

AutoMARS: Automatic and Scalable Multi-Agent Large Language Model Framework for Supply Chain Risk Assessment

Yinzhu Quan^{a,*}, Zefang Liu^b, Frederick Benaben^{a,c} and Benoit Montreuil^a

^aPhysical Internet Center, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, 30332, GA, USA

^bSchool of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, 30332, GA, USA

^cCentre Genie Industriel, IMT Mines Albi, Albi, France

ARTICLE INFO

Keywords:

Generative AI

Large Language Models

Multi-Agent System

Supply Chain Risk Assessment

Automation

Scalability

ABSTRACT

We propose AutoMARS (Automatic Multi-Agent Risk Assessment), an agentic framework for automatic and scalable comparative risk assessment in supply chain networks. The framework enables end-to-end execution of analytical workflows by automatically planning tasks, coordinating specialized agents, invoking external analytical tools, and synthesizing domain-specific results into integrated risk evaluations. AutoMARS further supports scalable risk comparison across multiple locations by parallelizing independent analytical tasks. The framework adopts a multi-domain, multi-agent, and multi-tool architecture that integrates heterogeneous risk factors from socioeconomic, disaster, environmental, and transportation domains. Through coordinated agent interactions and tool integration, AutoMARS enables structured, systematic, and interpretable comparative analysis to support supply chain risk-aware location evaluation.

1. Introduction

Modern supply chains operate in increasingly volatile, uncertain, complex, and ambiguous (VUCA) environments. In such environments, companies increasingly depend on geographically dispersed suppliers, complex logistics networks, and digitally coordinated operations to sustain production and distribution activities. While these configurations improve efficiency and global reach, they also expand the exposure of supply chains to diverse disruption sources. Events such as natural hazards, environmental disruptions, transportation accidents, and socioeconomic vulnerabilities can interrupt material flows, delay logistics operations, and create instability across multiple stages of supply chains. Consequently, managing disruption risk has become a critical challenge for organizations seeking to maintain supply continuity and operational resilience in increasingly uncertain environments (Ivanov, 2022; Katsaliaki et al., 2022).

To address the above challenge, extensive research has focused on developing analytical frameworks and decision-support methods for assessing and managing risks in interconnected supply chain networks (Montreuil, 2011; Montreuil et al., 2012). Existing studies have explored various approaches to support risk identification, evaluation, and mitigation in complex supply chain environments, including optimization-based models (Ivanov et al., 2018), simulation methods (Cerabona et al., 2024), and multi-agent coordination mechanisms (Giannakis and Louis, 2011). These approaches aim to improve the ability of supply chain systems to anticipate disruptions, evaluate potential impacts, and support informed operational and strategic decision-making. More recently, advances in artificial intelligence (AI) and large language models (LLMs) have opened new opportunities for enhancing risk assessment capabilities in complex supply networks. For example, Quan et al. (2026) propose a novel LLM-enhanced framework, MARS, that integrates both structured and unstructured factors to support risk-aware logistic hub site selection.

Despite the advantages of existing approaches, several key limitations remain. First, fully integrated end-to-end automation is still limited, as many multi-agent systems (MAS) require partial human intervention and manual coordination across analytical steps (Ambati, 2025), preventing seamless execution from task decomposition to final assessment. Second, scalability remains a challenge in large-scale risk assessment scenarios (Katsaliaki et al., 2022), where sequential interactions with data sources and analytical agents lead to significant computational overhead,

*Corresponding author

✉ yquan9@gatech.edu (Y. Quan); liuzefang@gatech.edu (Z. Liu); frederick.benaben@mines-albi.fr (F. Benaben); benoit.montreuil@isye.gatech.edu (B. Montreuil)

ORCID(s): 0009-0003-6119-9009 (Y. Quan); 0000-0003-1947-3249 (Z. Liu); 0000-0002-1676-6524 (F. Benaben); 0000-0001-9909-9212 (B. Montreuil)

highlighting the need for parallel execution. Third, reliance on absolute risk scores can reduce stability when LLM-based agents are involved, as small output variations may affect final results (Aqlan and Lam, 2015; Babu and Yadav, 2023). Comparative evaluation based on relative ordering provides a more robust alternative, improving consistency and interpretability across repeated assessments.

In this paper, we propose AutoMARS (Automatic Multi-Agent Risk Assessment), an agentic framework designed to support automatic and scalable comparative risk assessment in supply chain networks. The framework enables end-to-end execution of the analytical workflow, allowing a user to submit a high-level request while the system automatically plans and coordinates multiple specialized agents to invoke external tools, perform domain-specific analyses, and synthesize the results into an integrated risk evaluation. By integrating multi-domain information, coordinating multiple agents, and invoking diverse analytical tools, AutoMARS provides a unified framework for complex risk analysis tasks. AutoMARS further supports scalable analysis across multiple locations by enabling parallel execution of independent analytical tasks. In addition, the framework emphasizes comparative risk evaluation across locations, enabling more consistent and reliable decision-making by focusing on relative risk ordering rather than absolute scores. This comparative perspective provides a more robust and interpretable basis for decision-making in supply chain location evaluation. Risk-conscious logistic location selection is essential for facility location and network design decisions, including hub siting and infrastructure investment. Overall, this work aims to develop an automatic, scalable, and reliable framework for comparative supply chain risk assessment across multiple locations.

In this study, we aim to address the following research questions (RQs):

- **RQ1:** How to enable end-to-end automation in supply chain risk assessment?
- **RQ2:** How to scale supply chain risk assessment while mitigating the computational complexity introduced by multi-agent coordination?
- **RQ3:** How to conduct comparative risk assessment across multiple locations?

Answering these research questions, our main contributions in this paper are as follows:

1. **End-to-end automatic framework for supply chain risk assessment:** We propose AutoMARS, an automatic multi-agent framework that enables end-to-end execution of supply chain risk assessment tasks. Given a high-level user request, the framework automatically plans and decomposes the analytical workflow, coordinates specialized agents, invokes relevant external tools, and synthesizes domain-specific results into a unified risk evaluation.
2. **Scalable risk comparison through parallel execution:** The proposed framework enables scalable comparative risk assessment across multiple locations by executing independent analytical tasks in parallel. This design significantly reduces computational time when evaluating many locations and supports efficient risk comparison in large-scale supply chain decision contexts.
3. **Multi-domain risk analysis through multi-agent multi-tool integration:** The framework integrates heterogeneous risk factors across multiple domains, including socioeconomic, disaster, environmental, and transportation domains, by coordinating specialized agents and invoking diverse external analytical tools. This multi-domain, multi-agent, and multi-tool architecture enables comprehensive risk analysis by combining information from multiple sources and analytical perspectives.

The rest of this paper is organized as follows. Section 2 reviews related work on supply chain risk management, LLM-based multi-agent orchestration, and LLM-as-a-judge evaluation. Section 3 defines the problem of comparative supply chain risk assessment. Section 4 presents the proposed AutoMARS framework, including system architecture, agent design, and execution strategies. Section 5 reports experimental results and evaluates the effectiveness, stability, and scalability of the framework. Section 6 concludes the paper and discusses future research directions.

2. Related Work

In this section, we review prior research relevant to this study. We first review existing studies on automation in supply chain risk management (SCRM). Next, we discuss recent developments in large language model (LLM)-based multi-agent orchestration frameworks. Finally, we discuss the emerging paradigm of LLM-as-a-Judge for evaluating multi-agent reasoning and outcomes.

2.1. Automation in Supply Chain Risk Management

SCRM involves continuous tasks such as risk identification, assessment, monitoring, and response (Fan et al., 2017), which are difficult to sustain under rapidly changing conditions. In practice, risk assessment often relies on expert judgment and static analyses, limiting scalability across regions and time (Sodhi et al., 2012; Ho et al., 2015). Meanwhile, risk information arises from heterogeneous, high-frequency sources, creating the need for automated processing and timely updates (Chu et al., 2020). This motivates artificial intelligence (AI)-driven automation in SCRM to improve efficiency, consistency, and scalability (Ganesh and Kalpana, 2022; Ziari and Taleizadeh, 2025; Chowdhury et al., 2026).

Existing automation approaches support the representation and modeling of temporal events from text, enabling standardized analysis of evolving risks (Liu and Quan, 2024, 2025b; Liu et al., 2025). They also integrate multi-source data for risk assessment and decision support in logistics contexts (Zhao et al., 2024; Quan et al., 2025b). More broadly, generative AI and automated decision systems enhance analytical capabilities and responsiveness (Oger et al., 2018; Sharma, 2022; Jackson et al., 2024), with applications in areas such as inventory management demonstrating efficiency gains (Saleem, 2020).

Despite these advances, automation in SCRM remains fragmented, with most approaches focusing on isolated functions rather than end-to-end workflows (Das and Perona, 2025) and offering limited support for systematic risk comparison across regions (Katsaliaki et al., 2022). Simulation-based and digital twin systems enable automated monitoring but rely on predefined models, limiting adaptability (Ivanov and Dolgui, 2021). Evidence from automated multi-agent systems (MAS) in other domains further suggests the need for more integrated frameworks (Liu, 2025). Overall, automation enhances efficiency, consistency, and responsiveness in SCRM by enabling data integration and supporting coordinated risk assessment.

2.2. Automated LLM-Based Multi-Agent Orchestration

Language interaction enables flexible coordination among decision-making entities in complex systems. Traditional MAS rely on predefined protocols, limiting adaptability to dynamic contexts and heterogeneous information. In contrast, LLM-based agents use natural language to coordinate roles, enabling orchestration through conversational workflows, role assignment, and task decomposition (Li et al., 2023; Wu et al., 2024; Hong et al., 2023; Guo et al., 2024).

Existing work spans domain-specific applications and general-purpose frameworks. Domain-oriented systems use LLM-driven agents coordinated through automated workflows for tasks such as inventory management and customer-facing processes (Quan and Liu, 2024b; Quan et al., 2025a). General approaches orchestrate multiple agents or tools via task decomposition to solve complex problems end to end (Shen et al., 2023; Schick et al., 2023).

In practice, multi-agent coordination is enabled at multiple levels. Frameworks such as LangChain¹ and LangGraph² support structured orchestration through pipeline- or graph-based workflows, while protocols such as Model Context Protocol³ (MCP) and Agent2Agent Protocol⁴ (A2A) standardize agent interoperability and tool integration. Building on these foundations, systems such as BMAD⁵, GSD⁶, and OpenCLAW⁷ further demonstrate specification-driven workflows and structured multi-agent orchestration.

Despite recent progress, LLM-based multi-agent orchestration faces challenges in stability and evaluation (Li et al., 2024b; Guo et al., 2024; Tran et al., 2025). Ensuring alignment, safety, and controllability in coordinated workflows remains critical for reliable deployment (Bai et al., 2022). Overall, automated multi-agent orchestration enables flexible coordination, task decomposition, and scalable reasoning across agents and tools, but further advances are required to ensure robustness, stability, and well-governed execution in real-world settings.

2.3. LLM-as-a-Judge in Multi-Agent Collaboration

Evaluating complex reasoning and collaborative outcomes remains a key challenge in LLM-driven MAS. Traditional approaches rely on handcrafted metrics or human annotation, which do not scale well to multi-step interactive settings. Recent work uses LLMs as evaluators to assess outputs, reasoning processes, and task performance, enabling

¹<https://www.langchain.com/>

²<https://www.langchain.com/langgraph>

³<https://www.anthropic.com/news/model-context-protocol>

⁴<https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>

⁵<https://github.com/bmad-code-org/BMAD-METHOD>

⁶<https://github.com/gsd-build/gsd-2>

⁷<https://openclaw.ai/>

flexible evaluation (Bai et al., 2022; Madaan et al., 2023). LLM-based judges further support preference comparison and feedback generation and have been proposed as scalable alternatives to human evaluation in multi-agent settings (Li et al., 2024a, 2025; Thakur et al., 2025). However, using the same LLM for both generation and evaluation may introduce shared biases, although empirical studies show reasonable alignment with human judgment.

LLM-as-a-Judge has been applied to evaluate agent behavior, reasoning quality, and task outcomes in interactive environments. Early studies demonstrate automatic assessment of outputs and reasoning traces (Yue et al., 2023; Zhang et al., 2023), while later work extends evaluation to multi-step interactions, tool usage, and task completion (Xu et al., 2023; Chiang et al., 2024). Recent benchmarks formalize this paradigm by comparing agent outputs across diverse tasks and interaction patterns (Tan et al., 2024). Domain-specific applications further evaluate sequential reasoning and agent performance in economic and web-based settings (Quan and Liu, 2024a; Liu and Quan, 2025a).

Despite its promise, LLM-as-a-Judge faces limitations in reliability and standardization. Prior work shows sensitivity to context, inconsistent judgments (Chen et al., 2025; Szymanski et al., 2025), and risks of bias amplification and feedback effects (Kiela et al., 2021; Saunders et al., 2022). The lack of standardized protocols and concerns about transparency and alignment with human judgment further motivate more robust and holistic evaluation frameworks (Liang et al., 2022; Chen et al., 2024; Bavaresco et al., 2025). To mitigate these issues, recent studies explore strategies such as multi-judge aggregation, cross-model evaluation, structured criteria, and hybrid human-AI evaluation to improve robustness (Gu et al., 2024). Overall, LLM-as-a-Judge offers a flexible paradigm for evaluating complex multi-agent reasoning, while its reliability can benefit from human oversight.

3. Problem Definition

Logistic location selection is a critical decision in supply chain networks, where evaluating and comparing risk levels across different geographic regions plays an important role in ensuring operational stability and resilience. Regions may differ significantly in terms of socioeconomic conditions, disaster exposure, environmental quality, and transportation safety, all of which can affect the reliability of supply chain operations. However, previous risk assessment approaches often require manual coordination across multiple analytical steps, provide limited capability to integrate heterogeneous risk information from multiple domains, and require substantial computational time when evaluating risks across many regions. These limitations restrict the ability to conduct efficient and comprehensive comparative risk assessment across regions, particularly when a large number of locations must be evaluated. As supply chain decision-making increasingly requires integrating heterogeneous risk information and comparing risks across multiple regions, a key challenge is to develop frameworks that can automatically coordinate multi-domain analyses, support parallel task execution, and efficiently synthesize risk indicators from diverse information sources. This highlights the need for intelligent systems capable of planning tasks, coordinating multiple analytical components, and synthesizing diverse forms of risk information to enable comprehensive and adaptive location risk evaluation.

Problem Description: Consider a set of candidate locations within a supply chain network. For each location, risk information can be obtained from multiple domains. To analyze these heterogeneous risk indicators, a set of specialized domain agents is available, where each agent is designed to handle a specific analytical task. Each agent can access a collection of external tools that provide domain-specific data retrieval, processing, and analysis capabilities. The objective is to enable Manager Agent to automatically plan and decompose the overall risk assessment workflow, coordinate the execution of Domain Agents and their associated tools, and integrate multi-domain information into a unified risk evaluation for each location. Reflection Agent evaluates the synthesized results and provides feedback, allowing Manager Agent to revise the analysis until a consistent comparative risk assessment is obtained. To further improve scalability, independent analytical tasks within the workflow can be executed in parallel.

In summary, the problem of comparative supply chain risk assessment requires a framework capable of coordinating domain-specific analyses and integrating heterogeneous risk information across multiple locations. To address this challenge, we propose an automatic multi-agent framework that supports dynamic task orchestration, parallel execution, and iterative reflection for scalable and adaptive risk evaluation.

4. Methodology

This section presents the methodology of AutoMARS, including its overall framework, agent architecture, and efficient execution strategies. We first describe the system architecture and workflow for structured multi-agent coordination. Next, we detail the design and interactions of specialized agents for cross-domain risk analysis. Finally,

we introduce execution strategies that improve efficiency, including parallelism, caching mechanisms, and a sorting algorithm.

4.1. Framework

AutoMARS is an automatic multi-agent framework for comparative risk assessment across multiple locations in supply chain networks. To enable end-to-end automation (RQ1), it supports coordinated planning, tool invocation, and cross-domain synthesis within a unified workflow. As illustrated in Figure 1, the framework adopts a modular architecture consisting of a Manager Agent, multiple Domain Agents with associated external tools, and a Reflection Agent. Manager Agent plans and orchestrates the overall assessment workflow, decomposing the task into domain-specific analyses and coordinating Domain Agents, which invoke external tools to retrieve and process heterogeneous risk information. The results are then integrated into a comparative risk evaluation. Reflection Agent subsequently evaluates the integrated results to verify the completeness and consistency of the reasoning process. If deficiencies are identified, feedback is returned to Manager Agent, which revises the assessment accordingly. This iterative interaction continues until a stable and reliable comparative risk assessment is obtained. The detailed workflow of AutoMARS framework is formally summarized in Algorithm 1.

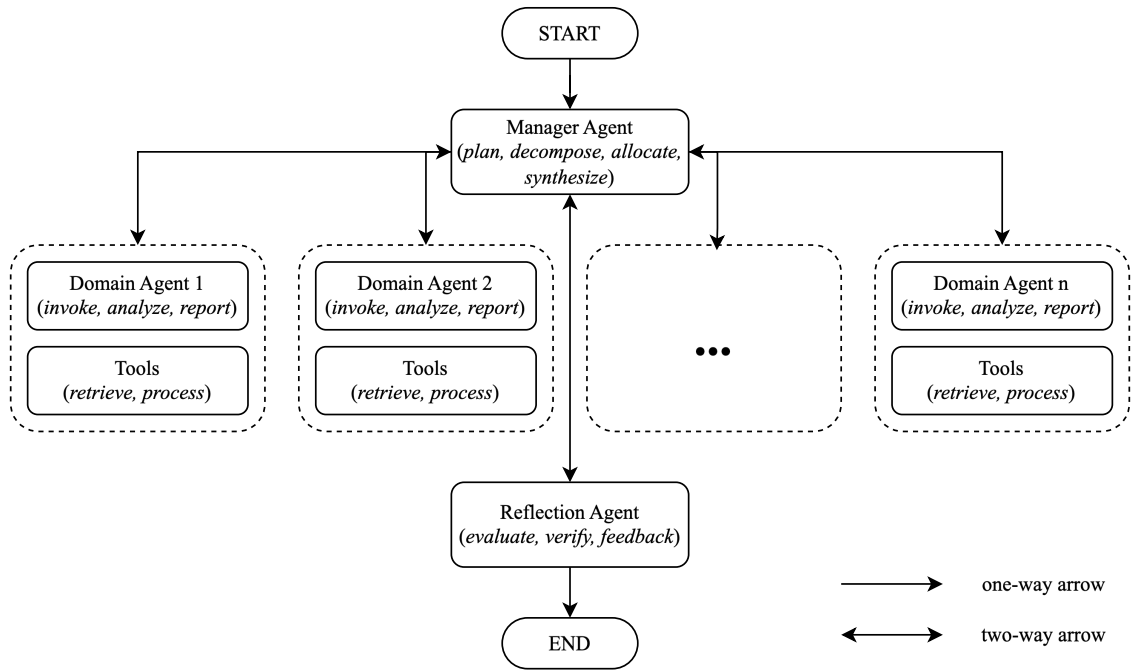


Figure 1: Framework of AutoMARS for supply chain risk assessment.

4.2. Agents

This section describes agents in AutoMARS and their roles in the overall workflow. The system includes Manager Agent for coordination and integration, Domain Agents for specialized analysis, and Reflection Agent for validating the final results.

4.2.1. Manager Agent

Manager Agent serves as the central coordinator of the AutoMARS framework and is responsible for planning, orchestrating, and integrating the overall risk assessment workflow. Given a high-level user query, Manager Agent generates an initial analysis plan that determines which Domain Agents should be activated and how the analytical process should proceed. The behavior of Manager Agent is guided by a structured system prompt that specifies its planning, execution, and integration responsibilities. A simplified version of the prompt is summarized (see “Prompt for Manager Agent (Condensed Version)”).

Algorithm 1 Parallel Multi-Agent Workflow

```
1: Input: User query  $q$ , manager agent  $M$ , reflection agent  $R$ , domain worker agents  $\{W_1, W_2, \dots, W_n\}$ , tool sets  $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n\}$ 
2: Output: Final response
3: Initialize shared state  $S \leftarrow \{\text{messages} : [q], \text{reports} : \emptyset\}$ 
4: function WORKERSUBGRAPH( $W_i, \mathcal{T}_i, S$ )
5:   Initialize private worker state  $S_i$ 
6:   repeat
7:     Generate worker response  $m_i \leftarrow W_i(S_i)$ 
8:     if  $m_i$  contains tool calls then
9:       Execute requested tools from  $\mathcal{T}_i$ 
10:      Append tool outputs to  $S_i$ 
11:    end if
12:  until  $m_i$  contains no further tool calls
13:  Set worker report  $r_i \leftarrow m_i$ 
14:  return  $r_i$ 
15: end function
16: repeat
17:   Generate manager response  $m_M \leftarrow M(S)$ 
18:   Extract selected worker set  $\mathcal{W}^*$  from  $m_M$ 
19:   if  $\mathcal{W}^*$  is not empty then
20:     Execute selected workers in parallel:  $\{r_i\}_{W_i \in \mathcal{W}^*} \leftarrow \text{WORKERSUBGRAPH}(W_i, \mathcal{T}_i, S)$ 
21:     Merge worker reports  $\{r_i\}$  into  $S$ 
22:   end if
23:   Generate reflection response  $m_R \leftarrow R(S)$ 
24: until  $m_R$  accepts the assessment
25: return final response
```

Prompt for Manager Agent (Condensed Version)

You are Manager Agent responsible for orchestrating multiple Domain Agents to perform a comparative risk assessment between Location A and Location B. Your task is to plan, execute, and integrate the analytical workflow. Domain Agents provide domain-specific evidence, while you control the workflow and decision logic.

Stage 1: Initial Planning. At the beginning of the workflow, Manager Agent constructs an initial analysis plan before consulting any Domain Agent. The plan specifies the domains to be analyzed, identifies which Domain Agents should be activated, and determines which analyses can be executed in parallel. It also defines the role of each Domain Agent and outlines how evidence from multiple domains will be integrated to support the final comparative risk judgment.

Stage 2: Evidence-Guided Execution. Manager Agent coordinates the execution of Domain Agents and collects domain-specific evidence as analyses are completed. During this process, Manager Agent summarizes the obtained results, evaluates whether the current plan remains appropriate, and revises the execution strategy when new evidence suggests that adjustments are necessary.

Stage 3: Reflection Feedback Handling. After synthesizing the comparative risk assessment, Manager Agent submits the result to Reflection Agent for validation. If Reflection Agent identifies deficiencies, Manager Agent revises the comparative assessment according to the feedback while preserving previously generated domain evidence. If additional domain analyses are required, updated evidence is incorporated into the synthesis. This revision loop continues until Reflection Agent accepts the final assessment.

Based on the above prompt specification, Manager Agent operates in a structured three-stage process. In the initial planning stage, it decomposes the comparative risk assessment task into multiple domain-specific analyses and determines the execution strategy, including the selection of Domain Agents and opportunities for parallel execution. In

the evidence-guided execution stage, it coordinates Domain Agents and incrementally collects domain-level evidence while evaluating and updating the execution plan as new information becomes available. In the reflection feedback stage, it synthesizes the multi-domain evidence into a unified comparative risk assessment and submits the result to Reflection Agent for validation. If deficiencies are identified, the assessment is revised by incorporating updated evidence until a consistent and reliable final judgment is obtained.

4.2.2. Domain Agents

AutoMARS incorporates multiple Domain Agents to capture heterogeneous risk factors across different dimensions of supply chain environments. Specifically, the framework includes Census Agent, Disaster Agent, Environment Agent, and Traffic Agent, which respectively focus on socioeconomic conditions, hazard exposure, environmental quality, and transportation safety. Each Domain Agent is equipped with a set of specialized external tools for retrieving and analyzing domain-specific data. Tools used by different Domain Agents, along with their corresponding data source types and functional roles, are summarized in Table 1.

Table 1

Tools used in AutoMARS, including their associated agents, data source types, and functional descriptions.

Tool	Agent	Type	Description
Labor Exposure Tool ⁸	Census	API	Extract labor composition indicators
Workforce Dynamics Tool ⁹	Census	API	Capture employment flow and turnover indicators
Socioeconomic Profile Tool ¹⁰	Census	API	Retrieve core socioeconomic indicators
Community Resilience Tool ¹¹	Census	API	Characterize population risk factor distributions
Industry Scale Tool ¹²	Census	API	Quantify industry-level economic scale indicators
Disaster Declaration Tool ¹³	Disaster	API	Identify federal disaster declaration events
Multi-Hazard Risk Index Tool ¹⁴	Disaster	API	Provide composite multi-hazard risk scores
Climate Tool ¹⁵	Disaster	API	Query historical climate observations
Severe Weather Tool ¹⁶	Disaster	API	Aggregate hazard event frequency and impact indicators
Air Quality Tool ¹⁷	Environment	API	Aggregate air quality exposure indicators
Environmental Facility Tool ¹⁸	Environment	API	Identify regulated environmental facilities
Social Vulnerability Index Tool ¹⁹	Environment	API	Assess relative social vulnerability levels
Traffic Accident Tool ²⁰	Traffic	Database	Aggregate traffic accident frequency and severity indicators

Census Agent. Census Agent is responsible for characterizing the socioeconomic and labor conditions of a region, which provide evidence on human and economic exposure, population vulnerability, and underlying socioeconomic conditions that influence how risks manifest across locations. This agent focuses on indicators describing population composition, labor market conditions, economic activity, and community characteristics.

To regulate the behavior of Census Agent, a structured prompt is designed to define its analytical scope, data retrieval strategy, and output constraints. The prompt ensures that the agent identifies relevant socioeconomic indicators, retrieves data efficiently using appropriate tools, and produces consistent, evidence-based comparative outputs without introducing subjective judgments or workflow decisions. A simplified version of the prompt is presented below for illustration purposes, using Census Agent as a representative example (see “Prompt for Census Agent (Condensed Version)”).

⁸<https://www.census.gov/programs-surveys/ces/data/restricted-use-data/lehd-data.html>

⁹<https://www.census.gov/data/developers/data-sets/qwi.html>

¹⁰<https://www.census.gov/programs-surveys/acs.html>

¹¹<https://www.census.gov/data/developers/data-sets/community-resilience-estimates.html>

¹²<https://www.census.gov/data/developers/data-sets/economic-census.html>

¹³<https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v2>

¹⁴<https://hazards.fema.gov/nri/data-resources>

¹⁵<https://www.ncdc.noaa.gov/cdo-web/webservices/v2>

¹⁶<https://www.ncdc.noaa.gov/stormevents/>

¹⁷<https://www.epa.gov/aqs>

¹⁸<https://www.epa.gov/frs>

¹⁹<https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html>

²⁰https://smoosavi.org/datasets/us_accidents

Prompt for Census Agent (Condensed Version)

You are Census Agent responsible for analyzing socioeconomic and demographic conditions for comparative risk assessment between Location A and Location B. Your role is to retrieve and compare indicators related to population structure, labor force characteristics, economic activity, and mobility patterns, in order to characterize differences in human and economic exposure across locations.

Census Agent must identify required indicators in advance and retrieve relevant datasets efficiently, prioritizing recent and complete summary data. The analysis should be strictly evidence-based, clearly distinguish the two locations, and produce structured outputs suitable for downstream comparative risk evaluation. The agent must not provide recommendations or influence workflow decisions.

Disaster Agent. Disaster Agent is responsible for characterizing historical hazard exposure and disaster-related risk at the regional level, which provides essential context for assessing the frequency, severity, and consequences of disruptive events. This agent focuses on hazard occurrence, disaster declarations, climate conditions, and multi-hazard risk indicators that capture both acute shocks and longer-term environmental threats.

Environment Agent. Environment Agent is responsible for characterizing environmental exposures and infrastructure-related conditions that influence regional risk and vulnerability. This agent focuses on ambient environmental quality, regulated facility presence, and social vulnerability patterns that reflect chronic exposure and underlying environmental burdens.

Traffic Agent. Traffic Agent is responsible for characterizing transportation-related safety risk and infrastructure exposure at the regional level. This agent focuses on the frequency and severity of traffic accidents as indicators of mobility-related risk and infrastructure vulnerability.

4.2.3. Reflection Agent

Reflection Agent serves as the final validation component of AutoMARS framework and is responsible for evaluating the completeness and consistency of the integrated risk assessment results. Given the outputs synthesized by Manager Agent from all Domain Agents, Reflection Agent determines whether the assessment is sufficient for decision support by verifying cross-domain evidence coverage, logical consistency, and the clarity of the final comparative conclusion. The behavior of Reflection Agent is guided by a structured system prompt that specifies its validation criteria and decision routing logic. For readability, a simplified version of the prompt is summarized (see “Prompt for Reflection Agent (Condensed Version)”).

Prompt for Reflection Agent (Condensed Version)

You are the Reflection Agent. Your task is to validate the final comparative risk assessment between Location A and Location B. You do not perform analysis, planning, or refinement; you only evaluate whether the result is sufficient for decision support.

The assessment is acceptable if:

1. All required domains are meaningfully referenced
2. A clear higher-risk location is explicitly identified
3. The conclusion is supported by consistent, domain-specific evidence

If the assessment is acceptable:

- Produce a concise final determination
- Summarize cross-domain evidence
- Output a structured result and terminate the workflow

If the assessment is not acceptable:

- Identify missing or inconsistent components
- Specify which domains require revision
- Route back to the Manager Agent

Based on the above prompt specification, Reflection Agent determines whether the integrated assessment is sufficiently complete and logically consistent for final decision support. The assessment is accepted when all required domains are meaningfully incorporated, cross-domain evidence supports a consistent comparative conclusion, and a clear higher-risk location is explicitly identified. In this case, Reflection Agent produces the final structured output and terminates the workflow. Otherwise, if missing domain evidence, unclear conclusions, or logical inconsistencies are detected, the agent identifies the problematic components and routes the workflow back to Manager Agent for further revision and coordination.

4.3. Efficient Execution

This section presents the key mechanisms that enable efficient execution in AutoMARS, including parallel execution for reducing latency, caching mechanisms for avoiding redundant computation, and a comparison-based sorting algorithm for scalable ranking across locations.

4.3.1. Parallel Execution

AutoMARS adopts a hierarchical parallel execution strategy to support efficient and scalable execution of multiple independent tasks (RQ2). The system is implemented as a graph-based execution model in which Manager Agent dynamically dispatches multiple Domain Agents as independent worker subgraphs that can execute concurrently. Within each subgraph, tool calls are asynchronously executed through a dedicated tool execution module to support parallel data retrieval from heterogeneous external sources. Intermediate outputs are stored in structured shared states indexed by unique keys and merged through synchronized associative updates before being passed back to Manager Agent for integration. By enabling parallel processing at both inter-agent and intra-agent levels, the system minimizes idle waiting time while maintaining modularity and supporting scalable extension to additional domains and data sources. This fan-out and fan-in execution pattern enables efficient multi-level parallel processing while ensuring deterministic aggregation of multi-domain evidence.

4.3.2. Caching Mechanism

AutoMARS incorporates a caching mechanism to reduce redundant computation and improve efficiency in repeated workflows involving external data retrieval and large language model (LLM) inference. The caching mechanism is implemented through multiple cache layers tailored to different system components. For external tools, the system maintains a disk-based cache in which final structured outputs are stored as persistent JSON files indexed by deterministic identifiers constructed from input locations. Upon receiving a request, the system performs a lookup using the corresponding identifier; matching entries are directly reused, while missing entries trigger new execution and storage for future reuse. For LLM calls, caching is handled separately through a key-based mechanism in which prompts and contextual inputs are normalized and serialized into a canonical representation before generating a hash-based cache key. Upon each invocation, the system computes the corresponding key and checks for its existence in the cache; matching keys enable direct reuse of previous responses and avoid repeated model inference. Together, these mechanisms enable efficient and consistent reuse of both external data retrieval and repeated LLM computation.

4.3.3. Sorting Algorithm

To rank locations by overall risk, AutoMARS adopts a comparison-based sorting procedure in which ordering decisions are determined through pairwise comparative risk assessments rather than predefined scalar scores. For each location pair, the system invokes AutoMARS graph to generate a structured comparison result identifying the higher-risk location, which is then mapped to an ordering relation used by the sorting algorithm. This design derives ranking decisions directly from cross-domain evidence synthesized during multi-agent analysis, avoiding fixed global weighting schemes while supporting interpretable comparative risk ranking. The underlying sorting algorithm follows a stable and adaptive strategy based on Timsort²¹, which combines merge-based and insertion-based techniques to efficiently handle partially ordered data, achieving a time complexity of $O(n \log n)$ in the average and worst cases and $O(n)$ in the best case (Auger et al., 2018).

²¹<https://en.wikipedia.org/wiki/Timsort>

5. Experiments

This section evaluates AutoMARS in terms of effectiveness, stability, and scalability for comparative risk assessment. We first introduce the experimental setup, followed by pairwise experimental results to demonstrate the workflow and comparative performance. Finally, we present ranking experimental results, analyzing the consistency and robustness of the framework in multi-location ranking scenarios.

5.1. Experimental Setup

Geographical Scope. We conduct experiments on counties within Atlanta metropolitan area at multiple spatial scales, including Atlanta Urban Core, Atlanta Regional Commission (ARC) region, and Metro Atlanta. Atlanta Urban Core is defined as a five-county subset representing the most densely populated and economically active areas. ARC region corresponds to an 11-county regional planning area, while Metro Atlanta is defined as Atlanta-Sandy Springs-Alpharetta Metropolitan Statistical Area, consisting of 29 counties. This multi-scale geographical design enables systematic evaluation of the proposed framework across different levels of spatial aggregation.

Model Configuration. All agents in AutoMARS are implemented as large language model (LLM)-driven modules using GPT-5²² as the underlying reasoning engine. The system uses a temperature value of 1 during response generation. All interactions with the model are conducted through API-based invocation with standardized prompts and input formats.

Framework Implementation. AutoMARS is implemented using LangGraph to support graph-based multi-agent orchestration. LangGraph provides the workflow structure for coordinating Manager Agent, Domain Agents, and Reflection Agent, while maintaining shared state across execution steps. It also supports parallel execution of independent agent branches and iterative routing for reflection-based revision.

5.2. Pairwise Experimental Results

This subsection presents pairwise comparative experiments to evaluate the effectiveness of AutoMARS in two-location risk assessment scenarios. We first provide a detailed case study to illustrate the end-to-end workflow and cross-domain reasoning process, followed by evaluations of result consistency and robustness across repeated runs.

5.2.1. Case Study

We present a case study of a comprehensive pairwise risk assessment between Paulding County, GA (FIPS: 13223) and Cobb County, GA (FIPS: 13067) to illustrate the application of AutoMARS. While Figure 1 shows the general multi-agent architecture, Figure 2 depicts the concrete execution flow of this specific case, in which the workflow follows the revision branch and is routed back from Reflection Agent to Manager Agent for iterative refinement.

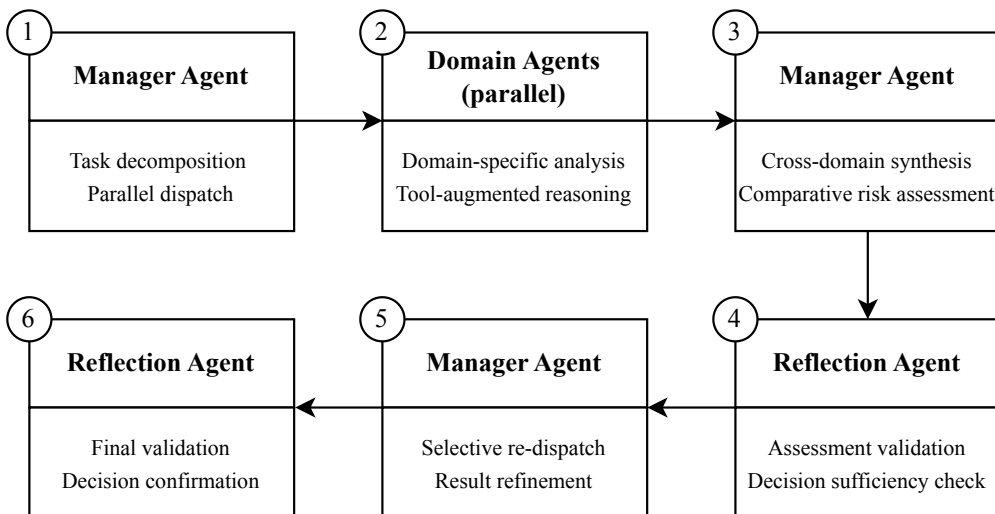


Figure 2: Workflow of AutoMARS case study for pairwise risk assessment.

²²<https://developers.openai.com/api/docs/models/gpt-5>

Manager Agent: Initial Planning and Task Decomposition. At the beginning of this run, Manager Agent produced a concrete execution plan and issued four routing directives in parallel. The plan explicitly specified the required evidence types for each domain (e.g., socioeconomic indicators, hazard and loss metrics, air quality and facility-based exposure, and crash statistics) and defined a synthesis strategy to align risk direction and compare relative differences across counties. These domain-level comparisons were then combined based on three factors: magnitude, consistency, and reinforcement. Magnitude reflects how large the differences are, consistency captures whether multiple domains point to the same conclusion, and reinforcement considers whether risks amplify each other across domains. Domains with stronger and more consistent indicators, especially in high-impact areas, carry more weight in the final judgment.

Domain Agents: Parallel Domain-Specific Analysis. Domain Agents execute in parallel to perform domain-specific analysis and generate structured evidence for comparative assessment. Each Domain Agent focuses on a distinct aspect of risk and applies tool-augmented reasoning by integrating authoritative datasets with LLM-based interpretation. All Domain Agents produce comparable, county-level evidence summaries that are aligned into a unified representation of directional risk indicators for subsequent cross-domain synthesis. This parallel execution ensures both efficiency and consistency in evidence generation across domains.

Census Agent. Table 2 highlights clear socioeconomic differences between two counties. Cobb County shows substantially higher overall exposure, reflected in its larger population scale and greater concentration of residents with multiple vulnerability factors. Its economic structure is more centralized, functioning as a regional employment hub with strong inward commuting patterns, whereas Paulding exhibits a smaller and more outward-oriented structure.

Table 2

Tools invoked by Census Agent and corresponding comparative socioeconomic indicators for Paulding and Cobb Counties.

Tool	Paulding County	Cobb County
Socioeconomic Profile Tool (2019-2023)	Population: ~174k; Labor Force: ~90k; Median Household Income: \$94.6k; Unemployment Rate: 4.46%	Population: ~769k; Labor Force: ~433k; Median Household Income: \$98.7k; Unemployment Rate: 4.55%
Community Resilience Tool (2023)	Population with 3+ Risk Factors: 12.13% (~22k)	Population with 3+ Risk Factors: 14.96% (~115k)
Labor Exposure Tool (2023)	In-county Jobs: ~29k; Commuting Pattern: Net outflow	In-county Jobs: ~414k; Commuting Pattern: Net inflow

Disaster Agent. Table 3 highlights key differences in hazard and disaster risk profiles between Paulding and Cobb Counties. Cobb County exhibits higher modeled multi-hazard risk and expected losses, along with lower community resilience, indicating a stronger baseline exposure to potential hazard impacts. In contrast, Paulding has experienced more frequent federally declared weather-related events in recent years, suggesting greater recent disaster occurrence. Observed severe weather impacts are broadly comparable in terms of economic losses, but differences emerge in event frequency and fatalities. Climate indicators further distinguish two counties, with Cobb showing conditions associated with greater heat-related exposure.

Environment Agent. Table 4 highlights clear differences in environmental exposure between Paulding and Cobb Counties. Cobb County shows observed air quality measurements, including PM2.5 and ozone levels, while no county-coded observations are available for Paulding during the study period. In addition, Cobb exhibits a substantially larger industrial and waste-related facility footprint, including more industrial plants, multiple landfills, and a major power plant, whereas Paulding has a more limited set of such facilities.

Traffic Agent. Table 5 highlights clear differences in transportation safety risk between Paulding and Cobb Counties. Cobb County exhibits a substantially higher level of traffic activity, reflected in a much larger number of reported crashes. In addition, severe or fatal incidents are observed in Cobb but not in Paulding during the study period, indicating the presence of higher-impact events. While average severity levels are broadly similar, the occurrence of more severe incidents in Cobb points to a more elevated overall safety risk profile.

Manager Agent: Cross-Domain Synthesis and Comparative Assessment. After receiving outputs from all Domain Agents, Manager Agent validates that the collected evidence is sufficient and consistent, and then performs a cross-domain synthesis integrating disaster, environmental, transportation, and socioeconomic indicators into a

Table 3

Tools invoked by Disaster Agent and corresponding comparative hazard and disaster risk indicators for Paulding and Cobb Counties.

Tool	Paulding County	Cobb County
Multi-Hazard Risk Index Tool	Composite Risk: 69.97; Expected Annual Loss: 79.52; Community Resilience: 44.12	Composite Risk: 94.37; Expected Annual Loss: 95.85; Community Resilience: 28.72
Disaster Declaration Tool (2020-Present)	6 declarations; includes Tropical Storm Zeta (2020), Hurricane Helene (2024), Winter Storm (2026)	4 declarations; includes Hurricane Helene (2024), Winter Storm (2026)
Severe Weather Tool (2023)	Events: 355; Damage: ~\$37.4M; Deaths: 2; Injuries: 8	Events: 319; Damage: ~\$37.3M; Deaths: 1; Injuries: 8
Climate Tool (2023)	Annual TMIN: ~10.2–10.6°C; Precipitation: ~1174–1179 mm	Annual TMIN: ~11.9°C; Precipitation: ~1080–1416 mm

Table 4

Tools invoked by Environment Agent and corresponding comparative environmental exposure indicators for Paulding and Cobb Counties.

Tool	Paulding County	Cobb County
Air Quality Tool (2023)	No county-coded AQS observations for PM2.5, Ozone, or CO	PM2.5: ~10.82 $\mu\text{g}/\text{m}^3$; Ozone: ~0.030 ppm
Environmental Facility Tool (2023)	~7 industrial plants; 1 landfill; ~1 chemical facility	~30 industrial plants; ~6 landfills; ~35 chemical facilities; includes a major natural gas power plant

Table 5

Tools invoked by Traffic Agent and corresponding comparative transportation safety indicators for Paulding and Cobb Counties.

Indicator	Paulding County	Cobb County
Traffic Accident Tool (2023)	Total Crashes: 13; Severe/Fatal Crashes: 0 (0%); Mean Severity: 2.00	Total Crashes: 606; Severe/Fatal Crashes: 18 (3.0%); Mean Severity: 2.06

unified comparative assessment. Across domains, the evidence consistently indicates higher overall risk in Cobb County. Disaster-related indicators suggest higher modeled hazard intensity and expected losses accompanied by lower resilience. Environmental analysis reveals elevated particulate pollution levels and a denser concentration of industrial and waste-related sources, indicating greater potential for chronic exposure. Transportation indicators show substantially higher crash frequency and the presence of severe or fatal incidents. Socioeconomic evidence further indicates a larger and more vulnerable population, increasing the potential impact of hazards and environmental burdens. Based on the combined magnitude and consistency of these indicators across domains, Manager Agent determines that Cobb County represents the higher-risk location relative to Paulding County.

Reflection Agent: Assessment Validation and Sufficiency Check. Upon receiving the integrated assessment, Reflection Agent evaluates whether the synthesized result is sufficiently supported for decision-making by examining domain coverage, structural completeness, and cross-domain consistency. During this process, minor inconsistencies are identified between modeled hazard risk metrics (e.g., NRI scores and expected annual loss) and interpretations based on disaster declarations and observed events. In addition, limitations are noted in the transportation safety analysis, which relies on a single-year snapshot and is not normalized by exposure factors such as population size or traffic volume. Although these issues do not invalidate the overall analytical framework, Reflection Agent determines that targeted refinement is necessary to improve the robustness of the supporting evidence. The workflow is therefore

routed back to Manager Agent with guidance to resolve inconsistencies in hazard interpretation through numerical comparisons and reassess the role of transportation indicators while preserving the overall analytical structure.

Manager Agent: Targeted Revision and Evidence Refinement. Following feedback from Reflection Agent, Manager Agent performs a targeted revision of the assessment while preserving previously validated evidence. In the disaster domain, categorical interpretations of FEMA National Risk Index (NRI) scores are removed, and the comparison is revised using only numerical indicators such as composite risk, expected annual loss, and community resilience. Although Paulding County reports more recent disaster declarations, Cobb County consistently exhibits higher numerical hazard risk indicators. In the traffic domain, Manager Agent acknowledges that crash statistics are based on a single-year snapshot and are not normalized by exposure factors such as population or traffic volume. As a result, transportation indicators are treated cautiously and excluded from primary decision weighting. The revised synthesis therefore places greater emphasis on more consistent evidence from socioeconomic, environmental, and hazard-related analyses, which collectively continue to indicate higher overall risk in Cobb County relative to Paulding County. The revised assessment is then forwarded to Reflection Agent for validation.

Reflection Agent: Final Validation and Acceptance. After receiving the revised assessment, Reflection Agent conducts a final validation and confirms that the previously identified issues have been sufficiently addressed, including inconsistencies in hazard interpretation and limitations in transportation data. The revised synthesis is supported by consistent cross-domain evidence, with socioeconomic exposure, modeled disaster risk, and environmental burden providing mutually reinforcing indicators. Based on the convergence and relative strength of the remaining evidence, Reflection Agent accepts the assessment and confirms that Cobb County exhibits higher overall risk relative to Paulding County. This conclusion is primarily driven by higher modeled hazard risk and expected losses, greater chronic environmental burden, and a substantially larger and more vulnerable population in Cobb County. The workflow is then terminated, indicating that the assessment is sufficiently robust for decision-making. To provide a holistic view of the execution process, Figure 3 presents the overall execution timeline of AutoMARS workflow.

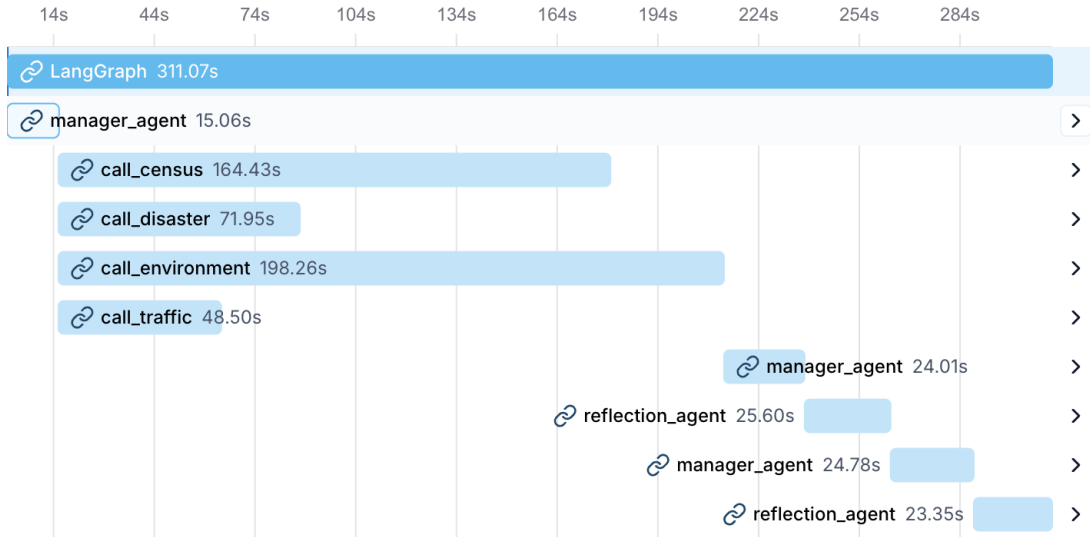


Figure 3: Execution timeline of AutoMARS for pairwise risk assessment.

5.2.2. Comparison Stability

We evaluate the stability of the pairwise comparative assessment by conducting repeated experiments for each county pair. Specifically, for each ordered pair (A, B), we perform 20 independent runs to measure the consistency of the model’s judgment. To examine the potential impact of input order, we also reverse the pair and conduct an additional 20 runs for (B, A). This design enables us to assess both the robustness of the comparative outcomes across repeated runs and the sensitivity of the model to prompt ordering.

We select three representative county pairs with varying levels of risk difference to conduct this analysis, including Fulton vs Cobb, Gwinnett vs Clayton, and DeKalb vs Henry. The results are summarized in Table 6. We observe that the comparative outcomes are highly consistent across repeated runs. For all three pairs, the same county is identified as more risky in nearly all cases, with most comparisons achieving 100% agreement and only minor variation observed in one pair. Importantly, reversing the input order does not change the overall conclusions, indicating that the model’s judgments are robust to prompt ordering. These findings demonstrate that the pairwise comparative framework produces stable and reliable risk assessments across repeated evaluations.

Table 6
Pairwise comparison stability across repeated runs.

County A	County B	More Risky	Percentage
Fulton	Cobb	Fulton	100%
Cobb	Fulton	Fulton	100%
Clayton	Gwinnett	Gwinnett	100%
Gwinnett	Clayton	Gwinnett	85%
DeKalb	Henry	DeKalb	100%
Henry	DeKalb	DeKalb	100%

5.3. Ranking Experimental Results

We evaluate AutoMARS under three experimental scenarios corresponding to different spatial scales within Atlanta metropolitan area, including Atlanta Urban Core (5 counties), ARC region (11 counties), and Metro Atlanta (29 counties). In all experiments, GPT-5 is used as the backbone LLM to support risk assessment and cross-domain reasoning. For each scenario, county rankings are derived based on pairwise comparative assessments.

5.3.1. Ranking Stability

This part evaluates the stability of ranking results across repeated runs under different spatial scales. For visualization purposes, the runs in each figure are arranged using a quantitative criterion that minimizes the total distance between consecutive rankings, where distance is defined as the sum of absolute position changes across counties. This ordering improves visual continuity by reducing abrupt transitions between adjacent runs, while leaving the underlying experimental results unchanged.

For Atlanta Urban Core (Figure 4a), rankings are highly stable across runs. Fulton County is consistently identified as the highest-risk location, while Clayton and Cobb Counties alternate as the lowest-risk locations. DeKalb and Gwinnett Counties remain in the middle positions with minor alternations, reflecting small variations between counties with relatively close risk levels. Overall, the ordering is highly consistent with only limited position swaps among intermediate counties.

For ARC region (Figure 4b), rankings remain largely stable as the number of counties increases. Most counties maintain consistent relative positions across runs, with only small variations observed. In particular, Clayton County exhibits slight fluctuations in its ranking, while the majority of other counties show stable ordering. This indicates that the framework preserves consistent comparative relationships even in moderately larger settings.

For Metro Atlanta region (Figure 4c), which represents the largest and most complex setting, rankings continue to exhibit strong stability at both extremes. The highest-risk and lowest-risk counties remain consistent across runs, while variations are primarily concentrated in the middle-ranked counties. These fluctuations can be attributed to the relatively small differences in risk levels among these counties, making them more difficult to distinguish consistently. Overall, the results demonstrate that AutoMARS produces robust and reliable ranking structures, with variability largely limited to cases where risk profiles are inherently similar.

5.3.2. Ranking Statistics

We further analyze the statistical properties of the ranking results by examining the average ranking and standard deviation across repeated runs, as shown in Figure 5. For Atlanta Urban Core (Figure 5a), the results exhibit clear and stable grouping patterns. Fulton County is consistently ranked as the highest-risk location with zero variance, indicating perfect stability. DeKalb and Gwinnett Counties form a middle group with similar average rankings and

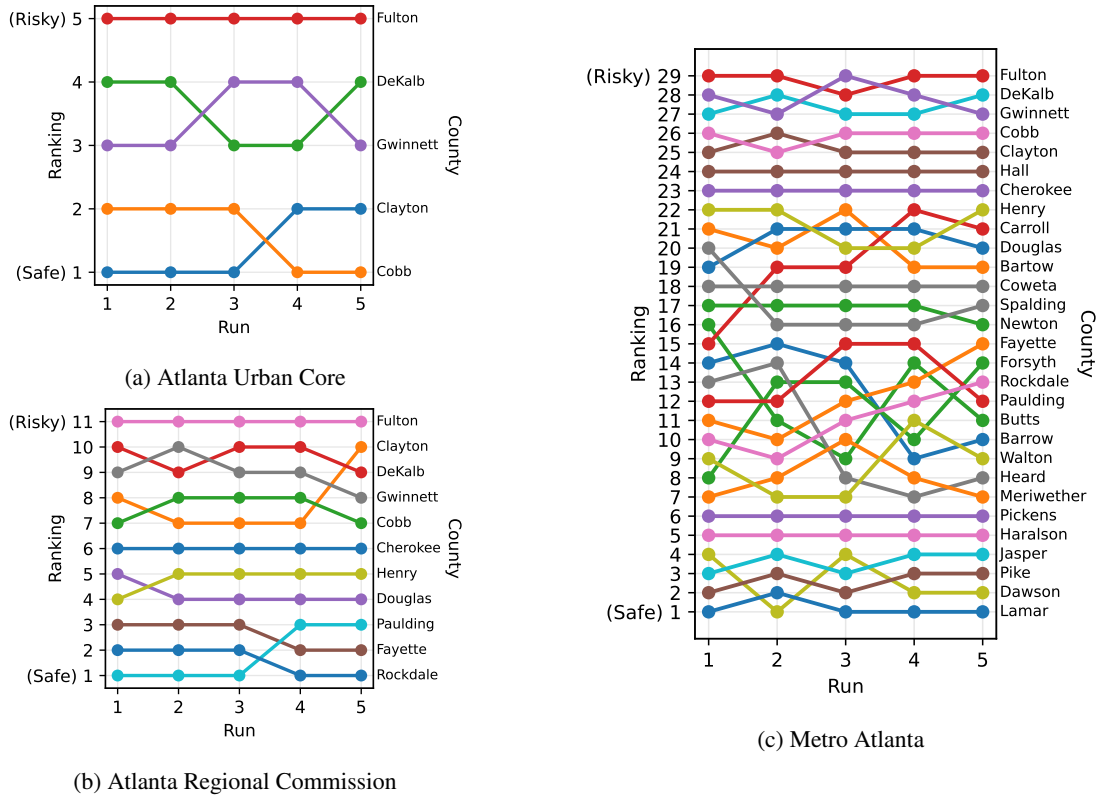


Figure 4: Ranking stability across repeated runs at different spatial scales.

small variations, while Clayton and Cobb Counties are consistently ranked as the lowest-risk locations with closely aligned mean positions. Overall, the small standard deviations confirm that ranking variability is minimal in this setting.

For ARC region (Figure 5b), the ranking statistics remain highly stable as the number of counties increases. Most counties exhibit standard deviations within one ranking position, indicating strong consistency across runs. Slightly larger variations are observed for a small subset of counties, such as Paulding and Clayton, but the overall ranking structure remains largely unchanged. In Metro Atlanta region (Figure 5c), a more structured pattern emerges despite the increased complexity. Counties can be broadly grouped into four tiers: the highest-risk and lowest-risk groups both show very small standard deviations and highly stable rankings, while the middle counties form two subgroups. The first subgroup (e.g., Meriwether to Paulding) remains relatively stable in the lower-to-middle risk range, whereas the second subgroup (e.g., Newton to Henry) occupies the middle-to-higher risk range with slightly greater variation. These patterns are consistent with the qualitative observations from the ranking results presented earlier.

To provide a spatial interpretation of ranking outcomes, Figure 6 visualizes the mean rankings across regions. The maps highlight geographic distribution of relative risk levels, where color intensity reflects average ranking positions. This visualization provides an intuitive overview of how risk is distributed spatially, complementing the statistical analysis with a geographic perspective. Risk assessment in this paper is not derived from a single metric, but is based on a comprehensive integration of more than a dozen indicators spanning socioeconomic, disaster, environmental, and traffic domains. This multi-indicator approach captures diverse aspects of exposure, vulnerability, and hazard, which can lead to noticeable differences in risk levels even between adjacent counties.

5.3.3. Model Configurations

Table 7 shows a highly consistent ranking pattern across five Atlanta Urban Core counties under different model configurations. Fulton consistently receives the highest risk ranking, followed by DeKalb and Gwinnett, while Cobb and especially Clayton tend to exhibit comparatively lower risk levels. Although minor ordering differences appear between configurations, the overall relative risk structure remains stable across all experiments. This indicates that

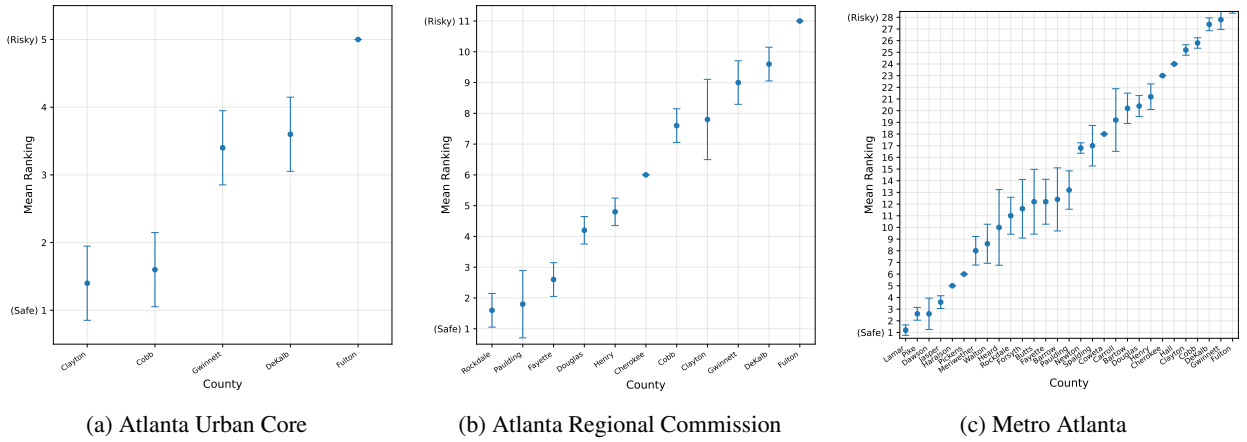


Figure 5: Average and standard deviation of county rankings across repeated runs.

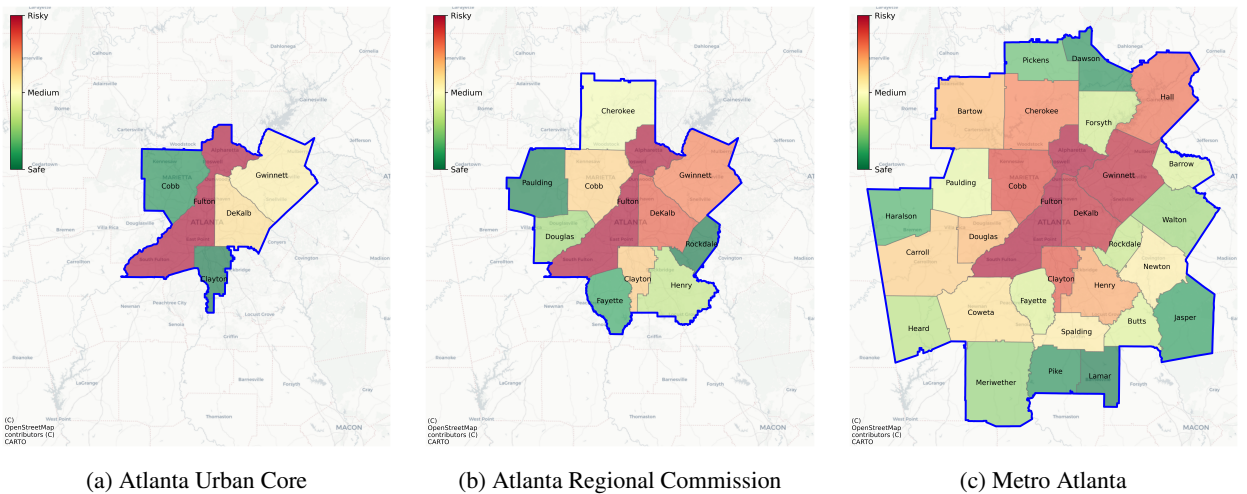


Figure 6: Spatial distribution of mean county rankings across Atlanta regions.

the framework captures persistent regional characteristics rather than producing rankings that are overly sensitive to individual model choices. In addition, the standard deviations are generally small across counties and model settings, with many values below 1.0 and several rankings showing zero variance. The limited variability suggests that the framework produces robust and repeatable outputs even when replacing the underlying LLMs with GPT-5, GPT-5 mini²³, or GPT-5 nano²⁴ variants. Overall, these results demonstrate that the multi-agent pipeline is relatively insensitive to model scale differences and maintains stable county-level prioritization across repeated runs.

Table 7
Performance comparison of end-to-end agent pipelines on Atlanta Urban Core with mean and standard deviation.

Manager	Domain	Reflection	Clayton	Cobb	Gwinnett	DeKalb	Fulton
GPT-5	GPT-5	GPT-5	1.4 ± 0.55	1.6 ± 0.55	3.4 ± 0.55	3.6 ± 0.55	5.0 ± 0.00
GPT-5 mini	GPT-5 mini	GPT-5 mini	1.8 ± 0.84	1.6 ± 0.89	2.6 ± 0.55	4.0 ± 0.00	5.0 ± 0.00
GPT-5 nano	GPT-5 nano	GPT-5 nano	1.0 ± 0.00	2.4 ± 0.55	2.8 ± 0.84	3.8 ± 0.45	5.0 ± 0.00

²³<https://developers.openai.com/api/docs/models/gpt-5-mini>

²⁴<https://developers.openai.com/api/docs/models/gpt-5-nano>

5.3.4. Computational Efficiency

Table 8 presents the impact of parallel execution on the overall runtime of AutoMARS framework. Both sequential and parallel settings were evaluated over 10 runs, and the reported results correspond to mean runtime and standard deviation. The results demonstrate that parallel execution significantly improves computational efficiency compared with sequential execution. By allowing independent domain-specific agents to operate concurrently, the framework effectively reduces idle waiting time and improves overall resource utilization. In addition, the relatively stable standard deviations across repeated experiments indicate that the framework maintains consistent and reproducible runtime performance under both execution settings. Overall, these findings highlight the effectiveness of asynchronous multi-agent coordination in improving the scalability and responsiveness of the proposed AutoMARS framework.

Table 8

Impact of parallel execution on overall runtime with mean and standard deviation.

Setting	Total Runtime (s)
Sequential execution	622.11 $\pm$ 59.84
Parallel execution	327.49 $\pm$ 64.70

Table 9 presents the impact of caching mechanism on tool invocation efficiency, using Air Quality Tool and Environmental Facility Tool as representative examples. For both caching and non-caching settings, the same query parameters were executed five times, and the reported values correspond to the average invocation time. The results show that caching substantially reduces repeated query latency for both tools. In particular, Air Quality Tool decreases from 76.68 seconds without caching to near-zero latency with caching, while the Environmental Facility Tool decreases from 37.98 seconds to 0.58 seconds on average. These results demonstrate that caching mechanism can effectively reduce redundant external API calls and significantly improve the computational efficiency of the framework.

Table 9

Impact of caching mechanism on tool invocation time.

Setting	Tool Invocation Time (s)
Air Quality Tool w/o caching	76.68
Air Quality Tool w caching	0.00
Environmental Facility Tool w/o caching	37.98
Environmental Facility Tool w caching	0.58

6. Conclusion

In this paper, we propose AutoMARS, an automatic multi-agent framework for scalable and comparative supply chain risk assessment. The framework enables end-to-end execution of analytical workflows by coordinating multiple specialized agents, integrating heterogeneous risk information across domains, and invoking external analytical tools. By combining structured task planning, multi-agent orchestration, and reflection-based validation, AutoMARS provides a systematic and interpretable approach to comparative risk evaluation. In addition, the framework supports parallel execution and a comparison-based sorting mechanism, enabling efficient analysis across multiple locations without relying on predefined scoring schemes.

Experimental results demonstrate effectiveness, robustness, and scalability of the proposed framework. Pairwise case study illustrates the ability of AutoMARS to synthesize cross-domain evidence and produces consistent comparative assessments. Stability analyses show that the framework generates reliable results across repeated runs, with strong consistency in both pairwise comparisons and ranking outcomes. Furthermore, ranking experiments across multiple spatial scales reveal that AutoMARS maintains stable ordering structures, with variability primarily limited to locations with similar risk profiles. Although adjacent counties appear geographically continuous, the observed red-green differences reflect underlying heterogeneity captured by more than a dozen indicators across socioeconomic, disaster, environmental, and traffic domains. Variations in population scale, vulnerability, economic structure, and exposure conditions can lead to substantial differences in risk even between neighboring counties. These boundaries

therefore represent aggregated multi-dimensional differences rather than isolated factors. Nevertheless, because the analysis is conducted at the county level, neighboring counties may occasionally exhibit sharp local contrasts due to administrative boundary granularity and heterogeneous domain indicators. Decision-makers should therefore interpret localized ranking differences with appropriate caution.

Despite the strengths discussed above, some limitations remain. The current framework relies on manually designed prompts for LLM-based agents, making the results sensitive to prompt formulation and potentially affecting consistency across different runs and settings. Future work could address this limitation by exploring automatic prompt tuning techniques to improve the robustness and consistency of LLM-based agents. By adapting prompts based on data characteristics and task requirements, the framework can achieve more stable and reliable performance across different settings.

Overall, AutoMARS provides a unified and extensible framework for automated, multi-domain risk assessment in complex supply chain environments. By shifting from score-based evaluation to comparison-based reasoning, the approach improves the robustness and interpretability of risk assessment outcomes, particularly in settings involving LLM-based analytical processes. The proposed framework demonstrates the potential of multi-agent LLM systems to support scalable and structured decision-making, offering a practical foundation for risk-aware location evaluation in real-world supply chain applications.

References

- Ambati, K.C., 2025. An event-driven architecture for autonomous supply chain risk detection and decision automation. *International Journal of Computer Technology and Electronics Communication* 8, 1202–1211.
- Aqlan, F., Lam, S.S., 2015. A fuzzy-based integrated framework for supply chain risk assessment. *International journal of production economics* 161, 54–63.
- Auger, N., Jugé, V., Nicaud, C., Pivoteau, C., 2018. On the worst-case complexity of timsort. *arXiv preprint arXiv:1805.08612*.
- Babu, H., Yadav, S., 2023. A supply chain risk assessment index for small and medium enterprises in post covid-19 era. *Supply Chain Analytics* 3, 100023.
- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al., 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., et al., 2025. Llm instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255.
- Cerabona, T., Benaben, F., Montreuil, B., Luras, M., Faugère, L., Campos, M.R., Jeany, J., 2024. The physics of decision approach: a physics-based vision to manage supply chain resilience. *International journal of production research* 62, 1783–1802.
- Chen, G.H., Chen, S., Liu, Z., Jiang, F., Wang, B., 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.
- Chen, J., Lu, Y., Wang, X., Zeng, H., Huang, J., Gesi, J., Xu, Y., Yao, B., Wang, D., 2025. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. *arXiv preprint arXiv:2507.21028*.
- Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., et al., 2024. Chatbot arena: An open platform for evaluating llms by human preference, in: *Forty-first International Conference on Machine Learning*.
- Chowdhury, P., Kam, B.H., Teoh, S.Y., Childerhouse, P., 2026. Navigating through storms: A review of supply chain resilience practices. *International Journal of Production Economics*, 110001.
- Chu, C.Y., Park, K., Kremer, G.E., 2020. A global supply chain risk management framework: An application of text-mining to identify region-specific supply chain risks. *Advanced Engineering Informatics* 45, 101053.
- Das, S.K., Perona, M., 2025. Supply chain risk management automation: A literature review. *Electronic Markets* 35, 104.
- Fan, H., Li, G., Sun, H., Cheng, T., 2017. An information processing perspective on supply chain risk management: Antecedents, mechanism, and consequences. *International Journal of Production Economics* 185, 63–75.
- Ganesh, A.D., Kalpana, P., 2022. Future of artificial intelligence and its influence on supply chain risk management—a systematic review. *Computers & Industrial Engineering* 169, 108206.
- Giannakis, M., Louis, M., 2011. A multi-agent based framework for supply chain risk management. *Journal of Purchasing and Supply Management* 17, 23–31.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al., 2024. A survey on llm-as-a-judge. *The Innovation*.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X., 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Ho, W., Zheng, T., Yildiz, H., Talluri, S., 2015. Supply chain risk management: a literature review. *International journal of production research* 53, 5031–5069.
- Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., et al., 2023. Metagtpt: Meta programming for a multi-agent collaborative framework, in: *The Twelfth International Conference on Learning Representations*.
- Ivanov, D., 2022. Viable supply chain model: integrating agility, resilience and sustainability perspectives—lessons from and thinking beyond the covid-19 pandemic. *Annals of operations research* 319, 1411–1431.

- Ivanov, D., Dolgui, A., 2021. A digital supply chain twin for managing the disruption risks and resilience in the era of industry 4.0. *Production Planning & Control* 32, 775–788.
- Ivanov, D., et al., 2018. Structural dynamics and resilience in supply chain risk management. volume 265. Springer.
- Jackson, I., Ivanov, D., Dolgui, A., Namdar, J., 2024. Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. *International Journal of Production Research* 62, 6120–6145.
- Katsaliaki, K., Galetsi, P., Kumar, S., 2022. Supply chain disruptions and resilience: a major review and future research agenda. *Annals of operations research* 319, 965–1002.
- Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al., 2021. Dynabench: Rethinking benchmarking in nlp, in: *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124.
- Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al., 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791.
- Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B., 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36, 51991–52008.
- Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y., 2024a. Llm-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.
- Li, X., Wang, S., Zeng, S., Wu, Y., Yang, Y., 2024b. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinearth* 1, 9.
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al., 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Liu, Z., 2025. Autotbn: Multi-agent incident response with large language models, in: *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE. pp. 1–6.
- Liu, Z., Nguyen, N.H., Quan, Y., Zhang, S.X., 2025. Temporal tokenization strategies for event sequence modeling with large language models. *arXiv preprint arXiv:2512.13618*.
- Liu, Z., Quan, Y., 2024. Tpp-llm: Modeling temporal point processes by efficiently fine-tuning large language models. *arXiv preprint arXiv:2410.02062*.
- Liu, Z., Quan, Y., 2025a. Econwebarena: Benchmarking autonomous agents on economic tasks in realistic web environments. *arXiv preprint arXiv:2506.08136*.
- Liu, Z., Quan, Y., 2025b. Retrieval of temporal event sequences from textual descriptions, in: *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 37–49.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoy, S., Yang, Y., et al., 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36, 46534–46594.
- Montreuil, B., 2011. Toward a physical internet: meeting the global logistics sustainability grand challenge. *Logistics Research* 3, 71–87.
- Montreuil, B., Meller, R.D., Ballot, E., 2012. Physical internet foundations. *IFAC Proceedings Volumes* 45, 26–30.
- Oger, R., Bénaben, F., Lauras, M., Montreuil, B., 2018. Towards decision support automation for supply chain risk management among logistics network stakeholders. *IFAC-PapersOnLine* 51, 1505–1510.
- Quan, Y., Li, X., Chen, Y., 2025a. Crmagent: A multi-agent llm system for e-commerce crm message template generation. *arXiv preprint arXiv:2507.08325*.
- Quan, Y., Liu, Z., 2024a. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. *arXiv preprint arXiv:2405.07938*.
- Quan, Y., Liu, Z., 2024b. Invagent: A large language model based multi-agent system for inventory management in supply chains. *arXiv preprint arXiv:2407.11384*.
- Quan, Y., Liu, Z., Benaben, F., Montreuil, B., 2026. Leveraging large language models to enhance multi-agent risk assessment in supply chain networks. *International Journal of Production Research*, 1–29.
- Quan, Y., Xu, Y., Chen, G., Benaben, F., Montreuil, B., 2025b. Leveraging large language models for risk assessment in hyperconnected logistic hub network deployment. *arXiv preprint arXiv:2503.21115*.
- Saleem, A., 2020. Automated inventory management systems and its impact on supply chain risk management in manufacturing firms of pakistan. *International Journal of Supply Chain Management* 9, 220–231.
- Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., Leike, J., 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36, 68539–68551.
- Sharma, A., 2022. The future of procurement: Integrating ai, automation, and blockchain for next-generation supply chains. *International Journal of Artificial Intelligence, Data Science, and Machine Learning* 3, 1–10.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y., 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36, 38154–38180.
- Sodhi, M.S., Son, B.G., Tang, C.S., 2012. Researchers' perspectives on supply chain risk management. *Production and operations management* 21, 1–13.
- Szymanski, A., Ziems, N., Eicher-Miller, H.A., Li, T.J.J., Jiang, M., Metoyer, R.A., 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, in: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 952–966.
- Tan, S., Zhuang, S., Montgomery, K., Tang, W.Y., Cuadron, A., Wang, C., Popa, R.A., Stoica, I., 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

- Thakur, A.S., Choudhary, K., Ramayapally, V.S., Vaidyanathan, S., Hupkes, D., 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, in: *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 404–430.
- Tran, K.T., Dao, D., Nguyen, M.D., Pham, Q.V., O’Sullivan, B., Nguyen, H.D., 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.
- 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., Wang, C., 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations, in: *First Conference on Language Modeling*.
- Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., Wu, Y., 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*.
- Yue, X., Wang, B., Chen, Z., Zhang, K., Su, Y., Sun, H., 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.
- Zhang, X., Yu, B., Yu, H., Lv, Y., Liu, T., Huang, F., Xu, H., Li, Y., 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.
- Zhao, M., Hussain, O., Zhang, Y., Saberi, M., 2024. Optimizing supply chain risk management: An integrated framework leveraging large language models, in: *2024 IEEE Conference on Artificial Intelligence (CAI)*, IEEE. pp. 1057–1062.
- Ziari, M., Taleizadeh, A.A., 2025. Integrated data-driven and artificial intelligence framework to develop digital twins in distribution system of supply chains: A real industrial case. *International Journal of Production Economics*, 109743.