT 4	0.43	0.26	0.27	0.25
Topic GPT	0.43	0.24	0.26	0.25
PriORE+GPT4	0.54	0.43	**0.47**	**0.42**

Evaluation of uniformity: PriORE leads to higher uniformity for all kinds of datasets

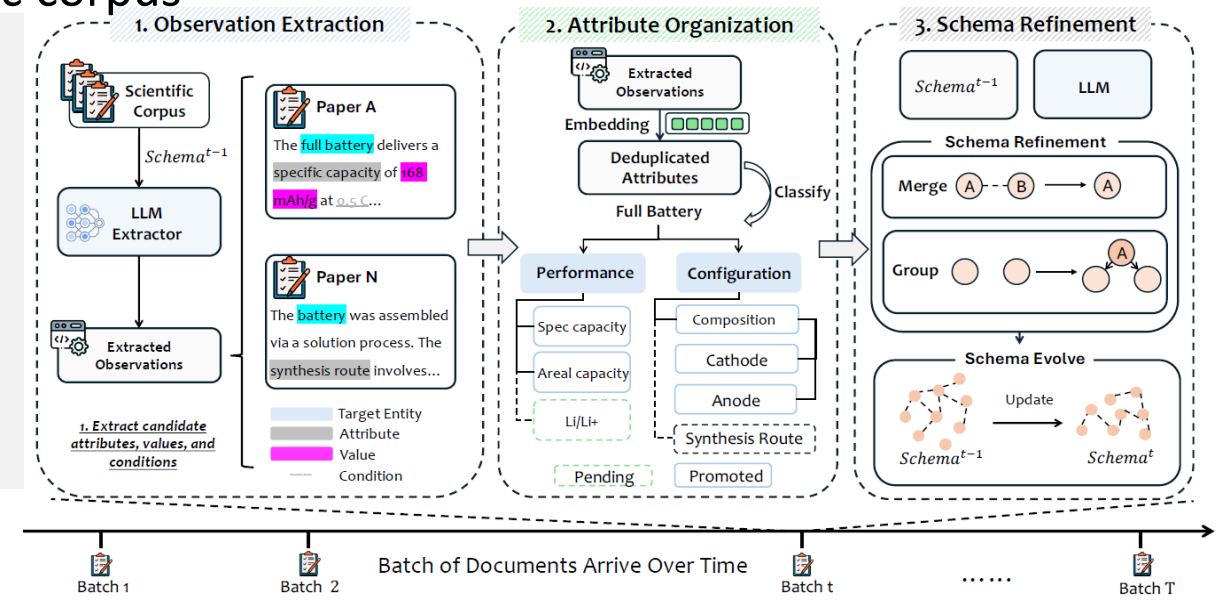
- PriORE is more effective on narrow topics compared to others
- PriORE has a large improvement on the informativeness

ObSchema: From Local Observations to Global Entity Attribute Schemas in Scientific Literature

- Local scientific observations mix attributes, values, units, and conditions, while the **desired schema** should capture recurring type-level attributes and organize them into coherent scientific dimensions
- **Closed-schema** extraction requires predefined attribute vocabularies that are hard to enumerate for property-rich scientific entities, while **open-schema** extraction often yields duplicate and condition-specific attribute labels
- ***Entity-centric attribute schema induction*** (**ObSchema**) constructs a hierarchical type level attribute schema from a scientific corpus



ObSchema incrementally induces a hierarchical attribute schema through three stages. **Observation Extraction** extracts n-ary observations binding the entity, candidate attribute, value, category, and measurement conditions, guided by the current schema S_{t-1}. **Attribute Organization** consolidates near-identical raw attributes, classifies them into the schema, and promotes them into the active schema only after accumulating sufficient corpus-level evidence. **Schema Refinement** applies conservative duplicate merging and adaptive grouping to update S_{t-1} into S_t.



Patrick Xu et al., ObSchema: From Local Observations to Global Entity Attribute Schemas in Scientific Literature , submitted

ObSchema: Performance Study

Data Statistics: Battery science & GeoScience

Domain	#Articles	#Paragraphs	Para/Art	Avg words/Para	Year Range	#Venues
Battery Science	185	927	5.0	240	2015–2025	13
Geoscience	118	378	3.2	267	2007–2025	24
Total	**303**	**1,305**	**4.3**	**247**	**2007–2025**	**37**

End-to-end **runtime** and **token consumption** for entity extraction and schema induction

Domain	Method	Runtime (min)	Tokens (M)
Battery	Chain-of-Layer	14.73	8.30
	TaxoAdapt	21.55	31.92
	Iterative LLM	60.42	7.20
	ObSchema	14.00	13.22
Geoscience	Chain-of-Layer	9.44	6.43
	TaxoAdapt	13.78	11.34
	Iterative LLM	23.71	3.92
	ObSchema	8.67	5.33

By converting local entity-property mentions into structured observations, aggregating corpus-level evidence, and refining promoted attributes into coherent hierarchies, ObSchema bridges the gap between open-schema extraction & global scientific schema construction.

Domain	Method	Attr. Validity	Path Gran.	Sibling Coh.	Uniqueness	Avg.
Base model: Qwen3-30B-A3B						
Battery	Chain-of-Layer	0.860	0.711	0.667	0.383	0.648
	TaxoAdapt	0.951	0.717	0.626	0.559	0.713
	Iterative LLM Induction	0.935	0.665	0.756	0.733	0.772
	ObSchema (ours)	**0.964**	**0.807**	**0.816**	**0.830**	**0.854**
Geoscience	Chain-of-Layer	0.950	0.703	0.708	0.594	0.739
	TaxoAdapt	0.964	**0.800**	0.630	0.690	0.796
	Iterative LLM Induction	0.891	0.589	0.556	0.724	0.740
	ObSchema (ours)	**0.991**	0.798	**0.818**	**0.807**	**0.853**
Base model: Qwen3-14B						
Battery	Chain-of-Layer	0.817	0.683	0.538	0.447	0.629
	TaxoAdapt	0.976	0.808	0.548	0.651	0.746
	Iterative LLM Induction	0.912	0.537	0.444	0.658	0.638
	ObSchema (ours)	**1.0**	**0.935**	**0.667**	**0.704**	**0.827**
Geoscience	Chain-of-Layer	0.920	0.802	0.606	0.564	0.723
	TaxoAdapt	0.919	0.787	0.539	0.679	0.731
	Iterative LLM Induction	0.754	0.521	0.333	0.771	0.595
	ObSchema (ours)	**1.0**	**0.864**	**0.800**	**0.911**	**0.893**

Schema induction quality under two base LLMs. Methods are evaluated against four metrics: **Attribute Validity**, **Path Granularity**, **Sibling Coherence**, and **Uniqueness**, with Avg. reporting their mean across five independent judges.

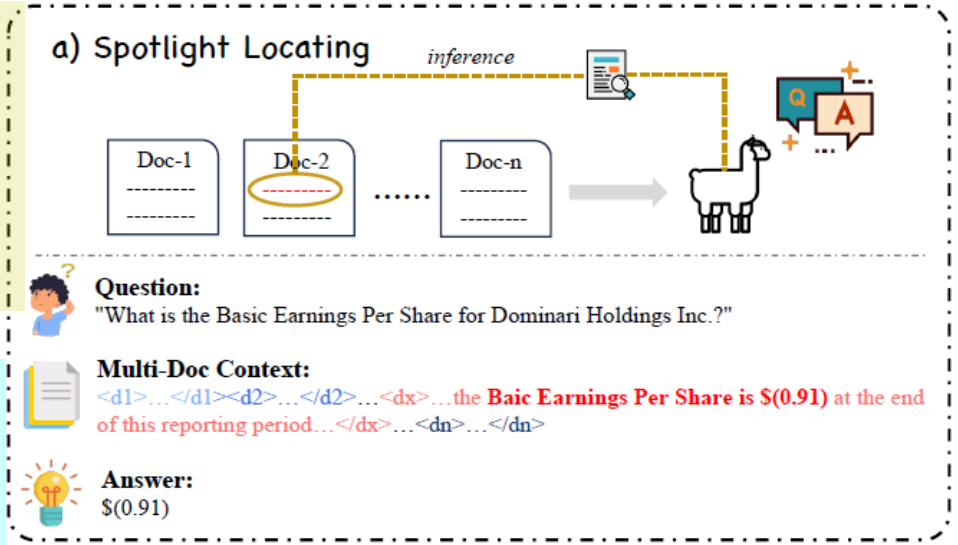
Outline

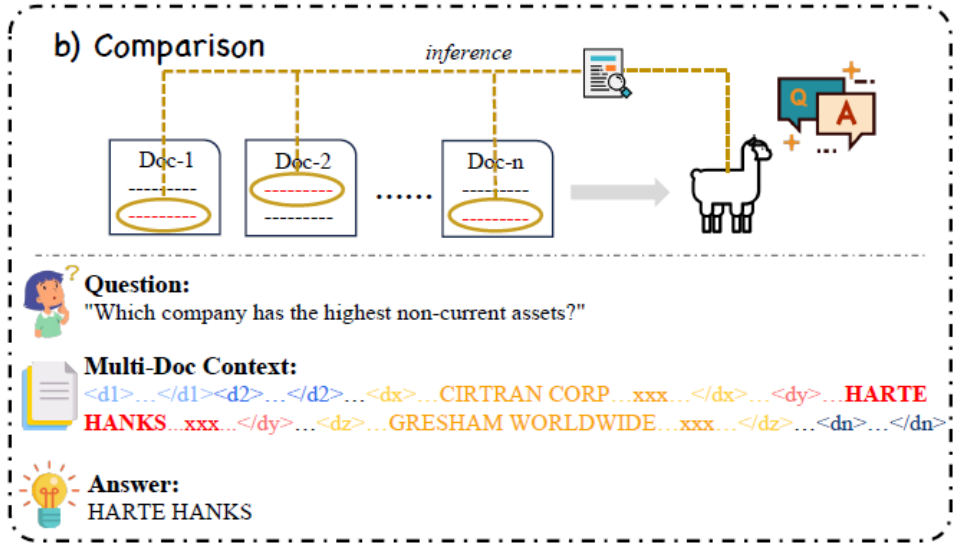
- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

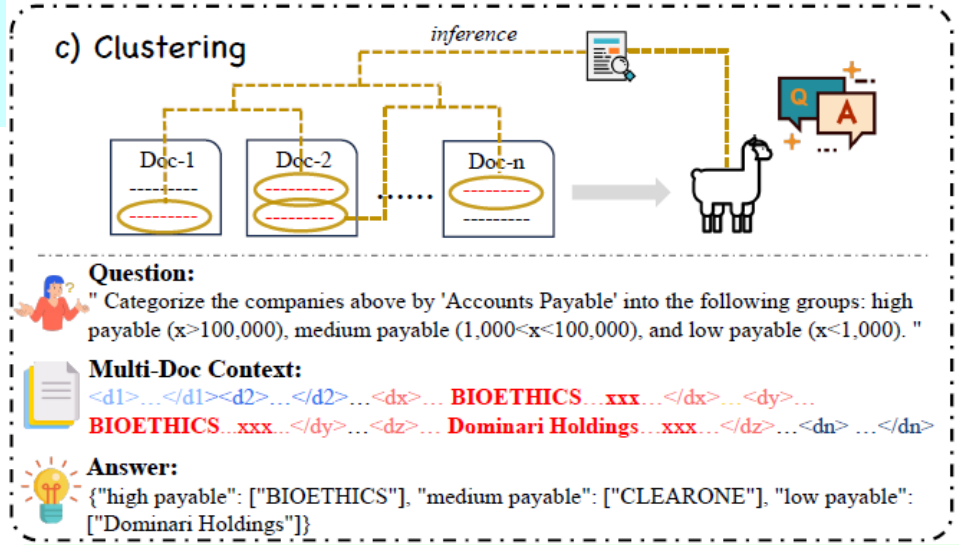
Long Document QA Challenge

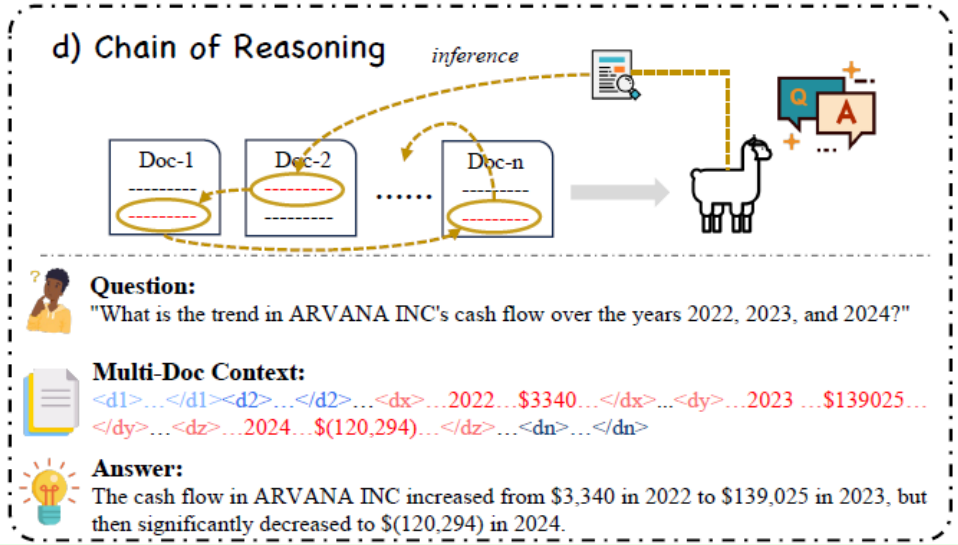
*** Evidence in Loong are scattered in different parts across multi-document long contexts, necessitating that no doc can be ignored for success.**

* Performance study shows that RAG achieves poor performance, demonstrating that Loong can reliably assess the model's long-context modeling capabilities.









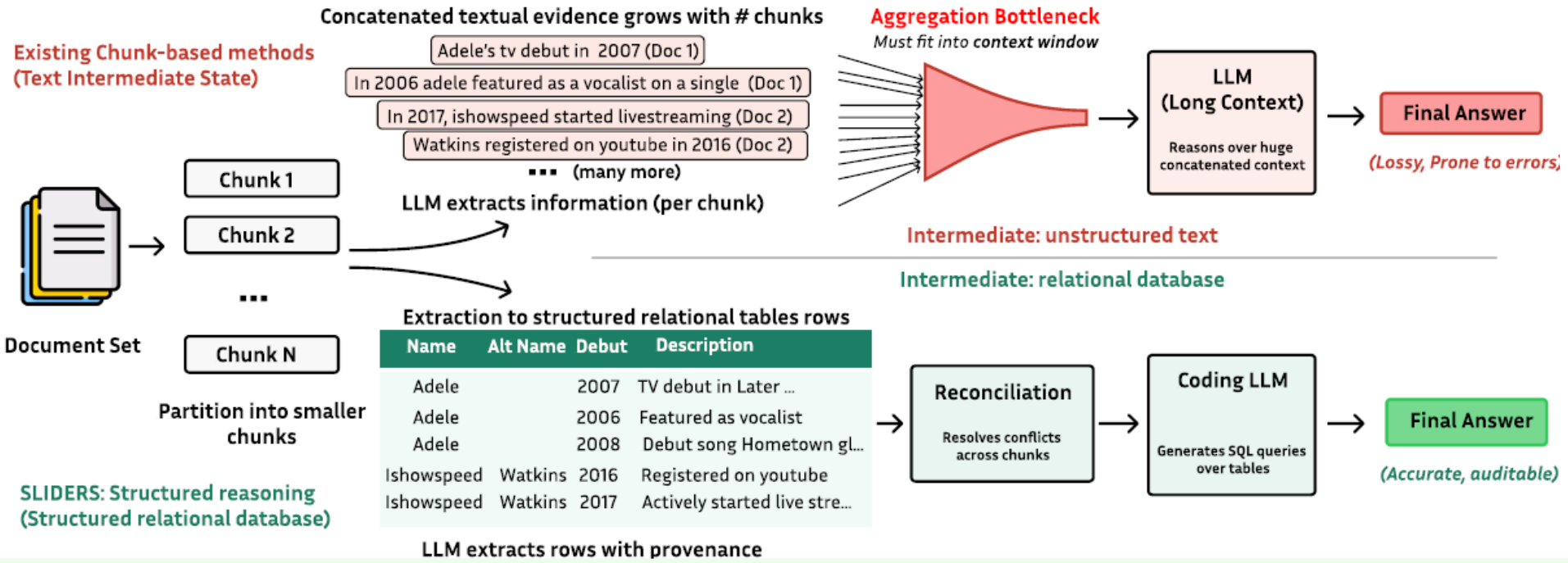
Showcase of 4 evaluation tasks in Loong (<di>...</di> marks the i-th doc). a) Spotlight Locating: Locate the evidence; b) Comparison: Locate & compare the evidence; c) Clustering: Locate & cluster the evidence into groups. d) Chain of Reasoning: Locate & reasoning along a logical chain

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo2, Yunshui Li, Min Yang, Fei Huang, Yongbin Li, "Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA", ArXiv: 2405-17419v2

Sliders: Scalable Long-document Integration through Decomposed Extraction and Reconciliation System

Real-world doc QA is challenging: Analysts must synthesize evidence across multiple documents and different parts of each doc.

Chucking introduces an aggregation bottleneck: as # of chunks grows, systems must combine & reason over an increasingly large body of extracted evidence.

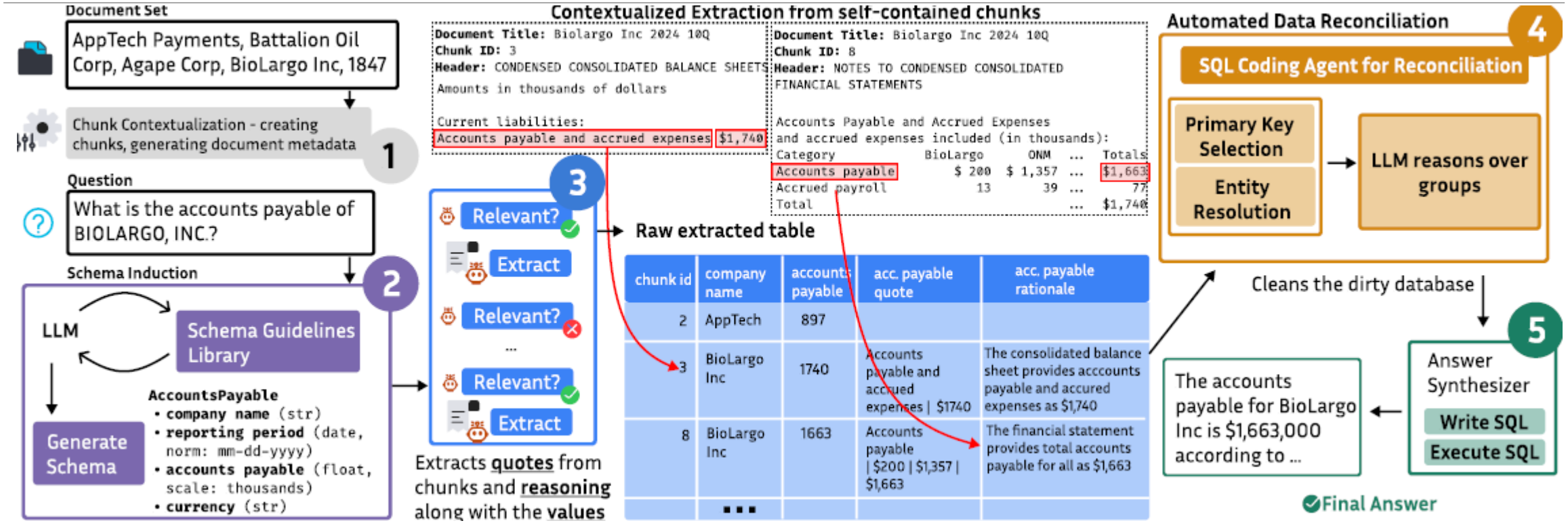


Chunking-based methods regenerate the long-context problem. SLIDERS mitigates it using structured reasoning

- SLIDERS extracts salient info into a relational DB, enabling scalable reasoning over persistent structured state via SQL rather than concatenated text
- To make the locally extracted representation globally coherent, SLIDERS introduces a data reconciliation stage that leverages provenance, extraction rationales, and metadata to detect and repair duplicated, inconsistent, and incomplete records.
- SLIDERS achieves high performance

Harshit Joshi, Priyank Shethia, Jadelynn Dao, and Monica S. Lam, "**Contexts are Never Long Enough: Structured Reasoning for Scalable Question Answering over Long Document Sets**", ArXiv:2604-22294

SLIDERS: Structured Reasoning for Long Document QA (I)



Processes: (1) Contextualization of chunks, (2) Inducing schema, (3) Structured Extraction, (4) Data Reconciliation, (5) QA over the final DB

- **Step 1. Contextualized Chunking**: Augment each doc d with metadata prior to chunking, where a generated title and brief document-level description to provide high-level context **shared across all chunks**; and the local part captures structural signals: section headers, tables, and figure captions, enabling chunk boundaries be formed that aligns with the document's natural layout
  - With this enriched represent., partition each doc into chunks, preserving semantic/structural coherence

SLIDERS: Structured Reasoning for Long Document QA (II)

Processes: (1) Contextualization of chunks, (2) Inducing schema, (3) Structured Extraction, (4) Data Reconciliation, (5) QA over the final DB

- **Step 2. Schema Induction**: Taking a query q and extracted doc metadata M and produces a relation schema with multiple tables $S = \{S_1, S_2, \ldots, S_k\}$; a table schema $S = \langle sn, f_1, \ldots f_n \rangle$ with schema name sn, and a set of fields f_i. A field is defined as a tuple $f = \langle fn, d, \tau, u, \sigma, \rho \rangle$ where fn is the field name; d is a semantic text description; τ is the data type (e.g., int, str); u is the unit of measure (e.g., USD, kg); σ is the scale (e.g., millions, thousands); and ρ represents normalization rules (e.g., currency conversion, date formatting)
- **Step 3. Contextualized Extraction with Relevance Gating:**
  - Relevance gating: Filter out the text if it does not contain evidence relevant to the schema entities
  - Structured extraction: taking query q, doc D, metadata M, schema S, produce a set of rows, capturing relationships mentioned in the chunk, in tables T conforming to schema S
- **Step 4. Data Reconciliation**: Each agent applies a distinct reconciliation operation over related records, using provenance and rationale to guide decisions and produce a coherent representation
  - Reconciliation operations consists of *Deduplication, Conflict Resolution, Consolidation*
- **Step 5. QA over the final DB:** The agent generates SQL programs that integrate the rows in the partition. It may issue intermediate queries (e.g., to compute distinct counts, isolate specific columns, or retrieve related rows, before producing a final update)

SLIDERS: Performance Study (I)

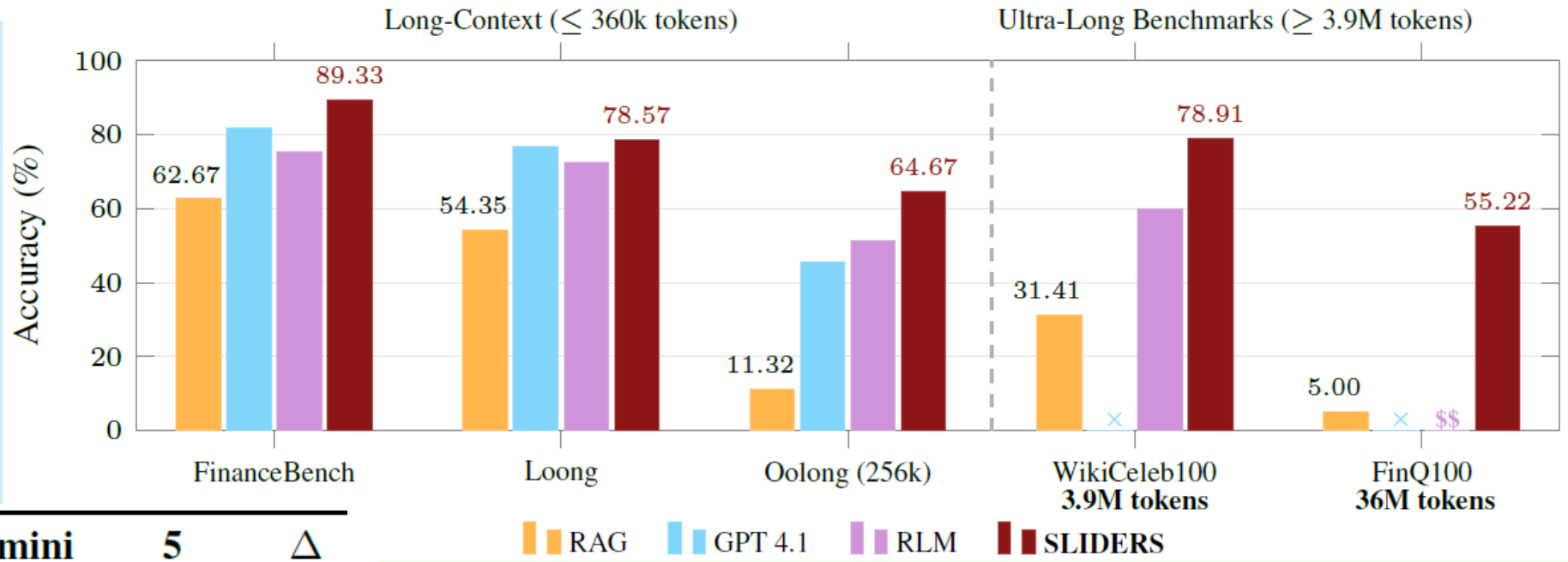
Evaluation metrics: (i) **LLM-as-a-judge**: Given the question, the gold answer with justification, and the model's predicted answer with its justification, the judge model determines whether the prediction is correct; (ii) for numerical aggregation questions, adopt Oolong's **metric**, assigning higher scores to predictions with smaller deviation.

		"Long-Context" Benchmark (<360k)				3.9M T	36M T
Models	**LLMs**	**FB**	**Loong**	**Oolong**	**Avg.**	**WC**	**FQ**
RAG	Qwen3-4B & GPT 4.1	62.67	54.35	11.32	42.77	31.41	5.00
LongRAG	Qwen3-4B & GPT 4.1	72.00	59.10	22.00	51.03	43.20	28.87
GraphRAG	Qwen3-4B & GPT 4.1	75.33	61.28	22.00	52.87	48.59	$$
Basemodel	GPT 4.1	82.00	76.74	45.56	68.69	N.A.	N.A.
Basemodel	Qwen3.5 122B-A10B	84.67	74.78	24.89	61.44	N.A.	N.A.
DocETL	GPT 4.1	63.33	75.03	49.00	62.44	54.26	$$
Chain of Agent	GPT 5 & GPT 5-mini	71.30	54.46	17.11	47.62	$$	$$
RLM	GPT 5 & GPT 5-mini	75.33	72.64	51.42	66.46	59.80	$$
SLIDERS	GPT 4.1 & GPT 4.1-mini	**89.33**	**78.57**	64.67	**75.56**	**78.91**	55.22
SLIDERS	Qwen3.5 122B-A10B	82.10	75.70	**68.00**	75.26	76.92	**60.18**

- **FinanceBench** [Islam et al.,'23] A single-document financial question-answering benchmark targeting realistic analyst style queries. It comprises of 150 questions about publicly traded companies, with evidence drawn from public filings.
- **Loong** [Wang et al.,'24] A multi-document question-answering benchmark where every doc provided is required to answer the query, and omitting any yields an incorrect answer. Covers three domains, two languages: finance, law, and academic research papers (English), with an average of ~11 docs per instance.
- **Oolong** [Bertsch et al.,'25] A long-context reasoning benchmark focused explicitly on aggregation. Each input consists of a datapoint from a large dataset and models must first classify and then aggregate those local predictions to produce a global answer.

SLIDERS: Performance Study (II)

Accuracy across long-context and ultra-long benchmarks. SLIDERS consistently outperforms all baselines and is the only system to scale beyond 3.9M tokens. × denotes baselines that exceeded the model's context window; $$ denotes prohibitively expensive runs



Dataset	4.1	4.1-mini	5	Δ
Loong Papers	**91.30**	89.96	88.00	3.30
Loong Legal	64.12	**68.34**	61.26	7.08
Loong Finance EN	**74.50**	68.10	73.10	6.40
Loong Finance ZH	**93.96**	90.46	93.20	3.50
Loong Avg	**80.97**	79.22	78.89	2.08
FinanceBench	76.71	**80.00**	**80.00**	3.33

Accuracy across schema-induction models. Bold: best per dataset; Δ: range (max − min).

- Cost of SLIDERS across benchmarks: average $0.76/query
  - ~40% by Entity Resolution (scan the entire table)
- Latent: end-to-end setting, a structured representation is built from scratch for a single query, average latency is 2.6 min on Loong and 3.0 min on FinanceBench
- SLIDERS's provenance tracking enhances auditability and interpretability, facilitating error analysis

Outline

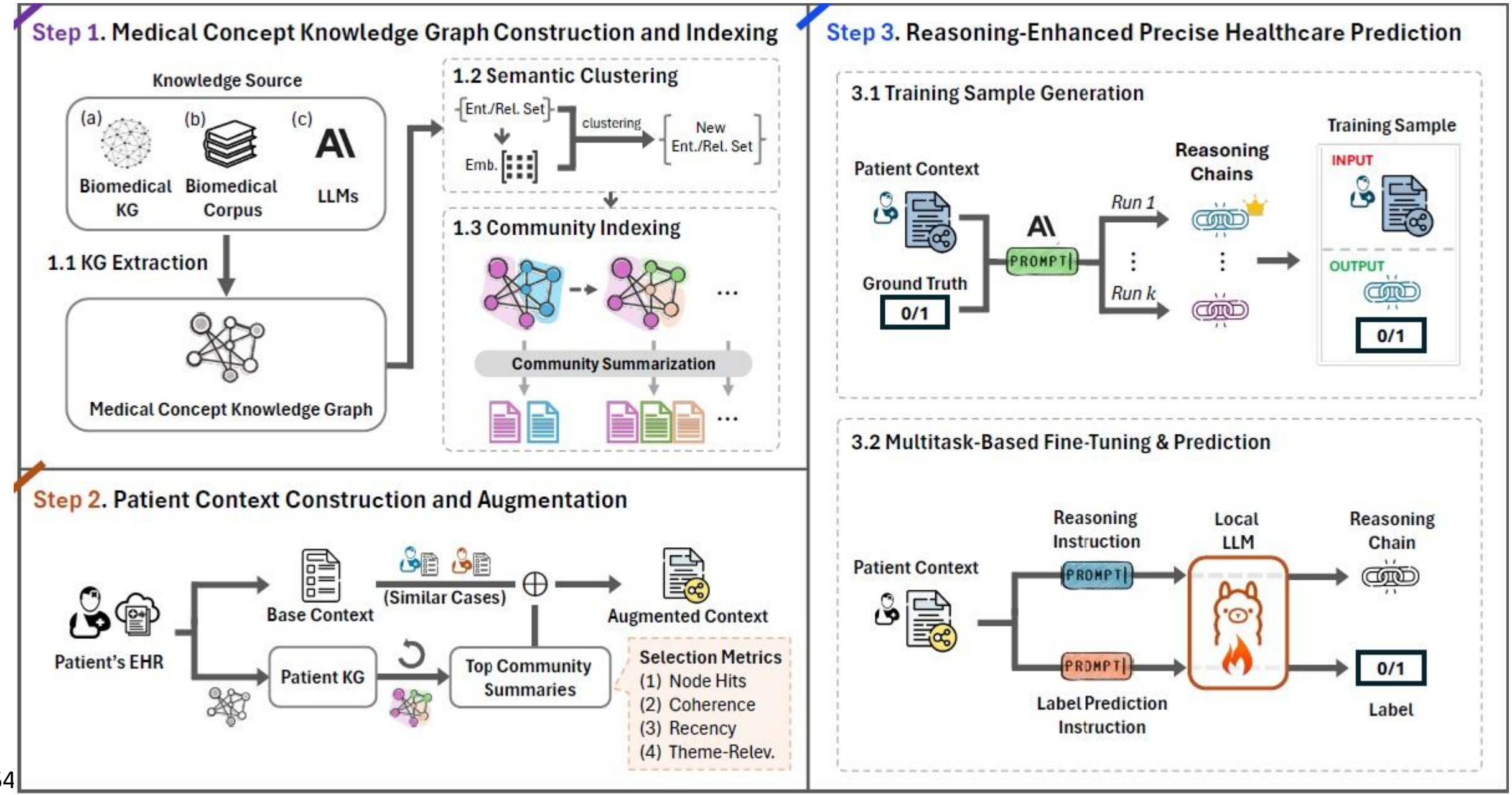
- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

Retrieval and Structuring for LLM-Empowered Reasoning

- KARE: Knowledge Aware Reasoning-Enhanced Framework
  - Improve healthcare predictions with **retrieval and LLM reasoning**
  - Integrate knowledge graph (KG) community-level retrieval with LLM reasoning to enhance healthcare predictions
- Three steps
  - Medical concept knowledge graph construction and indexing
    - A dense medical knowledge structuring approach enables accurate retrieval of relevant information
  - Patient context construction and augmentation
    - A dynamic knowledge retrieval mechanism enriches patient contexts with focused, multi-faceted medical insights
  - Reasoning-enhanced precise healthcare prediction
    - A reasoning-enhanced prediction framework leverages these enriched contexts to produce

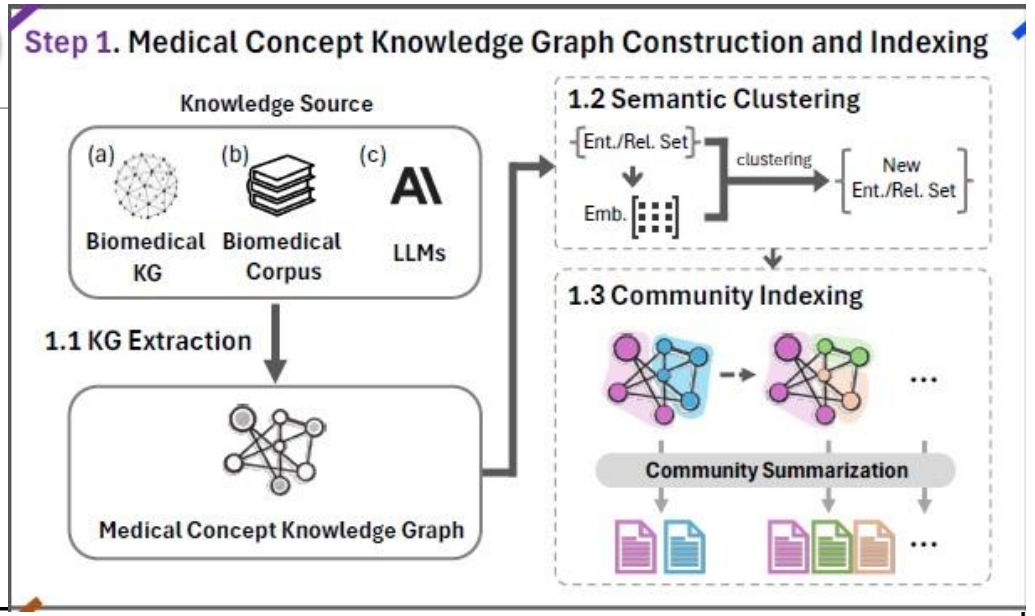
Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, Jiawei Han, "Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval", Int. Conf. on Learning Representation (ICLR'2025)

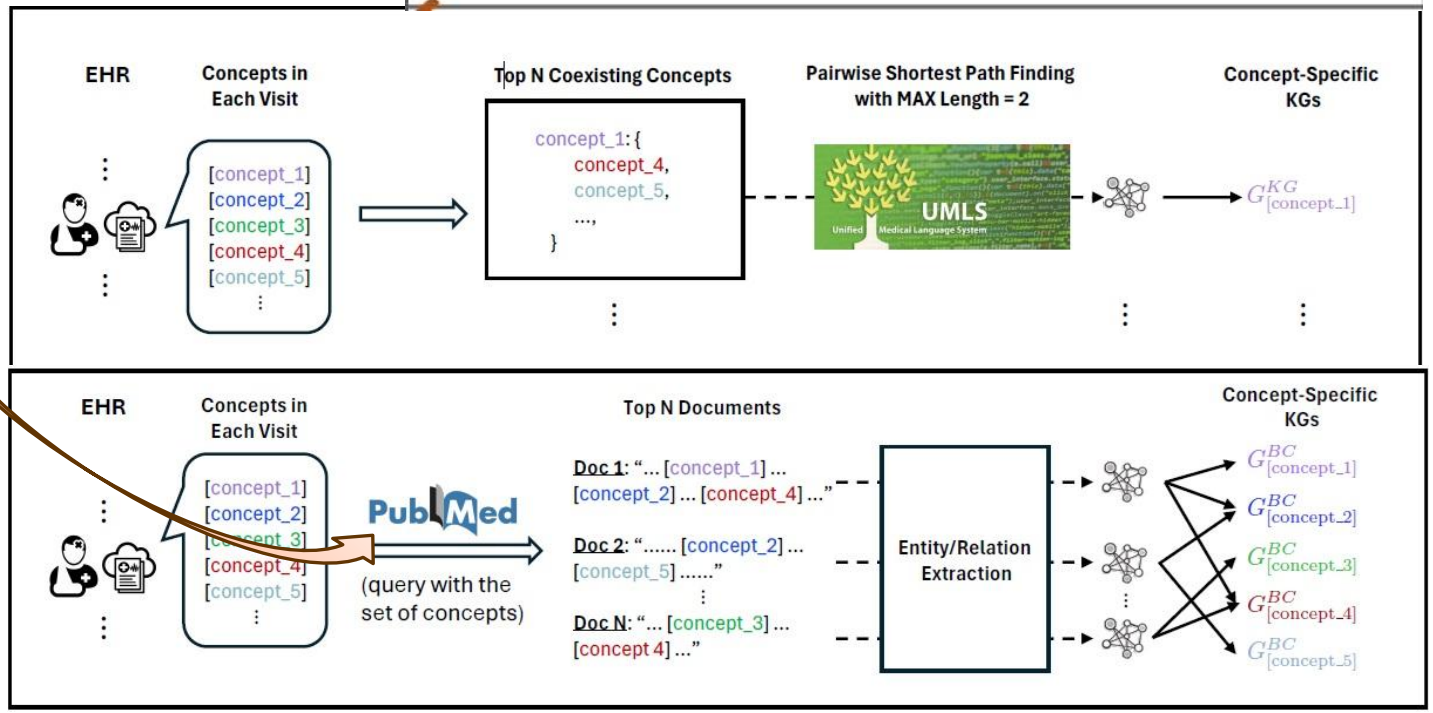
KARE: The General Framework



Step 1: Medical Concept Knowledge Graph Construction and Indexing (1)

- Constructs a comprehensive *medical concept knowledge graph* by *integrating information from multiple sources*, *organizing it into a hierarchical community structure*
  - Allows for the generation of community summaries that facilitate precise knowledge retrieval
- For each medical concept c_i in HER system, extract a c_i -specific KG $Gc_i = (Vc_i, Ec_i)$ from 3 sources:
  - Biomedical KG (e.g., UMLS)
  - Biomedical Corpus (e.g., PubMed)
  - LLMs: Prompt the LLM to identify the relationships among the concepts that are helpful to the clinical predictions
    - Allow LLM to add intermediate relationships between two concepts



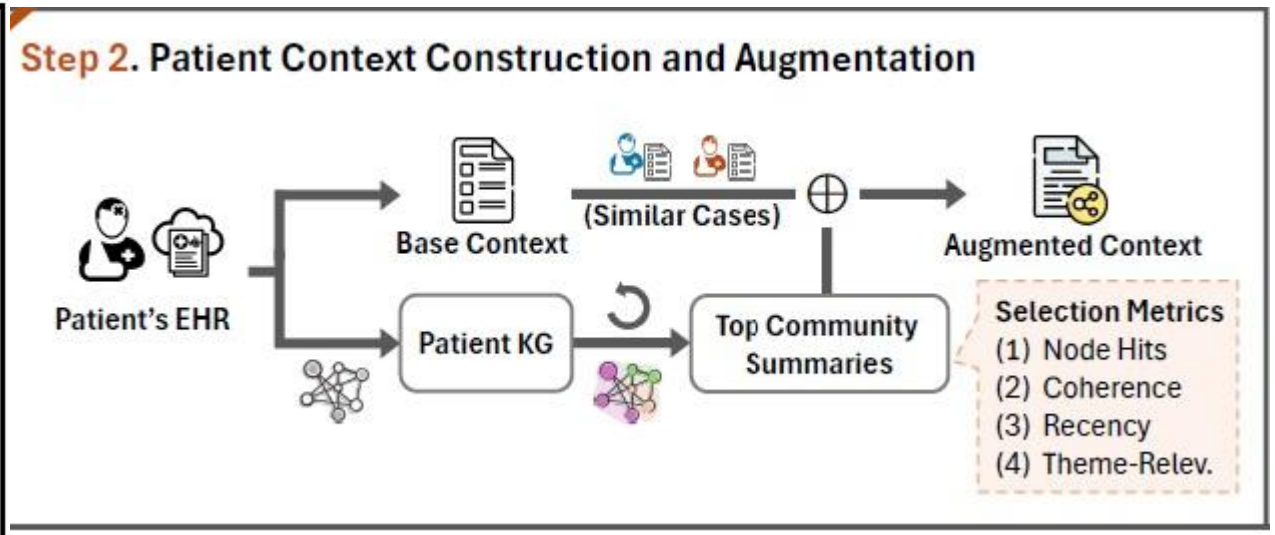


Step 2: Patient Context Construction and Augmentation

```
=====================================
# Task #

Readmission Prediction Task:
Objective: Predict if the patient will be
readmitted to the hospital within 15 days of
discharge.
Labels: 1 = readmission within 15 days, 0 = no
readmission within 15 days

Note: Analyze the information comprehensively to
determine the likelihood of readmission. The goal
is to accurately distinguish between patients who
are likely to be readmitted and those who are not


=====================================
# Patient EHR Context #

Patient ID: xxxxxxxx

Visit 0:
Conditions:
1. pleurisy; pneumothorax; pulmonary collapse
2. pneumonia
3. congestive heart failure; nonhypertensive
4. nutritional deficiencies
5. other circulatory disease
6. cardiac dysrhythmias
7. screening and history of mental health and
substance abuse codes
8. heart valve disorders
                    ⋮
Procedures:
1. incision of pleura; thoracentesis; chest
drainage

Medications:
1. high-ceiling diuretics
2. adrenergics, inhalants
3. potassium supplements
4. other drugs for obstructive airway diseases,
inhalants in atc
5. beta blocking agents
6. viral vaccines
7. other analgesics and antipyretics in atc
8. drugs for constipation
9. antithrombotic agents
10. opioid analgesics
11. antiemetics and antinauseants
12. other mineral supplements in atc
13. i.v. solution additives
14. other diagnostic agents in atc
15. antiarrhythmics, class i and iii
                    ⋮
                                    (continue)
```

```
=====================================
# Similar Patients #

Patient ID: yyyyyyyy

Visit 0:
Conditions:
1. pleurisy; pneumothorax; pulmonary collapse
2. cardiac dysrhythmias
3. congestive heart failure; nonhypertensive
4. other liver diseases
                    ⋮
Procedures:
1. incision of pleura; thoracentesis; chest
drainage
2. other or heart procedures
3. other diagnostic cardiovascular procedures

Medications:
1. antithrombotic agents
2. other analgesics and antipyretics in atc
3. high-ceiling diuretics
4. antacids
5. drugs for constipation
                    ⋮
Label:
1

-------------------------------------
Patient ID: zzzzzzzz

Visit 0:
Conditions:
1. pleurisy; pneumothorax; pulmonary collapse
2. congestive heart failure; nonhypertensive
3. complications of surgical procedures or
medical care
4. screening and history of mental health and
substance abuse codes
5. cardiac dysrhythmias
                    ⋮
Procedures:
1. other diagnostic procedures of respiratory
tract and mediastinum
2. other non-or therapeutic procedures on
respiratory system
3. incision of pleura; thoracentesis; chest
drainage

Medications:
1. other agents acting on the renin-
angiotensin system in atc
2. antiadrenergic agents, peripherally acting
                    ⋮
Label:
0
```

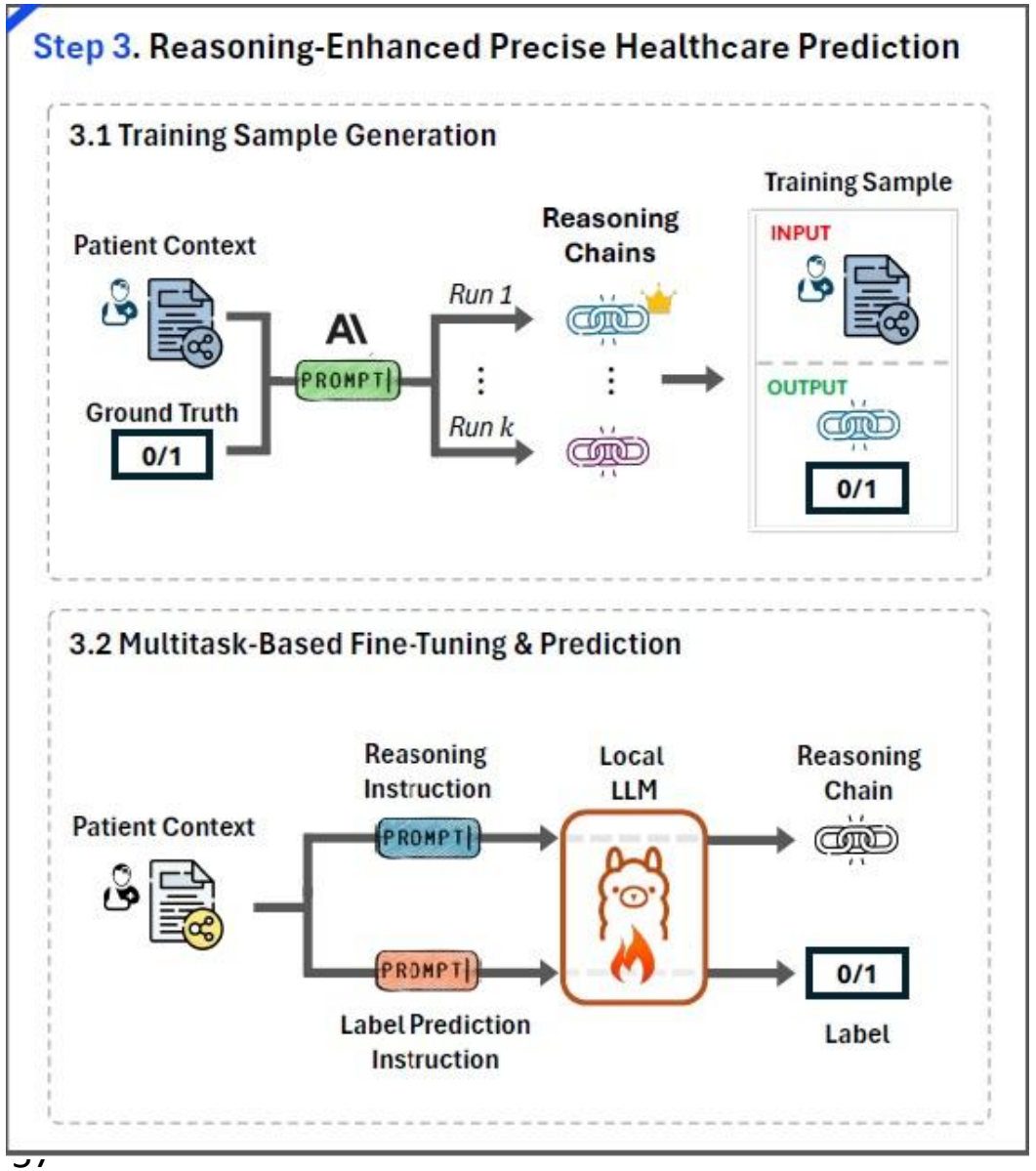
Patient Base Context



- Base Context Construction: For a patient p, construct a base context: (1) task description, (2) the patient's conditions, procedures, and medications, and (3) similar patients: one has the same label as patient p and the other has a different label
- Context Augmentation: Enrich p's base context with relevant info from the KG and select the most relevant summaries for context augmentation

Ensure the augmented context includes the most relevant and diverse info from the KG, tailored to the patient's specific conditions and the prediction task

Step 3: Reasoning-Enhanced Precise Healthcare Prediction



- Training Sample Generation: Generate reasoning chains in a unified format for each patient p and task τ .
  - Entering (1) the task description, (2) the augmented patient context, and (3) the corresponding ground truth label
  - The LLM generates K reasoning chains along with confidence levels.
  - We select the reasoning chain with the highest confidence, ensuring that only the most reliable explanations are used
- Multitask-based Fine-tuning and Prediction
  - We fine-tune a relatively small local LLM (7B) to perform both reasoning chain generation and label prediction
  - The model is trained using (i) task description and (ii) the augmented patient context, with a prepended instruction
  - Prediction: Given a new patient and task, we provide the appropriate instruction to the fine-tuned model to generate the reasoning chain or predict the label

Experiment Setting: Task, Data and Metrics

- Tasks: EHR-based prediction
  - Mortality Prediction: Estimates mortality outcome for next visit (Patient's survival status during visit x_t)
  - Readmission Prediction: Predicts if patient will be readmitted within σ days (σ is set to 15 in this study)

Table 1: Statistics of pre-processed EHR datasets. "#": "the number of", "/ patient": "per patient".

	MIMIC-III-Mort.			MIMIC-III-Read.			MIMIC-IV-Mort.			MIMIC-IV-Read.		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
# Patients (Samples)	7730	991	996	7730	991	996	8018	996	986	8029	958	1013
# Visits / Patient	1.56	1.60	1.61	1.56	1.60	1.61	1.26	1.30	1.21	1.26	1.28	1.25
# Conditions / Patient	23.27	23.92	25.89	23.27	23.92	25.89	14.34	15.30	13.59	13.62	14.21	13.21
# Procedures / Patient	6.22	6.56	7.17	6.22	6.56	7.17	2.96	3.08	2.84	2.89	2.96	2.81
# Medications / Patient	54.79	55.77	63.73	54.79	55.77	63.73	30.66	32.86	28.40	28.74	30.61	27.59

- Datasets: Use the publicly available MIMIC-III (v1.4) and MIMIC-IV (v2.0) EHR datasets
  - Use PyHealth (Yang et al., 2023a) for preprocessing, …
- Evaluation Metrics: Four key metrics:
  - Accuracy: Overall correct predictions across both outcomes
  - Macro-F1: A balanced measure, crucial for the imbalanced datasets
  - Sensitivity: Model's ability to identify patients at risk of mortality or readmission
  - Specificity: Identify patients unlikely to experience these outcomes, helping avoid unnecessary measures

Performance Comparison on MIMIC-III Dataset

Results are averaged by multiple runs. asterisk (∗): important for handling imbalanced datasets.

		MIMIC-III							
		Mortality Prediction (pos = 5.42%)				Readmission Prediction (pos = 54.82%)			
Type	Models	Accuracy	Macro F1*	Sensitivity*	Specificity	Accuracy	Macro F1	Sensitivity	Specificity
ML	GRU (Chung et al., 2014)	92.7	50.7	3.7	97.8	62.2	61.5	68.9	54.0
	Transformer (Vaswani et al., 2017)	92.7	51.9	5.6	97.6	58.8	58.2	65.0	51.3
	RETAIN (Choi et al., 2016)	92.4	50.6	3.7	97.6	59.1	56.9	74.9	40.0
	GRAM (Choi et al., 2017)	92.4	50.2	5.2	95.2	61.8	60.4	74.9	46.4
	Deepr (Nguyen et al., 2016)	91.9	51.0	3.7	98.2	62.6	62.1	66.7	57.6
	TCN (Bai et al., 2018)	91.6	53.2	9.3	96.4	63.4	62.7	70.7	54.7
	ConCare (Ma et al., 2020b)	94.6	48.6	0.0	**100.0**	59.2	59.0	61.5	56.4
	AdaCare (Ma et al., 2020a)	90.6	54.1	9.1	97.6	61.6	60.5	70.8	50.3
	GRASP (Zhang et al., 2021)	93.7	49.9	1.9	98.9	61.3	59.5	74.9	44.8
	StageNet (Gao et al., 2020)	90.5	50.5	5.6	95.4	60.5	60.0	65.1	54.9
	KerPrint (Yang et al., 2023b)	92.4	52.2	9.8	94.7	63.5	62.1	68.0	56.1
LM+ML	GraphCare (Jiang et al., 2024a)	94.9	58.3	17.2	97.1	65.4	64.1	70.3	57.8
	RAM-EHR (Xu et al., 2024)	94.4	59.6	14.8	98.9	64.8	63.5	74.7	52.4
	EMERGE (Zhu et al., 2024a)	94.1	57.7	13.2	98.4	63.7	62.0	68.0	55.9
LLM	Zero-shot (LLM: Claude 3.5 Sonnet)								
	w/ EHR context only	89.5	50.4	6.4	94.4	54.3	35.4	98.9	0.2
	w/ Classic RAG[a]	89.9	51.2	10.2	92.8	53.2	34.6	91.2	1.4
	w/ KARE-augmented context[b]	92.3	54.6	14.2	94.6	56.3	43.8	93.9	10.6
	Few-Shot (LLM: Claude 3.5 Sonnet)								
	w/ exemplar only (N=2)[c]	88.7	49.5	5.6	93.4	52.7	42.2	87.0	11.1
	w/ exemplar only (N=4)	88.0	49.2	5.6	92.7	53.6	44.7	84.0	15.7
	w/ EHR-CoAgent[d] (Cui et al., 2024)	87.4	51.7	13.0	91.8	55.2	46.1	78.2	20.1
	w/ KARE-augmented context	91.5	53.5	13.7	94.0	57.1	49.3	75.5	27.2
	KARE (ours)	**95.3**	**64.6**	**24.7**	98.3	**73.9**	**73.7**	**76.7**	**70.7**

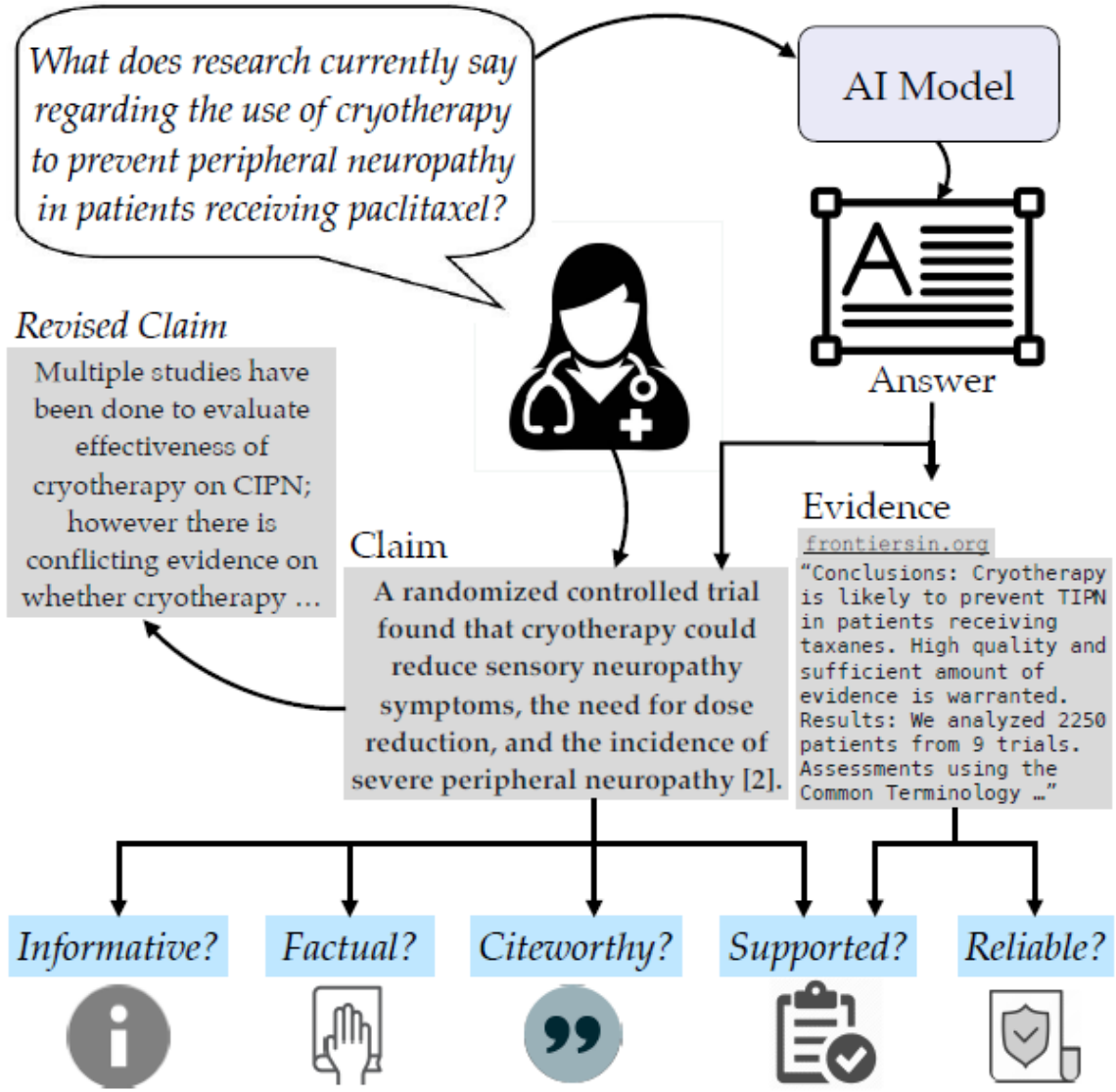
Outline

- Reasoning in LLMs: Prompting, Chain of Thought, Tools
- Structure-Augmented Reasoning Generation
- LLM-Enhanced Knowledge Structuring
- Structuring for Reasoning with Long Documents
- Theme-Based Knowledge Graphs for LLM Reasoning
- Looking forward: Theme-Based, LLM-Enhanced Reasoning Generation

Towards Answering Expert Questions: ExpertQA

- The use of LLMs in specialized domains has many potential benefits, but it also carries significant risks
- Collect expert-curated questions from 484 participants across 32 fields of study, and ask the same experts to evaluate generated responses to their own questions
- Ask experts to improve upon responses from language models. The output of our analysis is EXPERTQA, a high-quality long-form QA dataset with 2177 questions spanning 32 fields, along with verified answers and attributions for claims in the answers.

Each claim-evidence pair in an answer is judged by experts for various properties such as the claim's informativeness, factuality, cite worthiness, whether the claim is supported by the evidence, and reliability of the evidence source. Further, experts revise the original claims to ensure they are factual and supported by trustworthy sources.



Chaitanya Malaviya, Subin Lee , Sihao Chen, Elizabeth Sieber, Mark Yatskar, Dan Roth, "ExpertQA : Expert-Curated Questions and Attributed Answers", NAACL'2024

Examples from ExpertQA

Field	Question	Types
Anthropology	*Why is it that Africa's representation is still a problem in modern day times regardless of the academic writings that state otherwise?*	II,VII
Architecture	*Suppose an architect decides to reuse an existing foundation of a demolished building, what is to be considered to ensure success of the project?*	IV
Biology	*Can you explain the mechanisms by which habitat fragmentation affects biodiversity and ecosystem functioning, and provide examples of effective strategies for mitigating these impacts?*	III,VI
Chemistry	*Why does gallic acid have an affinity with trivalent iron ions?*	I
Engineering & Technology	*How different will licensing a small modular reactor be as compared to licensing traditional large nuclear power plants?*	VII
Healthcare/Medicine	*If a 48 year old woman is found to have an esophageal carcinoma that invades the muscularis propria and has regional lymph node metastases but no distant metastasis, what is her stage of cancer and what are possible recommended treatments?*	I,III
Law	*Can direct evidence in a case that has been obtained illegally be considered by the court in some cases if it directly points to the defendant's guilt?*	I
Music	*What exercises would you do in a singing class with a teenager with puberphonia?*	IV
Physics & Astronomy	*Standard Model does not contain enough CP violating phenomena in order to explain baryon asymmetry. Suppose the existence of such phenomena. Can you propose a way to experimentally observe them?*	V
Political Science	*Despite the fact that IPCC was formed in 1988, several studies have showed that argubaly more than 50% of all carbon emissions in history have been released since 1988. What does this show about IPCC and developed countries' efforts?*	VII
Visual Arts	*Tell me the step by step process of recycling a canvas.*	III

Towards Theme-Specific Knowledge Graph and Reasoning

- LLMs: Power and limitations
  - LLMs are trained from massive general data
  - But specific problem solving often needs to go deep and current
  - Theme-specific retrieval should be obtained by task-specific retrieval
  - Will RAG (retrieval augmented generation) be sufficient for complex reasoning/problem solving?
- Structures can help problem solving
  - Structures have helped human learning, understanding, reasoning, and discovery
  - The general KGs could be too general to help solving specific problems
  - We need to use theme-specific and task-specific structures/graphs
- Research questions
  - How can we construct theme-specific and task-specific graphs automatically?
  - How can we use such graphs efficiently for structure-augmented LLM reasoning?

A Retrieving-Structuring-Reasoning Framework

LLM is important in each step but the whole framework enhances LLM's performance!

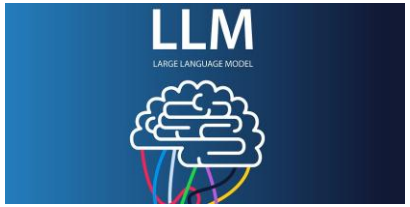
User Query/Task

Text & Multimodal Data

Retrieving

Query/task-guided Theme-focused Information Retrieval

General KB



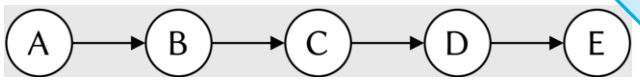
LLMs

Structuring

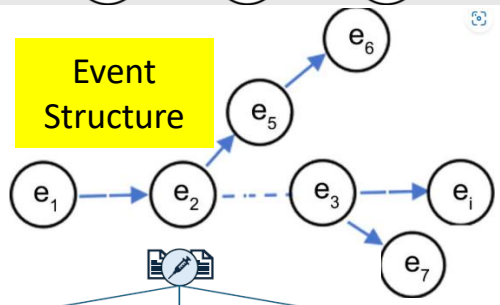
Task-specific Structure Mining & Graph Construction

Selected, Distilled, Relevant Documents

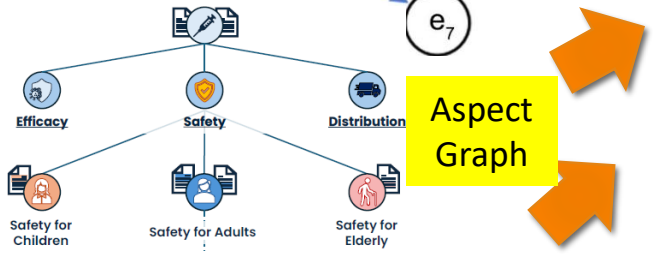
Causal Graph



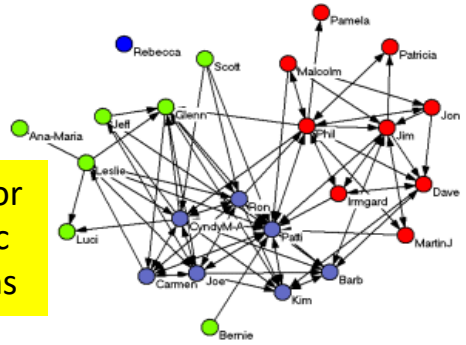
Event Structure



Aspect Graph



Multiple Theme- or Function- Specific Knowledge Graphs



Reasoning

Task- and Structure-based Augmentation for LLM Generation

Knowledge with Quality Reasoning

References I

- Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, Huajun Chen, “KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction”, WWW’22
- Yew Ken Chia, Lidong Bing, Soujanya Poria, Luo Si, “RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction”, ACL’22 Findings
- Linyi Ding, Jinfeng Xiao, Sizhe Zhou, Chaoqi Yang, Jiawei Han, “Topic-Oriented Open Relation Extraction with A Priori Seed Generation”, EMNLP’24
- Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. ACL
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. ICML
- Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, Jiawei Han, "Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval", ICLR’2025
- Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, “Active retrieval augmented generation,” 2023.
- Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, Jiawei Han. "Open-vocabulary argument role prediction for event extraction," EMNLP’22 Findings
- Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. “Instruct and Extract: Instruction Tuning for On-Demand Information Extraction.” EMNLP 2023.
- Yizhu Jiao, Sha Li, Sizhe Zhou, Heng Ji, and Jiawei Han. “ Text2DB: Integration-Aware Information Extraction with Large Language Model Agents .” ACL 2024 findings.
- Priyanka Kargupta, Yunyi Zhang, Yizhu Jiao, Siru Ouyang, Jiawei Han, "Synergizing Unsupervised Episode Detection with LLMs for Large-Scale News Events", ACL 2025
- Priyanka Kargupta, Runchu Tian, Jiawei Han, "Beyond True or False: Retrieval-Augmented Hierarchical Analysis of Nuanced Claims", ACL 2025
- Priyanka Kargupta, Ishika Agarwal, Tal August, Jiawei Han, "Tree-of-Debate: Multi-Persona Debate Trees Elicit Critical Thinking for Scientific Comparative Analysis", ACL 2025
- Seoyeon Kim, Kwangwook Seo, Hyungjoo Chae, Jinyoung Yeo, Dongha Lee, “VerifiNER: Verification-augmented NER via Knowledge-grounded Reasoning with Large Language Models”, ACL’24

References II

- Le Scao, T., & Rush, A. M. (2021). How many data points is a prompt worth? NAACL
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. ACL
- P. Manakul, A. Liusie, and M. J. F. Gales, "SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models," 2023
- Yubo Ma, Yixin Cao, YongChing Hong, Aixin Sun, "Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!", EMNLP'23 Findings
- S. Min, K. Krishna, X. Lyu, M. Lewis, W. tau Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," 2023
- O. Ovadia, et al (2023), "Fine-tuning or retrieval? comparing knowledge injection in LLMs," arXiv:2312.05934
- B. Paranjape, S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro, "Art: Automatic multi-step reasoning and tool-use for large language models," 2023
- Jash Parekh, Pengcheng Jiang, Jiawei Han, "CC-RAG: Structured Multi-Hop Reasoning via Theme-Based Causal Graphs", arXiv:2506.08364
- X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai and X. Huang, Pre-trained models for natural language processing: A survey, Science China-Technological Sciences, 2020
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI blog.
- Schick, T., & Schütze, H. (2021). Exploiting cloze questions for few shot text classification and natural language inference. EACL
- Omar Shaikh, et al. (2022), "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning" arXiv.2212.08061
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., Wei, J. (2023). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ACL Findings.
- N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," 2023.
- Chengen Wang and Murat Kantarcioglu, "A Review of DeepSeek Models' Key Innovative Techniques", arXiv:2503.11486
- Wei, J., et al. (2022). Emergent Abilities of Large Language Models. TMLR.
- T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai, "PromptChainer: Chaining large language model prompts through visual programming," 2022

References III

- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS.
- Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. “FolkScope: Intention Knowledge Graph Construction for E-commerce Commonsense Discovery”, ACL’23 Findings
- S. J. Zhang, et al., “Exploring the MIT mathematics and EECS curriculum using large language models,” 2023
- Zhang, Z., Zhang, A., Li, M., Smola, A. (2023). Automatic Chain of Thought Prompting in LLMs. ICLR
- Kai Zhang, Bernal Jiménez Gutiérrez, Yu Su, “Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors”, ACL'23 Findings
- Yu Zhang, Yu Meng, Xuan Wang, Sheng Wang, Jiawei Han. "Seed-guided topic discovery with out-of-vocabulary seeds." NAACL’22.
- Yu Zhang, Yunyi Zhang, Yucheng Jiang, Martin Michalski, Yu Deng, Lucian Popa, ChengXiang Zhai, Jiawei Han. "Entity Set Co-Expansion in StackOverflow." IEEE Big Data 2022
- Ming Zhong, Siru Ouyang, Minhao Jiang, Vivian Hu, Yizhu Jiao, Xuan Wang, Jiawei Han, “ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision”, ACL’23 Findings
- Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao and Jiawei Han, "Reaction Miner: An Integrated System for Chemical Reaction Extraction from Textual Data", EMNLP’23
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: Less Is More for Alignment.
- Sizhe Zhou, Suyu Ge, Jiawei Han, “Corpus-Based Relation Extraction by Identifying and Refining Relation Patterns”, ECMLPKDD’23
- Sizhe Zhou, Yu Meng, Bowen Jin, Jiawei Han, “Grasping the Essentials: Tailoring Large Language Models for Zero-Shot Relation Extraction”, arXiv’24
- Wenxuan Zhou, Muhao Chen, “An Improved Baseline for Sentence-level Relation Extraction”, AACL’22
- Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” 2023