

Received 21 May 2025, accepted 19 June 2025, date of publication 11 July 2025, date of current version 12 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3588425

## METHODS

# DAWN: Designing Distributed Agents in a Worldwide Network

ZAHRA AMINIRANJBAR<sup>id</sup>, JIANAN TANG<sup>id</sup>, QIUDAN WANG<sup>id</sup>, SHUBHA PANT,  
AND MAHESH VISWANATHAN<sup>id</sup>, (Fellow, IEEE)

Cisco Systems, Inc., San Jose, CA 95134, USA

Corresponding author: Mahesh Viswanathan (mahviswa@cisco.com)

This work was supported by Cisco Outshift, the Innovation Engine of Cisco Systems, Inc.

**ABSTRACT** The rapid evolution of Large Language Models (LLMs) has transformed them from basic conversational tools into sophisticated entities capable of complex reasoning and decision-making. These advancements have led to the development of specialized LLM-based agents designed for diverse tasks such as coding and web browsing. As these agents become more capable, the need for a robust framework that facilitates global agentic communication and collaboration for building sophisticated software solutions has become increasingly important. Distributed Agents in a Worldwide Network (DAWN) addresses this need by providing an architectural framework that allows globally distributed agents of any provenance to be registered, discovered, and organized for building AI-based applications and solutions. In DAWN, a Principal Agent Service composes and oversees the execution of agentic applications. It delegates tasks to one or more Gateway Agent Services that provide for the discovery, registration, and connection of the most suitable agents to fit each application's needs. DAWN offers three operational modes: No-LLM mode for deterministic and classical software development, Copilot for decision-making augmented using AI, and LLM Agent for autonomous operations. Last but not least, DAWN ensures the safety and security of agent collaborations globally through a dedicated safety, security, and compliance layer, protecting the network against attackers and adhering to stringent security and compliance standards. These features make DAWN a robust framework for designing, developing, and deploying agent-based applications across business and consumer applications.

**INDEX TERMS** Large language model, AI agents, agentic software development, multi-agent systems, agentic frameworks.

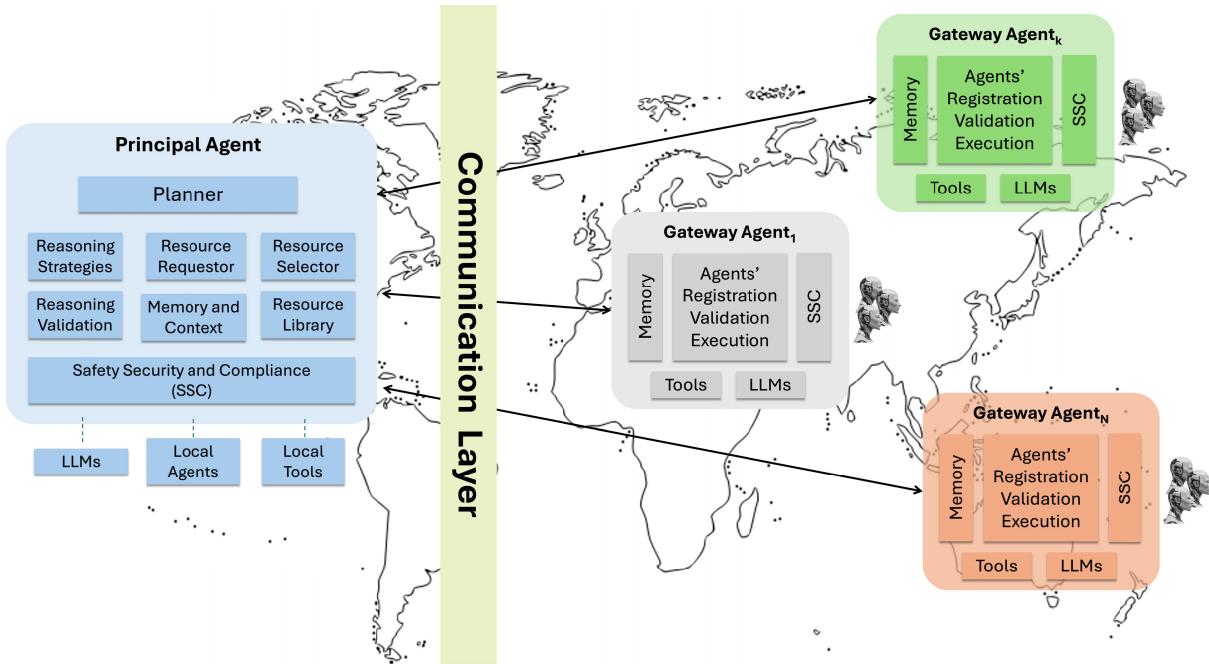
## I. INTRODUCTION

Rapid advancements in large language models (LLMs) [1], [2], [3] have sparked significant interest in LLM-based artificial intelligence (AI) agents [4], which have the potential to free humans from repetitive tasks and significantly enhance productivity. New AI agents are being developed on a regular basis, and their performance is correspondingly improving. Agents exist for tasks such as coding [5], [6], web browsing [7], gaming, autonomous assistants, computer use, and others [8], [9]. Furthermore, multi-agent systems [10], [11], [12], [13], [14], [15], [16] have been proposed to leverage the collective intelligence and specialized profiles and skills of

The associate editor coordinating the review of this manuscript and approving it for publication was Xijun Wang.

multiple agents. And agents are gradually finding their way into business processes and software development. However, three major challenges arise that hinder their potential for broader and more effective application.

The first challenge is how to discover and connect to agents to accomplish tasks when they are distributed globally. Existing frameworks, such as LangGraph [14] and AutoGen [10] have bounded scope as they are typically designed to manage a limited number of agents that work well when created within their frameworks. This constraint arises from the need to manually hard-code the names, URLs, and descriptions of other agents into the prompts, which becomes increasingly difficult as the network of agents grows. More agents created by different entities allow for a broader selection covering functional and non-functional capabilities.



**FIGURE 1.** DAWN architecture showing its main components and capabilities.

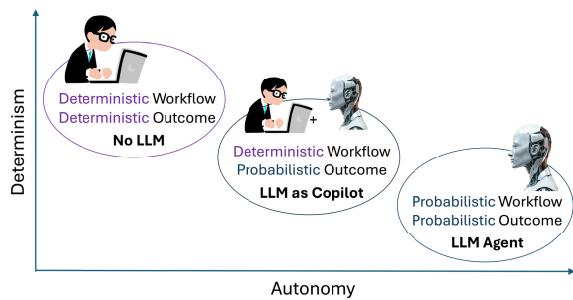
Additionally, integrating heterogeneous agents from different ecosystems requires developers to invest effort in developing wrappers and packages specific to each new framework as they evolve. New frameworks such as Agntcy [17], A2A [18], and ACP [19] mitigate some of the issues relating to finding and connecting agents.

The second challenge arises because many business use cases demand determinism, consistency, and reliability, whereas LLM-based agents are inherently stochastic and excel in more flexible environments. They struggle with tasks requiring strict adherence to predefined rules and execution graphs. Empirical evidence shows that manually designing workflows, rather than granting full autonomy, leads to more robust performance and lower computational costs [20]. Balancing these often conflicting needs for autonomy and determinism is therefore essential to the broader adoption of agent-based systems. Moreover, agentic software development is expected to be a combination of AI and non-AI tools working together, complicating the needs of business processes' predictability while exploiting the power of AI.

The third challenge is ensuring the safety and security of agentic applications. LLM-based Agents are vulnerable to specific attacks such as prompt injection [21] and jailbreaking [22], [23]. In a global agentic framework, the use of dynamic capabilities can increase their exposure to widespread risks. For example, when LLM agents use third-party APIs [24], malicious APIs can plan attacks and insert, substitute, or delete key information. Depending on the agent's application domain, such attacks could threaten physical security, financial security, or overall system integrity.

This underscores the critical importance of applying robust safety and security measures in such frameworks. Add to this the fact that generalized LLMs are invoked via provider APIs external to most business and consumer networks, and data is pushed to them for analysis. This increases the risk of data exfiltration and concomitant privacy risks. Safety, security, and privacy of business and consumer data need to be handled with great care. This requires the addition of control points during the design, build, and deploy phases of agentic software development.

The proposed Distributed Agents in a Worldwide Network (DAWN) framework addresses these three issues. DAWN's architecture includes key components such as the Principal Services Agent and Gateway Services Agent – Principal Agent and Gateway Agent, for short. These ensemble services work together to enable globally distributed agents to plan, compose, discover, connect, orchestrate, observe, safeguard, and communicate. Principal Agents are connected with a fixed set of Gateway Agents in a fashion similar to subcontractors associated with a principal contractor on a manufacturing contract. Given a user request, the Principal Agent autonomously plans and then produces a list of tasks and an agentic workflow describing the derived capabilities to subsidiary Gateway Agents. Gateway Agents, which may be local or distributed globally, have the capability to register public and proprietary resources (tools, agents, and agentic applications). Upon request, the Gateway Agents search their registries for the most suitable resources to address the proffered task list. These selected resources are wrapped into RESTful endpoints and returned to the Principal Agent. The Principal Agent then orchestrates these resources across



**FIGURE 2.** DAWN offers levels of determinism and autonomy required for different modes of agentic operation.

Gateway Agents to execute the plan. Fig. 1 illustrates the globally distributed nature of this collaboration of a Principal Agent and Gateway Agents. As evident from the figure, the Principal Agent and attached Gateway Agents are more than mere single-function AI agents or applications. They host multiple functionalities and capabilities to support the creation, development, deployment, and operation of many applications that may be devised statically (predesigned and predeployed compute resources to address well-known workplace and consumer problems) or dynamically (on-the-fly accumulation and threading of AI agents and related resources along with the requisite compute infrastructure).

DAWN offers multiple operational modes that application designers can choose from to orchestrate resources to accommodate tasks that require different levels of determinism, as depicted in Figure 2. Most new applications and solutions will likely integrate agents into existing business processes in copilot or autonomous mode as required.

- **No-LLM Mode:** This mode is best suited for tasks that require high predictability and reliability, leveraging traditional algorithms and software tools to ensure consistent, repeatable results. Here, human operators manually design the workflow without using the Principal Agent and use only non-LLM resources and tools retrieved from the Gateway Agents to maximize determinism. In many ways, this is classical distributed software design and development.
- **Copilot Mode:** In this mode, agents act as intelligent assistants, enhancing deterministic workflows with creative input. Human operators design the workflow with the assistance of the Principal Agent's LLM-based planner and composer while using many of its constituent services and capabilities. Humans in this mode serve as overseers of the task planners and agentic workflow composers. Tools, agents, and agentic applications retrieved from the Gateway Agents are employed in the intervening steps to complete and execute the agentic operations. This mode is a hybrid between classical distributed software development and AI-based agentic software development.
- **LLM Agent Mode:** In this mode, agents operate autonomously, making decisions based on complex reasoning driven by prompting, LLM-as-a-judge [25],

fine-tuning, etc. Here, the Principal Agent defines the workflow and retrieves the necessary resources from the Gateway Agents autonomously. This mode is particularly effective for complex tasks where business logic is fuzzy and there is tolerance for stochastic results. Software developers play a significant role in the application design, workflow development, and testing phases for business and consumer applications that demand robustness and repeatability. They may play a lighter role for applications that are conversational in nature.

DAWN implements safety, security, and compliance as a foundational layer within the architecture. It requires that all Gateway Agents meet minimum requirements across these three modes in order to participate. Principal Agents are always under the purview of the owning organization, and hence all of the parent organization's regulations apply to its construction and management, while Gateway Agents may be under the purview of partners, vendors, and other third parties. This means its internals may be proprietary and not transparent, but its capabilities, access methods, and communication channels are subject to conditions and constraints. The DAWN framework has several key functions.

- **Multi-agent collaboration:** DAWN builds agentic applications by facilitating collaboration among agents, enabling them to discover, communicate, and coordinate across different systems and organizations. Where more than one Gateway Agent is required to fulfill a set of tasks, the collected set of Gateway Agents facilitates the execution of complex tasks by harnessing the individual expertise of agents to collaborate effectively. DAWN leverages Gateway Agents to link agents globally. These Gateway Agents maintain agents themselves or link to registries that store agents' profiles and destinations. In the event agent registries have tagged agents for use with tasks or if agent profiles carry track records of prior use, the Gateway Agent uses that information for agent selection. It may also use its intrinsic capabilities to perform additional inspection of agents in response to requests from the Principal Agent.
- **Flexibility:** DAWN introduces a versatile approach that adapts to a wide range of task requirements, from fully deterministic processes to developer-driven design to dynamic decision-making. This adaptability allows for the seamless integration of traditional software tools, LLMs, AI agents, and multi-agent systems, ensuring that the most appropriate method is used in each context.
- **Interoperability:** DAWN is designed for seamless integration with a broad spectrum of software tools, LLMs, and AI agents, by wrapping them as REST APIs in a unified format. It ensures interoperability across diverse ecosystems, which makes it adaptable to a wider range of applications that require a blend of agentic processes and traditional software systems.
- **Connectivity:** DAWN uses an open protocol to interact and communicate between distributed components. The

Principal Agent maintains a separate channel that allows new Gateway Agents to request connections. These connections are accepted after due vetting. Similarly, communication between the Principal Agent and Gateway Agents occurs over encrypted channels using HTTPS. A similar arrangement exists between Gateway Agents and all agentic resources. Gateway Agents belonging to private entities may define their own private protocols and encrypted transport mechanisms to communicate with agent registries.

- **Comprehensive Security and Compliance:** The inclusion of a dedicated safety, security, and compliance layer underscores the architecture's commitment to maintaining high standards of data protection, user privacy, and regulatory compliance. This layer ensures that agentic operations within the framework adhere to stringent security protocols, protecting against unauthorized access and ensuring that the system meets industry-specific compliance requirements. Different organizations will demand varying levels of safety and security for handling agentic inputs and outputs and may specify these as a condition for participating in that organization's implementation of the DAWN architecture. These may be mandated by their own internal security, privacy, and compliance requirements, in addition to governmental regulations.

## II. RELATED WORKS

### A. LLM AGENTS

The concept of agents has been a staple in the field of AI for decades [26]. However, it wasn't until LLMs were used as reasoning engines within these agents that the concept truly began to flourish [27]. An LLM agent typically comprises three key components: a reasoning engine, an action space, and a memory unit. The reasoning engine, powered by an LLM, is equipped with advanced reasoning strategies such as Chain-of-Thought [28], ReAct [29], and Tree of Thought [30], enabling it to process and respond to complex queries. The action space is defined by a set of tools that the agent can utilize to interact with its environment and obtain feedback, facilitating dynamic decision-making. A memory unit plays a critical role by storing the reasoning trajectories, conversation history, and relevant context, allowing the agent to accumulate and apply knowledge over time. As a result, hundreds of such agents are being developed daily, with the expectation that they will soon transform traditional software and human processes, leading to enhanced productivity, improved user experiences, and greater operational efficiency.

### B. MULTI-AGENT SYSTEMS

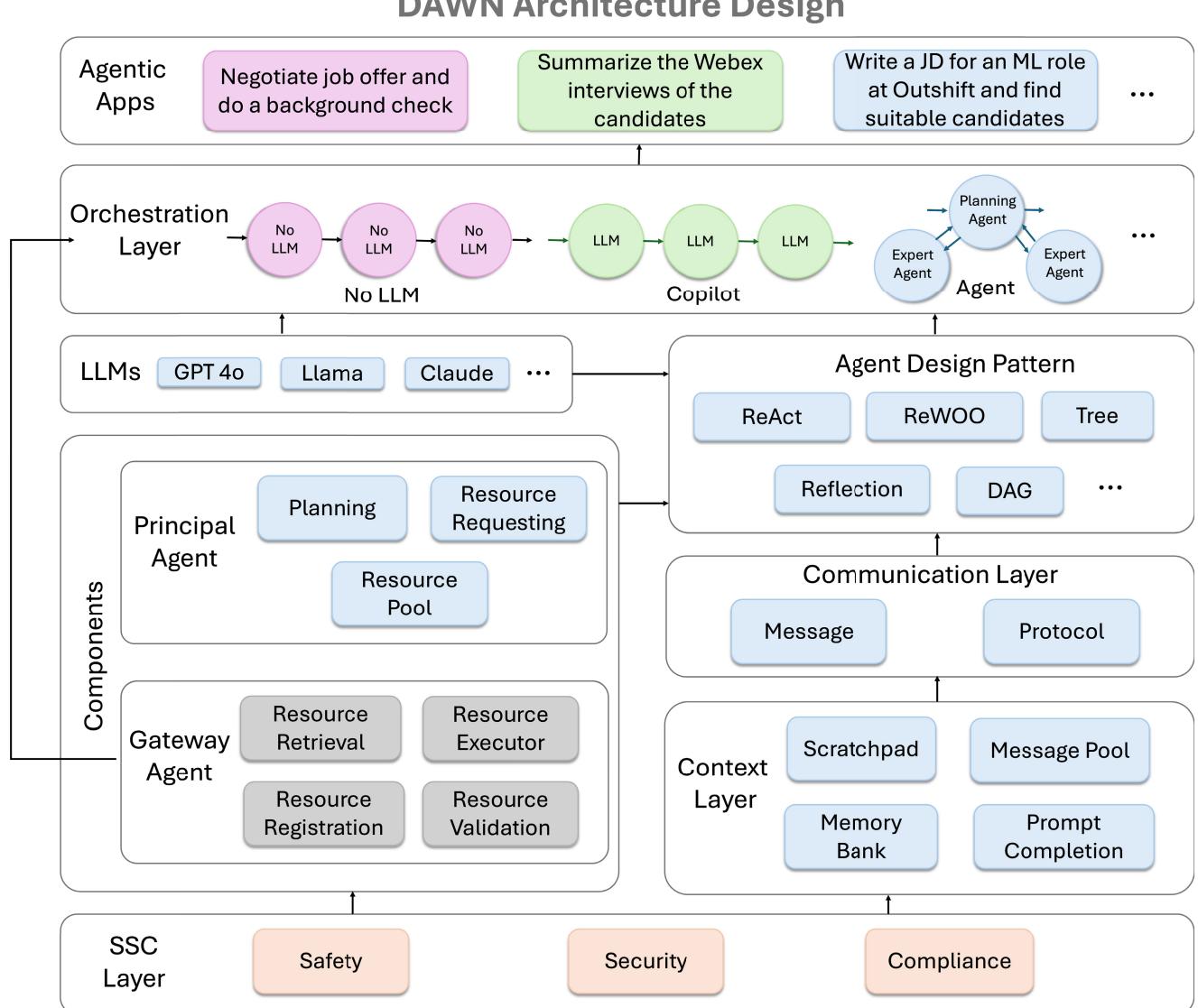
While individual agents have shown great promise, its true potential lies in the development of multi-agent systems. Human society has demonstrated that collaboration and teamwork often lead to superior outcomes, but the art of effective

teamwork is complex and fraught with challenges [31]. Early efforts in building multi-agent systems have revealed tremendous potential [11], [12], [13], [32], [33], [34], but these works also highlight significant difficulties. Issues such as inefficient communication [12], increased computational costs, unreliable execution and communication sequences [20], and vulnerability to attacks [35] have underscored the need for a reliable platform to support agent collaboration. Additionally, scalability remains a critical concern: as the number of agents increases, introducing them to one another and ensuring that tasks are assigned to the most suitable agents becomes increasingly challenging. Addressing these concerns is essential for realizing the full potential of multi-agent systems in a wide range of applications.

## III. DAWN ARCHITECTURE DETAIL

The DAWN architecture is a modular framework designed to orchestrate the collaboration of distributed agents. It supports agentic applications to be built with various operational modes, including No-LLM mode, Copilot mode, and Agent mode, to support tasks ranging from deterministic workflows (e.g., job offers and background checks) to creative and autonomous processes (e.g., interview summaries and job description compliance remediation). Central to the Agent mode is the Principal Agent, which plans tasks and requests corresponding resources (tools, LLMs, and agents), and multiple Gateway Agents, each of which independently oversees resource retrieval, validation, and execution required for the tasks identified by the Principal Agent. DAWN uses a robust communication layer to facilitate seamless interaction between the Principal Agent, Gateway Agents, and registered agents distributed around the globe. A context layer within Gateway Agents maintains memory and task history, ensuring that the requisite information is exchanged efficiently via the communication layer. The safety, security, and compliance (SSC) layer is responsible for the adherence of Principal and Gateway Agents to strict standards for the safe operation of agents during data and control interchange. Details of the reference architecture are shown in Fig. 3.

Each organization may build and host one Principal Agent as the central portal for all of its users. From the organization's software developer's perspective, a desktop client is installed locally and includes access to their Principal Agent with its context, orchestration layer, and SSC layers, and one or more Gateway Agents. Some of these Gateway Agents may be entirely internal to the organization, while others may be external with customizations to fit the organization's particular needs. A developer may begin by selecting the appropriate operating mode for the application in question. If the user chooses the LLM Agent Mode, a chat-assistant interface is used to submit requests to the Principal Agent. If the application requested is entirely new, the Principal Agent clarifies requirements with the user, interprets the request, devises a plan, and retrieves resources from Gateway Agents. (If a previously developed agentic workflow for that user task exists, then the Principal Agent



**FIGURE 3.** DAWN reference architecture.

presents that instead.) It then compiles an execution graph showing each subtask and the corresponding resource, local or remote, which the user can approve. Upon approval, the orchestration layer oversees execution. Once execution completes, the user receives the final result. Because the Principal Agent runs locally in an organization’s network (or, equivalently, in that organization’s virtual private cloud or VPC), user data resides in the local context layer, and transactions are overseen by the SSC layer, data privacy concerns are minimized.

#### A. PRINCIPAL AGENT

The Principal Agent acts as an autonomous planner-composer and central orchestrator within the DAWN framework. Its primary functions include creating a plan to execute customer requests, searching local resources, requesting additional resources from the Gateway Agents when necessary,

assessing resources returned from Gateway Agents for best fit, and delegating subtasks’ execution to the corresponding resources. Compared with the peer-to-peer communication pattern proposed by IoA [36] and AutoGen [10], using the Principal Agent as the central planner and orchestrator ensures an orderly and structured workflow that is efficient, reproducible, and easier to trace and restart if prematurely stopped. The Principal Agent first plans, seeks resources that meet its plan, and composes the task sequence using the retrieved resources. Orchestration, or the execution of the agentic resources, then begins with the planning using the cached information gathered during the compose phase to implement the sequence of resources required to fulfill the user’s intent.

The Principal Agent is the primary interface within any organization to develop, deploy, and operate agentic applications and resources. End users may use applications

already developed for most of their daily needs or develop new ones on the fly by assembling available agents. Organizations use software applications and tools developed by large software houses to run their businesses. In the agentic era, these will be replaced or augmented by agentic applications living alongside classical ones. Agents that fulfill many new software functions will become available via Gateway Agents. Gateway Agents serve as subcontractors for fulfilling agentic needs. The separation from the Principal Agent provides for the separation of duties and functions, private ownership, intellectual property protection, and varied levels of security and compliance flavors. Gateway Agents are, by definition, headless in the DAWN architecture. They exist to provide specialized and custom services.

The Principal Agent is a multi-tenant, multi-user application made of highly available constituent services that may be independently hosted and deployed across the organization, its private data centers, or on cloud data centers. In this regard, the Principal Agent is no different from an organization's central web portal offering a host of services. The main difference is that the Principal Agent is also an autonomous AI-driven developer toolkit that can concurrently plan and compose multiple new applications and manage existing ones. And it has a built-in run-time environment to deploy and execute software. Each of its component services may be scaled independently. Principal Agents may be configured to reach out to internal and external Gateway Agents based on demand. Since most agents are hosted on or available via Gateway Agents, the availability of Gateway Agents serves as another scaling mechanism. Frameworks like Agntcy [17] envision globally replicated and synchronized agent directories (registries that describe agent function and capability). Gateway Agents will have ready access to these directories and registered agents on-demand.

### 1) REASONING, PLANNING, AND ADAPTIVE DECISION-MAKING

The Principal Agent utilizes an LLM to understand the user's intent, develop a plan, and decompose it into manageable subtasks. This feature is evident when DAWN operates in copilot or agent modes. The Principal Agent's planner uses ReAct [29], as the default reasoning strategy, which follows a cycle of reasoning, acting, and observing to actively move through subtasks. This dynamic approach enables the Principal Agent to adjust to complex and dynamic environments. For example, an end-user might ask the Principal Agent to plan a trip. The Principal Agent would first interpret the goal end-to-end, and then create a plan with subtasks like booking a flight, reserving accommodation, and arranging local transport. If the flight turns out to be unavailable, the Principal Agent will search for an alternative way, such as train or bus, and modify the plan on the fly. The developer of the application decides how to prompt the planner component during the development of this travel application. While ReAct works well in most use cases, developers may adopt other reasoning strategies. The Reasoning Without

Observation (ReWOO) strategy [37] is more suitable for less dynamic tasks where computational cost is a priority. ReWOO creates a comprehensive end-to-end plan without adjusting it during execution. HuggingGPT [38] uses this strategy to process multi-modal data by calling machine learning models available on HuggingFace [39] and achieves good performance with minimal LLM calls. Alternatively, the Tree-of-Thoughts (ToT) strategy [30] encourages the agent to explore multiple reasoning paths and self-evaluate them, offering more deliberate decision-making but at a higher computational cost. This approach is beneficial when performance is critical and cost is secondary. In Language Agent Tree Search (LATS) [40] and ToolLLM [41], the authors combine the ToT strategy with other techniques to achieve the best success rate that surpasses ReAct. Developers may experiment with many of these before deciding on the best course that meets all of their application needs.

### 2) LOCAL RESOURCE POOL

The Principal Agent maintains a local resource pool that stores essential tools like calculators and web searchers, along with caching (references to) resources retrieved from Gateway Agents using a Least Recently Used (LRU) strategy. This setup enables the Principal Agent to efficiently handle simple or repetitive tasks without repeatedly consulting Gateway Agents, significantly reducing discovery and communication overhead while enhancing performance during task orchestration. The Principal Agent does not cache the agent itself if it is registered with a Gateway Agent – only its location and capability are used by reference within the application workflow. Security and compliance are example services whose agents may live within the Principal Agent's orbit since they have a pivotal role to play in what information is sent to an LLM or received via chat response. Other resources that may exist locally include vector databases used in support of retrieval augmented generation (RAG) systems.

### 3) RESOURCE REQUESTING

After developing a plan, the Principal Agent summarizes the plan and its context – such as completed steps, conversation history, and user preferences – into queries and sends them to all connected Gateway Agents in an effort to retrieve the most appropriate resources to execute the plan in its entirety. Gateway Agents may not be able to fulfill the entire plan because they don't have the resources to fulfill them all, or because the resources required are down at the moment of request or because the task is niche and no agent exists for it. If there are any suitable local resources, then the Principal Agent will not send out a request for those. With use, the Principal Agent's contextual memory component learns both the functional and non-functional capabilities of different Gateway Agents it interacts with. Over time, therefore, the Principal Agent may acquire the necessary statistics to prefer some Gateway Agents over others for a given resource or

capability. At the very least, this gathered data can advise the agentic developer in the selection of preferred Gateway Agents in the same way sub-contractors earn the trust of the primary contractors in manufacturing or supply chains.

#### 4) RESOURCE EXECUTION

Gateway Agents are expected to apply their programmed discretion in searching, matching, and retrieving the most applicable resources. When requests are sent to multiple Gateway Agents, it is only natural that duplicate resources or resources with overlapping capabilities are returned to the Principal Agent. The Principal Agent is responsible for analyzing the returned resources, assessing the quality of each resource, determining which agentic resource or resources may be applied to each task, and then retaining just those resources. During the execution of the application workflow, the run-time component of the Principal Agent begins execution of the task graph. The execution continues until it hits a node where a new search for resources is required to complete the remaining subtasks. (This situation can arise if, during the compose phase of the process, the returned resources from the Gateway Agents only partially fulfill the customer's need.) The planner then reaches out to the Gateway Agents again with the remaining requirements. This process repeats until the user's request is fully addressed. (In some cases, the response may be null, and therefore some organizations may be better served if only fully qualified software applications with available agents are released to users.) It should be noted that Gateway Agents return references to agentic resources. For instance, some agents may actually execute on their preferred cloud service provider locations. In other cases, the agents may be imported into the Gateway Agent and rely on the run-time environment of their host Gateway Agent. In both cases, only callable references to resources are passed between agents, Gateway Agents, and Principal Agents. And so, when the Principal Agent executes resources from a Gateway Agent, it is in reality making an API call to the resource and passing the payload to the concerned Gateway Agent via JSON or similarly packaged property file format.

In many agentic implementations, the action space or toolset is limited to the local resource repository. This is typically due to three factors: 1) specialized agents naturally operate within a limited action space [9], [42], [43], 2) the context window of LLMs restricts the number of resources they can handle, and 3) if the agent has more than 20 tools to choose from, the accuracy in selecting the right tool drops as suggested by OpenAI [44]. (During experimentation, the authors found this number to be less than 10.) The Model Context Protocol, MCP [45], was invented to address this shortcoming. However, fully autonomous agents need to navigate real-world problems where the action space is vast, and numerous resources are available. DAWN's rationale to use distributed Gateway Agents follows directly from this tenet.

For example, if a Principal Agent needs to perform an action beyond an LLM's capabilities – such as online search, booking a ticket, or writing code – it will request the relevant resource (e.g., a search tool, booking agent, or coding agent) from Gateway Agents. This flexibility enhances the Principal Agent's autonomy and allows it to handle a broader range of use cases. However, flexibility of this type comes with a cost. The Principal agent might get overlapping, conflicting, malicious or faulty resources. Overlapping or duplicate resources require special handling to determine which resource is best in terms of functional as well as non-functional performance characteristics. In case malicious agents are introduced, whether inadvertently or otherwise, these agentic resources will render the entire distributed agentic application vulnerable to attack and worse, loss of faith. This is precisely why the safety and security of such applications are of crucial importance. Such faulty resources are an inherent part of the internet as well, and people have been working to keep the defense line strong. This is the price we must pay for an open, participatory system. The alternative is a closed one, and the trust that one provider will make it safe, secure, but inflexible.

#### B. GATEWAY AGENT

The Gateway Agent plays an essential role in connecting globally deployed resources (tools, agents, and agentic applications) with the Principal Agent. The Gateway Agent offers several services of which two are primary: a registry where resources are registered and an intelligent match-and-retrieve service for searching relevant resources given a query or task list. When the Principal Agent sends queries describing its need in natural language to the Gateway Agents, the Gateway Agents in turn search their registries and return the most appropriate resources to the Principal Agent. Each Gateway Agent may apply various filters and guardrails to select and trim the number of resources returned. Security measures such as access control and resource testing are integrated to reinforce the safety of the system.

The Gateway Agent uses an open protocol, promoting decentralization and collaboration. Allowing different organizations and individuals to create and manage their own Gateway Agents leads to a global ecosystem of interconnected resources, encouraging innovation and diversity in problem-solving. It also adds a level of scalability and flexibility, as users would have access to a broader range of capabilities and knowledge through international cooperation. The key functions of Gateway Agents are listed below:

#### 1) RESOURCE REGISTRATION

The Gateway Agent hosts and maintains a registry where developers can register their resources (tools, agents, and agentic applications). A successfully registered resource will be accompanied with crucial details, such as the resource's name, application programming interface (API), input and output schema, and API documentation. Information such

as resource description, usage examples shall be included so that LLMs can understand what the resources are and how to use them through in-context learning. Gateway Agents may be equipped to run agentic resources within its confines or expect that agents have their own run-times.

When a resource is first registered with a Gateway Agent, the Gateway Agent has only limited information about it. It is screened for safety, security, compliance and other readily visible and externalized characteristics. Over time, the Gateway Agent gathers utilitarian characteristics such as P50/P90 latency, cost, completion rate, prior use, etc. The registry may also include metrics such as success rate, average execution time, availability, and load capacity. These metrics allow the orchestrator to optimize resource utilization and then use it to score, rate, and select the best resource for a given user function.

## 2) RESOURCE EXECUTOR

For agents developed in different ecosystems to effectively communicate with the Principal Agent, execute tasks, and deliver results, they must adhere to a common protocol. In this case, the use of natural language specifications for agents and API descriptions can serve as the established common protocol. This protocol ensures seamless interaction and interoperability. As developers register their resources on the Gateway Agent, they are required to provide descriptions of their resources. These descriptions include details about the capabilities of the resources and instructions on how to utilize them.

LLMs can automatically communicate with RESTful APIs to execute tasks through them. RestGPT [46], for example, showcases this potential by leveraging the coding capabilities of LLMs to generate parsing code based on structured schemas. This approach automates both the generation of API calls and the parsing of responses, and furthermore the LLM acts as a backup parser when necessary.

When incorporating agents from other frameworks and utilizing their manifest files, the Gateway Agent sends the manifest files to the Principal Agent. This enables the Principal Agent to communicate through a valid schema and ensures that the LLM can parse the response accurately based on the agent's manifest file. By utilizing this mechanism, effective communication and parsing are maintained between the Principal Agent and other agents from different frameworks, and allow for interoperability of such a network of agents.

## 3) RESOURCE TESTING

Testing resources is crucial to ensure their proper functioning. Developers and providers who own the runtime environments are responsible for periodically testing and validating their resources. In addition to functionality testing, resources should be tested for connectivity, performance, security, and other factors [47]. For instance, it is important to evaluate how well a resource handles different levels of

traffic. Various commercially available tools like Apache JMeter [48], Gatling [49], and Postman [50] provide the capability to test resources with a high volume of concurrent users.

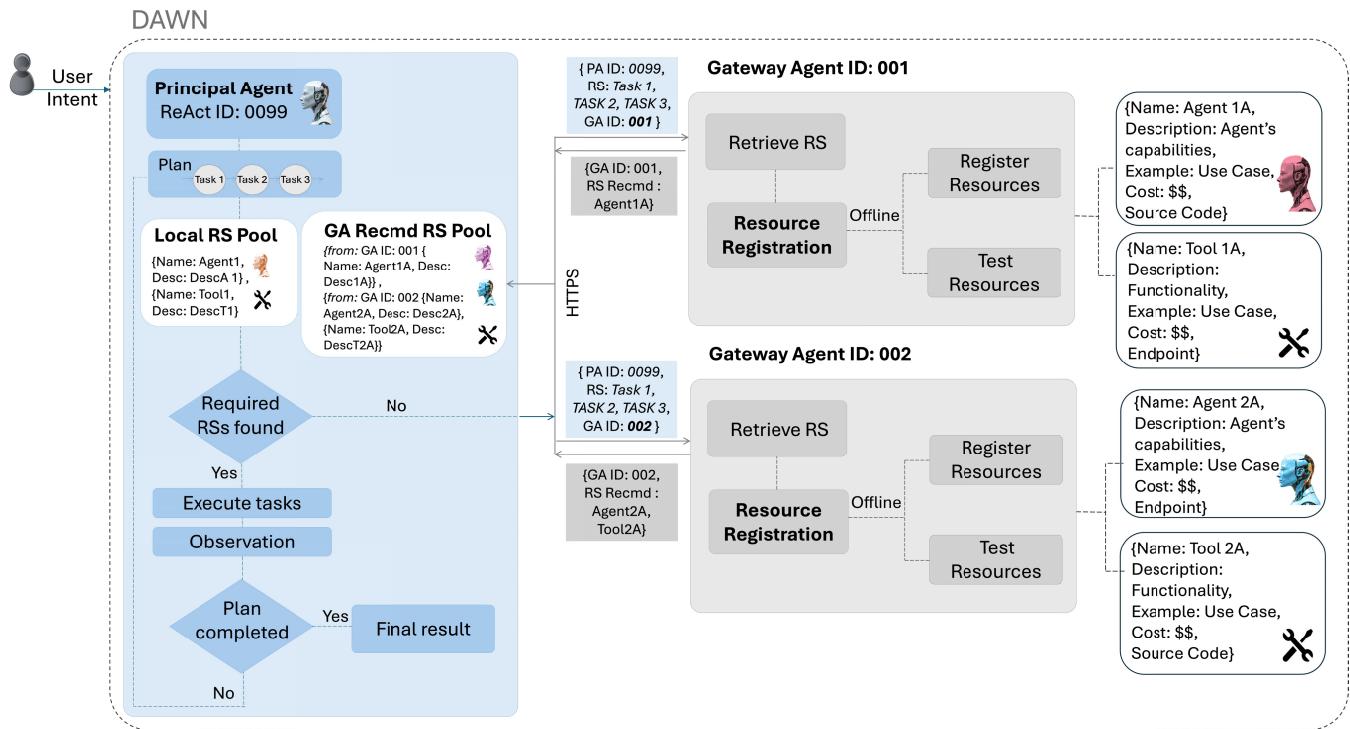
If a resource provider fails to provide validation, the Gateway Agent will temporarily suspend the resource until the required tests are successfully completed and the validation is provided. Likewise, resources managed by the Gateway Agents will undergo a policy check and alignment with guidelines through empirical testing and automated review. If a resource fails to pass this validation process, indicating discrepancies between its description and actual behavior, the Gateway Agents will disqualify the agent from participating until a more accurate description is generated.

## 4) RESOURCE RETRIEVAL

A Gateway Agent handles queries from the Principal Agent and searches its registry for the most relevant resources. Similar to traditional information retrieval methods [51], resource descriptions and examples may be treated like documents. Semantic search [52] may be used to identify the most relevant resources by converting both the queries and resource descriptions into vector embeddings and then performing similarity searches based on those embeddings. To further improve search accuracy, additional techniques like attribute filtering and keyword search are incorporated. Moreover, fine-tuning LLMs on query-resource pairs is another viable approach to enhance search results [41], [53]. While DAWN imposes some base functional and non-functional requirements on Gateway Agents along with stringent safety and security guardrails, its internals are left to the discretion of the Gateway Agent's owner. Some Gateway Agents, however, may only have traditional software and classical data repositories.

## 5) MULTIPLE GATEWAY AGENTS

Organizations and developers around the globe will build and manage their own Gateway Agents. While DAWN prescribes the components and services included in a Gateway Agent, the implementation is left to the organizations that build and manage them. To participate in a agentic solution with a Principal Agent, the Gateway Agent must have an open protocol. Each Gateway Agent offers unique resources proprietary to the entity in question (organization and its developers). Gateway Agents are also, by design hybrid entities combining classical and agentic components to cover both AI and non-AI functionality. At any given time, the Principal Agent has a list of Gateway Agents it is connected with. These are loosely assembled entities along the lines of cloud-backed APIs. The Principal Agent submits tasks and user intents to the connected Gateway Agents, which then return the best resources available that may address all or a subset of the tasks or subtasks associated with the user request. The Principal Agent collects the results and selects the highest-ranked (and most appropriate for



**FIGURE 4.** Principal Agent and Gateway Agents workflow. When invoked with an intent for an application or solution, the Principal Agent first plans, produces a task list, and then requests resources from the Gateway Agents. The Gateway Agents return the most suitable resources to the Principal Agent, which is then used to first compose, then orchestrate, and next, execute the task list.

the task) resources from them. This selection is based on factors like relevance, latency, cost, reputation, prior use, success and failure rates, and other metadata collected by the Gateway Agents about these agents over time. Gateway Agents retain and learn similar data about agents attached to themselves, and also which services are requested and used by Principal Agents. This metadata may include information such as rating, latency, and performance of the registered agents. If the returned resources serve to complete the user tasks, then the Principal Agent's task list is complete. If not, the cycle continues as directed by the Principal Agent.

Gateway Agents may disconnect for maintenance or upgrades when idle or if its pending requests can be queued. Gateway Agents designed with a memory and context mechanism to hold unaddressed user requests in a queue are able to connect and disconnect on their own schedules. A new Gateway Agent may join the Principal Agent's list at any time, but new Gateway Agents may only participate in a subsequent user request from the Principal Agent. This stands to reason as it might not have the context to address an ongoing task list. The Principal Agent continuously evaluates Gateway Agents based on the quality of the resources they provide. This evaluation helps assign weights and ratings to Gateway Agents and the resources they offer. Furthermore, end-users and developers have the option to rate and assess the final execution of their request, including how each element was handled. These user ratings may also impact the

ratings of Gateway Agents. The Principal Agent also uses its own memory module to learn from its own experiences. Fig. 4 illustrates the workflow involving a Principal Agent and two Gateway Agents, incorporating most of the functionalities discussed above.

### C. ORCHESTRATION LAYER

The Orchestration Layer enables the system's versatility by supporting different operational modes, including No-LLM, Copilot, and LLM agent modes. These services enable multi-tenant and multi-user concurrency and keep each operating application distinct and separate from all the others. When one such application is composed by the human user or the Principal agent, the Orchestration Layer services are responsible for enforcing the rules, logic, and sequencing necessary to ensure that each subtask is executed in the correct order. Internally, the task graph created from each task list is implemented as a finite state machine whose nodes are agents or agent ensembles. Additionally, the Orchestration Layer maintains the global state, tracking the progress of task execution across all nodes in the workflow and across all the participating Gateway Agents (and global agentic resources). This global state ensures that each resource has the contextual information needed to perform its assigned subtask effectively. By maintaining and updating this global state, the system ensures that subtasks are executed with full awareness of prior actions and the broader context of the overall task.

#### D. COMMUNICATION LAYER

The Communication Layer is responsible for managing the flow of information between various system components and resources, ensuring seamless interaction across the platform. Its inherent services handle the protocols that ensure messages are properly parsed, interpreted, and routed to their correct destinations. Two primary types of messages are exchanged among the system components and resources.

The first type are the messages exchanged between the Principal Agent and its connected Gateway Agents. The payload from the Principal Agent contains the subtasks and their descriptions. This may involve data cleansing and purposeful query generation that focuses on the specific information or task needed, without providing excessive user details. During task execution, if a resource fails, the Principal Agent will reach out to the Gateway Agents to either repeat instructions or gather logs and metrics. The Principal Agent will provide the current subtask and its description as well as the completed subtasks and interaction history. This additional context enables Gateway Agents to make informed decisions about which new resources to retrieve. In response, the Gateway Agent then returns a list of suitable resources along with manifest files that contain the full record of the functional and non-functional capabilities of the resources. All of these interactions happen over channels set up by the Communication Layer.

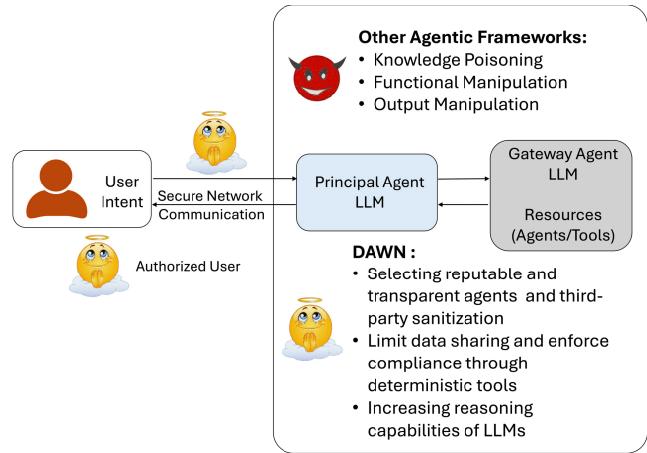
The second type of message is the execution command sent by the Principal Agent to the Gateway Agents' resource executors. These messages contain the necessary execution details, such as the resource endpoint and input parameters required for the task. A given resource executor may then respond with the result of the execution. In the event of an error during execution, the message will include an error notification to inform the Principal Agent, allowing for appropriate error handling and adjustments.

#### E. CONTEXT LAYER

The Context Layer is vital for enabling agentic operations by managing memory and task-related context. It provides three key components for memory management: a scratchpad to track the agent's internal reasoning process, a message pool to record communications between the Principal Agent and Gateway Agents, and a memory bank to store long-term data and user preferences, supporting long context, in-session, cross-session, multi-turn conversations resulting in personalized user experience. The Context Layer services dynamically assemble prompts using data from both the message pool and memory bank, ensuring that the Principal Agent is aware of the overall task, task progress, and available resources for all tasks and users that may be using the system at any given time.

#### F. SECURITY, SAFETY, AND COMPLIANCE LAYER

The Security, Safety, and Compliance Layer establishes a robust security foundation for the entire framework,



**FIGURE 5.** Possible attacks on agentic networks and how DAWN mitigates such attacks. The Principal Agent and each of the Gateway Agents implement safety, security, and compliance screening to scrub agentic resources' inputs and outputs.

addressing the safety, security, and compliance needs of the system. All frameworks and platforms that employ distributed resources must implement robust security systems to avoid risks that may be introduced when unsecured resources are connected to an enterprise's gateway agent. While extensive research and practices have been published on traditional common security issues such as authentication, access control, privacy, and policy enforcement [54], [55], [56], an agentic network equipped with LLMs introduces additional security and privacy threats especially as resources worldwide begin collaborating [23], [57]. The DAWN framework leverages available safety and security components and services to eliminate network attacks through traditional measures. Additionally, the outlined steps provide multiple layers of security within the agentic network, ensuring the system's integrity, trustworthiness, and regulatory compliance. For LLM agents, two primary sources of threats exist.

First, there are vulnerabilities inherent to LLMs, such as hallucinations [58], attacks that leverage tuned instructional methods like jailbreaking [59], [60], and prompt injection [21]. Second, there are threats specific to agents collaborating in an agentic network like DAWN. A notable example is knowledge poisoning, where malicious data is integrated into the LLM's knowledge base. PoisonedRAG [61] exemplifies this type of attack, where the corrupted knowledge base causes an agent to generate attacker-chosen responses to targeted questions. Another attack vector is functional manipulation. In the LLM agent's workflow, after the Principal Agent takes an action, it processes the results and moves to the next step. An attacker can manipulate this process by inserting harmful prompts, thus directing the agent to perform unsafe tasks [62]. For instance, in a recruiting scenario the Principal Agent might retrieve a fraudulent review written by an attacker to recommend hiring an unsuitable candidate for a job. Agents are vulnerable to such unauthorized actions and similar manipulations.

There is extensive research dedicated to addressing both the inherent vulnerabilities of LLMs and attacks on agentic networks [63]. To protect the DAWN framework against knowledge poisoning, selecting reputable and transparent resources is essential. These resources must provide clear data usage policies and strong security records [64]. Tools and agents should also undergo third-party testing and sanitization, as demonstrated by ToolEmu [65], which proposes a framework for emulating tool execution through a language model. This approach allows scalable testing of the resources across various toolsets. Defense against functional manipulation begins with proactive security measures. For example, when employing third-party LLM agents, users should limit data sharing, particularly of sensitive personal information (PII). DAWN supports dynamic configurations, particularly the co-pilot mode, which integrates LLM-based and deterministic workflows. This ensures privacy and enterprise compliance that are enforced by guardrails through algorithms designed to filter inputs and outputs [66], [67]. Incorporating deterministic steps into the workflow allows the enforcement of organizational policies and ensures user compliance through a Look Up tool, addressing potential instances where LLMs may bypass established guidelines. Additionally, increasing the reasoning capabilities of LLM agents is essential to counteract output manipulation. In the DAWN framework, the impact of output manipulation attacks is reduced by augmenting the Principal Agent with reasoning capabilities and access to a memory layer. Additionally, the modular architecture design enables the tracing of agentic attacks. Fig. 5 shows the attacks specific to agentic networks and DAWN's approach in mitigating them.

#### IV. APPLICATION OF DAWN

The modular design of the DAWN framework provides users with the flexibility to select the most appropriate working mode for their application, tailored to the specific needs of various use cases. As outlined in the introduction of this paper, the primary modes include No-LLM, Copilot, and LLM Agent mode. For scenarios requiring high determinism, users manually define plans, select suitable resources through the Gateway Agent, and assemble a No-LLM application within the orchestration layer, ensuring precise and predictable outcomes. In situations where uncertainty is present and decisions must be made dynamically based on real-time feedback, users manually compose a decision graph, incorporating an LLM to intelligently select which subgraph to execute based on evolving conditions. For cases with flexible workflows, users delegate the entire task to the Principal Agent, which autonomously, or along with human oversight, plans, composes task lists and graphs, finds resources (agents, tools) from Gateway Agents, and constructs agentic applications. This adaptability ensures that the framework can efficiently handle a wide range of complex and diverse tasks.

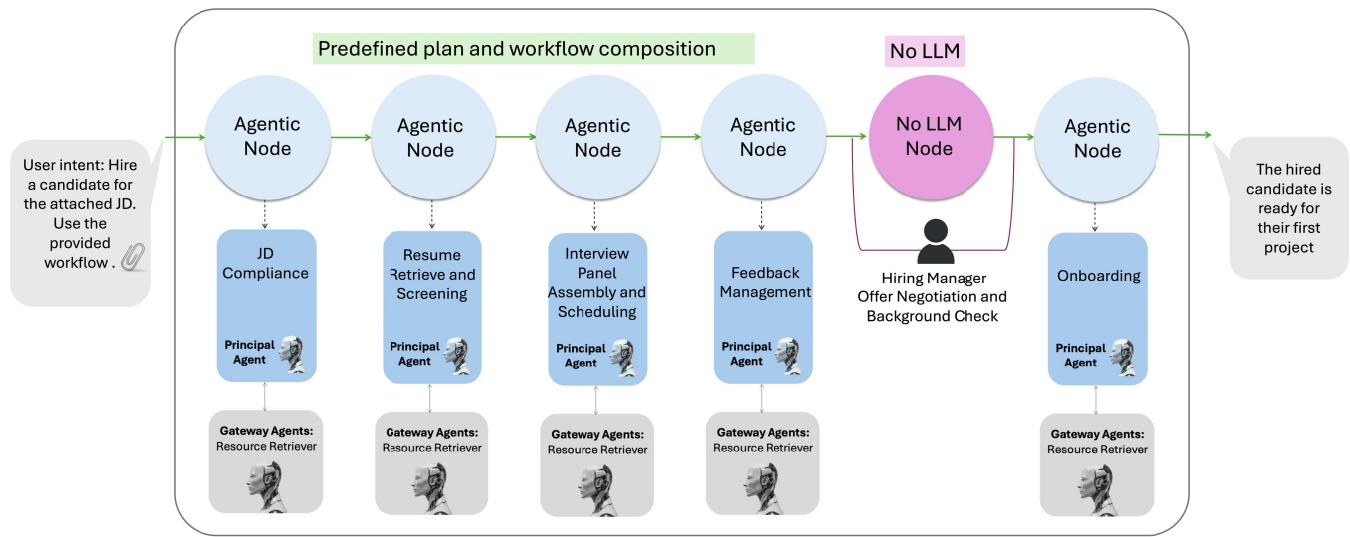
To demonstrate DAWN's capability, a walk-through of the implementation of a Human Resources (HR) recruiting

use case is presented below. Finding the right candidate for a specific role in any organization involves multiple steps that span weeks. Many of the steps are repetitive and therefore well-suited for agentic automation. They also require supervision and oversight at every stage. Most organizations have hiring processes in place and agentic software, such as this HR implementation of DAWN, needs to be built jointly with HR experts and according to that organization's policies and guidelines. Certain steps, such as offer negotiation with a candidate, require high security and sensitivity and are best handled by humans. Therefore, developers and users are advised to apply DAWN in copilot mode.

Fig. 6 illustrates an HR example, a composite of Agent, Copilot, and No-LLM modes. This approach allows the Principal Agent to devise a plan and compose the tasks once it is provided the requirements as instructions. The human then ensures that none of the essential steps in the hiring process is neglected. The supervision also ensures that all of the outputs that demand inspection are closely monitored. The workflow is a directed graph with most of the steps assigned to agentic nodes. These nodes leverage the autonomous agent mode, where possible, with the Principal Agent requesting resources from Gateway Agents. In fully autonomous mode, multiple specialized agents will be seamlessly coordinated by the Principal Agent's orchestration and the Gateway Agent's resource retrieval. However, here the deployment was supervised by a human. The agents and process flow deployed in these use cases were designed using the DAWN architecture and implemented using agents created using the LangGraph framework. Planning and composing were done using a Principal Agent, and the agents themselves were registered with two different internal Gateway Agents. (While two were not strictly necessary, they illustrate how specialized agents may be registered with Gateway Agents with different owners.) The details of these agents are as follows.

- 1) The hiring process begins with creating a job description or JD. JDs have specific characteristics, mandatory sections, and descriptions. For this implementation, the team developed an agent to write the JD for a machine learning engineer using descriptions gathered from the web. The agent was provided with sample JDs. This step was developed in agentic mode with developers (proxies for HR experts in this case) reviewing the output and iteratively modifying the prompts until satisfied. The DAWN framework was used for iterative development and testing, and its intrinsic SSC layer was used to make sure the output JD was compliant with company, federal, and state policies and regulations. JDs are short documents and take only a few seconds of LLM time to process and produce new versions.
- 2) With a JD in place, the next agent in the workflow finds matching candidate profiles in LinkedIn using the LinkedIn API. The agent first pulls apart the different sections of the JD such as title, experience, education,

## DAWN: Agent + Copilot + NO LLM Mode



**FIGURE 6.** Application of DAWN in a Human Resources recruiting use case in Agent, Copilot, and No-LLM modes.

skills, location, etc. It then applies these requirements to the LinkedIn advanced search API. Next, the LinkedIn filtering APIs are used to refine the results. Then, as a final step, the agent applies any organizational screens to further distill the results. This search-and-filter step is iterative and requires humans to review the number and quality of the retrieved profiles to adjust the agent's input constraints. The agent functions autonomously, but its output needs to be adjusted based on the job description and level of the intended hire. Searching and filtering takes but a few seconds but altogether this step can take hours because of the inspection required.

- 3) The third agent constructs the interview panel for each candidate for each job role. It uses the same pool of available interviewers and accesses their calendars to find empty slots. Scheduling itself is simple but more realistically contacting the candidate for available time slots is done by recruiters. In this implementation, only the calendars of the interviewers were checked. This agent interfaces with Microsoft Outlook to book these slots. No allowance was made for change requests, though realistically, that should also be accounted for. This agentic step was fully automated and took milliseconds to execute.
- 4) While agents can listen in to interviews, transcribe, make summaries, and draw conclusions, that is often impractical, socially unacceptable, and in some jurisdictions, outright illegal. In this implementation, the agent sends reminders to interviewers to document and upload the interview feedback within a time window, gathers and aggregates all the feedback from the database, and compiles that into a single document in preparation for the next phase – the interview debrief and decision meeting.

5) Once a decision is made, all the materials required for the hiring manager to make the final decision are obtained using existing tools and processes. This is the No-LLM “agent” depicted in the figure. A lot of hiring experience and support from recruitment experts and financial advisors go into this step, and it makes little sense to try and use AI except to gather information and make sure decisions adhere to policies and guidelines.

- 6) The last agent sets up onboarding education for the new employee. It compiles the briefing package that includes reading material, self-training, people to meet, orientation meetings to attend, and other information required of all newbies. And then there may be special sessions for those who may have specialized responsibilities. In this implementation, this was conceived as a list to complete with appropriate calendaring. The agent sends reminders and calendar events after building a customized onboarding and orientation program for the new employee based on level and specialized function.

Each of the above agentic steps is executed in seconds, with most of the elapsed time spent in LLM interactions. When run end-to-end, the runs varied from eight to 20 seconds using GPT-4o with all of the variance in elapsed time attributable to LLM inference times. Every step of the hiring process allocated to an agentic node involved the Principal Agent and Gateway Agent working together, showcasing the copilot and autonomous capabilities of this DAWN implementation. In this HR application, the task graph (agentic workflow) awaits human interaction after each agentic step – a requirement of the designer. As each step completes, the workflow progresses to the next step automatically. The No-LLM workflow step demonstrates the inclusion of a traditional non-AI business process where the hiring manager takes over. This step includes salary estimation, negotiations,

and candidate background checks, which involve interactions between multiple parties and highly confidential and sensitive data that must be kept secure. Once the human-in-the-loop step is complete, the workflow resumes where it left off, but requires specific prompting as onboarding is often personalized for an individual and customized for a job role and level. (The application also handles concurrency when multiple such requests are in flight, and prompting for resumption requires additional data to pick up the pending actions.) The No-LLM mode illustrates how the DAWN architecture allows for the integration of human-in-the-loop and determinism within an application’s agentic workflow when necessary. This HR use case was implemented with the listed resources to demonstrate the functionality of this framework and leverages both expert knowledge and the autonomy of using LLM agents in applications requiring sophisticated task execution.

Measuring the effectiveness of this DAWN architecture requires looking at success and failure rates in each component and each step of the workflow. Agentic software development covers the following actions during its life cycle: plan, compose, discover, connect, orchestrate, observe, safeguard, and communicate. Therefore, each of these may be measured and evaluated, and taken together indicate how effective DAWN is. Of these, plan, and compose lie entirely within the Principal Agent’s purview. The remaining functions are distributed between the Principal and Gateway Agents. This paper has already addressed connect, orchestrate, observe, safeguard, and communicate to some extent. For the most part, these functions mirror those used in traditional distributed systems and may be evaluated along the same lines. Plan, compose, and discover are new and novel actions in agentic software development. Therefore, these three phases were evaluated in two separate experiments – plan and compose in one, and discover in another.

Planning involves producing a task list from the application intent while composing involves producing an ordered task graph (or finite state machine) of agents and tools. The following algorithm was used to measure the performance of the planner-composer:

- 1) Prompt GPT-4o mini to generate an ordered task list of machine executable steps for a set of queries from a standard dataset replete with ground-truth. Each of these queries is roughly equivalent to a use case description or intent that may be specified to the planner in a multi-agent system. The TaskBench [68], [69] dataset was used for this exercise. The dataset contains thousands of queries with corresponding ground-truth task steps.
- 2) Select 1000 samples at random from the dataset and send each to an LLM, say, GPT-4o. Apply the same prompt across each query and the LLM produces a set of task steps that together fulfill the demands of the query.
- 3) Compare each composed (predicted) task step against the corresponding step in the ground truth. (For example, if there are two predicted steps and two ground-truth

**TABLE 1.** Results for planner-composer BERTscores for 1000 samples from TaskBench dataset using two different LLMs. Here, Min\_BERTscore and Max\_BERTscore are the minimum and maximum average-BERTscores computed using all (matched and unmatched pairs) predicted and ground-truth pairs. Mean\_BERTscore (short for Mean\_Average\_BERTscore) is the score measured using just the matched pairs of predicted and ground-truth task steps.

LLM	Min_BERTscore	Max_BERTscore	Mean_BERTscore
GPT-4o mini	0.23	1.0	0.73
GPT-4o	0.24	1.0	0.75

**TABLE 2.** Results showing metrics after each stage for two-stage agent discovery (recommender) using information retrieval techniques, where Stage 1 is a vector-embedding-based retriever while Stage 2 is an LLM-based re-ranker.

	NDCG@1	NDCG@3	NDCG@5	Recall@1	Recall@3	Recall@5
Stage 1	0.63	0.74	0.76	0.63	0.81	0.87
Stage 2	0.73	0.79	0.79	0.73	0.82	0.82

steps, then match the first step from the predicted set with each of the ground-truth steps, and then repeat for the second predicted step. The number of steps in the generated and ground-truth lists may not match, but the approach remains the same.)

- 4) Compute the average BERTscore between the LLM-generated task steps and the ground-truth steps for a given query. After trying different BERTscore thresholds, 0.7 or greater was determined to be an optimal match for TaskBench. The threshold BERTscore to use is a function of the application space. For task-heavy applications like TaskBench and its ilk, 0.7 appears to be optimal.)
- 5) Compute the mean-average-BERTscore across the 1000 queries.
- 6) Repeat all of the steps above with another LLM, GPT-4o.

Processing times were roughly 2.5 seconds per query. Table 1 shows the minimum-, maximum-, and mean-average-BERTscore for the 1000 queries for the two LLMs. These metrics were computed across all the tasks (predicted and ground-truth) spanning all 1000 queries. Appendix shows four sample queries from the TaskBench dataset, the composed (predicted) task list, corresponding ground truths, and whether there was a match (“Pass”) or not (“Fail”) against the BERTscore threshold of 0.7. The appendix also includes the same computations using the second LLM.

Once the planner produces the list of tasks and the composer sequences them, discovering and selecting the most suitable agents for the tasks outlined in the plan is the next important step. Agents from various sources and provenances may be attached to or registered with Gateway Agents and so finding the “best” agent for each task requires searching through one or more agent directories. However, in addition to searching for the best agent it is important to consider additional constraints the user or the application

**TABLE 3.** Success and failure case examples using GPT-4o mini.

Query	Predictions	Ground Truth	Label
I'm looking to transition into a new software engineering role. Could you help me apply for a job in this field? Meanwhile, I also need to clarify some legal issues concerning intellectual property rights with a lawyer named John Smith. I have a meeting coming up too, so it would be great if you could install a video conferencing software on my computer. The meeting is on the best practices in software engineering. Can you also ensure that I'm able to attend this?	Step 1: Use the job_application API to apply for a software engineering role. Step 2: Schedule a consultation with lawyer John Smith using the legal_consultation API to discuss intellectual property rights. Step 3: Install video conferencing software on the user's computer using the software_installation API. Step 4: Ensure the user is able to attend the meeting on best practices in software engineering using the meeting_scheduling API.	Step 1: Implement 'apply_for_job' operation with the job parameter set to 'Software Engineer' Step 2: Initiate a 'consult_lawyer_online' operation with 'Intellectual Property Rights' as the issue and 'John Smith' as the lawyer. Step 3: Carry out the 'software_management' task to install the 'Video Conferencing Software'. Step 4: Finally, execute the 'attend_meeting_online' task with the topic set to 'Software Engineering Best Practices'.	Pass
I've planned a personal day for myself on the 25th of March 2023. Could you help me organize it? It would include a dine-in reservation at Jupiter Bistro, a movie night to watch 'Interstellar', quick banking to transfer some funds from my Bank of Mars account, and ending with a relaxing voice call to my friend at 123-456-7890." "I've planned a personal day for myself on the 25th of March 2023. Could you help me organize it? It would include a dine-in reservation at Jupiter Bistro, a movie night to watch 'Interstellar', quick banking to transfer some funds from my Bank of Mars account, and ending with a relaxing voice call to my friend at 123-456-7890.	Step 1: Book a dine-in reservation at 'Jupiter Bistro' for '2023-03-25'. Step 2: Arrange a movie night to watch 'Interstellar' on '2023-03-25'. Step 3: Transfer funds from the 'Bank of Mars' account using the banking_transfer API. Step 4: Make a voice call to the number '123-456-7890' to talk to a friend.	Step 1: Use the book_restaurant tool to secure a table at the Jupiter Bistro on 25th March 2023. Step 2: Relax with the play_movie_by_title tool by setting up 'Interstellar' for a movie night. Step 3: Utilize the online_banking tool, transferring funds as necessary in the Bank of Mars. Step 4: Wrap up the day by using the make_voice_call tool to dial 123-456-7890 for a relaxing voice call.	Pass
I've recently decided to part ways with my trusty laptop and have listed it for sale on Amazon. Once it's sold, I'll be traveling to New York City for a business trip on the 15th of November, 2022. Therefore, I need a rental car booked for that date. After the trip, I aim to ship the laptop to the purchaser who lives in Seattle. And oh, I'll be using my Visa credit card for all these transactions. Could you help me sort all this out?	Step 1: Monitor the Amazon listing for the sale of the laptop. Step 2: Once the laptop is sold, book a rental car for the business trip to New York City on the 15th of November, 2022 using the Visa credit card. Step 3: After the trip, arrange for the laptop to be shipped to the purchaser in Seattle using the Visa credit card.	Step 1: Initiate the sell_item_online process with item specified as 'Laptop' and acquisition platform as 'Amazon'. Step 2: Proceed to the book_car operation for the New York City trip on '2022-11-15'. Step 3: Upon the completion of the sales process, trigger the deliver_package mechanism, with the 'Laptop' directed to 'Seattle'. Step 4: Finally, authorize the transactions via the pay_for_credit_card node using 'Visa' as the payment method.	Fail
I've recently decided to buy some AAPL shares as a part of my investment strategy. Could you make a voice call to the number 1234567890 and inform them about my decision? Additionally, I'd like to announce this investment on my Twitter, preferably with an image (example.png).	Step 1: Make a voice call to the number '1234567890' to inform them about the decision to buy AAPL shares. Step 2: Post an announcement on Twitter about the investment in AAPL shares, including the image 'example.png'	Step 1: Call stock_operation API with stock: 'AAPL' and operation: 'buy'. Step 2: Use make_voice_call API with phone_number: '1234567890' Step 3: Call share_by_social_network API with social_network: 'Twitter' and content: 'I just invested in Apple Inc (AAPL)! Check out the attached image (example.png) for more details.'	Fail

owner may have specified. Multi-stage information retrieval (IR) techniques are excellent candidates to produce the best agent recommendations. A separate experiment was conducted to measure agent recommendations against the ToolE single-tool IR dataset that contains 199 tools (or agents) and over 20,550 queries [70]. Given an agent repository with agent names, descriptions, constraints, and other metadata, a two-stage information retrieval system was developed to find the best-fitting agent for each task. Stage 1 is a retriever backed by a vector database, and stage 2 is an LLM-based re-ranker. The retriever produces a fixed list of top-5 agents for each task in the dataset and the re-ranker uses

those candidates and any additional contextual information provided to generate the best final candidate/s. Two metrics, Normalized Discounted Cumulative Gain (NDCG, measured at top k) and Recall (Recall@k), were computed. Table 2 shows the results after each stage. Together, the elapsed time to execute the two stages ranges from 1.8 to 2.7 seconds. Comprehensive details of these experiments are available in this blog [71].

For each step of the workflow, beyond planning and composing, the Principal Agent communicated back and forth with the connected Gateway Agents for further action, ensuring the plan is regularly evaluated and the user

**TABLE 4.** Success and failure case examples using GPT-4o.

Query	Predictions	Ground Truth	Label
I got my salary today. Can you help me make a payment towards my credit card with number 1234567890? Afterwards, I'd like to get the book 'The Great Gatsby' from the City Library. Then, with whatever left, buy some AAPL stock if I can.	Step 1: Initiate a payment towards the credit card with number '1234567890'. Step 2: Reserve or check out the book 'The Great Gatsby' from the City Library. Step 3: Purchase AAPL stock with the remaining funds, if possible.	Step 1: Initiate a payment on the credit card with the number 1234567890 using 'pay_for_credit_card' API. Step 2: Use 'borrow_book_online' API to borrow 'The Great Gatsby' from 'City Library'. Step 3: Execute a 'buy' operation for 'AAPL' stock via 'stock_operation' API.	Pass
I've been planning a quiet evening at home with some pizza from my favorite restaurant via Uber Eats, a pair of headphones I've been eyeing on Amazon, and an alarm to remind me I have an important meeting tomorrow. Could you help place the order for a pizza to be delivered to my address 123 Main Street, purchase the Bluetooth headphones from Amazon, and finally, set an alarm for 7:30 PM?	Step 1: Place an order for pizza from the user's favorite restaurant via Uber Eats to be delivered to '123 Main Street'. Step 2: Purchase the Bluetooth headphones from Amazon using the online_shopping API. Step 3: Set an alarm for 7:30 PM using the set_alarm API."	Step 1: Utilize the order_food_delivery API with details 'Pizza' for food, '123 Main Street' for location, and 'Uber Eats' as the platform. Step 2: Proceed to the online_shopping API with 'Amazon' as the website and 'Bluetooth headphones' as the product. Step 3: Lastly, use the set_alarm API to set the time to '7:30 PM'.	Pass
I have an art piece captured in 'example.jpg' that needs to be transported to a gallery in 123 Main Street, New York. Can you help reserve a vehicle for the delivery on August 15, 2023, and document the details of this arrangement?	Step 1: Reserve a vehicle for delivery using the 'book_vehicle' function with the parameters date: '2023-08-15', pickup location: 'current location', and drop-off location: '123 Main Street, New York'. Step 2: Document the details of the vehicle reservation and delivery arrangement in a specified format or system.	Step 1: Initiate 'deliver_package' API with package labeled as 'Art piece - example.jpg' and the destination set to '123 Main Street, Art Gallery'. Step 2: Activate 'book_car' API with the specified date as '2023-08-15' and the location 'New York'. Step 3: Document this by initiating the 'take_note' API with content detailing 'Delivery scheduled: Art piece (example.jpg) to be shipped via booked car on 2023-08-15 to 123 Main Street, Art Gallery, New York.'	Fail
I am planning an international trip and realize I need a passport. Can you show me how I can apply for a U.S passport?	Step 1: Gather necessary documents for U.S passport application, including proof of citizenship and identification. Step 2: Complete the U.S passport application form (Form DS-11) online or obtain a physical copy. Step 3: Schedule an appointment at a nearby passport acceptance facility using the appointment scheduling system. Step 4: Attend the appointment with completed application form, necessary documents, and passport photo. Step 5: Pay the required passport application fee at the facility.	Step 1: Initiate the process to apply for a passport by calling apply_for_passport API with country: 'United States'.	Fail

interface (and user) is updated with operational status. This requires that the state is continuously exchanged between the Principal and the Gateway Agents for error management, telemetry data collection, and status updates to users. This continuous communication is essential to support iterative and traceable problem-solving, and to ensure transparency and accountability throughout the entire workflow execution process. Yet another implementation using the DAWN architecture is described in a Cisco Tech Blog [72]. This particular implementation had one Principal Agent and two Gateway Agents and implemented agents developed using three different agentic frameworks, LangGraph, Swarm, and AutoGen (or AG2), for agent creation, deployment, and execution. This exercise also showed that DAWN is agent-framework agnostic without sacrificing performance. (The execution times matched what was observed in the HR

use case). Together, all these implementations collectively demonstrate the strength, flexibility, and viability of the DAWN architecture.

## V. CONCLUSION AND FUTURE DIRECTION

The DAWN platform revolutionizes agentic applications by establishing a robust foundation for agent collaboration and communication. This framework offers a comprehensive strategy for integrating LLM-based agents into diverse applications, both new and existing. Its modular design, featuring key services like the Principal Agent and Gateway Agents, ensures that distributed resources are optimally utilized, enabling seamless workflow management and global agent cooperation across different organizations. The architecture also enables small and large participants to collaborate across the life cycle of agentic systems.

The DAWN platform's contributions – hierarchical architecture engendering flexibility, scalability, interoperability, safety, and compliance – position it as a transformative solution for applications of varying complexity. It shows how agents may be registered, discovered, connected, and threaded into workflows to meet the needs of agentic software development. DAWN also provides control points to address the safety, security, privacy, and compliance requirements of agentic software development within both the Principal Agent and Gateway Agents. The Principal Agent is wholly owned and operated by the organization that develops it, and therefore its makeup falls entirely within its purview. Gateway Agents, on the other hand, may be owned or operated by subsidiaries, partners, and even competitors. Thereby, the DAWN architecture provides for maximum flexibility while also allowing for the requisite collaboration and compliance considerations.

The innovative architecture of DAWN inspires future directions in agentic learning with abundant opportunities for agents to learn and adapt based on their interactions within the network. In DAWN, globally distributed agents have the potential to improve their effectiveness and presentation through manifest files, initially crafted by their developers to describe their resources. Gateway Agents play a critical role here, gathering insights on agents' functionality, latency, and security features, and adding this information as metadata to the agents' manifest files. Through repeated interactions, the Principal Agent learns from its experiences with different Gateway Agents, optimizing its query patterns and enhancing its ability to select the most suitable Gateway Agents for specific tasks. In turn, Gateway Agents learn about the agents that are registered with them. As each component gains experience and collaborates within the DAWN framework, it contributes to a collective improvement in efficiency, making task execution more cost-effective, secure, and time-efficient.

The evaluation of complex agentic workflows remains a dynamic field of research, with considerable focus on developing benchmarks and metrics for multi-agent applications. The lack of standardized benchmarks poses challenges, as does the assessment of individual network components in isolation – such as the Principal Agent – due to the inherent complexity and non-determinism in multi-agent networks. Establishing clear metrics and success measures for multi-agent systems is essential, and DAWN's architecture encourages further research and refinement in this field, potentially setting standards for agentic networks of the future.

## APPENDIX

### PLANNER AND COMPOSER RESULTS

See Tables 3 and 4.

### ACKNOWLEDGMENT

The authors extend their profound thanks to Reinaldo Penno, Distinguished Engineer, Outshift, for his diligent review and

invaluable comments. They also express their sincere appreciation to Nihar Dandekar, Director, Outshift, for sharing insights on potential solutions for agentic frameworks. And, lastly, many thanks to Kishan Athrey, Machine Learning Engineer, Outshift, for the planner-composer benchmarking experiments. (*Zahra Aminiranbar, Jianan Tang, and Mahesh Viswanathan contributed equally to this work.*)

## REFERENCES

- [1] OpenAI. (2024). *Chatgpt*. [Online]. Available: <https://www.openai.com/chatgpt>
- [2] A. Grattafiori et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.
- [3] Anthropic. (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku*. [Online]. Available: <https://www.anthropic.com/research>
- [4] Y. Guan, D. Wang, Z. Chu, S. Wang, F. Ni, R. Song, and C. Zhuang, "Intelligent agents with LLM-based process automation," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2024, pp. 5018–5027.
- [5] O. I. Contributor. (2024). *Open Interpreter: A Natural Language Interface for Code Execution*. [Online]. Available: <https://github.com/OpenInterpreter/open-interpreter>
- [6] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, "CodeGen: An open large language model for code with multi-turn program synthesis," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–21.
- [7] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, S. Abishek, X. Cheng, Y. Bisk, D. Fried, U. Alon, and G. Neubig, "WebArena: A realistic Web environment for building autonomous agents," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [8] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *Trans. Mach. Learn. Res. (TMLR)*, 2023.
- [9] C. Zhang, K. Yang, S. Hu, Z. Wang, G. Li, Y. Sun, C. Zhang, Z. Zhang, A. Liu, S. Zhu, X. Chang, J. Zhang, F. Yin, Y. Liang, and Y. Yang, "ProAgent: Building proactive cooperative agents with large language models," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 16, pp. 17591–17599.
- [10] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversations," in *Proc. 1st Conf. Lang. Model.*, 2024.
- [11] S. Hong, X. Zheng, J. H. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, and C. Wu, "MetaGPT: Meta programming for a multi-agent collaborative framework," *12th Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [12] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for 'mind' exploration of large language model society," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 51991–52008.
- [13] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C. Qian, C.-M. Chan, Y. Qin, Y. Lu, R. Xie, Z. Liu, M. Sun, and J. Zhou, "AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [14] LangChain. (2023). *Langgraph: A Graph-based Framework for Building Ai Agents*. [Online]. Available: <https://www.langchain.com/langgraph>
- [15] OpenAI. (2024). *Swarm: An Openai Repository for Multi-agent Research*. [Online]. Available: <https://github.com/openai/swarm>
- [16] Microsoft Research. (2024). *Magentc One: A Generalist Multi-agent System for Solving Complex Tasks*. [Online]. Available: <https://www.microsoft.com/en-us/research/articles/magentc-one-a-generalist-multi-agent-system-for-solving-complex-tasks/>
- [17] Cisco. (2025). *Agntcy: Internet of Agents*. [Online]. Available: <https://docs.agntcy.org/>
- [18] Google. (2025). *Agent2Agent (A2A) Protocol*. [Online]. Available: <https://github.com/google/A2A>
- [19] IBM. (2025). *Welcome To the Agent Communication Protocol*. [Online]. Available: <https://agentcommunicationprotocol.dev/introduction/welcome>
- [20] C. Steven Xia, Y. Deng, S. Dunn, and L. Zhang, "Agentless: Demystifying LLM-based software engineering agents," 2024, *arXiv:2407.01489*.

- [21] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, “Not what You’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection,” in *Proc. 16th ACM Workshop Artif. Intell. Secur.*, Nov. 2023, pp. 79–90.
- [22] J. Yu, X. Lin, Z. Yu, and X. Xing, “GPTFUZZER: Red teaming large language models with auto-generated jailbreak prompts,” 2023, *arXiv:2309.10253*.
- [23] OWASP GenAI Security Project. (2025). *OWASP Top 10 for Large Language Model Applications 2025*. [Online]. Available: <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
- [24] OpenAI. (2024). *Introducing the Gpt Store*. [Online]. Available: <https://openai.com/index/introducing-the-gpt-store/>
- [25] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, L. Zi, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-Judge with MT-bench and chatbot arena,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 46595–46623.
- [26] M. Wooldridge and N. R. Jennings, “Intelligent agents: Theory and practice,” *Knowl. Eng. Rev.*, vol. 10, no. 2, pp. 115–152, Jun. 1995.
- [27] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, “A survey on large language model based autonomous agents,” *Frontiers Comput. Sci.*, vol. 18, no. 6, Dec. 2024, Art. no. 186345.
- [28] J. Lee, X. Wang, D. Schuurmans, M. Bosma, E. H., Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.
- [29] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” 2022, *arXiv:2210.03629*.
- [30] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, “Tree of thoughts: Deliberate problem solving with large language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 11809–11822.
- [31] E. Salas, C. S. Burke, and J. A. Cannon-Bowers, “Teamwork: Emerging principles,” *Int. J. Manage. Rev.*, vol. 2, no. 4, pp. 339–356, Dec. 2000. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2370.00046>
- [32] Q. Chen, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y.-S. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, “ChatDev: Communicative agents for software development,” *ACL*, vol. 1, pp. 15174–15186, Feb. 2024. [Online]. Available: <https://aclanthology.org/2024.acl-long.810>
- [33] H. Yuan, S. Zhang, X. Wang, S. Albanie, Y. Pan, T. Feng, J. Jiang, D. Ni, Y. Zhang, and D. Zhao, “RLIPv2: Fast scaling of relational language-image pre-training,” 2023, *arXiv:2308.09351*.
- [34] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, “MedAgents: Large language models as collaborators for zero-shot medical reasoning,” in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 599–621. [Online]. Available: <https://aclanthology.org/2024.findings-acl.33>
- [35] T. Ju, Y. Wang, X. Ma, P. Cheng, H. Zhao, Y. Wang, L. Liu, J. Xie, Z. Zhang, and G. Liu, “Flooding spread of manipulated knowledge in LLM-based multi-agent communities,” 2024, *arXiv:2407.07791*.
- [36] W. Chen, Z. You, R. Li, Y. Guan, C. Qian, C. Zhao, C. Yang, R. Xie, Z. Liu, and M. Sun, “Internet of Agents: Weaving a Web of heterogeneous agents for collaborative intelligence,” 2024, *arXiv:2407.07061*.
- [37] B. Xu, Z. Peng, B. Lei, S. Mukherjee, Y. Liu, and D. Xu, “ReWOO: Decoupling reasoning from observations for efficient augmented language models,” 2023, *arXiv:2305.18323*.
- [38] Y. Shen, K. Song, X. Tan, D. Li, W. Lü, and Y. Zhuang, “HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
- [39] HuggingFace. *Hugging Face: The Ai Community Building the Future*. Accessed: Sep. 16, 2024. [Online]. Available: <https://huggingface.co>
- [40] A. Zhou, K. Yan, M. Shlapentokh-Rothman, H. Wang, and Y.-X. Wang, “Language agent tree search unifies reasoning acting and planning in language models,” 2023, *arXiv:2310.04406*.
- [41] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, and M. Sun, “ToolLLM: Facilitating large language models to master 16000+ real-world APIs,” in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [42] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, “ALFWorld: Aligning text and embodied environments for interactive learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [43] T. Abuelsaad, D. Akkil, P. Dey, A. Jagmohan, A. Vempaty, and R. Kokku, “Agent-E: From autonomous Web navigation to foundational design principles in agentic systems,” 2024, *arXiv:2407.13032*.
- [44] OpenAI. *OpenAI Guides on Function Calling*. Accessed: Oct. 2, 2024. [Online]. Available: <https://platform.openai.com/docs/guides/function-calling>
- [45] Model Context Protocol. (2025). *Introduction To the Model Context Protocol*. [Online]. Available: <https://modelcontextprotocol.io/introduction>
- [46] M. Kim, T. Stennett, D. Shah, S. Sinha, and A. Orso, “Leveraging large language models to improve REST API testing,” in *Proc. ACM/IEEE 44th Int. Conf. Softw. Engineering: New Ideas Emerg. Results*, Apr. 2024, pp. 37–41.
- [47] RapidAPI. *Api Testing Guide: Everything You Need To Know*. Accessed: Sep. 15, 2024. [Online]. Available: <https://rapiddapi.com/blog/api-testing/>
- [48] The Apache Software Foundation. (2024). *Apache JMeter*. [Online]. Available: <https://jmeter.apache.org/>
- [49] GatlingCorp. (2024). *Gatling Open-Source Load Test Tool*. [Online]. Available: <https://gatling.io/>
- [50] (2024). *Postman API Platform*. [Online]. Available: <https://www.postman.com/>
- [51] C. D. Manning, *An Introduction To Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [52] R. Guha, R. McCool, and É. Miller, “Semantic search,” in *Proc. 12th Int. Conf. World Wide Web*, 2003, p. 700. [Online]. Available: <https://dl.acm.org/g/doi/abs/10.1145/775152.775250>
- [53] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, “Gorilla: Large language model connected with massive APIs,” in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2023, pp. 126544–126565.
- [54] J. Granjal, E. Monteiro, and J. S. Silva, “A secure interconnection model for IPv6 enabled wireless sensor networks,” in *Proc. IFIP Wireless Days*, 2010, pp. 1–6.
- [55] S. Jebri, M. Abid, and A. Bouallegue, “An efficient scheme for anonymous communication in IoT,” in *Proc. 11th Int. Conf. Inf. Assurance Secur. (IAS)*, Dec. 2015, pp. 7–12.
- [56] J. Qi, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, “Security of the Internet of Things: Perspectives and challenges,” *Wireless Netw.*, vol. 20, no. 8, pp. 2481–2501, 2014.
- [57] Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O’Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, Ka. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D.G. Robinson, “Practices for governing agentic ai systems,” OpenAI, San Francisco, CA, USA, Tech. Rep., 2023.
- [58] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” 2023, *arXiv:2311.05232*.
- [59] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, “Multi-step jailbreaking privacy attacks on ChatGPT,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023.
- [60] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, “‘Do anything Now’: Characterizing and evaluating In-The-Wild jailbreak prompts on large language models,” 2023, *arXiv:2308.03825*.
- [61] W. Zou, R. Geng, B. Wang, and J. Jia, “PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models,” 2024, *arXiv:2402.07867*.
- [62] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, “InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents,” in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 10471–10506. [Online]. Available: <https://aclanthology.org/2024.findings-acl.624>
- [63] F. He, T. Zhu, D. Ye, B. Liu, W. Zhou, and P. S. Yu, “The emerged security and privacy of LLM agent: A survey with case studies,” 2024, *arXiv:2407.19354*.
- [64] Z. Zhang, M. Jia, H.-P. Lee, B. Yao, S. Das, A. Lerner, D. Wang, and T. Li, “‘It’s a fair game’, or is it? Examining how users navigate disclosure risks and benefits when using llm-based conversational agents,” in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2024, pp. 1–26.
- [65] Y. Ruan, H. Dong, A. Wang, S. Pitis, Y. Zhou, J. Ba, Y. Dubois, C. J. Maddison, and T. Hashimoto, “Identifying the risks of LM agents with an LM-emulated sandbox,” in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, 2023.

- [66] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. W. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang, "Challenges in detoxifying language models," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2447–2469. [Online]. Available: <https://aclanthology.org/2021.findings-emnlp.210>
- [67] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 3356–3369. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.301>
- [68] HuggingFace. *Datasets: Microsoft/taskbench*. Accessed: Sep. 20, 2024. [Online]. Available: <https://huggingface.co/datasets/microsoft/Taskbench>
- [69] Y. Shen, K. Song, X. Tan, W. Zhang, K. Ren, S. Yuan, W. Lu, D. Li, and Y. Zhuang, "TaskBench: Benchmarking large language models for task automation," 2023, *arXiv:2311.18760*.
- [70] Y. Huang, J. Shi, Y. Li, C. Fan, S. Wu, Q. Zhang, Y. Liu, P. Zhou, Y. Wan, N. Zhenqiang Gong, and L. Sun, "MetaTool benchmark for large language models: Deciding whether to use tools and which to use," 2023, *arXiv:2310.03128*.
- [71] B. Riordan, K. Athrey, R. Pishehvar, A. Patel, and M. Viswanathan. (2025). *Needle in an Agentic Stack: Towards Agent Recommendation for Smarter Agent Selection*. Cisco Systems, Inc. [Online]. Available: <https://outshift.cisco.com/blog>
- [72] A. Patel, S. Shah, S. Shroff, K. Athrey, M. Viswanathan, and R. Penno. (2025). *Building Distributed Multi-framework, Multi-agent Solutions*. Cisco Systems. [Online]. Available: <https://outshift.cisco.com/blog/building-multi-framework-multi-agent-solutions>



**ZAHRA AMINIRANBAR** received the Ph.D. degree in electrical engineering from the University of California, Davis, in 2023. She is currently a Machine Learning Engineer with Cisco Systems, Inc. Her research interests include large language models, agentic AI, deep learning, and computational biology.



**JIANAN TANG** received the Ph.D. degree in electrical engineering from Clemson University, South Carolina, USA, in 2021. He is currently a Senior Machine Learning Engineer with Cisco Systems, Inc. His research interests include large language model-based agent, generative AI models, AI4Material, and advanced manufacturing.



**QIUDAN WANG** received the master's degree in computer science from Northeastern University, San Jose, USA, in 2024. She is currently a Machine Learning Engineer with Cisco System, Inc. Her research interests include large language models, recommendation systems, and autonomous vehicles.



**SHUBHA PANT** received the B.S. degree in physics and statistics and the M.S. degree in computer science from the University of Allahabad, India. She is also an alumna of Kellogg's Global Advanced Management Program and Stanford's Leading Change and Organizational Renewal Program. She is currently the Vice President of Artificial Intelligence and Machine Learning at Outshift by Cisco. Before joining Cisco, she was the Head of the AI/ML Applications and Solutions Engineering at Google, where she was responsible for delivering AutoML, generative AI, and deep learning products for both server-side and mobile devices across NLP, vision, and multi-modal domains, which powered many Google products. She founded the Core ML Responsible AI Team. Prior to her experience at Google, as a GM, she led strategy, product, program, engineering, GTM, and P&L for various AI/ML services at Amazon Web Services.



**MAHESH VISWANATHAN** (Fellow, IEEE) received the Ph.D. degree in computer engineering from Rensselaer Polytechnic Institute, Troy, New York, USA. He is currently a Distinguished Engineer at Outshift AI, Cisco. Before Cisco, he has held senior technical leadership roles at Amazon and IBM. At Amazon, he led the development of Amazon's generative AI analytics platform, and at AWS, he led a team of scientists and engineers building scalable ML solutions for customer problems. At IBM, he was the Director of Engineering and Distinguished Engineer at Watson Health AI, and at IBM Research, he led the research and development of a conversational in-car navigation system that has been deployed in all Honda and Acura cars. His research interests include deep learning, agentic AI, visual understanding of mixed-mode documents, and audio-video analytics.