KG-MASD:Knowledge Graph-guided Multi-agent System Distillation

Anonymous ACL submission

Abstract

Recent research has focused on optimizing multi-agent LLMs for complex reasoning tasks, revealing that such architectures can significantly enhance reasoning abilities. Nevertheless, there are critical challenges, such as the uncontrollable hallucinations caused by the multi-model and multi-round iteration mechanism. The current paradigm also fails to effectively distill the collaborative reasoning power of distributed multi-agent systems into a single deployable model, which limits its reasoning efficiency in practical application scenarios. To address these issues, we propose a solution that combines knowledge graphs with multi-agent systems. Focusing on industrial-field intelligent QA systems, we design a **Knowledge Graph-guided Multi-Agent System Distillation(KG-MASD)** . The framework makes three main contributions. First, it constructs an industrial field knowledge graph to provide prior information. Second, it establishes a collaborative reasoning mechanism for a multi-teacher model. Third, it develops a multi-agent distillation methodology. To verify its effectiveness, this study introduces the first standard industrial production instruction dataset. It comprises approximately 52k domain-specific question-and-answer pairs and an industrial knowledge graph encompassing around 36k entities and 131k relationships. Experimental results indicate that the KG-MASD framework may offer a potential domain adaptation advantage over existing single-model and multi-agent distillation frameworks, with a possible improvement ranging from 2.4% and 20.1% in domain adaptation.

1 Introduction

Knowledge distillation offers a viable solution for using large language models (LLMs) in scenarios with limited production resources and can effectively inject domain-specific datasets to fine-tune LLMs for particular domains. However, when domain problems are complex, a single large model serving as a teacher model struggles to effectively transfer the label distribution to the student model via data augmentation methods (Wang et al., 2021; Liu et al., 2022; Yao et al., 2023a). Multi-agent systems can address this issue by increasing the depth of reasoning of the teacher model. However, they may suffer from uncontrollable structure and low iteration efficiency, leading to the failure of data augmentation and low credibility of the results. The integration of knowledge graphs and LLMs has emerged as a new and reliable paradigm to address these challenges (Li et al., 2023; Deng et al., 2023; Zhang et al., 2024).

Based on this, to tackle these challenges, we propose the **Knowledge Graph-guided Multi-Agent System Distillation(KG-MASD)** framework, a method that combines multi-agent systems and local knowledge graphs for knowledge distillation. For example, this framework addresses the problem scenarios illustrated in Figure 1. The KG-MASD framework primarily focuses on the key issues in knowledge distillation: how to ensure that the multi-agent system can assist the teacher model in knowledge distillation, and how to guarantee the reliability of augmented data under a multi-agent system. Our contributions are as follows.

- We introduce a multi-agent system to assist knowledge distillation, efficiently extracting and integrating industrial knowledge. By integrating local knowledge graphs into the distilled instruction dataset, KG-MASD can effectively distill the student LLM.
- We introduce an industrial QA dataset with vertical annotations and the corresponding knowledge graph instruction set, which enriches practical application scenarios.
- Experimental results show that the KG-MASD framework outperforms single-teacher

LLMs and other multi-agent distillation methods.

2 Related Work

2.1 Knowledge Distillation

The explosive growth of parameters in LLMs has led to prohibitively high computational costs during inference, creating a significant bottleneck for their widespread application. Knowledge distillation, which transfers knowledge from LLMs to lightweight models through a teacher-student architecture, has emerged as a key technology to overcome this limitation (Hu et al., 2024; Du et al., 2024; Zhao et al., 2024).

Compared to traditional methods, knowledge distillation of LLMs pays more attention to extracting knowledge from core components such as the multi-head attention mechanism and intermediate layer hidden states in the Transformer architecture. Knowledge distillation enables student models to learn the semantic focus patterns of the teacher by aligning attention mechanisms, constructing prompt templates, and designing reinforcement learning rewards. It also helps unify the semantic space and dynamically optimize outputs. This effectively improves the learning efficiency and performance of the models (Yang and Liu, 2024; Liu et al., 2024b; Yang et al., 2024). Despite its significant achievements, knowledge distillation of LLMs still faces challenges. The complex structure of knowledge makes it difficult to extract and transfer key knowledge, and model compression can easily lead to performance loss. Therefore, it is necessary to balance performance and compression ratios. Additionally, the contradiction between the need for large-scale data training and limited computational resources also urgently requires more efficient distillation algorithms to be resolved.

2.2 Multi-agent Systems

As an important branch of distributed artificial intelligence, Multi-agent systems (MAS) have shown strong application potential in many fields in recent years. In complex task scenarios, multiple agents can solve problems more efficiently through interactions such as collaboration and competition. Their flexibility and adaptability far exceed those of single-agent models (Liang et al., 2023; Wang et al., 2025; Kang et al., 2024).

With the deepening of research, a series of advanced MAS frameworks have emerged successively. Multi-agent Debate (MAD) promotes deeper understanding and more accurate answer generation by having multiple independent agents debate on the same topic (Smit et al., 2023). The Self-Reflect framework consists of three models: participants, evaluators, and self-reflection. Each agent can adjust its reasoning strategy through self-reflection to improve overall task performance (Yuan et al., 2025; Shinn et al., 2023). When agents process large amounts of data, Rerank optimizes the information retrieval process by removing irrelevant nodes and reordering relevant nodes (Loem et al., 2023; He et al., 2024). MAS Protocol (MAPS) designs communication protocols, collaboration strategies, and resource allocation methods among agents to achieve optimal performance in Multi-agent systems within specific task domains (He et al., 2024; Thelasingha et al., 2025). Self-Consistency establishes a consistency verification mechanism. When an agent makes a decision or takes an action, the framework checks whether the decision or action is consistent with the overall system goal and the states of other agents to avoid conflicts or inconsistencies (Li et al., 2024; Wang et al., 2022).

In the field of knowledge distillation, although a

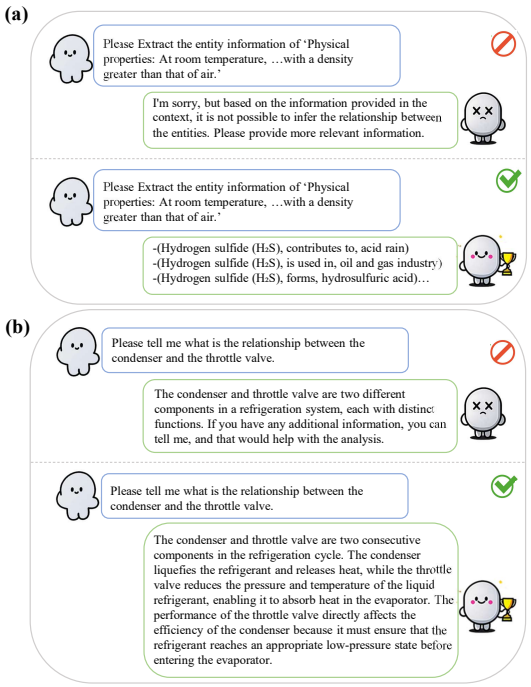

Figure 1: (a) represents the issues and expectations of the edgeside model when tasked with knowledge graph extraction, while (b) illustrates the problems and anticipated outcomes of the edge-side model in the context of question-and-answer formats.

mature MAS application system has not yet been formed, some explorations have been made. Some studies have attempted to use multiple agents as auxiliary roles for both teacher and student models. Agents assist the teacher model in more accurately extracting key knowledge and help the student model more efficiently receive and understand the content of knowledge transfer (Bo et al., 2024).

2.3 Industrial Question-Answering Knowledge

Natural language processing technology has become the core driving force behind industrial question-answering systems. Through the construction of knowledge graphs, fusion of multimodal data, and optimization of large language models, it has promoted intelligent interaction and decision support in industrial scenarios (Jiao et al., 2025; Kojima et al., 2022).

Large language models are used for industrial equipment fault diagnosis. For example, when a stamping machine on an automobile production line malfunctions, an LLM can analyze logs and maintenance records and quickly locate the fault by combining the experience database, thereby reducing downtime (Meng et al., 2024). Knowledge distillation enables the lightweight deployment of industrial question-answering models by training lightweight models for edge devices, balancing accuracy and computational resources (Alam et al., 2020). Wang Peng et al. proposed a multimodal large language model for the transportation field called TransGPT (Wang et al., 2024). It is fine-tuned based on transportation text/multimodal data and provides NLP technical support for intelligent transportation systems.

3 Dataset and Augmentation

3.1 Dataset Compilation

To create a basic knowledge graph and instruction-tuning dataset tailored for the industrial domain, we crawled question-answer data provided by human experts from publicly available online publications, as well as a large amount of unsupervised text obtained from public sources. The question-answer data from human experts were first manually filtered and organized into an instruction-tuning data format. Based on the data collection channels, these topics can be directly categorized into eight distinct thematic groups: Transportation, Health, Environment, Equipment, Production, Electricity, Disaster Prevention and Other. The distribution is shown in Appendix A.1.

For the unsupervised information data obtained, we first utilized Sentence-BERT to embed the original text sentences into vectors and calculated the cosine similarity between different sample vectors. If the similarity was greater than 0.5, the samples were classified as different segments; otherwise, they were considered as the same segment.We subsequently employed prompts derived from large language models (Cui et al., 2025; Sahoo et al., 2024; Trivedi et al., 2025) (as shown in Appendix A.2 for details) to transform the segmented information into an instruction-tuning format. The statistical information of the data collected by the two methods is shown in Appendix A.3.

3.2 Dataset Augmentation Directions

The process of transforming knowledge graph-extracted information into instruction-tuning data involves two primary augmentation directions: Relation Triple Extraction (RTE) and Knowledge Graph Completion (KGC). Specifically, we guide the Multi-agent system to explicitly express knowledge graph-related content before data output, which serves as a condition for autoregressive generation of instruction-tuning commands.

- Relation Triple Extraction: This task involves automatically extracting semantic relationships between entities from text descriptions and representing them in the form of triples. For example, given the description "Hydrogen sulfide is a colorless gas", the system should extract the triple: "Subject": "Hydrogen sulfide", "Predicate": "i", "Object": "colorless gas".

- Knowledge Graph Completion: The goal of KGC is to infer and fill in missing relationships in an existing knowledge graph. This task typically relies on contextual information and learns from existing relationships to predict missing entity relationships. For example, given the entity "insulation resistance meter" and its "purpose" relationship, the model should infer the corresponding action entity and associated context.

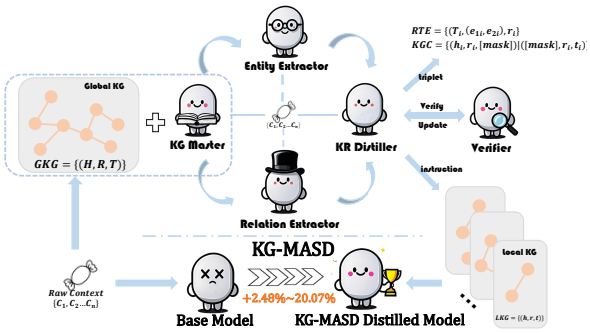

Figure 2: It illustrates the overall process of extracting the Global Knowledge Graph (GKG) from raw data using GraphRAG technology, and conducting entity and relation extraction, local knowledge graph generation, and instruction fine-tuning via a Multi-agent system.

4 Framework and Methodology

4.1 Overall Process of the KG-MASD Framework

The overall process of the KG-MASD framework is depicted in Figure 2. To begin with, we employ GraphRAG technology (Han et al., 2025) to extract a global knowledge graph $\{H, R, T\}$ for industrial decision-making from the raw data $\{C_1, C_2, \ldots, C_n\}$. In this context, C_n signifies the n-th information fragment, while H and T correspond to the head and tail entities within the knowledge graph, and R indicates the connecting relationships.

Subsequently, we define identity instructions for the five agents in the Multi-agent assisted distillation system, each focusing on different aspects. These agents include the Knowledge Graph Master (KG Master), Entity Extractor, Relation Extractor, Knowledge Relation Distiller (KR Distiller), and Verifier. The KG Master is responsible for decomposing the original question and expanding knowledge relationships using the global knowledge graph and RAG technology. Then, one agent is randomly selected from the Entity Extractor and Relation Extractor to provide raw information and extract entities and potential relationships from the data stream. These extracted entities and relationships are subsequently aggregated into a local knowledge graph in the KR Distiller, which combines the raw information to extract small-scale local triples (h_i, r_i, t_i) (following the knowledge graph generation paradigm). The Verifier reads and judges whether the extracted information is correct and valid. If the information fails the verification, it is sent back to the KR Distiller to update and iteratively optimize the extraction path. Once the update converges, the KR Distiller generates small local knowledge graphs and an instruction-tuning dataset $I = \{(I_1, I_2, \ldots, I_n) \mid (h, r, t)\}$ in an autoregressive manner. Finally, we apply LoRA (Hu et al., 2021) to distill and fine-tune the student model, transferring the reasoning capabilities of the global knowledge graph and the teacher model to the smaller student model.

4.2 Multi-agent System Design

4.2.1 Definition of the Agent System

Considering the responsiveness and tool diversity required for industrial task deployment, we define the roles of the five agents in the MAS-assisted distillation system using identity and chain-of-thought methods. Each agent is assigned specific roles and responsibilities to enhance focus and specialization. For example, the KG Master is responsible for database read operations, while the KR Distiller is equipped with short-term memory to achieve contextual understanding. The specific role prompt words for these agents are shown in Appendix A.4.

To verify the structural controllability of the KG-MASD system (Lin, 1974), we construct the system's Laplacian matrix L and evaluate the system's controllability by analyzing the spectral properties of L based on the concept of structural controllability proposed by Lin. Specifically, we describe the KG-MASD system as the following linear time-invariant model:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

where $x(t) \in R^5$ is the state vector, $A = -L - BK \in R^{5\times5}$ is the system matrix, $B \in R^{5\times1}$ is the control input matrix, and $u(t) \in R^1$ represents the control input. According to the research on structural controllability of Multi-agent systems by Zamani and Lin(Zamani and Lin, 2009), we compute the controllability matrix Q_c and check whether its rank equals the system dimension to determine the system's complete controllability. The controllability matrix Q_c is expressed as:

$$Q_c = [B\ AB\ A^2B\ \ldots\ A^{n-1}B] \tag{2}$$

The conclusion indicates that the KG-MASD system has good structural controllability, which mainly depends on the eigenvalue distribution of the Laplacian matrix L (e.g., eigenvalues of 0, 1, 2, $\pm\sqrt{2}$), thereby ensuring the non-singularity of the system matrix A and the full-rank condition of the

controllability matrix Q_c. Furthermore, referring to the research on structural controllability of directed signed networks by Ong et al (Guan et al., 2021), we further explore the impact of different network topologies on system controllability, thereby enhancing the system's adaptability and robustness. Subsequently, we introduce a decentralized control strategy (specifically, different agents can access the raw data stream), enabling each agent to self-regulate based on local information.

Additionally, through theoretical analysis and numerical simulations, we verify the effectiveness of the proposed KG-MASD framework and demonstrate the system's stability and controllability under different initial conditions and network topologies. As shown in Appendix A.5, as the number of iterations increases, the KG-MASD system can rapidly achieve self-regulation, and the performance of the Verifier also tends to stabilize.

4.2.2 Global Knowledge Graph Generation

In the global knowledge graph generation phase, our objective is to extract entities and relationships with semantic associations from a vast amount of raw data to construct a comprehensive knowledge representation framework. Specifically, we utilize GraphRAG technology to extract the global knowledge graph (GKG) from the raw dataset $\{C_1, C_2, \ldots, C_n\}$, with its structure represented as $\{(H, R, T)\}$, where H, R, and T denote the head entities, relationships, and tail entities, respectively. Through this process, we are not only able to capture explicit relationships in the data but also leverage the scalability of knowledge graphs to discover potential semantic connections. We provide a visual representation of the GKG in Appendix A.6.

To further enhance the completeness and accuracy of the knowledge graph, we introduce the KG Master module, which is based on the global knowledge graph and incorporates a retrieval-augmented generation mechanism in the Multi-agent system. This module can decompose the original question and expand knowledge relationships, providing a richer semantic context for subsequent entity and relationship extraction.

4.2.3 Construction of Local Knowledge Graphs

Further, based on the well-constructed GKG and Graph RAG technology, we perform retrieval augmentation on different text fragments to enhance the quality of answers and strengthen the connections between data and nodes in the latent space as much as possible. Meanwhile, local retrieval can be quickly conducted from the original labeled dataset (e.g., the production department in Appendix A.7), and they can be combined to form local heterogeneous knowledge graphs. The generation process is illustrated in Appendix A.7.

4.2.4 Self-Verification and Self-Update

Based on the well-constructed GKG, we process the relation extraction and integration of local knowledge graphs using the defined distiller module. As shown in Appendix A.8, the Verifier iteratively verifies and refines the information until it is judged to be reliable true information. This process generates a verified instruction dataset $\{(I_1, I_2, \ldots, I_n) \mid (h, r, t)\}$

4.2.5 Knowledge-Based Instruction Tuning

To improve tuning efficiency and reduce computational resource consumption, the KG-MASD framework employs Low-Rank Adaptation (LoRA) technology for efficient fine-tuning of large models (Hu et al., 2021; Touvron et al., 2023). During the model fine-tuning phase, we utilize the heterogeneous knowledge instruction set $I = \{(I_1, I_2, \ldots, I_n) \mid (h, r, t)\}$ to perform knowledge-based instruction tuning on the student model using LoRA. Specifically, the optimization process aims to minimize the loss of the language model M, as shown below:

$$L = E_{(x,y)\sim D_{RTE}} \left[\log \left(P_M(y \mid x) P_{LKG}\right)\right]$$

$$+ \mathrm{E}_{(x,y)\sim D_{KGC}} \left[\log \left(P_M(y \mid x) P_{LKG}\right)\right] \quad (3)$$

where x and y represent the instruction input and output in the trajectory, respectively.

5 Experiment

5.1 Experimental Setup

5.1.1 Dataset

As described above, our experiments are conducted on the domain-specific datasets we constructed. The first dataset, built by human experts, contains question-answer data from specialized fields, with thematic labels explicitly specified for the instructions. In contrast, the synthetic data generated by LLMs are derived from a large amount of unlabeled open-world text. Additionally, we conduct experiments by mixing the synthetic data generated by LLMs with the data extracted from human experts.

5.1.2 Baselines

MAS-assisted Knowledge Distillation Baselines:

MAD (Multi-agent Debate): Multiple agents express opinions in a debate format, managed by a referee to reach a final solution. This method encourages divergent thinking, corrects errors, supplements insufficient robustness, and obtains external feedback.

Self-Reflect: Uses self-reflection to iteratively improve answers through feedback. However, it suffers from "thought degeneration", where the model struggles to generate new ideas once it becomes confident in its answers. Despite this, it is suitable for model self-optimization tasks.

MAPS: Employs a Multi-agent path search method, mimicking the human process of summarization and refinement. It first analyzes and then refines through multi-step reasoning, suitable for tasks requiring complex decision-making and problem-solving.

Self-Consistency: Generates multiple outputs and determines the final answer through majority voting. This method effectively reduces randomness and inconsistency in model outputs, enhancing stability and accuracy.

Single LLM as Teacher Model Knowledge Distillation Baselines:

Vanilla Fine-tuning (Yao et al., 2023b): Directly fine-tunes the edge-side LLM using question-answer pairs constructed from the self-instruction dataset, then prompts the LLM with basic task definitions without providing examples.

Step-by-Step Distillation (Hsieh et al., 2023): Compared with several widely adopted advanced methods to demonstrate the validity of our results.

Gradient-Free Learning Methods: Instruction prompting techniques, including context learning and zero-shot inference, are also compared to showcase the experimental results.

5.1.3 Implementation and Detailed Settings

In the experiments, we select DeepSeek-V2 (Liu et al., 2024a) as the backbone LLM in the Multi-agent system, and Qwen2-7B (Yang et al., 2024) and LLama3.1-8B (Deroy and Maity, 2024) as the tuning models. All experiments are conducted on two NVIDIA 3090 GPUs. To precisely evaluate the models' performance on standard open-world datasets (Papineni et al., 2002), we introduce BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L metrics to provide data support for optimizing edge-side inference capabilities.

5.2 Experimental Results

5.2.1 Comparative experiments of MAS-assisted distillation

Table 1 shows the experimental results of different methods on the LLama3.1-8B and Qwen2-7B models. The results indicate that KG-MASD outperforms other MAS-assisted distillation methods in all evaluation metrics. Specifically, KG-MASD achieves BLEU-4(Papineni et al., 2002), ROUGE-1, ROUGE-2, and ROUGE-L(Lin and Hovy, 2003) scores of 66.812, 65.539, 51.573, and 49.524 on the LLama3.1-8B model, significantly higher than MAD, Self-Reflect, MAPS, and Self-Consistency methods. This demonstrates that KG-MASD can more effectively utilize the Multi-agent system for knowledge distillation, reducing model hallucination and enhancing student model performance. On the Qwen2-7B model, KG-MASD also performs remarkably well, with BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L scores of 68.148, 66.855, 52.605, and 50.474, further proving the effectiveness of the KG-MASD framework.

5.2.2 Comparative experiments of single LLM distillation

Table 1 shows the experimental results of different methods on the LLama3.1-8B and Qwen2-7B models. The results indicate that KG-MASD outperforms other single LLM distillation methods in all evaluation metrics. Specifically, KG-MASD achieves BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L scores of 66.812, 65.539, 51.573, and 49.524 on the LLama3.1-8B model, significantly higher than Step-by-Step, Vanilla Fine-tuning, In-Context Learning, and Zero-shot Reasoning methods. This demonstrates that KG-MASD can more effectively utilize knowledge graph signals for distillation, reducing model hallucination and enhancing student model performance. On the Qwen2-7B model, KG-MASD also performs remarkably well, with BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L scores of 68.148, 66.855, 52.605, and 50.474, further proving the effectiveness of the KG-MASD framework.

The experimental results demonstrate that the KG-MASD framework excels in both MAS-assisted distillation and single LLM distillation comparisons. KG-MASD effectively leverages domain knowledge graph signals to reduce model hallucination and enhance student model performance. The superior performance in BLEU-4, ROUGE-

Base Model	Method	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
		Multi-Model			
LLama3.1-8B	Self-Reflect	63.764	63.048	48.789	47.712
	MAPS	58.915$\downarrow^{-4.849}$	55.337$\downarrow^{-7.711}$	40.765$\downarrow^{-8.024}$	39.401$\downarrow^{-8.311}$
	Self-Consistency	58.630$\downarrow^{-5.134}$	55.549$\downarrow^{-7.499}$	41.165$\downarrow^{-7.624}$	39.736$\downarrow^{-7.976}$
	MAD	65.401$\uparrow^{+1.637}$	65.042$\uparrow^{+1.994}$	49.821$\uparrow^{+1.032}$	48.436$\uparrow^{+0.724}$
	KG-MASD	**66.812**$\uparrow^{+3.048}$	**65.539**$\uparrow^{+2.491}$	**51.573**$\uparrow^{+2.784}$	**49.524**$\uparrow^{+1.812}$
Qwen2-7B	Self-Reflect	64.762	63.974	49.572	48.489
	MAPS	60.009$\downarrow^{-4.753}$	56.239$\downarrow^{-7.735}$	41.301$\downarrow^{-8.271}$	40.085$\downarrow^{-8.404}$
	Self-Consistency	59.563$\downarrow^{-5.199}$	56.485$\downarrow^{-7.489}$	41.803$\downarrow^{-7.769}$	40.421$\downarrow^{-8.068}$
	MAD	66.437$\uparrow^{+1.675}$	66.098$\uparrow^{+2.124}$	50.523$\uparrow^{+0.951}$	49.250$\uparrow^{+0.761}$
	KG-MASD	**68.148**$\uparrow^{+3.386}$	**66.855**$\uparrow^{+2.881}$	**52.605**$\uparrow^{+3.033}$	**50.474**$\uparrow^{+1.985}$
		Single-Model			
LLama3.1-8B	Vanilla Fine-tuning	56.331	53.406	39.296	38.064
	In-Context Learning	47.804$\downarrow^{-8.527}$	44.955$\downarrow^{-8.451}$	33.439$\downarrow^{-5.857}$	32.625$\downarrow^{-5.439}$
	Zero-shot Reasoning	32.031$\downarrow^{-24.300}$	30.442$\downarrow^{-22.964}$	22.303$\downarrow^{-16.993}$	21.994$\downarrow^{-16.07}$
	Step-by-Step	60.739$\uparrow^{+4.408}$	57.376$\uparrow^{+3.970}$	43.299$\uparrow^{+4.003}$	41.364$\uparrow^{+3.300}$
	KG-MASD	**66.812**$\uparrow^{+10.481}$	**65.539**$\uparrow^{+12.133}$	**51.573**$\uparrow^{+12.277}$	**49.524**$\uparrow^{+11.460}$
Qwen2-7B	Vanilla Fine-tuning	57.458	54.477	40.092	38.806
	In-Context Learning	48.760$\downarrow^{-8.698}$	45.854$\downarrow^{-8.623}$	34.108$\downarrow^{-5.984}$	33.277$\downarrow^{-5.529}$
	Zero-shot Reasoning	32.671$\downarrow^{-24.787}$	31.045$\downarrow^{-23.432}$	22.749$\downarrow^{-17.343}$	22.434$\downarrow^{-16.372}$
	Step-by-Step	62.353$\uparrow^{+4.895}$	58.623$\uparrow^{+4.146}$	44.198$\uparrow^{+4.106}$	42.191$\uparrow^{+3.385}$
	KG-MASD	**68.148**$\uparrow^{+10.69}$	**66.855**$\uparrow^{+12.378}$	**52.605**$\uparrow^{+12.513}$	**50.474**$\uparrow^{+11.668}$

Table 1: Performance comparison of different distillation methods and models across various datasets

1, ROUGE-2, and ROUGE-L metrics proves its superiority in industrial question-answering tasks.

6 Module Analysis

The following analyses are all based on experiments with Qwen2-7B.

6.1 Ablation Study

To verify the roles of different modules in the KG-MASD framework, we conducted ablation studies. The results are shown in Table 2. It can be seen from Table 2 that both the Global Knowledge Graph (GlobalKG) and Local Knowledge Graph (LocalKG) outperform the complete KG-MASD framework in terms of performance. This may be because the removal of certain modules in the ablation experiments reduces the complexity of the model, thereby resulting in better performance on some metrics. However, the overall design of the KG-MASD framework aims to integrate the advantages of both global and local knowledge graphs, enhancing the credibility and completeness of data through the collaboration of the Multi-agent system. Although slightly lower than the ablation versions on some metrics, KG-MASD remains optimal in terms of overall performance, especially in reducing model hallucination and improving the accuracy of question-answering.

Method	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
GlobalKG	64.607	63.436	51.707	47.857
LocalKG	64.129	62.001	49.462	47.583
KG-MASD	**62.858**	**61.596**	**48.544**	**46.628**

Table 2: Statistical information on small heterogeneous knowledge graph generation under multiple methods comparison.

6.2 Credibility Analysis of Enhancing Data

To further demonstrate the effectiveness of the KG-MASD framework in extracting local knowledge graphs, we designed comparative experiments of two conventional extraction methods on knowledge graph evaluation tasks. These tasks include Relation Triple Extraction (RTE) and Knowledge Graph Completion (KGC), aimed at assessing the framework's capability in generating credible data. Our evaluation criteria are based on human evaluation and GPT evaluation (I and B, 2025; Min et al., 2023).

Figure 3 illustrates the performance comparison of different methods on RTE and KGC tasks, where "human evaluation accuracy scores" and "GPT evaluation accuracy scores" are represented by line graphs and bar charts, respectively, with inconsis-

tent left and right y-axes to more intuitively display the differences in results.

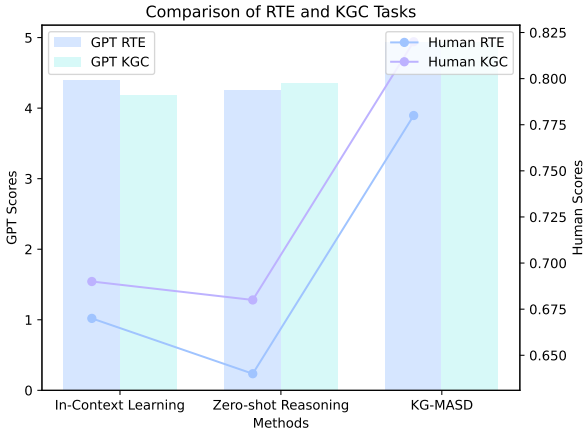

Figure 3: Performance Comparison of Different Methods in RTE (Relation Triple Extraction) and KGC (Knowledge Graph Completion) Tasks by GPT and Human Evaluations

KG-MASD-KGC generated 2,468 relation triples with 1,160 unique relations and 3,686 unique entities, while KG-MASD-RTE produced 2,454 relation triples with 1,148 unique relations and 3,694 unique entities. Other methods generated fewer of these elements, indicating KG-MASD's superior knowledge extraction and integration capabilities.

As shown in Section 6.2, KG-MASD not only enhances data credibility but also enriches the knowledge graph's content. Compared to similar algorithms, it more effectively achieves its design functionality, improving both credibility and completeness. This dual capability is vital for developing robust and comprehensive knowledge bases to support diverse AI applications.

6.3 Completeness Analysis of Enhancing Data

To validate the advantages of the KG-MASD framework in enhancing data completeness, we compared the performance of various methods in knowledge graph construction. Figure 4 presents the statistical results of different types of methods regarding the total number of relation triples, unique relations, and unique entities. As shown in Figure 4, KG-MASD extracts more comprehensive and extensive information in both enhancement modes (KGC and RTE). Specifically, KG-MASD-KGC generated 2468 relation triples, encompassing 1160 unique relations and 3686 unique entities; KG-MASD-RTE produced 2454 relation triples, including 1148 unique relations and 3694 unique

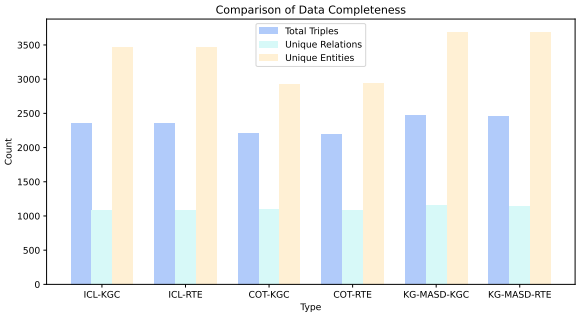

Figure 4: Comparison of Data Completeness Among Different Methods in Knowledge Graph Construction Including Total Triples, Unique Relations, and Unique Entities

entities. In comparison, other methods yielded fewer relation triples, unique relations, and unique entities than KG-MASD. This indicates that KG-MASD is more effective at extracting and integrating knowledge from data, thereby enriching the content of the knowledge graph.

As per Section 6.2, KG-MASD not only boosts data credibility but also enriches the knowledge graph's content. It surpasses similar algorithms in achieving its design goals, enhancing both credibility and completeness by generating credible data and expanding entities/relations. This dual capability is key for building robust and comprehensive AI knowledge bases.

7 Conclusion

This study introduces the KG-MASD framework, which integrates Multi-agent Systems (MAS) and domain-specific knowledge graphs to enhance inference efficiency and accuracy for edge-side models in industrial domains. KG-MASD outperforms traditional methods such as step-by-step distillation, vanilla fine-tuning, in-context learning, and zero-shot reasoning in tasks like RTE and KGC.

The framework can construct local heterogeneous knowledge graphs on edge devices efficiently, reducing computational resource demands while improving model generalization. To support further research, we have open-sourced an annotated industrial question-answering dataset with labeled questions, answers, tags, and context. The KG-MASD framework provides new insights and methods for industrial question-answering systems and supports the application of MAS and domain-specific knowledge graphs in industry.

Limitations

Although the multi-agent system has played a significant role in enhancing the depth and accuracy of model reasoning, there is still room for improvement in the coordination and communication efficiency among agents during complex industrial problem reasoning. For example, multiple iterations and interactions can lead to increased reasoning delays, and the information sharing mechanism still needs to be improved to reduce information loss or misinterpretation during transmission.

Improve the coordination and communication strategies within the multi-agent system to enhance the collaboration efficiency among agents. For example, advanced multi-agent reinforcement learning algorithms can be introduced to enable agents to more quickly learn optimal collaborative strategies during interactions. In addition, the information sharing mechanism between agents can be optimized to ensure that important information is accurately and promptly conveyed, reducing information loss and misinterpretation.

References

F. Alam, H. Sajjad, M. Imran, and F. Ofli. 2020. Standardizing and benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint*.

C. Bo, S. Liu, Y. Liu, Z. Guo, J. Wang, and J. Xu. 2024. Research on isomorphic task transfer algorithm based on knowledge distillation in multi-agent collaborative systems. *Sensors*, 24:4741.

C. Cui, Z. Liu, S. Gong, L. Zhu, C. Zhang, and H. Liu. 2025. When adversarial training meets prompt tuning: Adversarial dual prompt tuning for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 34:1427–1440.

C. Deng, T. Zhang, Z. He, and 1 others. 2023. Learning a foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint*.

A. Deroy and S. Maity. 2024. Code generation and algorithmic problem solving using llama 3.1 405b. *arXiv preprint*. Under Review.

Y. Du, Z. Sun, Z. Wang, H. Chua, J. Zhang, and Y.S. Ong. 2024. Active large language model-based knowledge distillation for session-based recommendation. *ArXiv*, abs/2502.15685.

Y. Guan, A. Li, and L. Wang. 2021. Structural controllability of directed signed networks. *IEEE Transactions on Control of Network Systems*, 8:1189–1200.

H. Han, Y. Wang, H. Shomer, and 1 others. 2025. Retrieval augmented generation with graphs (graphrag). *arXiv preprint*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, and 1 others. 2023. Distilling step by step! outperforming larger language models with less training data and smaller model sizes. *Findings of the Association for Computational Linguistics: ACL 2023*. Accepted to Findings of ACL 2023.

E.J. Hu, Y. Shen, P. Wallis, and 1 others. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint*.

S. Hu, G. Zou, S. Yang, Y. Gan, B. Zhang, and Y. Chen. 2024. Large language model meets graph neural network in knowledge distillation. *ArXiv*, abs/2402.05894.

Mese I and Kocak B. 2025. Chatgpt as an effective tool for quality evaluation of radiomics research. *European Radiology*, 35(4):2030–2042.

J. Jiao, S. Afroogh, K. Chen, D. Atkinson, and A. Dhurandhar. 2025. Generative ai and llms in industry: A text-mining analysis and critical evaluation of guidelines and policy statements across fourteen industrial sectors. *arXiv preprint*.

B. Kang, P. Saha, S. Sharma, B. Chakraborty, and S. Mukhopadhyay. 2024. Online relational inference for evolving multi-agent interacting systems. *arXiv preprint*.

T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint*.

Z. Li, W. Zhou, Y.-Y. Chiang, and M. Chen. 2023. Geolm: Empowering language models for geospatially grounded language understanding. pages 5227–5240.

Z. Li and 1 others. 2024. Efficient masked autoencoders with self-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8743–8757.

T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint*.

C.-T. Lin. 1974. Structural controllability. *IEEE Transactions on Automatic Control*, 19:201–208.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 707–712, Sapporo, Japan.

A. Liu, B. Feng, and 1 others. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint*.

C. Liu, S. He, Q. Zhou, S. Li, and W. Meng. 2024b. Large language model guided knowledge distillation for time series anomaly detection. *International Joint Conference on Artificial Intelligence*.

Y. Liu, J. Ding, and Y. Li. 2022. Developing knowledge graph based system for urban computing. pages 3–7.

M. Loem, M. Kaneko, and N. Okazaki. 2023. Saie framework: Support alone isn't enough - advancing llm training with adversarial remarks. In *European Conference on Artificial Intelligence*.

X. L. Meng, F. Jin, J. Zhao, and W. Wang. 2024. An improved-knowledge-distillation based method for working condition recognition of hot rolling heating furnace in steel industry. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

P. Sahoo, A.K. Singh, S. Saha, V. Jain, S.S. Mondal, and A. Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint*.

N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan, and S. Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Neural Information Processing Systems*.

A.P. Smit, P. Duckworth, N. Grinsztajn, K. Tessera, T.D. Barrett, and A. Pretorius. 2023. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*.

N. Thelasingha, A. Agung Julius, J. Humann, J.-P. Reddinger, J. Dotterweich, and M. Childers. 2025. Iterative planning for multi-agent systems: An application in energy-aware uav-ugv cooperative task site assignments. *IEEE Transactions on Automation Science and Engineering*, 22:3685–3703.

H. Touvron, L. Martin, K. Stone, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

P. Trivedi, S. Chakraborty, A. Reddy, V. Aggarwal, A.S. Bedi, and G.K. Atia. 2025. Align-pro: A principled approach to prompt optimization for llm alignment. In *AAAI Conference on Artificial Intelligence*.

H. Wang, Q. Yu, Y. Liu, D. Jin, and Y. Li. 2021. Spatio temporal urban knowledge graph enabled mobility prediction. *Unknown Journal*, pages 184:1–184:24.

J. Wang, J. Liu, F. Xiao, and Y. Zheng. 2025. Robustness and scalability of consensus networks: The role of memory information. *IEEE Transactions on Automatic Control*.

P. Wang, X. Wei, F. Hu, and W. Han. 2024. Transgpt: Multi-modal generative pre-trained transformer for transportation. In *2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP)*, pages 96–100.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E.H. Chi, and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint*.

D. Yang and Y. Liu. 2024. Active object detection with knowledge aggregation and distillation from large models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16624–16633.

M. Yang, Y. Chen, Y. Liu, and L. Shi. 2024. Distillseq: A framework for safety alignment testing in large language models using knowledge distillation. In *International Symposium on Software Testing and Analysis*.

L. Yao, J. Peng, C. Mao, and Y. Luo. 2023a. Exploring large language models for knowledge graph completion. *arXiv preprint*.

L. Yao, J. Peng, C. Mao, and Y. Luo. 2023b. Exploring large language models for knowledge graph completion. *arXiv preprint*. Work in progress.

P. Yuan, A. Ma, Y. Yao, H. Yao, M. Tomizuka, and M. Ding. 2025. Remac: Self-reflective and self-evolving multi-agent collaboration for long-horizon robot manipulation. *arXiv preprint*.

M. Zamani and H. Lin. 2009. Structural controllability of multi-agent systems. In *2009 American Control Conference*, pages 5743–5748.

Y. Zhang, Z. Chen, L. Liang, H. Chen, and W. Zhang. 2024. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17120–17130.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay-Yoon Lee, Mohit Iyyer, and Andrew McCallum. 2024. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14201–14214.

A Appendix

A.1 Analysis of Thematic Category Distributions

Figure 5 illustrates the proportion distribution of eight manually extracted thematic categories. These categories have been classified manually according to specific criteria to facilitate analysis and understanding of the main focal points within the dataset. The proportions provide an intuitive understanding of the significance of different themes within the dataset, aiding further analysis and decision-making. Notably, the high proportion of the “Others" category may suggest the need for further subdivision or reclassification of this data to enhance the precision and relevance of the analysis.

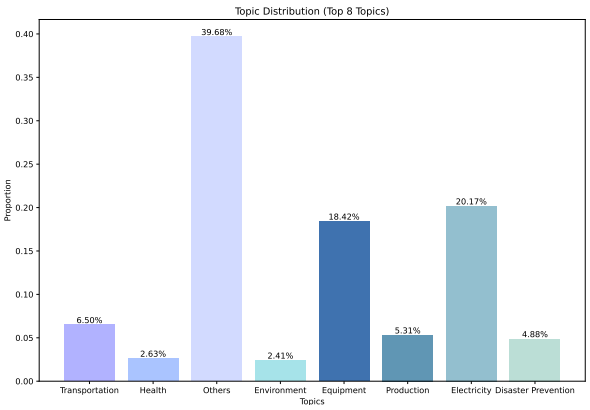

Figure 5: The proportions of the eight manually extracted thematic categories are as follows: Translation accounts for 6.5%, Health for 2.63%, Others for 39.68%, Environment for 2.41%, Equipment for 18.42%, Production for 5.31%, Electricity for 20.17%, and Disaster Prevention for 4%.

A.2 Prompts for Synthetic Data Generation

This appendix provides the prompts used to transform raw text fragments into instruction-tuning data, as shown in Figure 6. These prompts are designed based on the needs of industrial question-answering scenarios to help models better understand and generate QA content related to the industrial domain. Through these prompts, unsupervised text data can be converted into structured instruction data for model fine-tuning and optimization.

A.3 Dataset Statistics

In this study, two sets of datasets were utilized: one generated by humans and the other by GPT. The detailed statistical information for these datasets is presented in Table 3 below:

- The Human Dataset consists of 37,426 samples, with 22,510 allocated for training, 7,381 for testing, and 7,535 for validation. These data, annotated by humans, are generally considered to have higher quality and reliability.
- The GPT Dataset comprises 15,424 samples, with 9,366 designated for training, 2,980 for testing, and 3,078 for validation. These data are generated by the GPT model to assess its capability in producing text similar to that of humans.

Dataset	Total	Train	Test	Val
Human	37426	22510	7381	7535
GPT	15424	9366	2980	3078

Table 3: Statistics of Human and GPT Datasets

The division of these datasets ensures the independence of model training, testing, and validation, thereby accurately evaluating the model’s performance. By comparing the human-generated and GPT-generated datasets, we can gain a deeper understanding of the model’s performance differences across various data sources.

A.4 Specific Prompts for Different Agents

This appendix lists the specific prompts for each agent in the KG-MASD framework, as shown in Figure 7. These prompts define the roles and responsibilities of each agent, ensuring their efficient collaboration within the Multi-agent system to complete tasks such as knowledge graph construction, relation extraction, and information verification.

A.5 Analysis of Agent Dynamics in the KG-MASD System

The Figure 8 illustrates the temporal evolution of states for various agents within the KG-MASD system. These agents include the Knowledge Graph Master, Entity Extractor, Relation Extractor,

Task Objectives:
Based on the known information fragments, it is converted into annotated data that can be used for instruction fine-tuning, so as to improve the understanding and generation ability of the model in industrial Q&A scenarios.
For example:
Input: "Polyethylene typically has a density between 0.91 - 0.96g/cm³ and is chemically stable. "
Output:
{
"instruction": "Extract material names and properties from chemical material descriptions",
"input": "Polyethylene typically has a density between 0.91 - 0.96g/cm³ and is chemically stable." ,
"output": "Material name: polyethylene; Characteristics: density range 0.91 - 0.96g/cm³, good chemical stability"
}
Input: "In the Haber process of ammonia synthesis, the reaction is carried out under the conditions of high temperature and high pressure and iron catalyst. "
Output:
{
"instruction": "Identify the key elements of the chemical production process",
"input": "In the Haper process of ammonia synthesis, the reaction needs to be carried out under the conditions of high temperature and high pressure and iron catalyst." ,
"output": "Production process: Haber method for ammonia synthesis; Key elements: high temperature and high pressure conditions, iron catalysts"
}
Input: "Acetic acid and ethanol can be esterified under acidic conditions to produce ethyl acetate and water." "
Output:
{
"instruction": "Analysis of reactants, products and conditions of chemical reaction equations",
"input": "Acetic acid and ethanol can be esterified under acidic conditions to form ethyl acetate and water." ,
"output": "Reactants: acetic acid, ethanol; Products: Ethyl acetate, water; Conditions: acidic conditions, esterification reactions"
}
Task requirements:
According to the example format above, convert small pieces of text with domain information into annotated instructions to fine-tune the data. Each data item consists of three fields: "instruction" (the instruction of the annotation task that is clearly relevant), "input" (the original text snippet), and "output" (the result of annotating the text according to the instruction), which are presented in JSON format. Replace the small fragment of chemical text that needs to be converted to the following position to generate the corresponding annotation data:
```json
[
{
"instruction": "Please annotate the following text according to your expertise",
"input": "Background information to be involved",
"output": "Annotated content"
}
]
```

Figure 6: Example of Prompt-Based Data Transformation

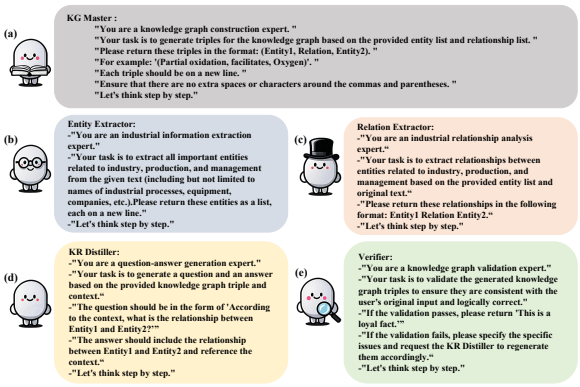

Figure 7: Example of Agent Prompts

Knowledge Relation Distiller, and Verifier. Below is a detailed analysis of the state changes for each agent:

- Knowledge Graph Master (Blue Curve): This agent's state rapidly decreases initially and then stabilizes, indicating a quick adjustment period followed by a stable state.
- Entity Extractor (Orange Curve): The state of this agent initially drops, then gradually increases and stabilizes, demonstrating adaptation during the system's self-adjustment phase.
- Relation Extractor (Green Curve): This agent's state spikes initially, then quickly declines and stabilizes, showing a rapid adjustment to a stable state.
- Knowledge Relation Distiller (Red Curve): The state of this agent also initially drops, then gradually increases and stabilizes, indicating adaptation during the system's self-adjustment.
- Verifier (Purple Curve): This agent's state rapidly decreases initially, then gradually increases and stabilizes, reflecting adaptation during the system's self-adjustment process.

As time progresses and the number of iterations increases, the KG-MASD system rapidly achieves self-stabilization. During this process, the Verifier also tends to stabilize with the increasing number of iterations. This indicates that the KG-MASD system possesses excellent self-adjustment and stabilization capabilities when handling knowledge graph construction tasks, effectively adapting to changing environments and task demands.

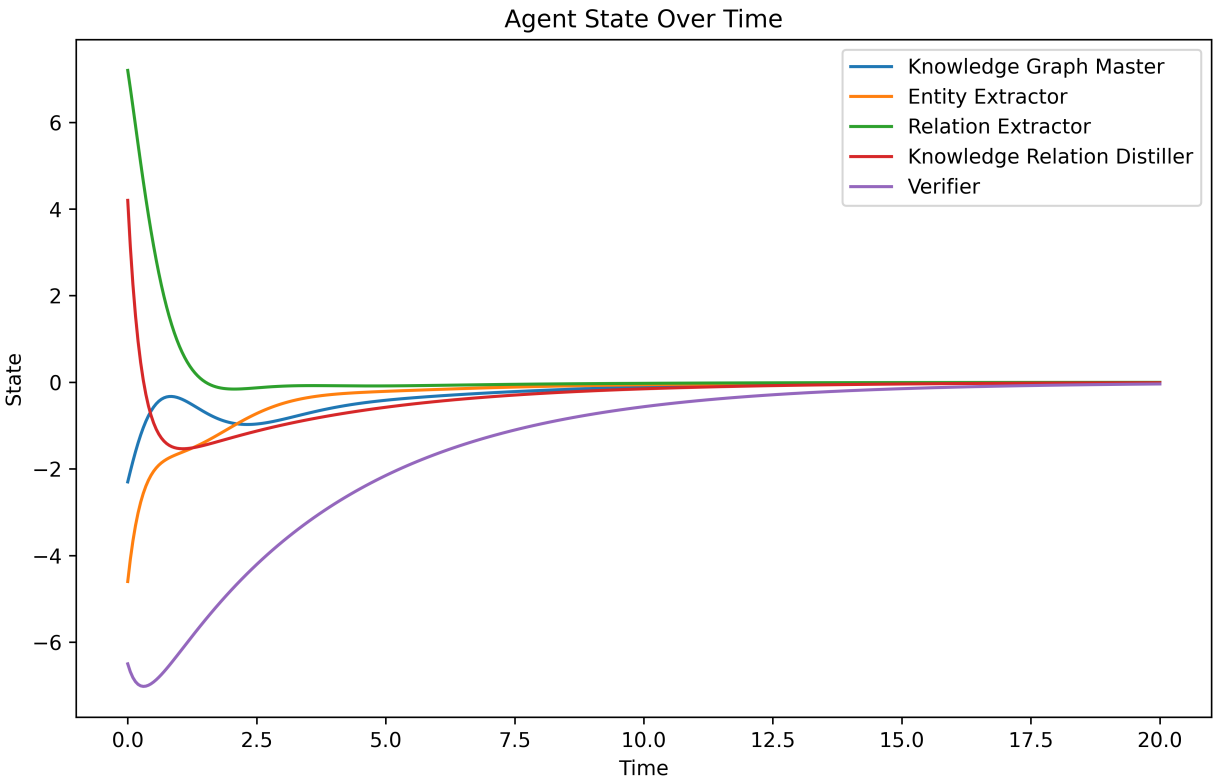

Figure 8: As time progresses and the number of iterations increases, the KG-MASD system can rapidly achieve self-stabilization. During this process, the Verifier also tends to stabilize with the increasing number of iterations.

A.6 Visualization of the Global Knowledge Graph (GKG)

This appendix provides a visual representation of the Global Knowledge Graph (GKG). As shown in Figure 9, the GKG is a global knowledge structure extracted from raw data, containing entities, relationships, and semantic information from the industrial domain. The visualization helps to intuitively understand the structure and content of the GKG.

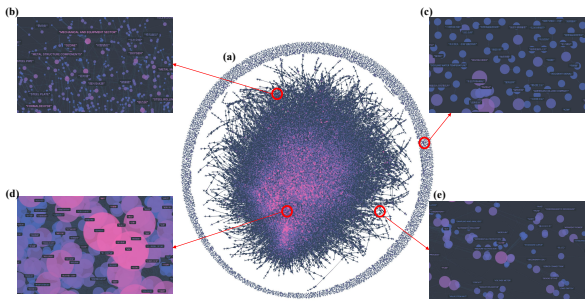

Figure 9: Visualization of the Global Knowledge Graph

A.7 Local Heterogeneous Knowledge Graph Construction

Figure 10 provides a visual representation of the process involved in constructing Local Heterogeneous Knowledge Graphs (LHKG) from a Global Knowledge Graph (GKG). The diagram illustrates how the input data stream is queried to perform retrieval and how it gradually searches for locally optimal association paths. This process continues until all possible paths have been explored, resulting in the formation of a local heterogeneous knowledge graph that is tailored to the current data.

Process Overview:

- Data Stream Input: The system receives the input data stream.
- Query Execution: A query retrieves relevant information from the Global Knowledge Graph (GKG).
- Path Exploration: The system explores paths to identify locally optimal associations.
- Graph Formation: A Local Heterogeneous Knowledge Graph (LHKG) is constructed based on the explored paths.

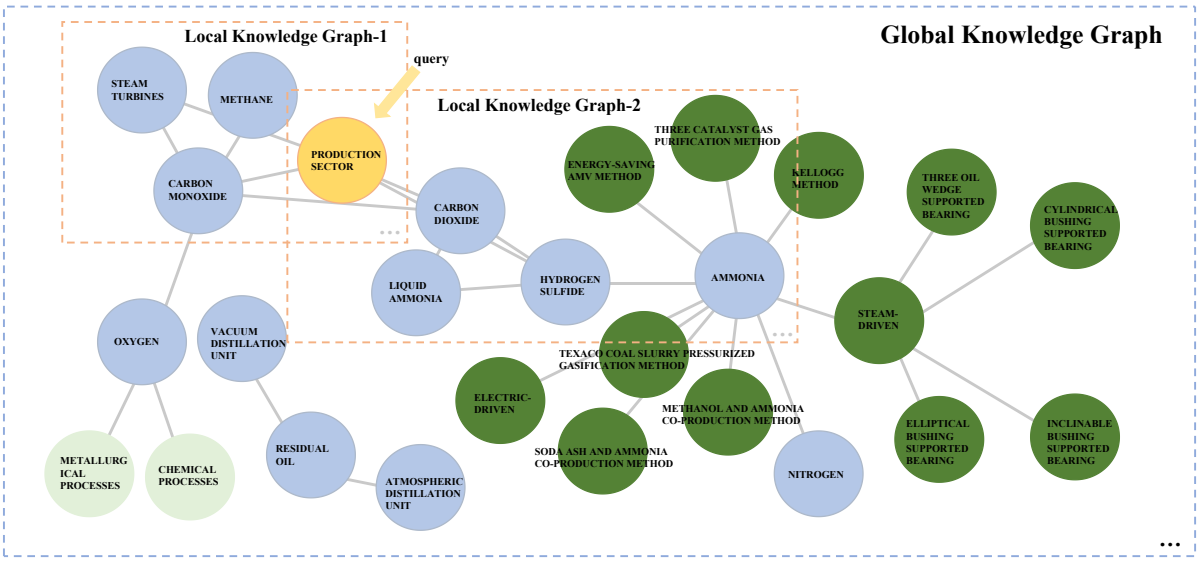

Figure 10: This figure illustrates examples of constructing Local Heterogeneous Knowledge Graphs (LHKG) from the Global Knowledge Graph (GKG). Initially, the input data stream enters through a query to perform retrieval and gradually searches for locally optimal association paths until all possible paths have been explored, forming a local heterogeneous knowledge graph for the current data.

This method of constructing LHKGs allows for a more focused and contextually relevant knowledge representation, which can be particularly useful in applications requiring detailed and specific insights from large and diverse datasets. The ability to form these graphs dynamically ensures that the knowledge base remains up-to-date and aligned with the latest data inputs.

A.8 Data Generation Process

This appendix details the data generation process in the KG-MASD framework, including the construction of local knowledge graphs from raw data and the generation of instruction-tuning data, as shown in Figure11. This process involves multiple steps, including Relation Triple Extraction (RTE), Knowledge Graph Completion (KGC), and collaboration within the Multi-agent system.

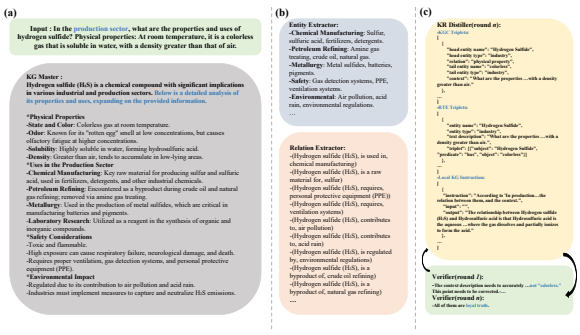

Figure 11: Data Generation Process Diagram