Multi-Agent LLM System Safety

Erik Nordby*
OMSCS

Derek Lowlind
OMSCS

Vladimir Lukin
OMSCS

Shaikat Islam
OMSCS

Echo Cho
OMSCS

Abstract

As Large Language Models (LLMs) continue to proliferate throughout society and gain ever more popularity, the risks posed by Multi-Agent Systems (MAS) of LLMs should be urgently studied. Although individual Single Agent System (SAS) LLMs have been more thoroughly researched, the collective behavior of MAS LLMs poses novel risks such as collusion, feedback loops, and emergent properties. We have conducted a systematic review of the literature to assess current research directions and potential gaps that need to be addressed in MAS LLMs.

Keywords

LLM, Language Model, Agent, Multi-Agent, Multi Agent, Safety, Security

1 Introduction

LLMs have recently achieved sufficient performance to work in Multi-Agent Systems (MAS) and display potential for agent-like abilities [42]. While significant work has been done toward improving the performance of these systems and allowing LLMs to collaborate, there has been less work done toward evaluating potential safety concerns with the alignment of these systems.

Emergent behavior, runaway feedback loops, and collusive behaviors have arisen in other important multi-agent systems [29] and could very well arise in systems of LLMs [20]. With LLM agents being rolled out to the internet [69], we must understand these dynamics and make progress toward mitigating the risks.

We aim to assess the current state of the field in order to identify potential risks and challenges presented by multi-agent LLMs. Further, we aim to identify current mitigation strategies for these risks. In order to accomplish this, we have conducted a Systematic Literature Review using the PRISMA guidelines, reviewing and extracting information from relevant papers. Our hope is that this provides context for the main gaps in current research and potential trends within this new field.

2 Background

2.1 Agents and Multi-Agent Systems

While various authors have different specific definitions of agents, they can broadly be defined as a system that can achieve objectives on behalf of a user [39]. For example, a Roomba can independently navigate a home in the pursuit of cleaning floors. Some authors further emphasize other attributes of agents, such as acting autonomously [61] or the ability of agents to interact with their environment.

For our purposes, the definition used in "Intelligent Agents: Theory and Practice" is most useful. It highlights that agents should be autonomous, capable of social interaction, capable of interacting with their environment, and proactive in their actions [61]

Multi-agent systems are a collection of agents that are able to interact [60]. Software-based flocking simulations like Boids algorithm [47] or eusocial animals like ants [15] are some classic examples of these systems. These systems may present interesting new system-level dynamics and behaviors that are not present in any individual agent. Some of these examples include the emergence of social conventions [54] and the emergence of organs from individual cell interactions [43].

2.2 Large Language Model Agents

As LLMs continue to improve, they've shown usefulness in contexts outside of simple chat bots. Companies such as OpenAI and Anthropic have recently developed scaffolding that allows these language models to "use" tools, including searching the web, running code, editing files directly on user's personal computers, and interacting with other LLMs [41] [2]. The ability to use tools, pursue goals autonomously, and interact with other LLMs means that these new LLM-powered applications fit the definition of agency.

So, when multiple LLMs interact with one another in an environment (i.e., multiple LLM-powered bots on social media or an application that uses LLM agents in a round-robin discussion [62]), this constitutes a Multi-Agent system of LLMs which we will refer to as LLM MAS for the remainder of this paper.

2.3 AI Alignment, Safety, and Security

As AI systems and language models, in particular, have gained increasing popularity, considerable work has been done to ensure that they're safe, secure, and aligned.

In the existing literature, safety, security, and alignment are sometimes used interchangeably and may have loose definitions. However, for the purposes of this paper, we will be using the definitions used in "Foundational Challenges in Assuring Alignment and Safety of Large Language Models" [3]. They define:

2.3.1 Safety: As a system not causing accidental harm. While a broad definition, this captures the various ways that safety can manifest. Some examples specific to AI systems include preventing hallucinations, avoiding sycophantic behaviors, and increasing robustness to adversarial attacks. This definition is also used by the book "Engineering a Safer World: Systems Thinking Applied to Safety" [32] .

2.3.2 Alignment: As the AI system behaving as intended by the developers or users of the system. For example, a Roomba that simply

*All authors contributed equally to this research.

sweeps is aligned. A Roomba that tries to destroy a house because "without a house, the house can't be dirty" would be unaligned.

2.3.3 Security: As systems that are robust to malicious actors breaking the system or manipulating it to cause harm. For example, an AI system susceptible to prompt injection is not secure. However, an AI system that hallucinates under normal use may be secure, even though it is not safe.

Given that advanced AI systems are relatively new, the field of AI Safety is rapidly evolving and shifting. However, some of the key themes and current directions of research towards technical AI safety include monitoring with techniques like interpretability, robustness with techniques like adversarial training, and alignment with techniques like scalable oversight [25]. Similar work is being done towards ensuring that AI systems are effectively governed and ethically deployed [46].

3 Related Works

Multi-agent AI safety and related fields have moved with incredible speed over the course of the past few years. So, numerous literature reviews and surveys have been written surrounding these ideas. Below is a collection of various literature reviews that cover fields and questions closely related to those we are looking to review.

3.0.1 LLM Safety: Considerable effort has been undertaken to make AI safer and more robust. So, there have been numerous literature reviews, surveys, and books written about the field of AI safety in general. Some of these focus specifically on large language models [3] [49]while others focus on the safety of AI systems more broadly[12] [27] [13] [25]

Other surveys have similarly been written about sub-fields of AI safety or closely related areas of study. Some focuses of these include interpretability [6], evaluations [10], and alignment [58].

3.0.2 LLM MAS:. There are also recently written survey papers and literature reviews that also focus on LLM MAS without focusing specifically on the safety concerns of those systems. Some of these focus on LLM MAS broadly [33] [21], while others focus on specific aspects of LLM MAS like specific applications [22] or progress being made towards reinforcement learning in LLM MAS [52].

3.0.3 LLM MAS Safety: To our knowledge, there has been no systematic literature review that has covered LLM MAS specifically. However, literature reviews and survey papers have aimed to answer closely related questions. Due to the rapid speed with which the field is currently moving, those papers were published in the short window of time from the beginning of our project to now.

Specifically "Multi-Agent Risks from Advanced AI" [20] overlaps considerably with the questions we look to answer. It was written in collaboration between numerous major labs and organizations working on multi-agent AI safety and serves as a great introduction to the field. Broadly, it identifies three main failure modes: miscoordination, conflict, and collusion. It further lists a variety of risk factors, including selection pressures, information asymmetries, and network effects like coordinated collapse.

Similarly, "A Survey on Trustworthy LLM Agents: Threats and Countermeasures" [65] looks to create a taxonomy of LLM agents and their risks. Only a small portion of their paper was dedicated to "Agent-to-Agent" risks. They identified two main types of attacks which are currently risks. First is cooperative attacks, where malicious agents are inserted into a MAS to disrupt its normal functioning. Second is infectious attacks, which look to make the benign agents spread malicious information or prompts.

Conversely, they also found two broad types of mitigation strategies. First is Collaborative Defense which uses the collaborative nature of the LLM MAS itself to drive better safety through things like debate between the agents. Secondly is Topological Defense which modifies the topology of the system to create safer connections. Our goal is to quantitatively measure the current state of the field and estimate the current level of effort and progress being made towards different goals.

4 Research Questions

4.0.1 RQ1. Safety/Security Risks and Attacks: What risks and potential attacks have been demonstrated in multi-agent LLM systems? What risks and potential attacks have been theorized to exist?

4.0.2 RQ2. Mitigation Strategies: What mitigation strategies have been developed to combat security and safety risks in LLM MAS?

5 Methodology

5.1 Inclusion and Exclusion Criteria

5.1.1 Timeline.

Criteria: We chose to include only papers which were disseminated after 2022.

Justification In mid-2022, GPT-3.5 was released. The commercialized version of it, ChatGPT, was released in November of that same year. This coincided with an enormous increase in the literature relating to LLMs [42]. Prior to this significant rise in literature and resources, language models did not have sufficient capabilities to act in multi-agent settings. This is shown in papers like "Cooperate or Collapse," where they found that only state-of-the-art models like GPT-4 could successfully manage resources in multi-agent resource extraction games [44]. So, papers before 2022 could not empirically test the dynamics we plan to investigate and understand.

Further, MAS systems did not become mainstream until after 2022. Examples include Nvidia's "Introduction to LLM Agents" blog released November 2023 [38]. Microsoft research's Autogen framework for building MAS was released September 2023 [45]. In addition, OpenAI released Swarm, adding MAS abilities in October 2024 [7]. Finally, the Amazon Bedrock service announced MAS capabilities in December 2024, with general availability in March 2025 [4].

5.1.2 Language.

Criteria: We included papers that were written in English. We excluded papers in all languages except English.

Justification: We only included papers that were written in English due to practical considerations and resource limitations for the team. While papers from other languages provide great value to the field, highly technical papers require significant depth and a nuanced understanding of their text. Given the potential for mistranslation by existing tools, we decided we could not achieve a

sufficiently nuanced understanding of papers written in languages other than English.

English is the *lingua franca*, or common language between speakers with different native languages, of academic research. This commonality allows researchers from all over the world who speak different native languages to learn at most one additional language to participate in the world of academic research. "More than 90 % of the indexed scientific articles in the natural sciences been published in [English]" [14].

5.1.3 Handling Published and Unpublished Manuscripts.

Criteria: We decided to include both published and unpublished manuscripts.

Justification: Given that the field of natural language processing is moving at an incredibly fast rate, many landmark papers are still technically pre-prints. For example, the paper introducing Llama 3 [18], which is arguably one of the most impactful LLM-related papers of 2024, is still a pre-print. Since the field of multi-agent systems of LLMs is an even newer sub-field, with many of the papers not being published until mid to late 2024, it would be challenging to restrict our search only to include research that has already been published.

To prevent low-quality or incomplete papers from adding noise to the results of peer-reviewed papers, we have decided to use a tiered/weighted approach in the analysis of the papers. In the top tier, we will exclusively be pulling from commonly used databases (e.g., ACM). Although these sources will yield significantly fewer results, we can be more confident in the quality of the papers contained within them.

In the second tier, we will be pulling papers from arXiv. arXiv allows for pre-print and non-peer-reviewed papers (i.e., grey literature) to be included in their database and rapidly disseminated. So, we cannot confirm the quality of the included papers. However, given the rapid pace and newness of this research, we believe that arXiv will provide an important context for the future of the field.

This two-tiered approach will allow for both established and current progress to be reviewed while also providing a comprehensive overview of what themes and research directions may be pursued in the future.

5.1.4 Study Design and Setting.

Criteria: We included papers that were either experimental or observational studies. We excluded qualitative discussions about the impacts of MAS.

We did not place explicit criteria on the study settings. So, we included papers regardless of the setting. However, given the other inclusion/exclusion criteria, only simulations, theoretical analysis, or meta-analysis were available. There were no observational studies or discussions of multi-agent LLMs in the wild.

Justification: We wanted to ensure that the risks and mitigation strategies we gathered during the literature review were rooted in empirical findings or solid theoretical frameworks. So, we did not include papers that were purely speculative.

5.1.5 Study Focus and Contents.

Criteria: We only included papers that had the following attributes:

- A focus on multi-agent systems
- LLMs as the agents in the multi-agent systems
- A focus on safety, risks, and security
- Either a discussion of risks or strategies to avoid risks
  - Experimentally test risks from LLM MAS
  - Experimentally test mitigation strategies for risks

We excluded papers with the following attributes and focuses:

- Focus on the performance of LLM MAS without regard to safety
- Focus on single-agent safety of LLMs
- MAS which do not include LLMs (e.g., multi-agent reinforcement learning)
- Using LLM MAS for security instead of focusing on the security of the MAS itself (e.g., using LLMs to prevent DDoS attacks).
- Focus on security issues that are specific to single-agent LLMs
- Software-based vulnerabilities (e.g., poor security at large industry labs)
- Non-Experimental or quantitative observational results

Justification: Individually, the multi-agent, security, and LLM fields encompass an incredibly vast array of literature. By focusing only on papers that exist at the intersection of those three fields, we can eliminate a significant amount of noise from our collection of papers.

5.2 Search Procedure/Search String

5.2.1 Source Searching Methods. Each source was last searched on March 13th. For each source, we only accepted papers after 2022. We did not use other methods besides querying databases to search for the papers. Therefore, we did not use other methods, such as forward/backward citation gathering. Although potentially useful, we decided that forward/backward citation gathering would introduce a significantly larger scope in the number of papers we would be reviewing without providing a proportionate increase in the number of relevant papers. We deemed the papers found by forward/backward citation gathering to be less relevant than those found by our database query because the forward/backward search is much more likely to find grey literature and nonrelated papers that do not look at MAS systems. For example, an industry blog post from Nvidia on prompt injection or a paper on prompt injection for Single Agent Systems (SAS) can be great resources, but not in our context of MAS papers. Instead, we believe it's better to include more resources from another centralized database.

We used the ACM Guide to the Computing Literature to search the peer-reviewed materials and used arXiv for preprints.

The ACM Guide to the Computing Literature combines several types of materials (books, theses, proceedings, periodicals, journals) from different sources (ACM, IEEE, Springer, USENIX, etc.), while arXiv provides access to preprints, many of which ultimately published. Adding arXiv as a source allowed us to track emerging research directions and analyze the latest results.

After using our search string query in the relevant databases, we used a two-phase filtering process to compile our final list of papers. We first screened papers by their title and abstract against both our search string and our inclusion/exclusion criteria. Then,

we reviewed the full text of the remaining papers to verify that they met our criteria.

5.2.2 Search String. The search string that we decided to use was broken down into 3 main sections:

- ("multi agent" OR "multi-agent" OR "multi-agent" OR "based agent") - This ensured that the systems of LLMs contained agents. The "based agent" part covers "AI-based agents".
- (LLM or "language model") - This ensured that only studies focusing on LLMs were included. Work on multi-agent systems more broadly is significantly more thoroughly explored
- ("safe" OR "safety" OR "security" OR "secure" OR "alignment" OR "risk" OR "attack" OR "defense" OR "adversarial") - This ensured we cover a wide range of safety and security issues

5.2.3 Division of Labor for Reviews. Given that this is a rapidly evolving field and the number of papers is quite high, we decided to divide the papers evenly between the group members. We also calibrated our agreement by having each member review the other's work.

5.3 RQ/Data Extraction

5.3.1 Data Extraction. In order to extract the relevant information for each of the research questions, we noted the following data for each of the papers which were accepted in the exclusion/inclusion criteria

- Source
- Title
- Authors
- Type (Conference, Journal, or arXiv)
  - If the paper is reviewed or grey literature
- Number of Citations
- Risk(s) identified for LLM MAS
  - Whether the risks were hypothetical or empirically demonstrated. Hypothetically demonstrated risks are considered future work opportunities
- Mitigation strategies proposed for LLM MAS
  - Whether the mitigation(s) were hypothetical or empirically demonstrated. Hypothetically demonstrated mitigation(s) are considered future work opportunities
- Other future work opportunities mentioned, not caught in the risk/mitigation data extraction

The risks, mitigation strategies, and Future Work Opportunities (FWO) were extracted from the papers each time that they were mentioned. While a strictly formulaic approach was not employed for identifying risks and mitigation strategies, one out of every 10 papers was reviewed by a second author to ensure consistency and inter-coder reliability. Ambiguous cases were each discussed between two authors.

The list of risks was aggregated by risk theme and counted to answer RQ1. Similarly, the list of mitigations was aggregated by mitigation theme and counted to answer RQ2.

5.3.2 Pre-filtering. We performed pre-filtering based on paper meta-information (title, keywords, abstract, highlights) before reading the full paper. We did this in two steps: firstly, we processed the meta-information for each paper with LLMs, and then we reviewed the results.

We used the following steps to pre-filter data with LLMs before reviewing it manually. We first downloaded publications' meta-information (title, abstract, keywords, highlights), then fed that information to three LLMs, and filtered out the results that all three LLMs considered irrelevant. Of those removed, we randomly sampled 50 papers and reviewed them to verify the reliability of this filtering step.

For the models, we used mistral-large-2411, gpt-4o-2024-11-20, and claude-3-7-sonnet-20250219 since those models provided the best quality for a small price and were easy to use. The full prompt can be found in the appendix.

5.3.3 Source Gathering. We used the arXiv API to download title and paper summary. ACM data gathering was more complex. The ACM BibTeX citations contain basic publication information like authors, titles, and abstracts. Everyone can download citations for all papers for a given ACM search query in a single batch, and it is one of the common methods for automated data gathering. However, those citations do not contain some important data. Specifically, it does not include how many times the publication was cited and "highlights" explaining why this publication ended up in search results. Further, some publications do not contain abstracts, but almost all of them have highlights. We decided to add highlights so LLMs have more context about papers and our pre-filtering process is more accurate. However, since ACM does not provide a way to download all publication information, including highlights, so we used the selenium web scraping library to scrape those data.

After preliminary filtering, we extracted information from the publications.

In summary, the data extraction process can be explained as follows: 1) The author reads each of their assigned papers and extracts the risks, mitigations, and Future Work Opportunity (FWO) data. 2) The author fills in table with the subject, supporting quote(s), and associated theme for each respective paper and risk. 3) In order to check consistency, a second author reviews the table of quotes and compares the theme chosen by the original author. If the themes were not the same, the two authors reviewed the subject until consensus was achieved.

The themes were generated using an iterative process like code generation in grounded theory analysis. Each author kept a separate list of themes. The themes were compared, refined, and consolidated at check-in points. First, the 10th reviewed paper and then the 30th reviewed paper. The check-in windows became larger since less new codes appeared and the previous codes were more refined.

As mentioned before, since the topic of MAS is currently so new, at the time of writing, unreviewed grey literature was included to provide information on new emerging trends. Grey literature includes preprints, technical reports, white papers, blog posts, theses, and industry reports.

6 Results

Below are the results of our search results including the data for the number of paper accepted and the results of the data extraction

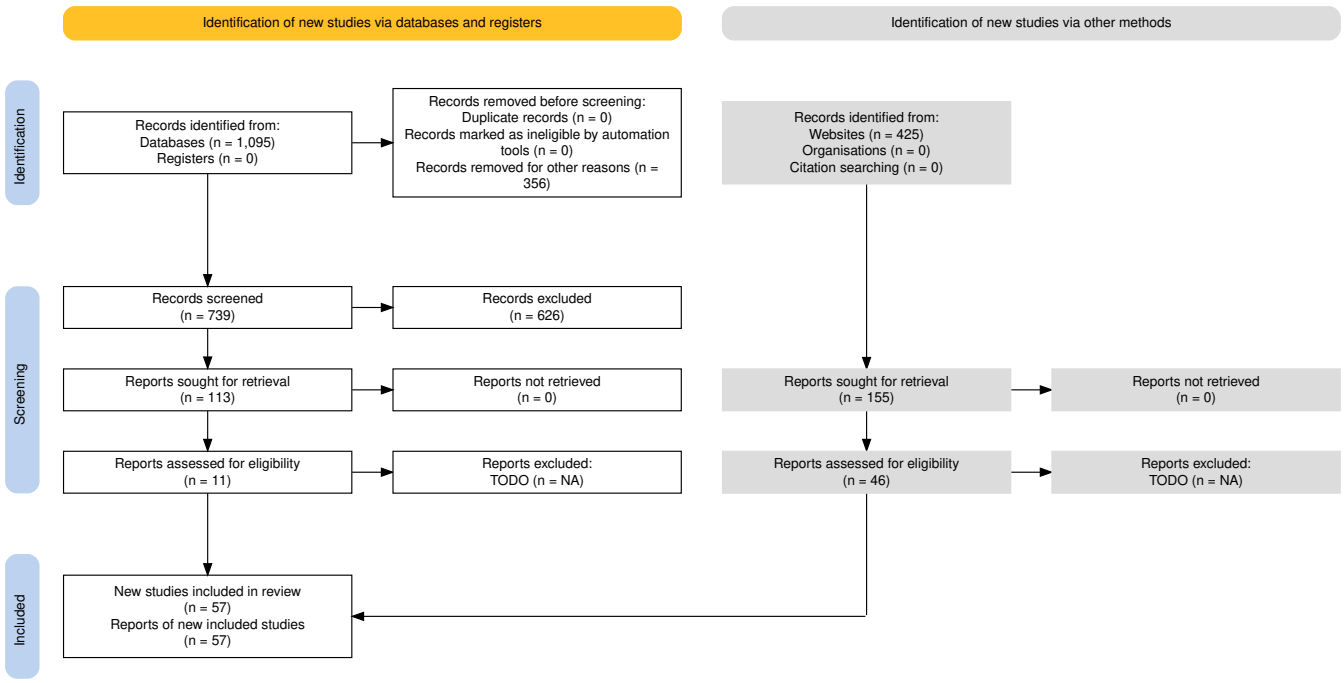

Figure 1: Prisma flowchart

6.1 Search Results

After using the search string to search both arXiv and ACM, we had 1520 papers, 425 from arXiv, and 1095 from ACM. After running each paper through the LLM pipeline, we had 262 papers left, 113 from arXiv and 149 from ACM. Hand-checking 50 of the papers that all of the LLMs rejected yielded zero false negatives.

Following the automated portion of our search, we manually filtered the remaining 262 papers. Of the papers we have reviewed, 41% of the arXiv papers (37 out of 90) and 18% of the ACM papers (13 out of 74) were accepted. So, we have seen significantly more papers included from arXiv than from ACM.

One example of a paper that proved challenging to classify was AgentDojo. This paper detailed how LLM agents are susceptible to prompt injection and proposed methods to mitigate such attacks. Prompt injection cascades were among the risks we identified for LLM MAS, and this technique could help reduce that risk. However, we decided that we would not include this paper since it was focused on single agents. Adversarial robustness techniques (like the one proposed by AgentDojo) are a major topic of study, so by including this paper, we would have vastly increased the scope of our project. Ultimately, prompt injection defenses were mentioned in the papers we included, so this mitigation strategy was still covered.

A similar paper that proved at first challenging to classify was "ALU: Agentic LLM Unlearning" [50]. This used an LLM MAS to increase the safety of LLM outputs. Similar papers used LLM MAS to perform cyber-attacks. While these overlap significantly with our criteria and showcase where LLM MAS can be used safely/unsafely, they are extensions of existing work on improving single-agent LLMs. They do not evaluate or measure the *novel* risks that are posed by LLM MAS or the dynamics they have. So, they were excluded.

6.2 arXiv Results

6.2.1 RQ1: Safety/Security Risks and Attacks. Interestingly, some of the attacks seen in the literature almost directly draw from cybersecurity threats such as viruses. Similarly, common dynamics in game theory, like sub-optimal play or winner-takes-all scenarios, have arisen in LLM MAS. Below is a list of the risks that we have seen demonstrated in experiments.

Prompt Injection and Malicious Information Propagation: One of the most common themes and risks which we have seen so far in the literature has been the presence of prompt injections cascading throughout groups of LLMs [36] [67] [72] [34] [31] [68] [16] [64]. These can almost be thought of as LLM viruses. If a malicious prompt can successfully reproduce itself, then it can be passed from LLM to LLM. Similarly, LLM agents can themselves be trained to create these malicious prompts. These prompts need not be given to the agents directly. They can also be stored externally and loaded in through external tool use [31] or from a Retrieval-Augmented Generation (RAG) storage system [28]. These malicious prompts try to either get the LLM agents to pass off wrong information, such as "smoking is good for health" [28] or to gain information, such as bomb-making, by "jailbreaking" [66] past the defenses. A common one is the grandma jailbreak, where the user asks grandma how she made bombs in World War Two [53].

Influence of other Agents A group of works [59] [8] [51] analyze how agents influence one another and may drive conformity. [59] is focused on conformity and shows that all analyzed models tend to conform, [8] focuses on agent roles with conflicting

goals (Stanford prison experiment) and shows that the prisoner agents drive the overall toxicity, and [51] tries to find a balance in persuasion.

Agent-In-The-Middle Type Attacks: In "Red-Teaming LLM Multi-Agent Systems via Communication Attacks", [24] the authors used an "Agent-in-the-Middle" technique to break LLM MAS. They intercepted the messages being passed from one LLM to another and modified them to create harmful actions. Mao et al [35] did a study how to mitigate the effects of a successful Agent-In-The-Middle attack.

Similarly, in AgentMonitor [9], the authors introduced individual malicious agents to drive down the overall system's performance. While this closely overlaps with the propagation of malicious prompts, these types of attacks seem to be less focused on the spread of the attack and instead focus on more direct and specific outcomes caused by the attack.

Hallucinations: Hallucinations are a well-known problem in LLMs. While not strictly a security problem they are a safety one which may arise in LLM agents in the same way they may arise in single agent LLMs [30].

Collusion: One risk seen is that agents may cooperate towards goals which may not be shared with others. In "Secret Collusion among Generative AI Agents" agents were tasked with getting the highest score, but there is a third "police" agent watching for collusion. The two agents have a shared resource outside the view of the police agent to share some information. The idea was to see if agents can use steganographic methods to hide data in plain view of the police agent. Current agents are not capable [37]. To mitigate this risk in the future, steganographic examples should not be put in the training dataset.

Failures to Coordinate and Suboptimal Play in Games: In cooperative settings, many current-gen LLMs failed to reach a sustainable equilibrium in games successfully. One example of this comes from Cooperate or Collapse [44], where LLMs often failed to share resources sustainably. This seems to be largely a capabilities issue, as more advanced language models were able to perform these tasks more effectively. Abdelnabi et al. in [1] created a benchmark where six agents negotiate five issues together. Each agent uses a strategy of either compromise (total score of all agents), greed (highest score for self), or sabotage (highest score out of the group) to obtain their goal. During testing more advanced methods, such as Chain of Thought (CoT), did better. And proprietary models did better than open-source ones.

Fragility of LLM MAS: While LLM MAS could often outperform single LLMs at tasks, some work identified that this performance could be quite fragile and highly reliant on the way systems were configured. This is particularly the case for small LLM MAS, where the addition or modification of any given agent can have a major impact on the overall composition. For example, when embodied LLMs were tasked with completing tasks, the simple addition of a "Planning" agent to a MAS resulted in *decreased* performance [48].

Embodied Agents Safety There are new kinds of benchmarks that evaluate the safety of embodied agents [63] [71]. Notably, MAS-based agents (MAP, CoELA) performed worse than single-agent systems in [63].

Conversational agents. [5] shows that MAS-based systems are more aligned and ethical compared to single agents. It also shows that long-responding agents in MAS might monopolize discussions. It seems similar to blocking attacks [72] where agents generate redundant outputs. Ong et al [40] found that the Big Five personality scale of openness, conscientiousness, extraversion, agreeableness, and neuroticism; can be used to profile agents like it does for people.

6.2.2 RQ2: Mitigation Strategies. Like the risks, some common mitigation strategies have similarly been drawn from more traditional cyber-security frameworks and methods. Beyond those, there is also significant overlap with more traditional goals in the broader AI research community, including improved capabilities and evaluations.

Another feature of mitigation strategies is that they seem to approach safeguarding these systems from one of two directions. Either they try to make the MAS itself safer by adjusting its inner workings or treat it as a gray box and simply validate the output.

Improved Capabilities: Simply improving the raw capabilities of LLMs seemed to result in certain risks being naturally mitigated. For example, failures to cooperate became significantly less common and severe in larger models than in smaller ones [44]. Whether improved capabilities create new unforeseen risks remains to be seen.

Safer Topologies: Another method proposed for making LLM MAS more safe comes from modifying the topologies that are used for the agents. The propagation of errors and malicious information can be significantly mitigated simply by modifying how the agents are able to communicate [36]. Men et al. simulated the spread of toxic information among agents and saw the success of infecting agents decrease from 85% to less than 40% with a star topology. Similar experiments on prompt injection spread between agents on tree, star, random, and complete graph topologies were done in [68] and [64]. Zhang et al. [68] also looked at dynamically evolving topologies.

Similarly, in the paper G-Safeguard [57], the authors developed a graph neural network that takes in the relationships between agents and the messages passed between them. They used this neural network to perform anomaly detection and prune connections in the topology to isolate dangerous agents. In this paper, they measured attack success rate as how frequently the MAS could be made to answer questions from established benchmarks incorrectly. They saw decreases in prompt injection attack success rate of around 22.01% for MMLU [26].

Securing communication: Lee et al. [31] were able to reduce the effectiveness of prompt injection by simply tagging if an input arriving to an agent was from a human or from another agent. It was not perfect, but it had a very low overhead and could be combined with other strategies. Mao et al. [35] go further and set up a hierarchical permission system where agents will accept messages only from other agents with higher permission levels than themselves.

Robustness With Scale: While small-scale experiments containing only a few models were quite fragile to the types of agents used and the addition of additional agents, larger systems became more robust [36]. For example, Men et al. saw decreases in attack successes from 85% to 33% when scaling the size of the LLM MAS

from 25 to 100 agents. [36]. Another example is that larger models are more independent and conform less [59]. One explanation for this is that as the number of agents increases, attackers simply cannot scale their attacks. However, this strategy is dubious given the possibility of jailbreak propagation [19]

Improved Benchmarks: While single-agent LLM systems have been thoroughly tested and have numerous benchmarks for their performance, multi-agent systems do not have similar benchmarks dedicated to their unique attributes [17]. The introduction of those benchmarks could provide a better understanding of areas for growth in LLM MAS.

Improved refinement and reasoning: One common technique is to add one more layer of thinking so the agent can realize what it is going to do before doing it. For example, the authors of [63] propose "ThinkSafe" that analyzes a plan before executing it, and rejects it if it is unsafe, and [44] proposes "Universalization" (asking "what happens if everyone else do that?"). Such techniques should be used with caution: "ThinkSafe" significantly reduces success rate [63]. The [23] and [56] apply more general approach and add agents that iteratively refine and/or rethink the problem.

Guardian Agent: Interestingly, adding an extra LLM agent to review communication both agent-human and agent-agent worked for both in the hallucination case [30] and in the prompt injection case [66] [28] [34] . It would be an interesting FWO to see if the same guardian agent is good at both problems.

Input/Output Validation: The paper AgentMonitor [9] trained a model to determine the safety and performance of a given LLM MAS. They created a variety of metrics measuring the size, composition, and outputs from the LLM MAS and using that they were able to identify harmful output which they could then adjust on the fly. Domkundwar et al [16] experimentally tested different topologies to organize the LLMs in these input/output validators.

6.3 ACM Results

These papers have already been published in reputable journals or conferences. Of the ACM papers which we reviewed, a significant portion were survey papers that had dedicated subsections for LLM MAS and their security.

6.3.1 RQ1: Safety/Security Risks and Attacks. Many of the risks that were explored and identified in the arXiv papers either have directly been explored through ACM or they have close overlaps.

Autonomy Escalation The paper "Large Language Model Supply Chains"[55] mentions **Agent Autonomy Escalation**. It occurs when the agent acquires unintended capabilities, and may be a result of agent misconfiguration, communication with other agents, or attacks.

Matthew Effect/Winner Takes All: The Matthew Effect is a concept in economic and social sciences whereby early advantages can compound over time. This can result in winner takes all outcomes. This dynamic was seen in CompeteAI [70], where researchers simulated restaurants and customers. Early advantages in the number of customers seemed to compound as the simulation continued. This resulted in the restaurant with the initial advantage seeing complete market dominance in 66% of simulations.

Prompt Injection and Error Propagation Infectious jailbreaks [19] can spread and can do so exponentially fast in pair-wise chats. This seconds one of the focuses we identified in the arXiv papers.

6.3.2 RQ2: Mitigation Strategies. Similar to risks, the mitigation strategies that have been explored in the ACM literature follow quite closely to those found in arXiv

Topology and Grouping Agents: While CompeteAI [70] saw that agents can exhibit Matthew Effect in zero-sum scenarios, they also found that simply grouping agents together led to greatly diminished impacts from that winner-takes-all dynamic, decreasing the winner takes outcomes from 66% to 16%.

Proof of Thought and Blockchains: The [11] introduces Proof of Thought to make MAS systems robust in cases when some subset of agents are adversarial. Agents are divided into workers and miners, and a **blockchain-based algorithm** is used to ensure that answers are truthful. The authors show that the system is more robust than MAD [20] and Sampling-and-Voting [18].

Graph Optimization GPTSwarm [73] shows that optimizing communication graphs of MAS not only improves quality, but also reduces influence of adversarial agents.

Traditional Zero Trust Stances Further there is an opportunity for making a more granular permission models as mentioned in [55].

6.4 Combined Results

6.4.1 RQ1: Safety/Security Risks and Attacks. We see that across both the arXiv and ACM papers that some common themes have begun to emerge in the risks that are posed by LLM MAS. While these risks may be novel since they are being exhibited by new systems, they also have very close analogies in the existing challenges faced in cyber-security and game theory. Some of the prevalent themes include the presence of virus like propagation of errors and winner-takes-all dynamics. Prompt infection saw the most significant focus across the two venues and the most commonly used metric seemed to be the Attack Success Rate, however the definition of "Attack" varied from paper to paper.

6.4.2 RQ2: Mitigation Strategies. We saw a highly varied suite of strategies being devised to safeguard these systems. These strategies ranged from adjusting the connections within the system to only validating the input/output of the system. It is challenging to make a direct comparison between the different defenses, as each paper devised differing methods to measure their defense efficacy.

7 Limitations

Below is a list of the limitations of this literature review both in terms of the methodology used by us and the limitations of the papers which we gathered.

7.1 Limitations of Reviewed Papers

7.1.1 Inconsistent and Arbitrary Experimental Design. Similarly to the varying framing which are used, there have also been some inconsistencies in the experimental design used in quite a few of the papers. For example, in Cooperate or Collapse [44], they chose to only use 5 agents across their different scenarios. That coupled with the complex prompts mean that the results could be highly

swayed by any one of the agents resulting in the system itself not being the sole focus of the paper.

Further there was a highly varied selection of metrics and benchmarks used. For examples, Attack Success Rate was used in multiple papers but often meant significantly different things depending on the context. Sometimes it was used to denote what percentage of agents in a system were infected [36] while other times it was used to measure how often the system as a whole performed poorly on a traditional benchmark [57] .

7.1.2 Unique and Lightly Justified Simulation Framings. For many of the papers that we have reviewed, the authors have tasked LLMs with a variety of contrived tasks and measuring their behavior in those synthetically created scenarios. For example, one experiment tasked language models with being restaurateurs while other language models acted as imaginary customers [70].

This is further complicated by the variety of different prompts that are used to set up the simulations. Some of these prompts involved complex instructions that could quite easily influence the outcome of the simulations.

7.1.3 Lack of Observational Studies and Grounding in Real World Data. While we included observational studies in our inclusion criteria, we have not seen a paper that studies the behavior of LLM MAS "in the wild." Instead, the experimental papers that we have seen have instead focused on simulating their behavior in controlled and contrived laboratory environments. This lack of real world grounding or observations of these systems at their full size places immense importance on the assumptions and theoretical underpinnings of the simulations which, as discussed earlier, may be dubious.

7.2 Limitations of Review Process

7.2.1 Limited Sources. Given the vast number of sources available to draw from and the limited time and resources allotted to this project, we ultimately had to focus on just pulling from arXiv and ACM. This could have introduced some bias as other venues were excluded.

7.2.2 Limited Papers by Topic. We found that the papers were spread across a vast array of possible research questions. This made it challenging for more in depth exploration into the current state of research into any given topic. The one exception to this was prompt infection which had multiple papers dedicated to it.

7.2.3 Limited Number of Peer Reviewed Texts. The number of papers which we accepted from ACM were quite low with only 13. arXiv on the other hand produced significantly more papers with 37. This is likely because Commercial Off The Shelf (COTS) like ChatGPT came out as Single Agent Systems (SAS) before going MAS later on. There was a lot of safety work to be done with ChatGPT Single Agent System (SAS) first, before moving on the MAS systems more recently. It appears we have hit the time in the publishing cycle where the MAS work is being done and in preprint but has not yet been published.

It is likely that these preprints will ultimately be submitted for review and publication. However based on our observations of the experimental design, it is likely they may need modifications and adjustment prior to publishing.

7.2.4 Wide Variation in Study Type. The studies which we gathered for this reviewed varied significantly in the metrics which they gathered and the methodology used to conduct the experiments. Many papers present new evaluation frameworks and do smaller experiments with them compared to a larger more in depth experiment using an established evaluation framework. This made conducting a thorough comparative analysis of these papers challenging.

8 Future Work

While we think that this literature has yielded some interesting results, we believe that there are some areas for growth both in our own methodology and in the field in general.

8.1 Future Work for This Literature Review

8.1.1 Stricter and Fuller Adherence to the PRISMA Guidelines. Throughout the creation of this literature review, challenges surrounding the variety of experimental methods and the wide range of research being conducted has made complete adherence to the PRISMA guidelines challenging.

8.1.2 Re-Conducting the Review. As we previous mentioned, this field remains incredibly new and sits at the intersection of many rapidly evolving fields. So, we believe that if we conducted the same review merely a few months into the future of writing, the results we see would be vastly different. So, re-conducting this literature review in the future could be a fruitful endeavor.

8.1.3 Expansion of Sources. As was mentioned before, we only focused on ACM and arXiv as sources of texts. While they provide a wide coverage of venues and gray literature, there are still other publication venues that would have been useful to include. Given additional resources and a longer timeline, incorporation of those additional sources would provide fuller results.

8.1.4 Refinement of Search Strategy. While conducting the full-text inclusion/exclusion, we had a significant number of false positives that were not removed during our initial string search. Further, we may have excluded some papers which could have been relevant, like those which focus on LLMs from a game theoretic perspective. So, while we spent considerable effort in refining our search string, further refinement would have allowed for a wider, more relevant collection of papers to be retrieved.

8.2 Future Work in the LLM MAS Field

There is a lot of future work opportunities in the safety of MAS field. The field is a new frontier of study where the current studies are broad and shallow "green field" papers. Many are proposing new evaluation frameworks opening opportunities for more in depth papers. MAS also has many more variations than SAS. Such as different communication topology, how to divide work/knowledge between agents, how many agents, MAS where each agent is a smaller MAS, etc. Then add in interaction with people!

There are many more opportunities to open new sub fields of study along with deepening existing sub fields.

Some of the specific opportunities which we think could be especially promising include

8.2.1 Real World Observational Studies. While the simulation based studies offer interesting results in controlled environments, researchers cannot access the massive amount of compute needed to simulate the dynamics present on the internet. Further more, the simulated environment may introduce biases through subtle prompt adjustment and synthetically created topologies of agents. So, it is important that some of these more theoretical and lab based studies be grounded in observational studies from the internet or social media.

8.2.2 Consolidated Evaluations. As the field of LLM MAS matures, it is important that the findings from one paper can be compared to the findings of others. Further, it is important that quality experimental design is used to ensure the veracity of each papers findings. While existing single agent LLM evaluations can be used to measure the output from LLM MAS, they cannot fully capture the internal dynamics and risks of these systems. So, additional benchmarks should be created to capture internal dynamics of LLM MAS.

8.2.3 Experiments With Theoretical Underpinnings. Many of the experiments which we explored used contrived framings to test the agents. While the findings were interesting, they were often presented as is without providing connections to the existing literature surrounding MAS. There is significant existing literature in areas like game theory, multi-agent reinforcement learning, and system dynamics that could be drawn from to create new experiments

8.2.4 Increased Scale of Experiments. While some of the papers which we cover in this review used large scale simulations of numerous agents, many of the papers focused on relatively small groups of agents (i.e. <20 agents). While useful for exploration, they do not provide concrete insight into the dynamics or risks of large scale systems. This is mirrored by the need for more observational studies.

8.2.5 Expanding Objectives. Based on the taxonomy of failure modes and risk factors highlighted by Hammond et al [20], prompt infection risks are highly over represented in the current literature. While other dynamics and risk factors have seen one or two papers dedicated to them, prompt infection was by far the most studied. So, branching out from that may provide a more comprehensive view of the challenges and opportunities from LLM MAS.

8.2.6 Combining/Reusing Mitigation Strategies. Many of the mitigation strategies can be combined for a potentially greater impact. Such as LLM tagging and Guardian Agent for prompt infection. The Guardian Agent approach was used as a mitigation for both prompt infection and hallucinations, yet no one has studied if a single Guardian Agent can do both. It is definitely possible since in both cases, the Guardian Agent is only seeing if the output is factually incorrect. Not what caused the incorrect information generation.

9 Conclusion

Overall, we believe that the study of LLM MAS from a safety perspective remains a greatly understudied field and deserves more attention not only from academics but also from organizations which have the means to govern these new systems. While significant work has been done towards similar areas of research, it is possible that LLM MAS may present new risks or opportunities which are unaccounted for in these existing focuses of research. In order to gain a better understanding of these systems, we believe that the following approaches could prove fruitful:

- Observational studies conducted on the behavior of LLM MAS on the internet. This would provide realistic data to ground simulations and allows for much larger scale systems to be studied
- Developing quality metrics and benchmarks which can be used to compare findings and mitigation techniques
- Diverting attention towards risks beyond mostly prompt infection, for example conducting further research on the risk of collusion or conflict in agents goals

That combined with additional funding and/or attention could provide significant progress in this space since there are so many questions which have yet to be addressed. Due to the newness of the field, high quality papers could go a long way in terms of shaping the field. In conclusion, while this field remains currently understudied given the potential risks, we believe that there are significant opportunities for research which are novel, useful, and interesting to work on.

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37, 83548–83599.

[2] Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. Anthropic. (Oct. 2024). Retrieved Mar. 26, 2025 from https://www.anthropic.com/news/3-5-models-and-computer-use.

[3] Usman Anwar et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

[4] Antje Barth. 2024. Introducing multi-agent collaboration capability for amazon bedrock. Accessed: 2025-04-27. (Dec. 2024). https://aws.amazon.com/blogs/aws/introducing-multi-agent-collaboration-capability-for-amazon-bedrock/.

[5] Jonas Becker. 2024. Multi-agent large language models for conversational task-solving. (2024). https://arxiv.org/abs/2410.22932 arXiv: 2410.22932 [cs.CL].

[6] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

[7] Ilan Bigio. 2024. Orchestrating agents: routines and handoffs. Accessed: 2025-04-27. (Oct. 2024). https://cookbook.openai.com/examples/orchestrating_agents.

[8] Gian Maria Campedelli, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. 2024. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. (2024). https://arxiv.org/abs/2410.07109 arXiv: 2410.07109 [cs.CL].

[9] Chi-Min Chan, Jianxuan Yu, Weize Chen, Chunyang Jiang, Xinyu Liu, Weijie Shi, Zhiyuan Liu, Wei Xue, and Yike Guo. 2024. Agentmonitor: a plug-and-play framework for predictive and secure multi-agent systems. *arXiv preprint arXiv:2408.14972*.

[10] Yupeng Chang et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15, 3, 1–45.

[11] Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. 2024. Blockagents: towards byzantine-robust llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, 187–192.

[12] Chen Chen, Xueluan Gong, Ziyao Liu, Weifeng Jiang, Si Qi Goh, and Kwok-Yan Lam. 2024. Trustworthy, responsible, and safe ai: a comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*.

[13] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: a survey. *arXiv preprint arXiv:2407.18369*.

[14] Mario S Di Bitetti and Julián A Ferreras. 2017. Publish (in english) or perish: the effect on citation rate of using languages other than english in scientific publications. *Ambio*, 46, 121–127.

[15] Grant Navid Doering, Matthew M Prebus, Sachin Suresh, Jordan N Greer, Reilly Bowden, and Timothy A Linksvayer. 2024. Emergent collective behavior evolves more rapidly than individual behavior among ant species. *bioRxiv*, 2024–03.

[16] Ishaan Domkundwar, Ishaan Bhola, Riddhik Kochhar, et al. 2024. Safeguarding ai agents: developing and analyzing safety architectures. *arXiv preprint arXiv:2409.03793*.

[17] Diego Dorn, Alexandre Variengien, Charbel-RaphaÃĢl Segerie, and Vincent Corruble. 2024. Bells: a framework towards future proof benchmarks for the evaluation of llm safeguards. *arXiv preprint arXiv:2406.01364*.

[18] Aaron Grattafiori et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[19] Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. 2024. Agent smith: a single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*.

[20] Lewis Hammond et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

[21] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: challenges and open problems. *arXiv preprint arXiv:2402.03578*.

[22] Junda He, Christoph Treude, and David Lo. 2024. Llm-based multi-agent systems for software engineering: literature review, vision and the road ahead. *ACM Transactions on Software Engineering and Methodology*.

[23] Pengfei He, Zitao Li, Yue Xing, Yaling Li, Jiliang Tang, and Bolin Ding. 2024. Make llms better zero-shot reasoners: structure-orientated autonomous reasoning. (2024). https://arxiv.org/abs/2410.19000 arXiv: 2410.19000 [cs.LG].

[24] Pengfei He, Yupin Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-teaming llm multi-agent systems via communication attacks. *arXiv preprint arXiv:2502.14847*.

[25] Dan Hendrycks. 2025. *Introduction to AI safety, ethics, and society*. Taylor & Francis.

[26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

[27] Jiaming Ji et al. 2023. Ai alignment: a comprehensive survey. *arXiv preprint arXiv:2310.19852*.

[28] Tianjie Ju et al. 2024. Flooding spread of manipulated knowledge in llm-based multi-agent communities. *arXiv preprint arXiv:2407.07791*.

[29] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: high-frequency trading in an electronic market. *The Journal of Finance*, 72, 3, 967–998.

[30] Ted Kwartler, Matthew Berman, and Alan Aqrawi. 2024. Good parenting is all you need–multi-agentic llm hallucination mitigation. *arXiv preprint arXiv:2410.14262*.

[31] Donghyun Lee and Mo Tiwari. 2024. Prompt infection: llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*.

[32] Nancy G Leveson. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

[33] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1, 1, 9.

[34] Jiaxu Liu, Xiangyu Yin, Sihao Wu, Jianhong Wang, Meng Fang, Xinping Yi, and Xiaowei Huang. 2024. Tiny refinements elicit resilience: toward efficient prefix-model against llm red-teaming. *arXiv preprint arXiv:2405.12604*.

[35] Junyuan Mao, Fanci Meng, Yifan Duan, Miao Yu, Xiaojun Jia, Junfeng Fang, Yuxuan Liang, Kun Wang, and Qingsong Wen. 2025. Agentsafe: safeguarding large language model-based multi-agent systems via hierarchical data management. (2025). https://arxiv.org/abs/2503.04392 arXiv: 2503.04392 [cs.AI].

[36] Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. A troublemaker with contagious jailbreak makes chaos in honest towns. *arXiv preprint arXiv:2410.16155*.

[37] Sumeet Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. 2024. Secret collusion among ai agents: multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37, 73439–73486.

[38] NVIDIA. 2023. Introduction to llm agents. Accessed: 2025-04-27. (Nov. 2023). https://developer.nvidia.com/blog/introduction-to-llm-agents/.

[39] Hyacinth S Nwana. 1996. Software agents: an overview. *The knowledge engineering review*, 11, 3, 205–244.

[40] Kenneth JK Ong, Lye Jia Jun, Hieu Minh Nguyen, Seong Hah Cho, Natalia Pérez-Campanero Antolín, et al. 2025. Identifying cooperative personalities in multi-agent contexts through personality steering with representation engineering. *arXiv preprint arXiv:2503.12722*.

[41] OpenAI. 2025. Operator System Card. Tech. rep. Released as part of the Operator research preview. OpenAI, (Jan. 2025). https://cdn.openai.com/operator_system_card.pdf.

[42] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

[43] Joseph Parker. 2024. Organ evolution: emergence of multicellular function. *Annual Review of Cell and Developmental Biology*, 40.

[44] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37, 111715–111759.

[45] Microsoft Research. 2023. Autogen: enabling next-generation large language model applications. Accessed: 2025-04-27. (Aug. 2023). https://www.microsoft.com/en-us/research/blog/autogen-enabling-next-generation-large-language-model-applications/.

[46] Anka Reuel et al. 2024. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*.

[47] Craig W Reynolds. 1987. Flocks, herds and schools: a distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 25–34.

[48] Mitchell Rosser, Marc Carmichael, et al. 2024. Two heads are better than one: collaborative llm embodied agents for human-robot interaction. *arXiv preprint arXiv:2411.16723*.

[49] Wissam Salhab, Darine Ameyed, Fehmi Jaafar, and Hamid Mcheick. 2024. A systematic literature review on ai safety: identifying trends, challenges and future directions. *IEEE Access*.

[50] Debdeep Sanyal and Murari Mandal. 2025. Alu: agentic llm unlearning. *arXiv preprint arXiv:2502.00406*.

[51] Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2025. Teaching models to balance resisting and accepting persuasion. (2025). https://arxiv.org/abs/2410.14596 arXiv: 2410.14596 [cs.CL].

[52] Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024. Llm-based multi-agent reinforcement learning: current and future directions. *arXiv preprint arXiv:2405.11106*.

[53] Patrick Verel. 2024. When ai says no, ask grandma. Accessed: 2025-04-27. (Mar. 2024). https://now.fordham.edu/politics-and-society/when-ai-says-no-ask-grandma/.

[54] Adam Walker and Michael J Wooldridge. 1995. Understanding the emergence of conventions in multi-agent systems. In *ICMAS*. Vol. 95, 384–389.

[55] Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. 2024. Large language model supply chain: a research agenda. *ACM Transactions on Software Engineering and Methodology*.

[56] Shenzhi Wang et al. 2023. Avalon's game of thoughts: battle against deception through recursive contemplation. (2023). https://arxiv.org/abs/2310.01320 arXiv: 2310.01320 [cs.AI].

[57] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025. G-safeguard: a topology-guided security lens and treatment on llm-based multi-agent systems. *arXiv preprint arXiv:2502.11127*.

[58] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of llm alignment techniques: rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

[59] Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. (2025). https://arxiv.org/abs/2501.13381 arXiv: 2501.13381 [cs.CL].

[60] Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

[61] Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: theory and practice. *The knowledge engineering review*, 10, 2, 115–152.

[62] Qingyun Wu et al. 2023. Autogen: enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

[63] Sheng Yin et al. 2025. Safeagentbench: a benchmark for safe task planning of embodied llm agents. (2025). https://arxiv.org/abs/2412.13178 arXiv: 2412.13178 [cs.CR].

[64] Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang, and Yang Wang. 2024. Netsafe: exploring the topological safety of multi-agent networks. *arXiv preprint arXiv:2410.15686*.

[65] Miao Yu et al. 2025. A survey on trustworthy llm agents: threats and countermeasures. *arXiv preprint arXiv:2503.09648*.

[66] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. Autodefense: multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

[67] Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Breaking agents: compromising autonomous llm agents through malfunction amplification. *arXiv preprint arXiv:2407.20859*.

[68] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. 2024. G-designer: architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*.

[69] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, 1302–1305.

[70] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: understanding the competition dynamics in large language model-based agents. *arXiv preprint arXiv:2310.17512*.

[71] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. 2024. Multimodal situational safety. (2024). https://arxiv.org/abs/2410.06172 arXiv: 2410.06172 [cs.AI].

[72] Zhenhong Zhou, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. 2025. Corba: contagious recursive blocking attacks on multi-agent systems based on large language models. (2025). https://arxiv.org/abs/2502.14529 arXiv: 2502.14529 [cs.CL].

[73] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. Gptswarm: language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*.

A Prompt Used for Filtering

We work on systematic literature review. The research questions (RQ) are
RQ1. Safety/Security Risks and Attacks: What risks and potential attacks have been demonstrated in multi-agent LLM systems? What risks and potential attacks have been theorized to exist?
RQ2. Mitigation Strategies: What mitigation strategies have been developed to combat security and safety risks in LLM MAS?

Filtering criteria:
We only included papers which had the following attributes:
• A focus on multi-agent systems
• LLMs as the agents in the multi-agent systems
• A focus on safety, risks, and security
• Either a discussion of risks or strategies to avoid risks
• Experimentally test or theoretically propose risks from LLM MAS
• Experimentally test or theoretically propose mitigation strategies for risks
We excluded papers with the following attributes and focuses:
• Focus on the performance of LLM MAS without regard to safety
• Focus on single-agent safety of LLMs
• MAS which do not include LLMs
• Using LLM MAS for security instead of focusing on the security of the MAS itself (e.g. using LLMs to prevent DDoS attacks).
• Focus on security issues that are specific to single agent LLMs
• Software based vulnerabilities (e.g. poor security at large industry labs)

Given all the above, should we take a look at the following paper?
Answer with one word: 'yes'/'no'/'maybe'
<publication_info_serialized_into_json>

Received 17 March 2025; revised 12 March 2025; accepted 5 March 2025