
Affordable AI Assistants with Knowledge Graph of Thoughts

Maciej Besta^{†1} Lorenzo Paleari¹ Jia Hao Andrea Jiang¹ Robert Gerstenberger¹ You Wu¹ Patrick Iff¹
Ales Kubicek¹ Piotr Nyczyk² Diana Khimey¹ Jón Gunnar Hannesson¹ Grzegorz Kwaśniewski¹
Marcin Copik¹ Hubert Niewiadomski² Torsten Hoefler¹

Abstract

Large Language Models (LLMs) are revolutionizing the development of AI assistants capable of performing diverse tasks across domains. However, current state-of-the-art LLM-driven agents face significant challenges, including high operational costs and limited success rates on complex benchmarks like GAIA. To address these issues, we propose the Knowledge Graph of Thoughts (KGoT), an innovative AI assistant architecture that integrates LLM reasoning with dynamically constructed knowledge graphs (KGs). KGoT extracts and structures task-relevant knowledge into a dynamic KG representation, iteratively enhanced through external tools such as math solvers, web crawlers, and Python scripts. Such structured representation of task-relevant knowledge enables low-cost models to solve complex tasks effectively. For example, KGoT achieves a 29% improvement in task success rates on the GAIA benchmark compared to Hugging Face Agents with GPT-4o mini, while reducing costs by over 36× compared to GPT-4o. Improvements for recent reasoning models are similar, e.g., 36% and 37.5% for Qwen2.5-32B and Deepseek-R1-70B, respectively. KGoT offers a scalable, affordable, and high-performing solution for AI assistants.

Website & code: <https://github.com/spcl/knowledge-graph-of-thoughts>

1. Introduction

Large Language Models (LLMs) are transforming the world. However, training LLMs is expensive, time-consuming, and resource-intensive. In order to democratize the access to generative AI, the landscape of agent systems has massively

evolved during the last two years (LangChain Inc., 2024; Rush, 2023; Kim et al., 2024; Summers et al., 2024; Hong et al., 2024; Guo et al., 2024; Edge et al., 2024; Besta et al., 2025b; Zhuge et al., 2024; Beurer-Kellner et al., 2024; Shinn et al., 2023; Kagaya et al., 2024; Zhao et al., 2024a; Stengel-Eskin et al., 2024; Wu et al., 2024). These schemes have been applied to numerous tasks in reasoning (Creswell et al., 2023; Bhattacharjya et al., 2024; Besta et al., 2025b), planning (Wang et al., 2023c; Prasad et al., 2024; Shen et al., 2023; Huang et al., 2023), software development (Tang et al., 2024), and many others (Xie et al., 2024; Schick et al., 2023; Beurer-Kellner et al., 2023).

Among the most impactful applications of LLM agents is the development of AI assistants capable of helping with a wide variety of tasks. These assistants promise to serve as versatile tools, enhancing productivity and decision-making across domains. From aiding researchers with complex problem-solving to managing day-to-day tasks for individuals, AI assistants are becoming an indispensable part of modern life. Developing such systems is highly relevant, but remains challenging, particularly in designing solutions that are both effective and economically viable.

The GAIA benchmark (Mialon et al., 2024) has emerged as a valuable standard for evaluating LLM-driven agent architectures in their capacity to function as general-purpose AI assistants. This benchmark rigorously tests these systems across diverse tasks (involving web navigation, code execution, image reasoning, scientific QA, and multimodal tasks), providing a clear measure of their competence. However, despite more than a year since its introduction, the top-performing solutions on GAIA still fail at many tasks. Furthermore, the cost of operating these systems is prohibitively high. For instance, executing all tasks from the validation set using Hugging Face Agents (Roucher & Petrov, 2024) with GPT-4o incurs costs of roughly \$200, illustrating the need for more cost-efficient alternatives. While deploying smaller models like GPT-4o mini offers significant cost reductions, it results in a substantial decline in task success rates, rendering it an insufficient solution. Alternatively, when using open models, maintaining the infrastructure for large models is costly and often prohibitive for an average user; small open models require inexpensive hardware but

¹ETH Zurich, Zurich, Switzerland ²Cledar, Wieliczka, Poland.
[†]Correspondence to: Maciej Besta <maciej.best@inf.ethz.ch>.

are less capable.

To address these challenges, we propose the Knowledge Graph of Thoughts (KGoT), a novel AI assistant architecture designed to significantly reduce task execution costs while maintaining a high success rate (**contribution #1**). The central innovation of KGoT lies in its use of a knowledge graph (KG) (Singhal, 2012; Besta et al., 2024b) to extract and structure knowledge relevant to a given task. A KG organizes information into triples, providing a structured representation of knowledge that small, cost-effective models can efficiently process. Hence, KGoT “turns the unstructured into the structured”, i.e., KGoT turns the often unstructured data such as website contents or PDF files into structured KG triples. This approach enhances the comprehension of task requirements, enabling even smaller models to achieve performance levels comparable to much larger counterparts, but at a fraction of the cost.

The KGoT architecture (**contribution #2**) implements this concept by iteratively constructing a KG from the task statement, incorporating tools as needed to gather relevant information. The constructed KG is kept in a graph store, serving as a repository of structured knowledge. Once sufficient information is gathered, the LLM attempts to solve the task by either directly embedding the KG in its context or querying the graph store for specific insights. This approach ensures that the LLM operates with a rich and structured knowledge base, improving its task-solving ability without incurring the high costs typically associated with large models. The architecture is modular and extensible towards different types of graph query languages and tools.

Our evaluation against top GAIA leaderboard baselines demonstrates its effectiveness and efficiency (**contribution #3**). KGoT solves $>2\times$ more tasks from the validation set than Hugging Face Agents with GPT-4o mini. Moreover, harnessing a smaller model dramatically reduces operational costs. Specifically, with GPT-4o mini instead of GPT-4o, KGoT *lowers task execution costs from \$187 to roughly \$5*. On top of that, KGoT reduces bias and improves fairness by externalizing reasoning into an explicit knowledge graph rather than relying solely on the LLM’s internal generation. This ensures that key steps when resolving tasks are grounded in transparent and auditable information. This highlights the potential of KGoT in developing affordable AI assistants capable of high performance across a diverse range of tasks.

2. Knowledge Graph of Thoughts

We first illustrate the key idea, namely, using a knowledge graph to encode *structurally* the task contents. Figure 1 shows an example task and its corresponding evolving KG.

2.1. What is a Knowledge Graph?

A knowledge graph (KG) is a structured representation of information that organizes knowledge into a graph-based format, allowing for efficient querying, reasoning, and retrieval. Formally, a KG consists of a set of triples, where each triple (s, p, o) represents a relationship between two entities s (subject) and o (object) through a predicate p . For example, the triple (“Earth”, “orbits”, “Sun”) captures the fact that Earth orbits the Sun.

Mathematically, a knowledge graph can be defined as a directed labeled graph $G = (V, E, L)$, where V is the set of vertices (entities), $E \subseteq V \times V$ is the set of edges (relationships), and L is the set of labels (predicates) assigned to the edges. Each entity or predicate may further include properties or attributes, enabling richer representation. Knowledge graphs are widely used in various domains, including search engines, recommendation systems, and AI reasoning, as they facilitate both efficient storage and complex queries.

2.2. Harnessing KGs for Effective Task Resolution

At the heart of KGoT is the process of transforming a task solution state into an evolving KG. The KG representation of the task is built from “thoughts” generated by the LLM in the iterative process of enhancing the KG. These “thoughts” are intermediate insights identified by the LLM as it works through the problem. Each thought contributes to refining or expanding the KG by adding new vertices, edges, or attributes that represent new information.

For example, consider the following Level 3 (i.e., highest difficulty) task from the GAIA benchmark: *“In the YouTube 360 VR video from March 2018 narrated by the voice actor of Lord of the Rings’ Gollum, what number was mentioned by the narrator directly after dinosaurs were first shown in the video?”* (see Figure 1). Here, the KG representation of the task solution state has a vertex *“Gollum (LotR)”*. Then, the thought *“Gollum from Lord of the Rings is interpreted by Andy Serkis”* results in adding a vertex for *“Andy Serkis”*, and linking *“Gollum (LotR)”* to *“Andy Serkis”* with the predicate *“interpreted by”*. Such integration of thought generation and KG construction creates a feedback loop where the KG continuously evolves as the task progresses, aligning the representation with problem requirements.

In order to evolve the KG task representation, KGoT interacts with tools and retrieves more information. For instance, the system might query the internet to identify videos narrated by Andy Serkis (e.g., *“The Silmarillion”* and *“We Are Stars”*). It can also use a YouTube transcriber tool to find their publication date, adding new vertices and edges to the graph. Intermediate results, such as the video type or comparisons, are incorporated back into the graph, creating a more complete and structured representation of the task.

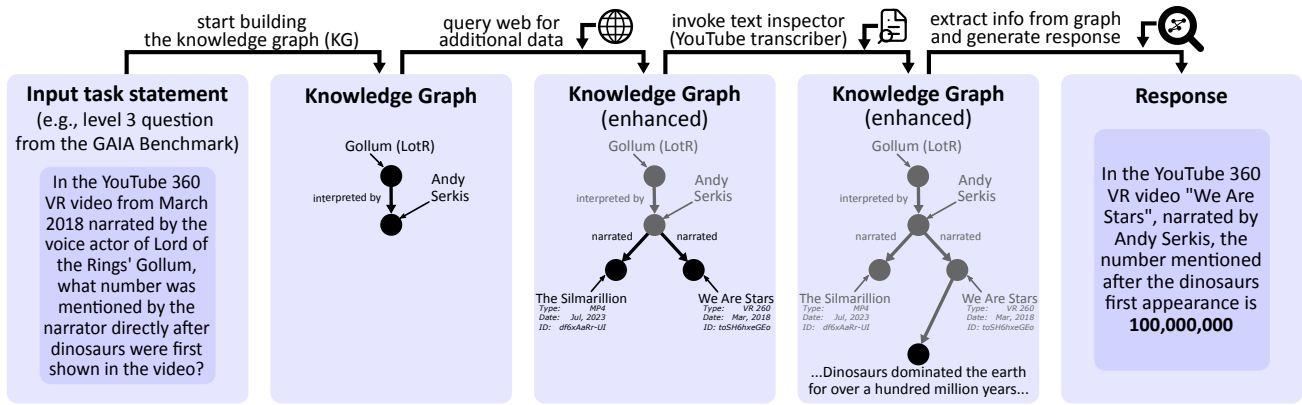


Figure 1. Illustration of the key idea behind Knowledge Graph of Thoughts (KGoT): transforming the representation of a task for an AI assistant from a textual form into a knowledge graph (KG). As an example, we use a Level-3 (i.e., highest difficulty) task from the GAIA benchmark. In order to solve the task, KGoT evolves this KG by adding relevant information that brings the task closer to completion. This is achieved by iteratively running various tools. Finally, the task is solved by extracting the relevant information from the KG, using – for example – a graph query, or an LLM’s inference process with the KG provided as a part of the input prompt.

This iterative refinement allows the KG to model the current “state” of the task at each step, bringing it closer to completion. The system’s dynamic nature enables it to address a wide range of tasks by adapting the graph’s structure and content in response to real-time interactions. For example, multi-step reasoning tasks, such as synthesizing data from different sources or performing calculations, are handled by adding relevant subgraphs or updating existing vertices based on the latest retrieved information. Once the KG has been sufficiently populated with task-specific knowledge, it serves as a robust resource for solving the problem.

2.3. Extracting Information from KG

To accommodate varying performance requirements and tasks, KGoT supports different ways to extract the information from the KG when solving a task. Currently, we offer graph query languages or general-purpose languages; each of them can be combined with the so-called Direct Retrieval.

Graph Query Languages First, to solve the task, one can use a graph query, prepared by the LLM in a language such as Cypher (Francis et al., 2018) or SPARQL (Pérez et al., 2009), to extract the answer to the task from the graph. This capability is particularly advantageous for tasks that require retrieving specific subgraphs, relationships, or patterns within the KG.

General-Purpose Languages Another way to extract the information from the KG that is needed to solve the task is to use a script prepared by the LLM in a general-purpose programming language such as Python. This approach, while not as effective as query languages for workloads such as pattern matching, offers greater flexibility and may outperform the latter when a task requires, for example, traversing a long path in the graph.

Direct Retrieval In certain cases, once enough information is gathered into the KG, it may be more effective to directly paste the KG into the LLM context and ask the LLM to solve the task, instead of preparing a dedicated query or script. We refer to this approach as Direct Retrieval.

Accuracy-Cost-Runtime Tradeoff The three above schemes offer a tradeoff between accuracy, cost, and runtime. For example, when low latency is of top priority, general-purpose languages and the corresponding frameworks such as NetworkX should be used, as they provide an efficient lightweight representation of the KG and offer rapid access and modification of graph data. When token cost is most important, one should avoid Direct Retrieval (which consumes many tokens as it directly embeds the KG into the LLM context) and focus on either query or general-purpose languages, with a certain preference for the former, because – based on our experience – generated queries tend to be shorter than scripts. Finally, when aiming for solving as many tasks as possible, one should experiment with all three schemes –As shown in the Evaluation section, **these methods have complementary strengths**—Direct Retrieval is effective for broad contextual understanding, while graph queries and scripts are better suited for structured reasoning.

3. System Architecture

The KGoT system, pictured in Figure 2, is designed as a modular and flexible framework that comprises three main components: the **Graph Store Module**, the **Controller**, and the **Integrated Tools**, each playing a critical role in the task-solving process. Below, we provide a detailed description of each component and its role in the system.

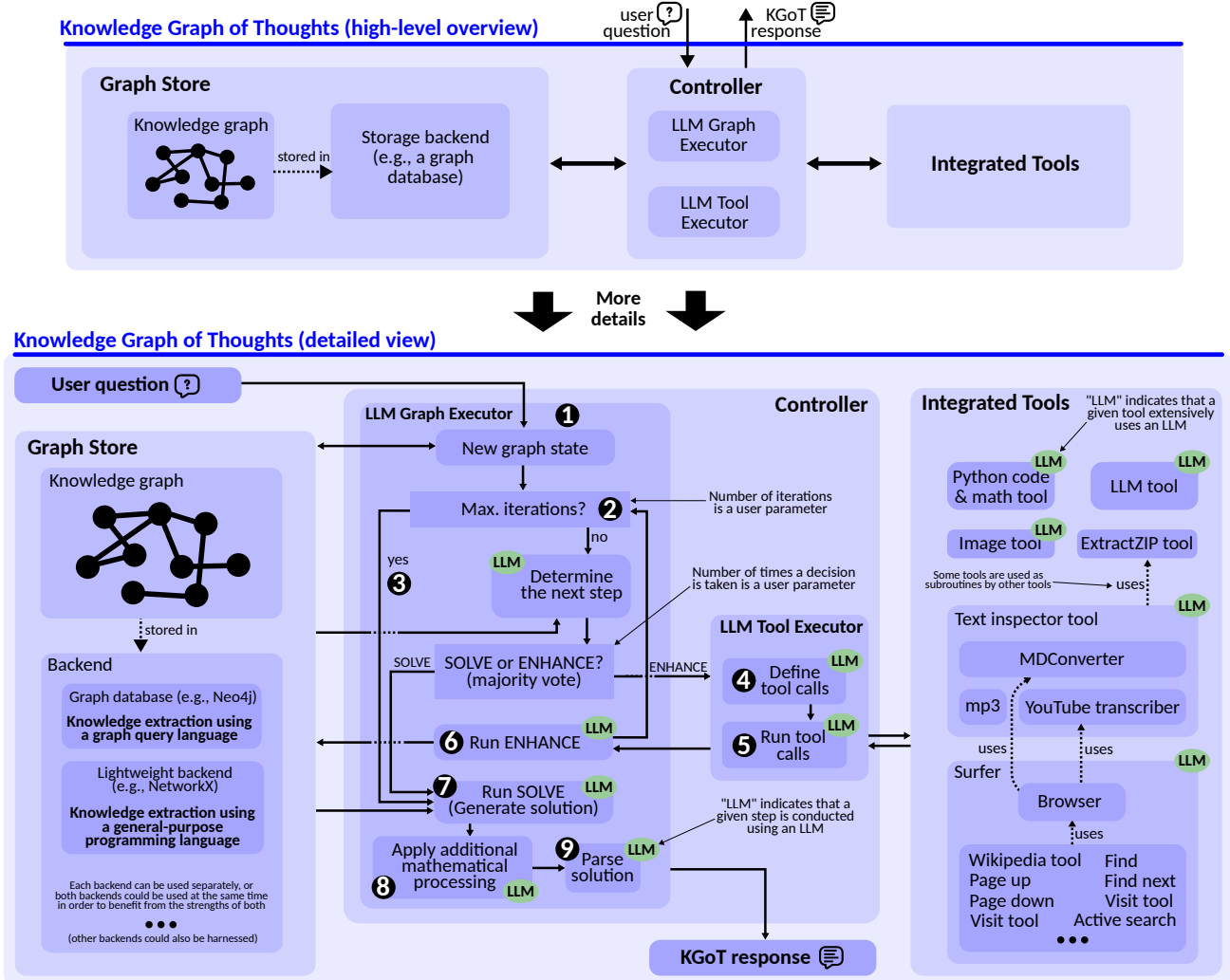


Figure 2. Architecture overview of KGoT (top part) and the design details combined with the workflow (bottom part).

3.1. Graph Store Module for Maintaining the KG

A key component of the KGoT system is the Graph Store Module, which manages the storage and retrieval of the dynamically evolving knowledge graph which represents the task state. In order to harness graph queries, we use a graph database backend; in the current KGoT implementation, it is Neo4j (Robinson et al., 2015), an established graph database (Besta et al., 2023c;b;d) (we selected Cypher and Neo4j after an analysis of the literature that indicated limitations for the LLM-based query generation of SPARQL queries (Emonet et al., 2025; Mecharnia & d’Aquin, 2025)). Then, in order to support graph accesses using a general-purpose language, KGoT harnesses the NetworkX library (NetworkX Developers, 2024) and Python. Note that the extensible design of KGoT enables seamless integration of any other backends and languages.

3.2. Controller for Workflow Management

The Controller is the central orchestrator of the KGoT system, responsible for managing the interaction between the knowledge graph and the integrated tools. When a user submits a query, the Controller initiates the reasoning process by interpreting the task and coordinating the steps required for resolution. It dynamically determines which tools to invoke based on the current state of the KG and the specific requirements of the task. As tools produce results, the Controller integrates these outputs back into the KG, updating its structure to reflect the new knowledge.

The KGoT Controller employs a dual-LLM architecture with a *clear separation of roles between constructing the KG (managed by the LLM Graph Executor) and interacting with tools (managed by the LLM Tool Executor)*.

The **LLM Graph Executor** determines the next steps after each iteration that constructs and evolves the KG. It identifies any missing information necessary to solve the

task, formulates appropriate queries for interacting with the graph store backend (retrieve/insert operations), and parses intermediate or final results for integration into the KG. It also prepares the final response to the user by synthesizing outputs from all previous steps and from the KG.

The **LLM Tool Executor** operates as the executor of the plan devised by the LLM Graph Executor. It identifies the most suitable tools for retrieving missing information, considering factors such as tool availability, relevance, and the outcome of previous tool invocation attempts. For example, if a web crawler fails to retrieve certain data, the LLM Tool Executor might prioritize a different retrieval mechanism or adjust its queries. The LLM Tool Executor manages the tool execution process, including interacting with APIs, performing calculations, or extracting information, and returns the results to the LLM Graph Executor for further reasoning and integration into the KG.

3.3. Integrated Tools for Evolving KG

The Integrated Tools Module in the KGoT system provides a hierarchical and diverse suite of specialized tools, each tailored to address specific task requirements. At the foundation is the **Python Code Tool**, enabling the generation and execution of dynamic scripts for complex computations and algorithmic tasks. The code tool is also used when solving math steps. Supplementing the controller’s reasoning, the **LLM Tool** integrates an additional language model to provide extended knowledge beyond the constrained capabilities of the controller’s LLM, ensuring robust reasoning with minimal risk of hallucinations. For multimodal tasks, the **Image Tool** facilitates image processing and information extraction. Web-based operations are handled by the **Surfer Agent**, based on the design by Hugging Face Agents (Roucher & Petrov, 2024), which interacts with the web through tools like the **Wikipedia Tool** and **granular navigation tools** (e.g., PageUp, PageDown, Find) while leveraging SerpApi (SerpApi LLM, 2025) for effective searches. Additional capabilities include the **ExtractZip Tool**, designed for processing compressed archives, and the **Text Inspector Tool**, which extracts and transforms text from diverse sources such as MP3 files, YouTube transcripts, and various formats into Markdown. This modular hierarchy ensures flexibility, extensibility, and adaptability in solving a wide range of complex tasks.

3.4. High-Performance & Scalability

KGoT uses various optimizations to enhance scalability and performance. They include (1) **asynchronous execution** using asyncio (Python Software Foundation, 2025b) to parallelize LLM tool invocations, mitigating I/O bottlenecks and reducing idle time, (2) **graph operation parallelism** by reformulating LLM-generated Cypher queries to enable

concurrent execution of independent operations in a graph database, and (3) MPI-based **distributed processing**, which decomposes workloads into atomic tasks distributed across ranks using a work-stealing algorithm to ensure balanced computational load and scalability.

3.5. System Robustness with Majority Voting

One of the key strategies employed to enhance robustness is the use of **majority voting**, also known as self-consistency (Wang et al., 2023b); using other strategies such as embedding-based approaches could also be possible (Besta et al., 2024d). In KGoT, majority voting is implemented by querying the LLM multiple times when deciding the next step, whether to insert more data into the knowledge graph or retrieve existing data, when deciding which tool to use, and when parsing the final solution. This approach reduces the impact of single-instance errors or inconsistencies in various parts of the KGoT architecture, ensuring that the decisions made reflect the LLM’s most consistent reasoning paths.

3.6. Layered Error Containment & Management

To manage **LLM-generated syntax errors**, KGoT includes LangChain’s JSON parsers that detect syntax issues. When a syntax error is detected, the system first attempts to correct it by adjusting the problematic syntax using different encoders, such as the “unicode escape” (Python Software Foundation, 2025a). If the issue persists, KGoT employs a retry mechanism (three attempts by default) that uses the LLM to rephrase the query/command and attempts to regenerate its output. If the error persists, the system logs it for further analysis, bypasses the problematic query, and continues with other iterations.

To manage **API & system related errors**, such as the OpenAI code 500, the primary strategy employed is exponential backoff, implemented using the tenacity library (Tenacity Developers, 2024). Additionally, KGoT includes comprehensive logging systems as part of its error management framework. These systems track the errors encountered during system operation, providing valuable data that can be easily parsed and analyzed (e.g., snapshots of the knowledge graphs or responses from third-party APIs).

The Python Executor tool, a key component of the system, is **containerized** to ensure **secure execution of LLM-generated code**. This tool is designed to run code with strict timeouts and safeguards, preventing potential misuse or resource overconsumption.

3.7. Implementation Details

Containerization with Docker and Sarus The KGoT system employs Docker (Docker Inc., 2024) and Sarus (Benedi-

cic et al., 2019) for containerization, enabling a consistent and isolated runtime environment for all components. We containerize critical modules such as the KGoT controller, the Neo4j knowledge graph, and integrated tools (e.g., the Python Executor tool for safely running LLM-generated code with timeouts). Here, **Docker** provides a widely adopted containerization platform that guarantees consistency between development and production environments. **Sarus**, a specialized container platform designed for high-performance computing (HPC) environments, extends KGoT’s portability to HPC settings where Docker is typically unavailable due to security constraints. This integration allows KGoT to operate efficiently in HPC environments, leveraging their computational power.

Adaptability with LangChain The KGoT system harnesses LangChain (LangChain Inc., 2024), an open-source platform specifically designed for creating and orchestrating LLM-driven applications. LangChain offers a comprehensive suite of tools and APIs that simplify the complexities of managing LLMs, including prompt engineering, tool integration, and the coordination of LLM outputs.

4. System Workflow

We show the workflow in the bottom part of Figure 2. The workflow begins when the user submits a problem to the system ①. The first step is to verify whether the maximum number of iterations allowed for solving the problem has been reached ②. If the iteration limit is exceeded, the system will no longer try to gather additional information and insert it into the KG, but instead will return a solution with the existing data in the KG ③. Otherwise, the majority vote (over several replies from the LLM) decides whether the system should proceed with the **Enhance** pathway (using tools to generate new knowledge) or directly proceed to the **Solve** pathway (gathering the existing knowledge in the KG and using it to deliver the task solution).

The Enhance Pathway If the majority vote indicates an Enhance pathway, the next step involves determining the tools necessary for completing the Enhance operation ④. The system then orchestrates the appropriate tool calls based on the KG state ⑤. Once the required data from the tools is collected, the system generates the Enhance query or queries to modify the KG appropriately. Each Enhance query is executed ⑥ and its output is validated. If an error or invalid value is returned, the system attempts to fix the query using a decoder or the LLM, retrying a specified number of times. If retries fail, the query is discarded, and the operation moves on. After processing the Enhance operation, the system increments the iteration count and continues until the KG is sufficiently expanded or the iteration limit is reached. This path ensures that the knowledge graph is enriched with relevant and accurate information, enabling the system to

progress toward a solution effectively.

The Solve Pathway If the majority vote directs the system to the Solve pathway, the system executes multiple solve operations iteratively ⑦. If an execution produces an invalid value or error three times in a row, the system asks the LLM to attempt to correct the issue by recreating the used query. The query is then re-executed. If errors persist after three such retries, the query is regenerated entirely, disregarding the faulty result, and the process restarts. After the Solve operation returns the result, final parsing is applied, which includes potential mathematical processing to resolve potential calculations ⑧ and refining the output (e.g., formatting the results appropriately) ⑨.

5. Evaluation

We now show advantages of KGoT over the state of the art. We focus on GAIA as this is currently the most comprehensive benchmark for general-purpose AI assistants, covering diverse domains such as web navigation, code execution, image reasoning, scientific QA, and multimodal tasks.

Comparison Baselines We focus on the **Hugging Face (HF) Agents**, the most competitive scheme in the GAIA benchmark for the hardest level 3 tasks with the GPT-4 class of models. We also compare to **Zero Shot** schemes where a model answers without any additional agent framework.

KGoT variants First, we vary the **approach for knowledge extraction (graph queries vs. general-purpose language, cf. Section 2.3)**. For each option, we vary how the Solve operation is executed, by either having the LLM send a request to the backend (a Python script for NetworkX and a Cypher query for Neo4j) or by directly asking the LLM to infer the answer based on the KG, which we termed **Direct Retrieval (DR)**. We also consider **“fusion” runs**, which simulate the effect from KGoT runs with both graph backends available simultaneously (or both Solve operation variants harnessed for each task). Fusion runs only incur negligible additional storage overhead because the generated KGs are small (up to several hundreds of nodes). Finally, we experiment with different **tool sets**. To focus on the differences coming from harnessing the KG, we reuse several utilities from AutoGen (Wu et al., 2024) such as Browser and MDConverter, and tools from HF Agents, such as Surfer Agent, web browsing tools, and Text Inspector.

Considered Metrics We focus primarily on the number of solved tasks as well as token costs (\$). Unless stated otherwise, we report single run results due to budget reasons.

Considered Dataset We use the GAIA benchmark (Mialon et al., 2024); focusing on the validation set (165 tasks) for budgetary reasons and also because it comes with the ground truth answers.

Affordable AI Assistants with Knowledge Graph of Thoughts

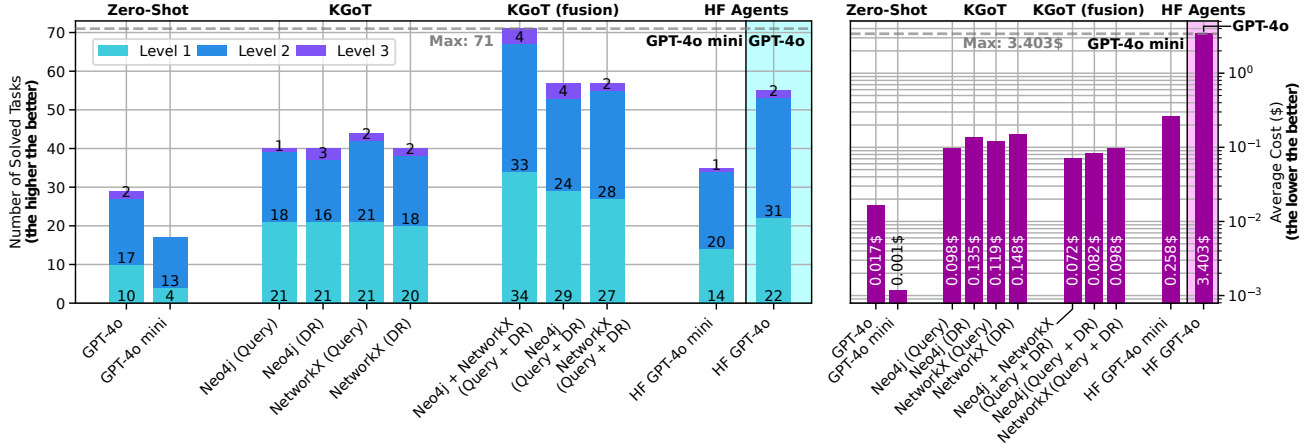


Figure 3. Comparison of different variants of KGoT with Hugging Face Agents and with Zero-Shot GPT-4o mini and GPT-4o. DR stands for Direct Retrieval. The used model is GPT-4o mini unless noted otherwise.

Scalability We verified that selecting Neo4j as the graph query backend is not the bottleneck (the majority of the time is spent on the tool usage, most importantly, web browsing and text parsing). Moreover, due to the effective knowledge extraction process and the nature of the tasks considered (i.e., AI assistance), none of the tasks require large KGs. The maximum graph size that we observed was 522 nodes. This is orders of magnitude below any scalability concerns.

5.1. Advantages of KGoT

Figure 3 shows the number of solved tasks (left side) as well as the average cost per solved task (right side) for different KGoT variants (explained above) as well as three baselines (HF Agents using both GPT-4o mini and GPT-4o, and the Zero-Shot GPT-4o mini and GPT-4o). Additionally, we show the Pareto front in Figure 4 for the multidimensional optimization problem of improving accuracy (i.e., reducing failed tasks) and lowering cost. All variants of KGoT solve a greater number of tasks (up to 9 more) compared to HF Agents while also being more cost-efficient (between 42% to 62% lower costs). The key reason for the KGoT advantages stems from harnessing the knowledge graph-based representation of the evolving task state.

The ideal fusion runs of Neo4j and NetworkX solve an even greater number of tasks (57 for both) than the single runs, and also have a lower average cost (up to 68% lower than Hugging Face Agents). The fusion of all combinations of backend and solver types would solve by far the highest number of tasks (71) – more than twice as much as Hugging Face Agents – while also exhibiting the lowest average cost per solved query (nearly 72% lower than Hugging Face Agents).

The direct zero-shot use of GPT-4o mini and GPT-4o has the lowest average cost per solved task (just \$0.0013 and \$0.0164 respectively), making it the most cost-effective,

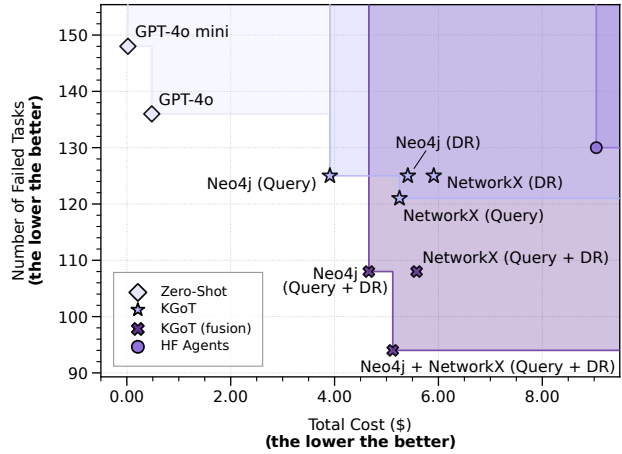


Figure 4. Pareto front plot of cost and error counts. We report results for answering 165 GAIA validation questions across different comparison targets, using the GPT-4o mini model with each baseline. For the Zero-Shot inference, we also include results for GPT-4o for comparison. DR means Direct Retrieval.

however this approach is only able to solve 17 and 29 tasks, respectively.

5.2. Impact from Various Design Decisions

We also analyze the impact of various design decisions.

We explore two different ways to extract knowledge for graph processing: graph queries (with Cypher and Neo4j) and a general-purpose language (with Python-based graph operations and NetworkX), each with distinct advantages. Graph queries and Neo4j excel at structured queries, such as counting patterns. However, Cypher queries can be difficult to generate correctly, especially for graphs with more nodes and edges. Python and NetworkX offers certain advantages over Neo4j by eliminating the need for a separate database

Configuration			Metrics		
Tools	ST	PF	Solved	Time (h)	Cost
HF	DR	XML	37	11.87	\$7.84
HF	GQ	MD	33	9.70	\$4.28
merged	GQ	XML	31	10.62	\$5.43
HF	GQ	XML	30	13.02	\$4.90
original KGoT	GQ	XML	27	27.57	\$6.85

Table 1. Analysis of different design decisions and tool sets in KGoT. “ST” stands for the type of the solve operation and pathway (“GQ”: graph query, “DR”: direct retrieval), “PF” for the prompt format (“MD”: Markdown) and “merged” stands for a combination of the original KGoT tools and the Hugging Face Agents tools.

server, making it a lightweight choice for the KG. Moreover, NetworkX computations are fast and efficient for small to medium-sized graphs without the overhead of database transactions. Unlike Neo4j, which requires writing Cypher queries, We observe that in cases where Neo4j-based implementations struggle, NetworkX-generated graphs tend to be more detailed and provided richer vertex properties and relationships. This is likely due to the greater flexibility of Python code over Cypher queries for graph insertion, enabling more fine-grained control over vertex attributes and relationships. Another reason may be the fact that the used model is more skilled with Python than Cypher queries.

We also compare the difference between **direct retrieval (DR)** and solving tasks using graph queries or general-purpose languages. A deeper analysis revealed that each approach exhibits distinct strengths and weaknesses, as evidenced by the complementary performance of backend fusions. Our analysis of failed tasks indicates that, in many cases, the knowledge graph contains the required data, but *the graph query fails to extract it*. In such scenarios, direct retrieval, where the entire graph is included in the model’s context, performs significantly better. This is because it bypasses query composition issues. However, direct retrieval demonstrates lower accuracy in cases requiring structured, multi-step reasoning.

We also found that direct retrieval excels at extracting dispersed information but struggles with structured queries, whereas graph queries are more effective for structured reasoning but can fail when the LLM generates incorrect query formulations. Although both Cypher and general-purpose queries occasionally are erroneous, Python scripts require more frequent corrections because they are often longer and more error-prone. However, despite the higher number of corrections, the LLM is able to fix Python code more easily than Cypher queries, often succeeding after a single attempt. During retrieval, the LLM frequently embeds necessary computations directly within the Python scripts

while annotating its reasoning through comments, improving transparency and interpretability.

We also explored **different tool sets**, with selected results presented in Table 1. Initially, we examined the limitations of our original tools and subsequently integrated the complete Hugging Face Agents tool set into the KGoT framework, which led to improvements in accuracy, runtime, and cost efficiency. A detailed analysis allowed us to merge the most effective components from both tool sets into an optimized hybrid tool set, further enhancing accuracy and runtime while only moderately increasing costs. Key improvements include a tighter integration between the ExtractZip tool and the Text Inspector tool, which now supports Markdown, as well as enhancements to the Surfer Agent, incorporating a Wikipedia tool and augmenting viewpoint segmentation with full-page summarization. This optimized tool set was used for all subsequent experiments.

We further evaluated **different prompt formats** in the initial iterations of KGoT. While our primary format was XML-based, we conducted additional tests using Markdown. Initial experiments with the Hugging Face Agents tool set (see Table 1) combined with Markdown and GPT-4o mini yielded improved accuracy, reduced runtime, and lower costs. However, these results were not consistently reproducible with GPT-4o. Moreover, Markdown-based prompts interfered with optimizations such as direct retrieval, ultimately leading us to retain the XML-based format.

We also analyzed the **advantages of KGoT on different open models**, see Figure 5. KGoT offers consistent advantages over HF Agents for nearly all considered models (Guo et al., 2025). Interestingly, certain sizes of DeepSeek-R1 offer high zero-shot performance that outperforms both KGoT and HF Agents, illustrating potential for further improvements specifically aimed at Reasoning Language Models (RLMs) (Besta et al., 2025a;b).

Finally, we investigate the **impact on performance coming from harnessing KGs**, vs. using no KGs at all (the “no KG” baseline), see Figure 6. Harnessing KGs has clear advantages, with up to nearly $2\times$ increase in the count of solved tasks. This confirms the positive impact from structuring the task related knowledge into a graph format.

6. Related Work

Our work is related to numerous LLM domains.

First, we use LangChain (LangChain Inc., 2024) to facilitate the integration of the LLM agents with the rest of the KGoT system. Other such **LLM integration frameworks**, such as MiniChain (Rush, 2023) or AutoChain (Forethought, 2023), could be used instead.

Agent collaboration frameworks are systems such as

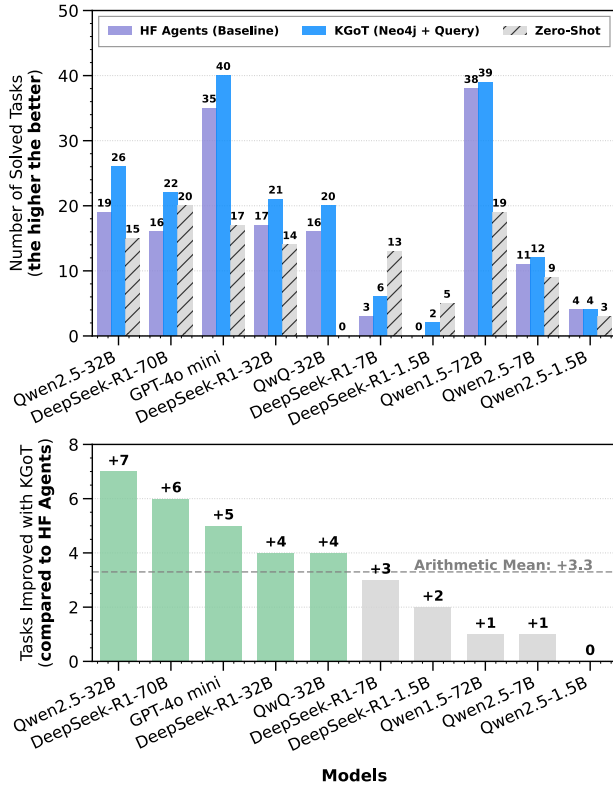


Figure 5. Performance on the GAIA validation set with KGoT (non-fusion) using various LLM models. For KGoT, we use Cypher graph queries for knowledge extraction from the Neo4j graph database.

MetaGPT (Hong et al., 2024), AutoAgents (Chen et al., 2024), and numerous others (Zhuge et al., 2024; Tang et al., 2024; Liu et al., 2024b; Li et al., 2024; Chu et al., 2024; Wu et al., 2024; Shinn et al., 2023; Zhu et al., 2024b; Kagaya et al., 2024; Zhao et al., 2024a; Stengel-Eskin et al., 2024; Significant Gravitas, 2025; Zhu et al., 2024a). The core KGoT idea can be applied to enhance such frameworks in that a KG can also be used as a common shared task representation for multiple agents solving a task together. Such a graph would be then updated by more than a single agent. Note that KGoT outperforms the highly competitive HF Agents baseline in the GAIA validation set, which means it offers more effective agent reasoning than other frameworks.

Many works exist in the domain of **general prompt engineering** (Beurer-Kellner et al., 2024; Besta et al., 2025b; Yao et al., 2023a; Besta et al., 2024a; Wei et al., 2022; Yao et al., 2023b; Chen et al., 2023; Creswell et al., 2023; Wang et al., 2023a; Hu et al., 2024; Dua et al., 2022; Jung et al., 2022; Ye et al., 2023). One could use such schemes to further enhance respective parts of the KGoT workflow. While we already use prompts that are suited for encoding

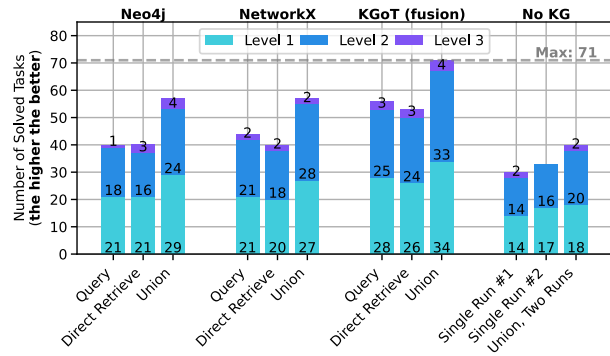


Figure 6. The impact coming from harnessing knowledge graphs (KGs) with different knowledge extraction methods (graph queries with Neo4j and Cypher, and general-purpose languages with Python and NetworkX), vs. using no KGs at all. DR stands for Direct Retrieval. Model: GPT-4o mini.

knowledge graphs, possibly harnessing other ideas from that domain could bring further benefits.

Task decomposition & planning increases the effectiveness of LLMs by dividing a task into subtasks. Examples include ADaPT (Prasad et al., 2024), ANPL (Huang et al., 2023), and others (Zhu et al., 2024a; Shen et al., 2023). Overall, the whole KGoT workflow already harnesses *recursive* task decomposition: the input task is divided into numerous steps, and many of these steps are further decomposed into sub steps by the LLM agent if necessary. For example, when solving a task based on the already constructed KG, the LLM agent may decide to decompose this step similarly to ADaPT. Other decomposition schemes could also be tried, we leave this as future work.

Retrieval-Augmented Generation (RAG) is an important part of the LLM ecosystem, with numerous designs being proposed (Edge et al., 2024; Gao et al., 2024; Besta et al., 2024c; Zhao et al., 2024b; Hu & Lu, 2024; Huang & Huang, 2024; Yu et al., 2024a; Mialon et al., 2023; Li et al., 2022; Abdallah & Jatowt, 2024; Delile et al., 2024; Manathunga & Illangasekara, 2023; Zeng et al., 2024; Wewer et al., 2021; Xu et al., 2024; Sarthi et al., 2024; Asai et al., 2024; Yu et al., 2024b). RAG has been used primarily to ensure data privacy and to reduce hallucinations. Using RAG is an orthogonal design choice; it could be combined with KGoT for further benefits.

Graph-Enhanced Agent Collaboration Frameworks There are works using graphs for more effective collaboration. Examples are GPTSwarm (Zhuge et al., 2024), MacNet (Qian et al., 2025), and AgentPrune (Zhang et al., 2024). These systems differ from KGoT in that they use a graph to model and manage *multiple agents* in a structured way, forming a hierarchy of tools. Contrarily, KGoT uses knowledge graphs to represent *the task itself*, including its

intermediate state. These two design choices are orthogonal and could be combined together. Moreover, while many of these systems require training, KGoT only relies on in-context learning.

Another increasingly important part of the LLM ecosystem is the **usage of tools** to augment the abilities of LLMs (Beurer-Kellner et al., 2023; Schick et al., 2023; Xie et al., 2024). For example, ToolNet (Liu et al., 2024a) uses a directed graph to model the application of multiple tools while solving a task, however focuses specifically on the iterative usage of tools at scale. KGoT harnesses a flexible and adaptable hierarchy of various tools, which can easily be extended, to solve a wide range of complex tasks.

7. Conclusion

In this paper, we introduce the Knowledge Graph of Thoughts (KGoT), an AI assistant architecture that enhances the reasoning capabilities of low-cost models while significantly reducing operational expenses. By dynamically constructing and evolving knowledge graphs (KGs) that encode the task and its resolution state, KGoT enables structured knowledge representation and retrieval, improving task success rates on complex benchmarks such as GAIA. Our extensive evaluation demonstrates that KGoT outperforms existing LLM-based agent solutions, achieving a substantial increase in task-solving efficiency of 29% or more over the competitive Hugging Face Agents baseline, while ensuring very low operational costs.

Beyond its current implementation, KGoT provides a flexible and scalable framework for AI assistant development, with potential applications in diverse fields such as automated research, data-driven decision-making, and multi-modal reasoning. Future work can explore integrating KGoT with additional external tools or with advanced graph predictive schemes for more robust KG construction (Besta et al., 2023a; 2024e), incorporating other classes of graph store backends such as neural graph databases (Besta et al., 2022) scaling KGoT to distributed-memory clusters (Blach et al., 2024), or refining its reasoning strategies by adapting more advanced task decomposition schemes.

Acknowledgements

We thank Hussein Harake, Colin McMurtrie, Mark Klein, Angelo Mangili, and the whole CSCS team granting access to the Ault and Daint machines, and for their excellent technical support. We thank Timo Schneider for help with infrastructure at SPCL. This project received funding from the European Research Council (Project PSAP, No. 101002047), and the European High-Performance Computing Joint Undertaking (JU) under grant agreement No. 955513 (MAELSTROM). This project was supported by the ETH Future

Computing Laboratory (EFCL), financed by a donation from Huawei Technologies. This project received funding from the European Union’s HE research and innovation programme under the grant agreement No. 101070141 (Project GLACIATION). We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017103.

References

- Abdallah, A. and Jatowt, A. Generator-Retriever-Generator Approach for Open-Domain Question Answering, March 2024. URL <https://arxiv.org/abs/2307.11278>. arXiv:2307.11278.
- Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the Twelfth International Conference on Learning Representations, ICLR '24*, Vienna, Austria, May 2024. OpenReview. URL <https://openreview.net/forum?id=hSyW5go0v8>.
- Benedicic, L., Cruz, F. A., Madonna, A., and Mariotti, K. Sarus: Highly Scalable Docker Containers for HPC Systems. In Weiland, M., Juckeland, G., Alam, S., and Jagode, H. (eds.), *High Performance Computing*, pp. 46–60. Springer, December 2019. ISBN 978-3-030-34356-9. doi: 10.1007/978-3-030-34356-9_5. URL https://link.springer.com/chapter/10.1007/978-3-030-34356-9_5.
- Besta, M., Iff, P., Scheidl, F., Osawa, K., Dryden, N., Podstawski, M., Chen, T., and Hoeffler, T. Neural Graph Databases. In Rieck, B. and Pascanu, R. (eds.), *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pp. 31:1–31:38, Virtual Event, December 2022. PMLR. URL <https://proceedings.mlr.press/v198/besta22a.html>.
- Besta, M., Catarino, A. C., Gianinazzi, L., Blach, N., Nyczyk, P., Niewiadomski, H., and Hoeffler, T. HOT: Higher-Order Dynamic Graph Representation Learning With Efficient Transformers. In Villar, S. and Chamberlain, B. (eds.), *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pp. 15:1–15:20, Virtual Event, November 2023a. PMLR. URL <https://proceedings.mlr.press/v231/besta24a.html>.
- Besta, M., Gerstenberger, R., Blach, N., Fischer, M., and Hoeffler, T. GDI: A Graph Database Interface Standard. Technical report, ETH Zurich, 2023b. Available at <http>

- [s://spcl.inf.ethz.ch/Research/Parallel_Programming/GDI/](https://spcl.inf.ethz.ch/Research/Parallel_Programming/GDI/).
- Besta, M., Gerstenberger, R., Fischer, M., Podstawski, M., Blach, N., Egel, B., Mitenkov, G., Chlapek, W., Michalewicz, M., Niewiadomski, H., et al. The Graph Database Interface: Scaling Online Transactional and Analytical Graph Workloads to Hundreds of Thousands of Cores. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '23, pp. 22:1–22:18, Denver, CO, USA, November 2023c. Association for Computing Machinery. ISBN 9798400701092. doi: 10.1145/3581784.3607068. URL <https://doi.org/10.1145/3581784.3607068>.
- Besta, M., Gerstenberger, R., Peter, E., Fischer, M., Podstawski, M., Barthels, C., Alonso, G., and Hoefler, T. Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries. *ACM Comput. Surv.*, 56(2), September 2023d. ISSN 0360-0300. doi: 10.1145/3604932. URL <https://doi.org/10.1145/3604932>.
- Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., and Hoefler, T. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, March 2024a. doi: 10.1609/aaai.v38i16.29720. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29720>.
- Besta, M., Gerstenberger, R., Iff, P., Sonawane, P., Luna, J. G., Kanakagiri, R., Min, R., Mutlu, O., Hoefler, T., Appuswamy, R., et al. Hardware Acceleration for Knowledge Graph Processing: Challenges & Recent Developments, November 2024b. URL <https://arxiv.org/abs/2408.12173>. arXiv:2408.12173.
- Besta, M., Kubicek, A., Niggli, R., Gerstenberger, R., Weitzendorf, L., Chi, M., Iff, P., Gajda, J., Nyczyk, P., Müller, J., et al. Multi-Head RAG: Solving Multi-Aspect Problems with LLMs, November 2024c. URL <https://arxiv.org/abs/2406.05085>. arXiv:2406.05085.
- Besta, M., Paleari, L., Kubicek, A., Nyczyk, P., Gerstenberger, R., Iff, P., Lehmann, T., Niewiadomski, H., and Hoefler, T. CheckEmbed: Effective Verification of LLM Solutions to Open-Ended Tasks, June 2024d. URL <https://arxiv.org/abs/2406.02524>. arXiv:2406.02524.
- Besta, M., Scheidl, F., Gianinazzi, L., Kwaśniewski, G., Klaiman, S., Müller, J., and Hoefler, T. Demystifying Higher-Order Graph Neural Networks, December 2024e. URL <https://arxiv.org/abs/2406.12841>. arXiv:2406.12841.
- Besta, M., Barth, J., Schreiber, E., Kubicek, A., Catarino, A., Gerstenberger, R., Nyczyk, P., Iff, P., Li, Y., Houliston, S., et al. Reasoning Language Models: A Blueprint, January 2025a. URL <https://arxiv.org/abs/2501.11223>. arXiv:2501.11223.
- Besta, M., Memedi, F., Zhang, Z., Gerstenberger, R., Piao, G., Blach, N., Nyczyk, P., Copik, M., Kwaśniewski, G., Müller, J., Gianinazzi, L., Kubicek, A., Niewiadomski, H., O'Mahony, A., Mutlu, O., and Hoefler, T. Demystifying Chains, Trees, and Graphs of Thoughts, February 2025b. URL <https://arxiv.org/abs/2401.14295>. arXiv:2401.14295.
- Beurer-Kellner, L., Fischer, M., and Vechev, M. Large Language Models are Zero-Shot Multi-Tool Users. In *Proceedings of the ICML Workshop on Knowledge and Logical Reasoning in the Era of Data-Driven Learning*, KLR '23, Honolulu, HI, USA, July 2023.
- Beurer-Kellner, L., Müller, M. N., Fischer, M., and Vechev, M. Prompt Sketching for Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3674–3706, Vienna, Austria, July 2024. PMLR. URL <https://proceedings.mlr.press/v235/beurer-kellner24b.html>.
- Bhattacharjya, D., Lee, J., Agravante, D. J., Ganesan, B., and Marinescu, R. Foundation Model Sherpas: Guiding Foundation Models through Knowledge and Reasoning, February 2024. URL <https://arxiv.org/abs/2402.01602>. arXiv:2402.01602.
- Blach, N., Besta, M., De Sensi, D., Domke, J., Harake, H., Li, S., Iff, P., Konieczny, M., Lakhotia, K., Kubicek, A., et al. A High-Performance Design, Implementation, Deployment, and Evaluation of the Slim Fly Network. In *Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation*, NSDI '24, pp. 1025–1044, Santa Clara, CA, USA, April 2024. USENIX Association. ISBN 978-1-939133-39-7. URL <https://www.usenix.org/conference/nsdi24/presentation/blach>.
- Chen, G., Dong, S., Shu, Y., Zhang, G., Sesay, J., Karlsson, B. F., Fu, J., and Shi, Y. AutoAgents: A Framework for Automatic Agent Generation. In Larson, K. (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, pp. 22–30, Jeju, South Korea, August 2024. International

- Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2024/3. URL <https://www.ijcai.org/proceedings/2024/3>.
- Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*, 11 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=YfZ4ZPt8zd>.
- Chu, Z., Wang, Y., Zhu, F., Yu, L., Li, L., and Gu, J. Professional Agents – Evolving Large Language Models into Autonomous Experts with Human-Level Competencies, February 2024. URL <https://arxiv.org/abs/2402.03628>. arXiv:2402.03628.
- Creswell, A., Shanahan, M., and Higgins, I. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR '23*, Kigali, Rwanda, May 2023. OpenReview. URL <https://openreview.net/forum?id=3Pf3Wg6o-A4>.
- Delile, J., Mukherjee, S., Pamel, A. V., and Zhukov, L. Graph-Based Retriever Captures the Long Tail of Biomedical Knowledge. In *Proceedings of the Workshop ML for Life and Material Science: From Theory to Industry Applications, ML4LMS '24*, Vienna, Austria, July 2024. URL <https://openreview.net/forum?id=RUwfsPWrv3>.
- Docker Inc. Docker: Accelerated Container Applications. <https://www.docker.com/>, December 2024. accessed 2025-01-27.
- Dua, D., Gupta, S., Singh, S., and Gardner, M. Successive Prompting for Decomposing Complex Questions. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP '22*, pp. 1251–1265, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.81. URL <https://aclanthology.org/2022.emnlp-main.81/>.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, April 2024. URL <https://arxiv.org/abs/2404.16130>. arXiv:2404.16130.
- Emonet, V., Bolleman, J., Duvaud, S., de Farias, T. M., and Sima, A. C. LLM-based SPARQL Query Generation from Natural Language over Federated Knowledge Graphs, February 2025. URL <https://arxiv.org/abs/2410.06062>. arXiv:2410.06062.
- Forethought. AutoChain. <https://autochain.forethought.ai/>, 2023. accessed 2025-01-27.
- Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., and Taylor, A. Cypher: An Evolving Query Language for Property Graphs. In *Proceedings of the International Conference on Management of Data, SIGMOD '18*, pp. 1433–1445, Houston, TX, USA, 2018. Association for Computing Machinery. ISBN 9781450347037. doi: 10.1145/3183713.3190657. URL <https://doi.org/10.1145/3183713.3190657>.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey, March 2024. URL <https://arxiv.org/abs/2312.10997>. arXiv:2312.10997.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. DeepSeek R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL <https://arxiv.org/abs/2501.12948>. arXiv:2501.12948.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. In Larson, K. (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*, pp. 8048–8057, Jeju, South Korea, August 2024. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2024/890. URL <https://www.ijcai.org/proceedings/2024/890>. Survey Track.
- Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. In *Proceedings of the Twelfth International Conference on Learning Representations, ICLR '24*, Vienna, Austria, May 2024. OpenReview. URL <https://openreview.net/forum?id=VtmBAGCN7o>.
- Hu, H., Lu, H., Zhang, H., Lam, W., and Zhang, Y. Chain-of-Symbol Prompting Elicits Planning in Large Language Models, August 2024. URL <https://arxiv.org/abs/2305.10276>. arXiv:2305.10276.
- Hu, Y. and Lu, Y. RAG and RAU: A Survey on Retrieval-Augmented Language Model in Natural Language Processing, April 2024. URL <https://arxiv.org/abs/2404.19543>. arXiv:2404.19543.
- Huang, D., Nan, Z., Hu, X., Jin, P., Peng, S., Wen, Y., Zhang, R., Du, Z., Guo, Q., Pu, Y., and Chen, Y. ANPL:

- Towards Natural Programming with Interactive Decomposition. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 69404–69440, New Orleans, LA, USA, December 2023. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dba8fa689ede9e56cbcd4f719def38fb-Abstract-Conference.html.
- Huang, Y. and Huang, J. A Survey on Retrieval-Augmented Text Generation for Large Language Models, August 2024. URL <https://arxiv.org/abs/2404.10981>. arXiv:2404.10981.
- Jung, J., Qin, L., Welleck, S., Brahman, F., Bhagavatula, C., Le Bras, R., and Choi, Y. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP '22*, pp. 1266–1279, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.82. URL <https://aclanthology.org/2022.emnlp-main.82/>.
- Kagaya, T., Yuan, T. J., Lou, Y., Karlekar, J., Pranata, S., Kinose, A., Oguri, K., Wick, F., and You, Y. RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents. In *Proceedings of the Workshop on Open-World Agents, OWA '24*, Vancouver, Canada, December 2024. OpenReview. URL <https://openreview.net/forum?id=Xf49Dpxuox>.
- Kim, S., Moon, S., Tabrizi, R., Lee, N., Mahoney, M. W., Keutzer, K., and Gholami, A. An LLM Compiler for Parallel Function Calling. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24370–24391, Vienna, Austria, July 2024. PMLR. URL <https://proceedings.mlr.press/v235/kim24y.html>.
- LangChain Inc. LangChain. <https://www.langchain.com/>, 2024. accessed 2025-01-27.
- Li, H., Su, Y., Cai, D., Wang, Y., and Liu, L. A Survey on Retrieval-Augmented Text Generation, February 2022. URL <https://arxiv.org/abs/2202.01110>. arXiv:2202.01110.
- Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More Agents Is All You Need. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=bgzUSZ8aeg>.
- Liu, X., Peng, Z., Yi, X., Xie, X., Xiang, L., Liu, Y., and Xu, D. ToolNet: Connecting Large Language Models with Massive Tools via Tool Graph, February 2024a. URL <https://arxiv.org/abs/2403.00839>. arXiv:2403.00839.
- Liu, Z., Zhang, Y., Li, P., Liu, Y., and Yang, D. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. In *Proceedings of the First Conference on Language Modeling, COLM '24*, Philadelphia, PA, USA, October 2024b. OpenReview. URL <https://openreview.net/forum?id=XII0Wp1XA9>.
- Manathunga, S. S. and Illangasekara, Y. A. Retrieval Augmented Generation and Representative Vector Summarization for Large Unstructured Textual Data in Medical Education, August 2023. URL <https://arxiv.org/abs/2308.00479>. arXiv:2308.00479.
- Mecharnia, T. and d’Aquin, M. Performance and Limitations of Fine-Tuned LLMs in SPARQL Query Generation. In Gesese, G. A., Sack, H., Paulheim, H., Merono-Penuela, A., and Chen, L. (eds.), *Proceedings of the Workshop on Generative AI and Knowledge Graphs, GenAIK '25*, pp. 69–77, Abu Dhabi, United Arab Emirates, January 2025. International Committee on Computational Linguistics. URL <https://aclanthology.org/2025.genaik-1.8/>.
- Mialon, G., Dessi, R., Lomeli, M., Nalmpantis, C., Pansuru, R., Raileanu, R., Roziere, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. Augmented Language Models: A Survey. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=jh7wH2AzKK>. Survey Certification.
- Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. GAIA: A Benchmark for General AI Assistants. In *Proceedings of the Twelfth International Conference on Learning Representations, ICLR '24*, Vienna, Austria, May 2024. OpenReview. URL <https://openreview.net/forum?id=fibxvahvs3>.
- NetworkX Developers. NetworkX Documentation. <https://networkx.org/>, October 2024. accessed 2025-01-27.
- Pérez, J., Arenas, M., and Gutierrez, C. Semantics and Complexity of SPARQL. *ACM Trans. Database Syst.*, 34(3):16:1–16:45, September 2009. ISSN 0362-5915. doi: 10.1145/1567274.1567278. URL <https://doi.org/10.1145/1567274.1567278>.

- Prasad, A., Koller, A., Hartmann, M., Clark, P., Sabharwal, A., Bansal, M., and Khot, T. ADaPT: As-Needed Decomposition and Planning with Language Models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4226–4252, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.264. URL <https://aclanthology.org/2024.findings-naacl.264/>.
- Python Software Foundation. Python Standard Library: codecs — Codec registry and base classes. <https://docs.python.org/3/library/codecs.html>, January 2025a. accessed 2025-01-27.
- Python Software Foundation. Python Standard Library: asyncio — Asynchronous I/O. <https://docs.python.org/3/library/asyncio.html>, January 2025b. accessed 2025-01-29.
- Qian, C., Xie, Z., Wang, Y., Liu, W., Zhu, K., Xia, H., Dang, Y., Du, Z., Chen, W., Yang, C., Liu, Z., and Sun, M. Scaling Large Language Model-Based Multi-Agent Collaboration. In *Proceedings of the Thirteenth International Conference on Learning Representations, ICLR '25*, Singapore, April 2025. OpenReview. URL <https://openreview.net/forum?id=K3n5jPkrU6>.
- Robinson, I., Webber, J., and Eifrem, E. Graph Database Internals. In *Graph Databases*, chapter 7, pp. 149–170. O’Reilly, 2nd edition, 2015. ISBN 9781491930892.
- Roucher, A. and Petrov, S. Beating GAIA with Transformers Agents. <https://github.com/aymeric-roucher/GAIA>, October 2024. accessed 2025-01-29.
- Rush, A. MiniChain: A Small Library for Coding with Large Language Models. In Feng, Y. and Lefever, E. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP '23*, pp. 311–317, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.27. URL <https://aclanthology.org/2023.emnlp-demo.27>.
- Sarathi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *Proceedings of the Twelfth International Conference on Learning Representations, ICLR '24*, Vienna, Austria, May 2024. OpenReview. URL <https://openreview.net/forum?id=GN921JHCRw>.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 68539–68551, New Orleans, LA, USA, December 2023. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.
- SerpApi LLM. SerpApi: Google Search API. <https://serpapi.com/>, 2025. accessed 2025-01-27.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 38154–38180, New Orleans, LA, USA, December 2023. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/77c33e6a367922d003ff102ffb92b658-Abstract-Conference.html.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 8634–8652, New Orleans, LA, USA, December 2023. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.
- Significant Gravitas. AutoGPT. <https://github.com/Significant-Gravitas/AutoGPT>, January 2025. accessed 2025-01-27.
- Singhal, A. Introducing the Knowledge Graph: things, not strings. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>, May 2012. accessed 2025-02-04.
- Stengel-Eskin, E., Prasad, A., and Bansal, M. ReGAL: Refactoring Programs to Discover Generalizable Abstractions. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, volume 235 of

- Proceedings of Machine Learning Research*, pp. 46605–46624, Vienna, Austria, July 2024. PMLR. URL <https://proceedings.mlr.press/v235/stengel-eskin24a.html>.
- Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*, February 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=1i6ZCvflQJ>. Survey Certification.
- Tang, X., Kim, K., Song, Y., Lothritz, C., Li, B., Ezzini, S., Tian, H., Klein, J., and Bissyandé, T. F. CodeAgent: Autonomous Communicative Agents for Code Review. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP '24*, pp. 11279–11313, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.632. URL <https://aclanthology.org/2024.emnlp-main.632/>.
- Tenacity Developers. Tenacity: Retrying Library. <https://github.com/jd/tenacity>, July 2024. accessed 2025-01-27.
- Wang, S., Liu, C., Zheng, Z., Qi, S., Chen, S., Yang, Q., Zhao, A., Wang, C., Song, S., and Huang, G. Avalon’s Game of Thoughts: Battle Against Deception through Recursive Contemplation, October 2023a. URL <https://arxiv.org/abs/2310.01320>. arXiv:2310.01320.
- Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR '23*, Kigali, Rwanda, May 2023b. OpenReview. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X. S., and Liang, Y. Describe, Explain, Plan and Select: Interactive Planning with LLMs Enables Open-World Multi-Task Agents. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 34153–34189, New Orleans, LA, USA, December 2023c. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6b8dfb8c0c12e6fafc6c256cb08a5ca7-A-abstract-Conference.html.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS '22)*, volume 35 of *Advances in Neural Information Processing Systems*, pp. 24824–24837, New Orleans, LA, USA, December 2022. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-A-abstract-Conference.html.
- Wewer, C., Lemmerich, F., and Cochez, M. Updating Embeddings for Dynamic Knowledge Graphs, September 2021. URL <https://arxiv.org/abs/2109.10896>. arXiv:2109.10896.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *Proceedings of the First Conference on Language Modeling, COLM '24*, Philadelphia, PA, USA, October 2024. OpenReview. URL <https://openreview.net/forum?id=BAakYlhNKS>.
- Xie, T., Zhou, F., Cheng, Z., Shi, P., Weng, L., Liu, Y., Hua, T. J., Zhao, J., Liu, Q., Liu, C., Liu, Z., Xu, Y., Su, H., Shin, D., Xiong, C., and Yu, T. OpenAgents: An Open Platform for Language Agents in the Wild. In *Proceedings of the First Conference on Language Modeling, COLM '24*, Philadelphia, PA, USA, October 2024. OpenReview. URL <https://openreview.net/forum?id=sKATR201Y0>.
- Xu, Z., Liu, Z., Yan, Y., Wang, S., Yu, S., Zeng, Z., Xiao, C., Liu, Z., Yu, G., and Xiong, C. ActiveRAG: Autonomously Knowledge Assimilation and Accommodation through Retrieval-Augmented Agents, October 2024. URL <https://arxiv.org/abs/2402.13547>. arXiv:2402.13547.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS '23)*, volume 36 of *Advances in Neural Information Processing Systems*, pp. 11809–11822, New Orleans, LA, USA, December 2023a. Curran Associates. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922.

- b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR '23*, Kigali, Rwanda, May 2023b. OpenReview. URL https://openreview.net/forum?id=WE_vluYUL-X.
- Ye, Y., Hui, B., Yang, M., Li, B., Huang, F., and Li, Y. Large Language Models are Versatile Decomposers: Decomposing Evidence and Questions for Table-based Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, pp. 174–184, Taipei, Taiwan, July 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591708. URL <https://doi.org/10.1145/3539618.3591708>.
- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., and Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. In Zhu, W., Xiong, H., Cheng, X., Cui, L., Dou, Z., Dong, J., Pang, S., Wang, L., Kong, L., and Chen, Z. (eds.), *Proceedings of the 12th CCF Conference, BigData*, volume 2301 of *Communications in Computer and Information Science*, pp. 102–120, Qingdao, China, August 2024a. Springer. ISBN 978-981-96-1024-2. doi: 10.1007/978-981-96-1024-2_8. URL https://link.springer.com/chapter/10.1007/978-981-96-1024-2_8.
- Yu, W., Zhang, H., Pan, X., Cao, P., Ma, K., Li, J., Wang, H., and Yu, D. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP '24*, pp. 14672–14685, Miami, FL, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.813. URL <https://aclanthology.org/2024.emnlp-main.813/>.
- Zeng, H., Yue, Z., Jiang, Q., and Wang, D. Federated Recommendation via Hybrid Retrieval Augmented Generation. In Ding, W., Lu, C.-T., Wang, F., Di, L., Wu, K., Huan, J., Nambiar, R., Li, J., Ilievski, F., Baeza-Yates, R., and Hu, X. (eds.), *Proceedings of the IEEE International Conference on Big Data, BigData '24*, pp. 8078–8087, Washington DC, USA, December 2024. doi: 10.1109/BigData62323.2024.10825302. URL <https://ieeexplore.ieee.org/document/10825302>.
- Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J. X., and Chen, T. Cut the Crap: An Economical Communication Pipeline for LLM-based Multi-Agent Systems, October 2024. URL <https://arxiv.org/abs/2410.02506>. arXiv:2410.02506.
- Zhao, A., Huang, D., Xu, Q., Lin, M., Liu, Y.-J., and Huang, G. ExpeL: LLM Agents Are Experiential Learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, March 2024a. doi: 10.1609/aaai.v38i17.29936. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29936>.
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., and Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey, June 2024b. URL <https://arxiv.org/abs/2402.19473>. arXiv:2402.19473.
- Zhu, Y., Qiao, S., Ou, Y., Deng, S., Zhang, N., Lyu, S., Shen, Y., Liang, L., Gu, J., and Chen, H. KnowAgent: Knowledge-Augmented Planning for LLM-Based Agents, March 2024a. URL <https://arxiv.org/abs/2403.03101>. arXiv:2403.03101.
- Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D., and Dai, H. Large Language Models Can Learn Rules, December 2024b. URL <https://arxiv.org/abs/2310.07064>. arXiv:2310.07064.
- Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., and Schmidhuber, J. GPTSwarm: Language Agents as Optimizable Graphs. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767, Vienna, Austria, July 2024. PMLR. URL <https://proceedings.mlr.press/v235/zhuge24a.html>.