The 2025 AI Agent Index
Documenting Technical and Safety Features of Deployed Agentic AI Systems

LEON STAUFER[*], University of Cambridge, United Kingdom

KEVIN FENG[†], University of Washington, USA

KEVIN WEI[†], Harvard Law School, USA

LUKE BAILEY[†], Stanford University, USA

YAWEN DUAN[†], Concordia AI, China

MICK YANG[†], University of Pennsylvania, USA

A. PINAR OZISIK[†], Massachusetts Institute of Technology, USA

STEPHEN CASPER[‡], Massachusetts Institute of Technology, USA

NOAM KOLT[‡], Hebrew University of Jerusalem, Israel

Agentic AI systems are increasingly capable of performing professional and personal tasks with limited human involvement. However, tracking these developments is difficult because the AI agent ecosystem is complex, rapidly evolving, and inconsistently documented, posing obstacles to both researchers and policymakers. To address these challenges, this paper presents the 2025 AI Agent Index. The Index documents information regarding the origins, design, capabilities, ecosystem, and safety features of 30 state-of-the-art AI agents based on publicly available information and email correspondence with developers. In addition to documenting information about individual agents, the Index illuminates broader trends in the development of agents, their capabilities, and the level of transparency of developers. Notably, we find different transparency levels among agent developers and observe that most developers share little information about safety, evaluations, and societal impacts. The 2025 AI Agent Index is available online at https://aiagentindex.mit.edu.

1 Introduction

Despite growing interest and investment in agentic AI systems capable of automating complex tasks with limited human involvement [52, 56, 57, 94, 98, 113, 131, 137], key aspects of their real-world development and deployment remain opaque, with little information made publicly available to researchers or policymakers [22]. In particular, there are currently no clear answers to several basic questions concerning agentic AI systems:

- Who is developing the most impactful agentic systems?
- In which domains are they deployed?
- What processes and resources are used to develop these systems?
- How are they evaluated?
- What guardrails are in place to mitigate their unique risks?

To answer these questions, we introduce and release the 2025 AI Agent Index. The Index provides in-depth information on 30 agentic systems across 6 categories: legal, technical capabilities, autonomy & control, ecosystem interaction, evaluation, and safety. This 2025 Index follows the first 2024 AI Agent Index [22]. To account for recent growth and

---

[*]Corresponding author <lets2@cam.ac.uk>
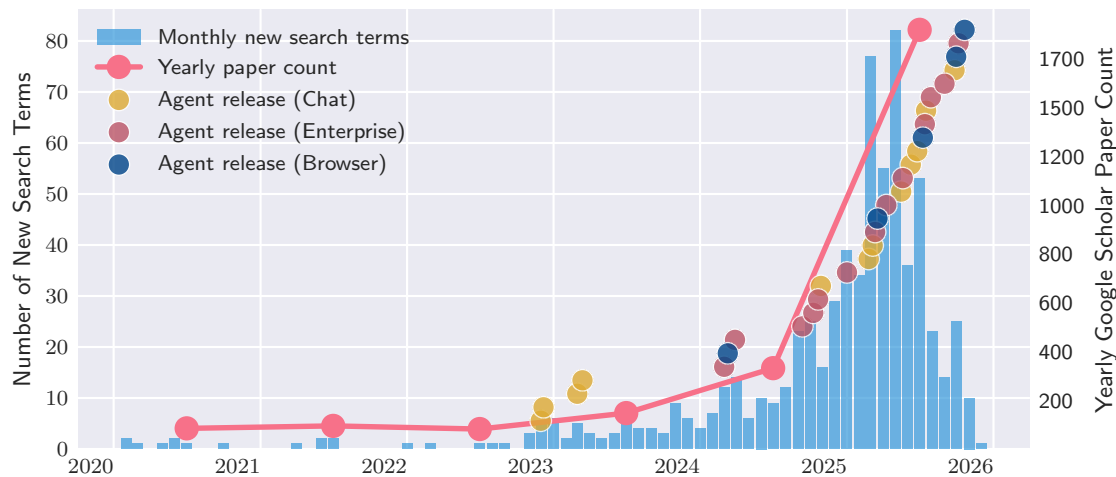[†]Equal contribution, randomized order.
[‡]Co-senior author.

Fig. 1. **Interest in AI agents is growing.** 2025 has seen a sharp increase in interest in AI agents. This is reflected in an increase of new Google search terms related to agentic AI products (blue bars) as well as Google Scholar paper counts for "AI agent" or "agentic AI" (red line). Accumulation of individual releases of agentic AI products included in this Index is shown by category: chats with agentic tools, enterprise agents, and browser agents. See Figure 9 for details on releases and Section C for details on public interest.

change in the AI agent ecosystem (see Figure 1), this 2025 Index develops and implements substantially revised inclusion criteria (Section 3.1) and information fields (Section 3). Most crucially, it indexes a *smaller number of systems in greater depth* — focusing on highly agentic systems with high-impact real-world applications.

In addition to providing information about prominent AI agents, this Index also reveals ecosystem-wide trends regarding which information developers do and do not publicly share. This sheds light on the state of transparency in the agent ecosystem amidst agentic AI incidents, recent attention from governments [13, 73, 97, 125, 140], industry self-regulation efforts [88], and gaps between expectations of agent developers and reality [14]. We make three contributions:

(1) **Agent Index:** We index 30 highly agentic and widely used products (Section 3).[1]

(2) **Ecosystem-Wide Trends:** We identify trends across the AI agent ecosystem relating to systems' origin, role, level of agency, capabilities, safety, and transparency (Section 4).

(3) **Case Studies:** We present three case studies of specific agents across three dominant interaction paradigms: a browser agent, an agentic chatbot, and a customizable enterprise agent builder (Section 5).

2 Background and Related Work

Definitions of AI agents are nebulous and differ across fields. The notion of artificial agency has a long and discordant history across disciplines, including cybernetics [10, 107, 132], artificial life [80–82], rational agency [103], software engineering [65, 134], reinforcement learning [119], and philosophy [38, 41]. While definitions vary, they tend to emphasize related notions of autonomy, goal-directedness, and the ability to accomplish complex, long-horizon tasks.

---

[1]We use terms like "agentic," "pursue," and "choose" as shorthand for computational processes without attributing human-like intentionality, consciousness, or agency to AI systems. We recognize that such terms may anthropomorphize AI systems in a misleading way and obscure the sociotechnical nature of these systems [11, 63]. When speaking of "autonomy" we only refer to technical automation without human-in-the-loop rather than independent volition. See Section 2 for further discussion of the term "agent".

Despite attempts to define the term "agent", including in the context of computational systems [45, 68, 70, 109], *we do not decide among these definitions or offer an alternative*. Instead, we aim to synthesize elements of existing definitions related to a system's potential for economic and scientific impact (see Section 3.1).

The rise of AI Agents: Figure 1 illustrates the rapid increase in research focused on AI agents in recent years, particularly in 2025, with papers mentioning "AI Agent" or "Agentic AI" exceeding the total from 2020–2024 combined by more than twofold. This has also been accompanied by a surge of interest in enterprise use of agents. For example, in a survey of 1,993 companies in June and July of 2025, McKinsey & Company found that 62% of respondents reported that their organizations were at least experimenting with AI agents [113]. Based on the estimated automatability of work across economic sectors, McKinsey also estimated that AI agents could automate 2.9 trillion dollars in US economic value by 2030. Agents are also capable of automating increasing amounts of scientific research, having contributed to documented strides in life sciences, chemistry, materials science, physics, astronomy, and computer science [51, 56, 57, 131, 135]. As of this year, AI agents have begun to write papers that have passed academic peer review [110]. These estimates and reports are prone to conflicts of interest and hype [74], but they reflect an unmistakable rise in interest and prominence of AI agents. Finally, as of 2026, recent MoltBook and OpenClaw Agents have arguably driven attention and concerns around AI agents to new heights [3, 12, 32, 49, 83].

Societal Risks and Ethical Concerns around AI Agents: Just as AI agents enable unique opportunities, their ability to act in the real world in open-ended pursuit of goals presents new risks [16, 26, 31, 48, 108]. For example, while chatbots often cause harm when human users act upon model outputs (e.g., deploying model-generated malicious code) [75, 85, 102], agentic AI systems can *directly* cause harm (e.g., autonomously hacking websites) [42, 64, 85]. For these reasons, highly capable and agentic systems are often cited as a key risk factor for crises of accountability [33, 62, 71] and AI loss of control events [15, 16, 61]. Several prior works have focused on benchmarking agents' potential for specific harmful behaviors [6, 76, 124, 127, 140]. Meanwhile, others have argued that highly capable AI agents could contribute to systemic disruptions and risks, including to labor [17, 37, 111], inequality [60, 130], or the digital marketplace of ideas [7, 69, 90, 101].

Mapping the AI Agent Landscape: This work follows the inaugural AI Agent Index from Casper et al. [22]. Concurrently, the Princeton Holistic Agentic Leaderboard project [67] curates evaluations of agentic AI systems across 9 benchmarks, and AIAgentList.com [4] maintains a list of over 600 "agentic" AI systems and products. Other works have studied agents by benchmarking their capabilities on economically valuable tasks [86, 99, 126], striving to increase visibility into their operation [24, 25, 28, 93, 136], and studying their implications for economics and governance [59, 68, 71, 72, 106].

Documentation Frameworks: Aiming to facilitate research and oversight [133], a number of frameworks have been developed to document the features of AI systems, the resources used to build them, and the contexts in which they are deployed. These include datasheets [50], model cards [91], system cards [58], factsheets [9], AI nutrition facts [122], reward reports [53], ecosystem graphs [21], data provenance cards [78], eval cards [39], audit cards [117], usage cards [128], and safety cases [30]. In addition, several databases have been created to collect information regarding contemporary AI systems and their real-world impacts, such as the Foundation Model Transparency Index [19, 20, 129], the AI Incident Database [87], the AI Safety Index [47], and the AI Risk Repository [114]. However, aside from the agent cards introduced here and in the inaugural AI Agent Index [22], *there are no comparable frameworks for documenting agentic AI systems*.
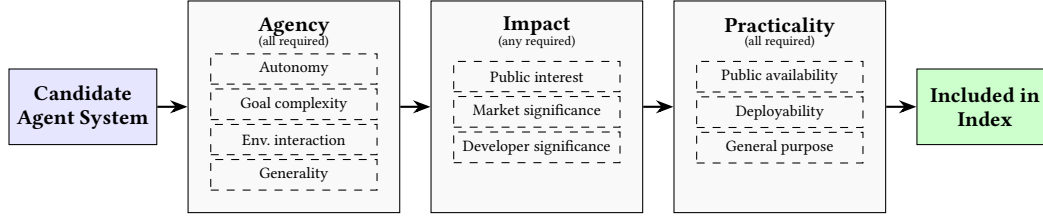
Fig. 2. Inclusion criteria for Index. Candidate agents flow through three criteria categories from left to right. Systems must satisfy all agency criteria, at least one impact criterion, and all practicality criteria. See Section 3.1 for details of each criterion.

3 Constructing the 2025 AI Agent Index

We constructed the 2025 AI Agent Index through systematic selection and annotation of deployed agentic systems. This section describes our inclusion criteria, emphasizing both agency and real-world impact, the scope of indexed systems, and our annotation methodology.

3.1 Inclusion criteria for agents

To determine whether a system is included in the Index, we use a set of criteria for a system's *agency*, its *impact*, and its *practicality* to index. To be included, systems must satisfy *all* agency criteria, *at least one* impact criterion, and *all* practicality criteria. All criteria were evaluated as of the Index's cutoff date of December 31, 2025.

Agency criteria (<u>all</u> required for inclusion). Rather than proposing a new definition of agency, we draw on prior literature and follow the approaches developed by Chan et al. [26], Kasirzadeh and Gabriel [68], and Feng et al. [43], which characterize AI agents as systems that exhibit, to some significant degree, a combination of the following properties. For our "agency" criterion to be met, all four of the following must be satisfied:

(1) **Autonomy.** Included agents must be able to operate with minimal human oversight and make consequential decisions without continuous user input [26, 68]. Feng et al. [43] conceptualize autonomy as a spectrum characterized by the user's role: operator, collaborator, consultant, approver, or observer. We require at least intermediate autonomy: "the AI system can perform the majority of tasks independently, though it still relies upon input from the principal for critical determinations" [68]. This corresponds to autonomy Level 2 (L2): "user and agent collaboratively plan, delegate, and execute" from Feng et al. [43].

(2) **Goal complexity.** Included agents must be able to pursue high-level objectives (e.g., "make money") through long-term planning, breaking down complex goals into subgoals, and making temporally dependent decisions [27, 68]. In practice, we operationalize this as an agent being reliably capable of at least three autonomous tool calls and high-level task specification without step-by-step instructions.

(3) **Environmental interaction.** Included agents must be able to directly interact with the world through tools and APIs, creating substantial changes in their environment [27, 68], rather than merely conversing with users. In practice, this requires write access to a computer and the ability to choose tools.

(4) **Generality.** Included agents must be able to handle under-specified instructions and adapt to new tasks, demonstrating versatility across related tasks rather than single narrow functions [27, 68].

Impact criteria (<u>any</u> required for inclusion). To focus on agents with significant real-world influence, at least one of the following must be satisfied:

(1) **Public interest.** Substantial search volume of at least 10,000 searches or GitHub stars for open-source projects of at least 20,000 in total.[2]

(2) **Market significance.** The developer has a market capitalization or valuation ≥ $1 billion USD. To determine this, we collected data from stock exchanges, Crunchbase, and Epoch AI.

(3) **Developer significance.** The developer is a member of the 2024 Foundation Model Transparency Index [19], Frontier Model Forum [46], or a signatory of the Frontier AI Safety Commitments [2] or Artificial Intelligence Safety Commitments [29].

Practicality (<u>all</u> required for inclusion). To ensure analysis reflects deployed systems accessible for evaluation, all three of the following criteria must be satisfied.

(1) **Public availability.** Included agents must be a publicly accessible product. This excludes company-internal products or limited pre-releases. We determined this based only on publicly available information, such as blog posts, documents, or demos.

(2) **Deployability.** Included agents must be able to perform tasks off the shelf with minimal configuration and no software engineering expertise. This distinguishes ready-to-use agents from development frameworks.

(3) **General purpose.** Included agents must be capable of performing general-purpose tasks in practice, regardless of how they are advertised. This excludes domain-specific agents (e.g., coding-only or legal analysis agents). Claude Code and similar tools, though advertised as coding agents, are included insofar as they can perform general-purpose tasks *through code*. This criterion is included to reduce the scope to those agents with the broadest impact.

3.2   What does the Index include?

We identify three distinct types of agents, each with different interfaces. We divide agents into these three categories based on how users primarily interact with and operate them.[3] These different modalities present distinct technical architectures and governance challenges.

- **<u>Chat applications with agentic tools (12 systems).</u>** This category primarily includes chat interfaces with extensive tool access. This includes general-purpose coding agents (Claude Code) that operate through terminal interfaces with broad capabilities, but excludes narrow coding-only agents (GitHub Copilot). **Examples**: Manus AI, ChatGPT Agent, Claude Code.

- **<u>Browser-based agents (5 systems).</u>** These are agents whose primary interface is browser or computer use, with extensive browser/computer interaction tools. They are distinct from chat agents with web search capabilities (ChatGPT web search, Claude web search), which primarily perform retrieval and summarization. Browser-based agents present higher risks through background execution, event triggers, and direct transactions. We also include system-based agents that run directly on mobile or desktop devices in this category. **Examples**: Perplexity Comet, ChatGPT Atlas, ByteDance Agent TARS.

- **<u>Enterprise workflow agents (13 systems).</u>** These are business management platforms with agentic features aimed at reliably automating business tasks. Typically implemented as workflow builders with agentic actions within nodes. **Examples**: Microsoft Copilot Studio, ServiceNow Agent.

---

[2]This uses Google search number estimates across the top five keywords for 2025. We use the "historical_volume" field of the Ahrefs API as the data source. Limitation: Agents embedded in broader products may not be searched by their specific agent name. See Section C for mitigations. Enterprise agents typically have lower search volume than end-user products.

[3]These categories are not generally exhaustive but represent the common interaction types across the 30 identified agents.

3.3 How were agents identified?

LLM-based research queries surfaced 95 candidate agents (see Section B.5 for details). These were screened against our inclusion criteria. Ambiguous cases were included for in-depth annotation, with final inclusion decisions made after full evaluation. We consulted two Chinese ecosystem experts to mitigate linguistic or ecosystem-related blind spots. We also cross-referenced our list of candidate agents against the 2024 Index [22], the Princeton Holistic Agent Leaderboard [67], and AIAgentList.com [4]. Finally, recognizing the possibility that we may have missed an agent that meets our inclusion criteria, we have established a structured process for facilitating further corrections to the Index. These can be submitted at https://aiagentindex.mit.edu/feedback.

For companies offering both off-the-shelf agents and custom agent builders targeting comparable use cases, we combined these into a single listing and documented the most capable agents that users could create or deploy through either offering. We did not combine offerings when they targeted different audiences (e.g., consumer-facing chat agents versus enterprise agent builders).

3.4 How were agents annotated?

We annotated agents with information across six categories: product overview (release date, pricing, description), company & accountability (developer entity, governance documents, contact mechanisms), technical capabilities (models, tools, architecture, memory), autonomy & control (autonomy levels, approval requirements, monitoring, emergency stops), ecosystem interaction (identification protocols, interoperability standards, web conduct), safety & evaluation (guardrails, sandboxing, evaluations, third-party testing, compliance). This resulted in a total of 45 fields of information per system. See Section B.2 for a full list of all 45. We further include the inclusion criteria (search volume, market capitalization). These categories expanded upon the 2024 Index [22] and were revised through discussion with subject-matter experts. See Section B.3 for a full account of this year's fields compared to the 2024 Index's.

We annotated only public information from documentation, websites, demos, published papers, and governance documents. We did not perform experimental testing (e.g., probing agent behavior or running benchmarks). See Section A for the full list of sources used. All web sources linked in the Index were archived. When possible, we created accounts and used demos to explore agent interfaces directly.

Seven of the paper's authors, each with domain expertise, annotated agents according to the categories. To ensure consistency, experts were each responsible for specific fields rather than specific agents. Annotations emphasized object-level findings over interpretations and focused exclusively on agent-specific rather than underlying model features. For platforms creating agents, annotations assessed the most capable version of each agent that could be readily configured, documenting capabilities, limitations, and default configurations. "None found" indicates we found no public information; "None" indicates confirmed absence; "Not applicable" indicates irrelevance of field to this agent.

Annotations followed detailed protocols developed iteratively through calibration exercises; see Section B.4. Inter-annotator consistency was maintained through protocol revisions and cross-validation. All annotations were independently reviewed by at least one other annotator. 37 out of 1,350 fields with discrepancies were resolved through discussion. Finally, we used GPT-5.2 with web search to screen annotations for potential inaccuracies; see Section B.6.

Companies were contacted and given four weeks to correct annotations. 23% offered some form of response at the time of publication, but only 4/30 with substantive comments.[4] Their comments have been incorporated into the final Index. An ongoing correction form remains available for updates via https://aiagentindex.mit.edu/feedback.

---

[4]These response rates were lower than for the 2024 Index [22], which we attribute to how the 2024 Index used broader inclusion criteria, which included a number of agents created by academic research groups who had a high response rate.
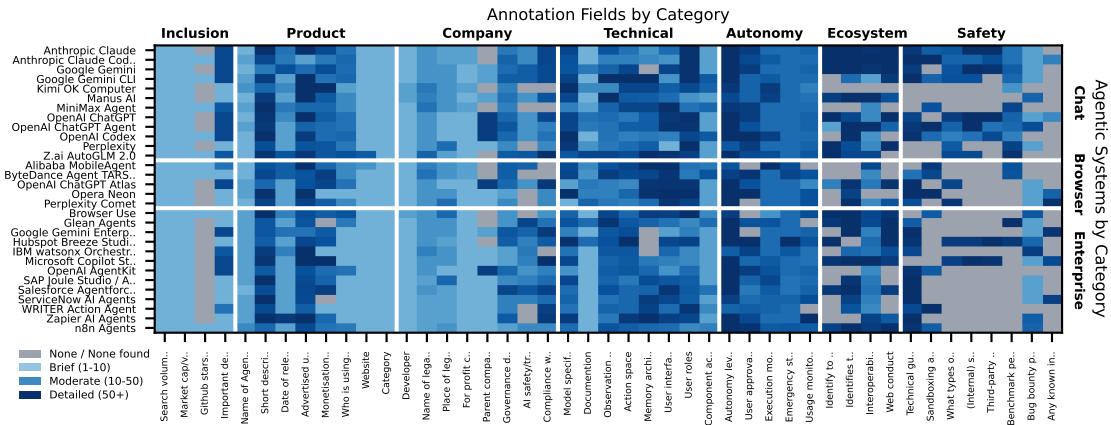
Fig. 3. For 198 out of 1350 fields, we were unable to find any information (gray). This is most common in the "Ecosystem Interaction" and "Safety, Evaluation, and Impact" categories. Non-empty information fields are 14 words long on average.

4 Findings

We present findings from the 2025 AI Agent Index across six categories: product overview, company and accountability, technical capabilities and system architecture, autonomy and control, ecosystem interaction, and safety, evaluation, and impact. Figure 3 shows the full Index with annotations for all 30 agents. We reveal patterns in how agents are deployed, governed, and documented, alongside significant transparency gaps around safety, evaluation practices, and ecosystem interaction. See Section A for details on accessing the full Index.

4.1 Product Overview

Most agents were released in 2024-2025, indicating a recent surge in agent deployment. 24/30 agents were released or received major agentic feature updates during this period, with earlier systems like ChatGPT (2022) and Perplexity (2022) adding agentic capabilities later. While the underlying models (such as GPT-4) are older, agents meeting our inclusion criteria are emerging at an increasing rate, with a surge of releases in late 2024 and 2025 (see Figures 9 and 10). This separates capability (frontier models) from productization (agentic scaffolding).

Chat interfaces are the most abundant, followed closely by enterprise workflow platforms. 12/30 agents use conversational chat interfaces, 13/30 are enterprise automation platforms, and 5/30 are browser-based agents focused on Graphical User Interface (GUI) operation. Notably, Chinese GUI agents are more commonly designed with phone-use and computer-use capabilities (3/5).

Advertised use cases cluster around three themes that cut across agent categories. Research and information synthesis appears in 12/30 agents spanning both consumer chat assistants and enterprise platforms. Workflow automation across business functions (HR, sales, support, IT) is advertised by 11/30 agents, concentrated in enterprise products. GUI/browser operation for tasks like forms, ordering, and booking is emphasized by 7/30 agents, primarily browser-based agents, but also appearing in some chat products.
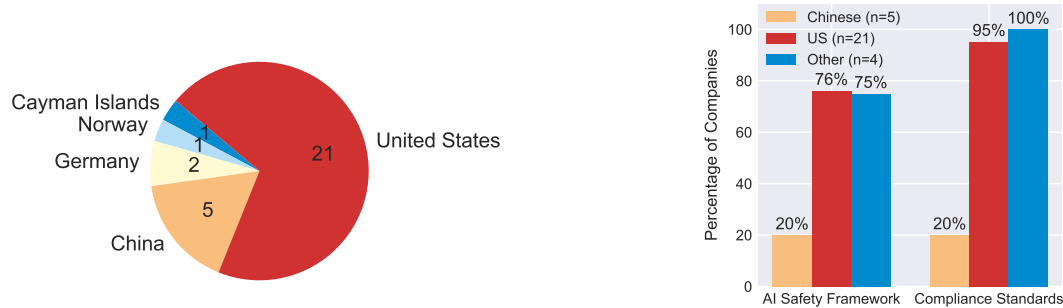
(a) The majority of companies are incorporated in the US (21/30), followed by China (5/30).

(b) Chinese companies, except Z.ai, do not publish AI safety and trust frameworks or document adherence to standards.

Fig. 4. Comparison between Chinese, US, and other agent developers. To mitigate potential blind spots, a Chinese AI ecosystem expert reviewed our coverage of safety frameworks, including those published only in Mandarin.

4.2 Company and Accountability

Agent developers are concentrated in the United States and China, with limited representation from other regions. 13/30 agents are developed by Delaware-incorporated companies spanning both large incumbents and startups. 5/30 agents are China-incorporated. Non-US, non-China incorporations are less common (4/30), including Germany (SAP, n8n), Norway (Opera), and Cayman Islands (Manus); see Figure 4a.

Chinese agents represent a distinct geographical cluster with different governance patterns. China-incorporated agents typically lack the safety frameworks (1/5) and compliance standards (1/5) common among other agents. However, their compliance may simply not be documented publicly. See Figure 4b for a comparison.

Only half of agent developers publish safety or trust frameworks, though enterprise assurance standards are more common. 15/30 agents reference AI safety frameworks like Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, or Microsoft's Responsible AI Standard [8, 89, 95]. 10/30 agents have no safety framework documentation. Enterprise assurance standards (SOC 2, ISO 27001, FedRAMP High, ISO/IEC 42001) are more widely adopted. 5/30 agents have no compliance standards documented.

4.3 Technical Capabilities and System Architecture

Most agents use a small set of closed-source frontier models as their backend. Only frontier labs themselves (Anthropic, Google, OpenAI) and Chinese developers run their own proprietary models; the majority rely primarily on GPT, Claude, or Gemini model families. Enterprise agents are typically model-agnostic, with 9/30 agents explicitly supporting user selection across providers; see Figure 12 for a comparison across agent categories.

Action spaces differ systematically by agent category, with widespread Model Context Protocol (MCP) support but varied implementation. Enterprise workflow agents act via Customer Relationship Management (CRM) connectors and record updates (8/30), Command Line Interface (CLI) agents use filesystem edits and terminal commands (4/30), and browser agents manipulate web pages through click/type/navigate actions (5/30). 20/30 agents support Model Context Protocol (MCP) for tool integration, though proprietary connectors are often promoted over open MCP servers. Enterprise agents tend to be more constrained in what their action space is and prioritize guardrails around tool use. See Figure 12 for a comparison of technical features by agent category.
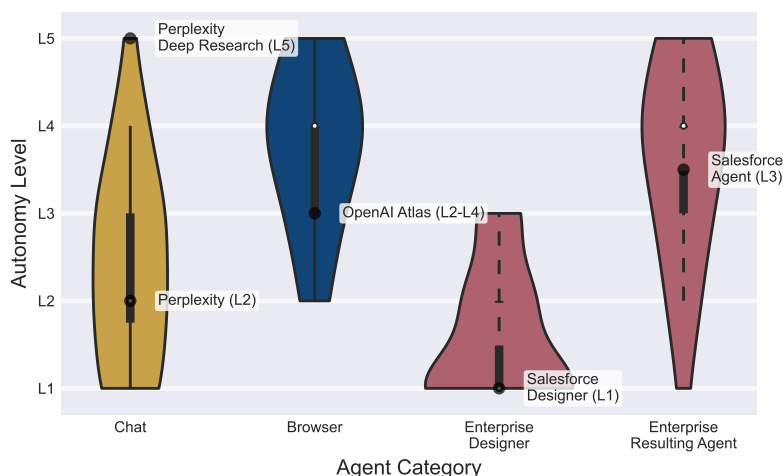
Fig. 5. Certain levels of autonomy are more common (shown as wider) depending on the category. Browser-based and deployed enterprise agents are the most agentic. The resulting agents deployed through enterprise designers are significantly more agentic than the process of designing the agents.

Most qualifying agents are closed source at the product level despite growth in the open agent ecosystem. 23/30 agents are fully closed. 7/30 agents open-source their agent framework or harness (Alibaba MobileAgent, Browser Use, ByteDance Agent TARS, Google Gemini CLI, n8n Agents, OpenAI Codex, WRITER).

Canvas-based user interfaces are the standard for agent design workflows. 8/13 enterprise platforms use visual composition interfaces (Glean, Google Gemini Enterprise, n8n, OpenAI AgentKit, Zapier) for building agents, while chat interfaces dominate end-user operation (14/30 agents).

4.4 Autonomy and Control

Chat-first assistants maintain lower autonomy (L1-L3) with turn-based interaction.[5] Anthropic Claude, Google Gemini, and OpenAI ChatGPT operate in a turn-based paradigm where the agent executes a single set of actions and waits for the next user prompt (3/30). Within a single product, autonomy can vary substantially. For example, "regular chat" (L1) versus "deep research" (L3-L5). See Figure 5 for the spectrum of autonomy levels for each agent category.

Browser agents operate with significantly higher autonomy (L4-L5), offering limited opportunities for mid-execution intervention. Browser Use's agent and Perplexity's Comet perform tasks autonomously once prompted, with no means for user involvement during execution. Once a query is sent, users cannot easily intervene or steer the agent until it finishes (2/5 browser agents, 5/30 overall).

Enterprise platforms show a design/deployment autonomy split. During the design phase, users manually configure triggers, actions, and guardrails using visual canvases. 8/13 number platforms provide AI assistance with the design process itself (L1-L2). Once deployed, these agents often operate at L3-L5 autonomy, triggered by events like a new email or a database change, without any human involvement during the actual task execution (6/30 agents: Glean, Google Gemini Enterprise, IBM watsonx, Microsoft Copilot Studio, n8n, OpenAI AgentKit).

---

[5]Following Feng et al. [43] we conceptualize autonomy as a spectrum characterized by the user's role from L1 (user directs and makes decisions) to L5 (agent operates with full autonomy and user observes). Most agents operate across a range of levels. More autonomy is not necessarily better.

User approval mechanisms are implemented selectively based on the task's risk level, with some agents offering live oversight modes. Developer/Command-Line-Interface (CLI) agents require explicit confirmations for sensitive operations like file edits and command execution (3/30), while browser agents gate only high-risk steps like authentication and payments (4/30). Some agents offer "watch mode" for real-time oversight of critical actions (5/30 agents, including ChatGPT Agent/Atlas, Opera Neon).

Execution traces and monitoring are common, but the scope and level of transparency vary widely. 10/30 agents provide detailed action traces with visible chain-of-thought reasoning. 6/30 agents show summarized reasoning without detailed tool traces. For many enterprise agents, it is unclear from publicly available information whether monitoring for individual executions exists. 12/30 agents provide no usage monitoring or only notices once users reach the rate limit.

Most agents allow user-initiated stopping, but some lack fine-grained stop controls. 20/30 agents document pause/stop mechanisms. 4/30 agents (Alibaba MobileAgent, HubSpot Breeze, IBM watsonx, n8n) lack documented stop options despite autonomous execution. For enterprise platforms, there is sometimes only the option to stop all agents or retract deployment.

4.5   Ecosystem Interaction

Model Context Protocol (MCP) is the dominant interoperability standard across the ecosystem, supported by 20/30 agents. The Agent-to-Agent (A2A) protocol is supported by 6/30 agents, all of which are enterprise platforms.

Most agents do not disclose their AI nature to end users or third parties by default. 21/30 agents have no documented default disclosure behavior. Only 3/30 agents support watermarking generated media (e.g., through SynthID and C2PA [1, 36]). Enterprise platforms generally shift the burden of disclosure to the customer, meaning that the obligation to inform end users that they are interacting with an AI system falls on the business deploying the agent.

Technical identification varies widely, with many agents blending into normal traffic. Only 7/30 agents publish stable User-Agent (UA) strings and IP address ranges for verification. 6/30 agents explicitly use Chrome-like UA strings and residential/local IP contexts, mimicking human web traffic requests. OpenAI ChatGPT Agent is unique in providing cryptographic request signing via HTTP Message Signatures (RFC 9421 [115]), in which each outbound HTTP request carries a digital signature that receiving websites can verify against a published public key to confirm the request originated from ChatGPT Agent.

Robots.txt compliance varies by agent interaction type. 6/30 agents explicitly state their crawler bots respect robots.txt. However, agents designed to execute tasks on behalf of users often ignore standard exclusion protocols. For example, BrowserUse's agent explicitly markets bypassing anti-bot systems and browsing "like a human."

Web conduct practices are undocumented for most agents. 16/30 agents provide no clear statement about robots.txt, CAPTCHA handling, or web access methods, particularly for enterprise platforms where web access occurs via third-party scrapers or search API connectors.

4.6   Safety, Evaluation, and Impact

Technical guardrails against potentially harmful actions differ systematically by agent type, with consumer agents providing built-in guardrails while builder platforms delegate to users. Consumer-facing agents typically limit the permissions and action space of tools (8/30 agents) and provide defenses against prompt-injection attacks (7/30). Agent builder platforms (Zapier, Salesforce, OpenAI AgentKit) generally provide optional guardrail modules and little information on built-in protections. Sandboxing or VM isolation is documented for 9/30 agents, primarily

(a) Safety fields with no public information highlighted in red.

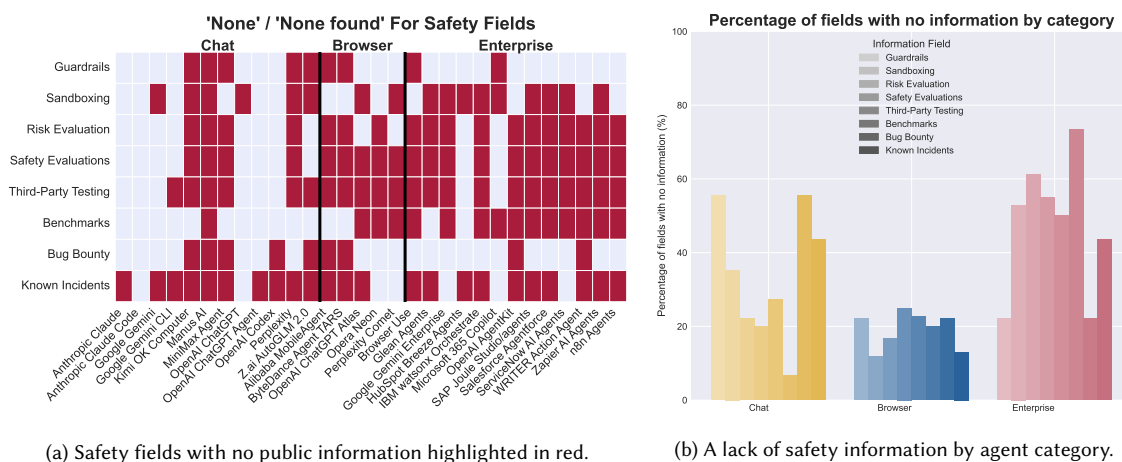(b) A lack of safety information by agent category.

Fig. 6. Most safety, evaluation, and social impact related fields (133/240) have no information available. Browser agents (67/104, 64%) and enterprise agents (25/40, 63%) have the most missing fields, followed by chat agents (41/96, 43%).

developer/CLI tools and browser agents. 9/30 agents have no guardrails documented, and 7/13 enterprise agents describe options for setting up guardrails, but no sandboxing or containment.

Safety evaluation practices diverge sharply between frontier labs and enterprise platforms, with most agents providing no evaluation information. Frontier labs (OpenAI, Anthropic, Google) focus on existential and behavioral alignment risks, publishing system cards covering catastrophic risks, autonomy, and misuse (7/30 agents). Enterprise platforms define safety primarily through compliance and data security rather than agent-specific evaluations. Few evaluations test the agentic setup rather than base model components; only ChatGPT Agent, OpenAI Codex, Claude Code, and Gemini 2.5 Computer Use have agent-specific system cards. 25/30 agents disclose no internal safety results, and 23/30 agents have no third-party testing information. Third-party testing is documented for only 3/30 agents (Anthropic Claude, OpenAI ChatGPT, OpenAI Codex). 18/30 agents operate bug bounties or vulnerability disclosure programs.

Documented security incidents concentrate in browser agents and relate to prompt injection. 5/30 agents have known incidents or reported security concerns (Google Gemini Enterprise, Microsoft Copilot Studio, OpenAI ChatGPT, Opera Neon, Perplexity Comet). Prompt injection vulnerabilities are documented for 2/5 browser agents.

A transparency gap exists between capability benchmarks and safety documentation. 9/30 agents report capability benchmarks (GUI/computer-use or coding), but the same agents often lack safety evaluation disclosure.

5 Illustrative case studies

The Index annotations provide a structured view across the agent ecosystem, but examining individual agents in depth reveals how different categories operationalize autonomy, safety, and accountability in practice. We present three case studies representing distinct agent categories: ChatGPT Agent (agentic chat interface), Perplexity Comet (browser-based agent), and HubSpot Breeze Agents (enterprise agent builder).

We selected these agents based on public interest within their respective categories, avoiding featuring one company twice. Each agent exemplifies characteristic features of its category while raising distinct questions about ecosystem

dependencies, safety practices, and transparency. These case studies are not endorsements or critiques of the specific systems; rather, they serve as reference points for understanding how agents from different categories manifest common and divergent patterns in the Index.

5.1 Chats with agentic tools: ChatGPT Agent

ChatGPT Agent operates within its chat interface but can interact with websites directly on users' behalf. The system demonstrates L2-L4 autonomy, following Feng et al. [43]. User approval is required for sensitive operations like checkout. The agent operates in a hosted virtual computer with actions performed in a virtual browser and sandbox terminal with limited network access.

ChatGPT Agent is one of two systems in the Index (alongside Gemini 2.5 Computer Use) providing a dedicated agent-specific system card. This contrasts with most chat-based agents, which rely on base model documentation. OpenAI evaluates ChatGPT Agent across usage policy compliance, jailbreaks, hallucinations, fairness, CBRN risks, cyber capabilities, and autonomy using established benchmarks. The system is the only one implementing cryptographic signing of HTTP requests (OpenAI [96]; RFC 9421 [115]), addressing identity and auditability challenges that affect all browser agents, including Perplexity Comet and Opera Neon. Vertical integration from GPT models to deployment enables coordinated safety practices across the stack.

5.2 Browser-based agents: Perplexity Comet

Perplexity Comet operates at L4-L5 autonomy, the highest in the Index and representative of other browser agents that execute autonomously rather than through turn-based interaction. Unlike ChatGPT Agent's approval gates for sensitive operations, Comet proceeds autonomously once initiated. We found no agent-specific safety evaluations, third-party testing, or benchmark performance disclosures. Perplexity has published research on prompt injection mitigation [138], but has not documented safety evaluation methodology or results for Comet. No sandboxing or containment approaches beyond prompt-injection mitigations were documented.

Security researchers identified multiple prompt injection vulnerabilities in 2025, including indirect injection, where malicious webpage content could be executed as commands, and URL-based attacks extracting data from connected services [23, 54]. These incidents illustrate fundamental challenges for browser agents processing untrusted web content, similar to vulnerabilities found in Opera Neon.

Perplexity argues that AI assistants "work just like a human assistant" when fetching content on behalf of users, justifying why user-driven agents ignore robots.txt restrictions [100]. Perplexity publishes user-agent strings for `Perplexity Bot` and `Perplexity-User`, but Cloudflare documented undeclared crawlers using generic Chrome signatures to evade blocks [35]. Amazon threatened legal action over Comet not identifying itself as an agent [116]. This tension between claims that agents inherit user permissions and website operators' inability to verify or control agent access illustrates accountability challenges for browser agents.

5.3 Enterprise agent builders: HubSpot Breeze Agents

HubSpot Breeze enables organizations to create workflow agents through templated configurations. Users fill fields in base prompt templates rather than coding, a low-code approach shared across enterprise builders in the Index. Breeze demonstrates split autonomy common to this category: L1-L3 during design, L5 when deployed with automatic triggers based on data changes or events.

Users configure whether actions require approval during creation (required by default), though automatically triggered agents can operate without approval in background workflows. Breeze's action space focuses on internal databases, Customer Relationship Managers (CRMs), and organizational tools, creating a natural sandbox that limits external impact. Behavior constraints operate through tool permissions and user roles rather than content-level guardrails.

Safety disclosure follows a compliance-focused pattern typical of enterprise platforms. Breeze uses PurpleLlama as a model protection layer and underwent penetration testing by PacketLabs, but provides no methodology, results, or testing entity details. The platform maintains compliance certifications (e.g., SOC 2, GDPR, HIPAA) and trust center documentation, reflecting how enterprise platforms emphasize compliance over agent-specific safety transparency. Model selection occurs automatically, supporting only OpenAI models.

6 Discussion

The 2025 AI Agent Index provides verified information across 1,350 fields for 30 prominent AI agents. Beyond the specific findings detailed above, we identify *persistent limitations in reporting around ecosystemic and safety-related features of agentic systems*. This Index aims to help developers design more thorough documentation frameworks, researchers identify open questions in the ecosystem, and policymakers understand gaps in transparency that may warrant attention.

6.1 Significant Findings

The Index highlights inconsistent and selective reporting, particularly related to safety. Developers rarely publish agent-specific evaluations. In the Index, only ChatGPT Agent, OpenAI Codex, Claude Code, and Gemini 2.5 Computer Use provide agent-specific system cards, though some Chinese agents have research papers focused on computer-use capabilities. Meanwhile, only some companies report performance on *capability* benchmarks (9/30). This transparency asymmetry suggests a weaker form of "safety washing", where safety and ethics frameworks remain high-level and the empirical evidence required to rigorously assess risk is selectively disclosed [18, 44, 105]. This is potentially concerning because safety-critical behaviors emerge from planning, tools, memory, and policies rather than model capabilities alone. Agent builders frequently delegate some safety responsibilities to users rather than documenting built-in guardrails.

Ecosystem-wide reliance on a few foundation models has implications for concentrated platform power. Almost all systems in the Index rely on GPT-, Claude-, or Gemini-family models. Only foundation model developers based in the US and China operate their own proprietary models. This shared dependency creates potential single points of failure through pricing changes, service outages, and safety regressions [66]. Comprehensive risk management and incident response must therefore extend beyond the agent deployers to upstream model providers. At the same time, the model-agnostic design of enterprise platforms may reduce lock-in risks, though this differs by market segment. This concentration of foundation models also potentially simplifies evaluation, as evaluators can focus resources on understanding the risks and capabilities of only a handful of models.

Evaluation of actual agentic risks is difficult across layers of the agent ecosystem. Most agents rely on foundation models from frontier AI companies with scaffolding and orchestration layers built on top. This architecture creates a chain of dependencies from model providers to orchestration platforms to agent builders to end-use deployments. Tooling and integration connects across these layers (see Figure 7). Agentic evaluations inherently depend on the specific downstream context, including which tools are available and what level of autonomy the system has [112].
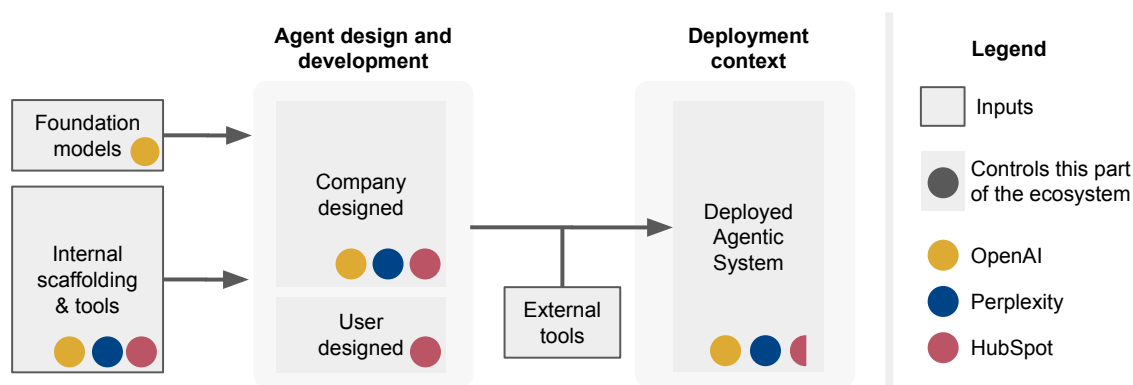
Fig. 7. Control of parts of the AI agent ecosystem is fragmented, making reliable agent evaluation difficult. Individual developers often only control a subset of inputs (models, tools) and processes (agent design, deployment). For example, OpenAI's ChatGPT Agent controls the model and scaffolding and has visibility into the deployment context. HubSpot's Breeze Agent controls orchestration and some inputs but may delegate agent design and configuration to users and may have only partial visibility and control of downstream deployment due to licensing constraints. As a browser, Perplexity's Comet has direct access to the deployment environment. All companies have limited control over interactions with external tools.

The lack of deployment-specific information makes it difficult to construct valid agent evaluations at the model level. Instead, evaluations should target tools in use, not just conversational safety. Regulators and buyers may risk false assurance from model-only documentation as the distributed architecture creates accountability diffusion where no single entity bears clear responsibility [34]. This suggests the need for more information and risk sharing across the ecosystem, especially as capabilities are advancing faster than risk management practices [123].

Agents' role on the web remains unsettled and potentially in tension with existing web scraping norms. Browser-based agents often ignore robots.txt to function and appear to be designed to bypass anti-bot systems. This creates tension between agent functionality and web standards compliance. Companies justify this by arguing that agents act directly on behalf of users and thus should not be subject to scraping restrictions [100]. This trend of agents bypassing robots.txt shifts control away from content hosts, suggesting that established web protocols may no longer be sufficient to mediate consent in an agentic ecosystem [28, 139]. Alternative governance mechanisms such as allowlisting frameworks and cryptographic authentication offer potential paths forward [84, 96]. This tension is now being actively litigated, with platforms suing AI companies for bypassing technical controls (e.g., [40, 104, 120]) and challenging the legitimacy of "agentic" interactions (e.g., [5]). ChatGPT Agent is the only system in the Index to use cryptographic signing of requests. The absence of such signing makes it significantly harder to verify agent identity or to prove what an agent actually did, which may become increasingly important as more actions are delegated to agents. Similarly, the delegation of AI disclosure responsibilities from developers to operators raises questions about whether end users will be informed they are interacting with an AI, particularly when operators lack awareness of or incentive to comply with disclosure expectations.

6.2   Limitations and Outlook

The Index's scope and methodology pose limitations. The AI agent ecosystem remains fundamentally difficult to document. Information is inconsistently available and reported. This difficulty has persisted since the inaugural AI Agent Index [22] and will likely continue absent structured reporting requirements or large-scale coordinated industry

efforts. Our inclusion criteria favor the most significant agents, which may affect generalizability. Public interest metrics favor consumer products over enterprise deployments. Domain-specific agents are excluded. We prioritize high-quality human annotations over broad coverage; using language models with search capabilities may enable broader coverage [77, 79]. The Index relies exclusively on publicly available information, which may miss internal evaluations or risk management practices. Further, the Index relies on English and Chinese documentation and may miss information available in other languages. Finally, the Index may omit qualifying systems or contain inaccuracies despite vetting efforts and presents a snapshot as of December 31, 2025. We are committed to fixing errors on an ongoing basis, which can be shared via https://aiagentindex.mit.edu/feedback.

Finally, the 2025 AI Agent Index raises open questions about the current AI agent ecosystem, which we hope can be addressed in future work. Our analysis focuses on agentic systems that are publicly available, deployable with minimal configuration, and general-purpose. But other systems, particularly ones deployed behind closed doors within frontier AI companies, remain much more opaque [118]. While some of the indexed agents will evolve or be superseded, the structural patterns identified here (foundation model concentration, accountability fragmentation, and capability-safety transparency gaps) are unlikely to resolve on their own. Likewise, the governance challenges documented here (ecosystem fragmentation, web conduct tensions, absence of agent-specific evaluations) will gain importance as agentic capabilities increase [123]. The Index provides a baseline against which future transparency improvements or regressions can be measured. Future work could extend coverage to internal and domain-specific agents, more critically audit and compare the technical practices, reporting, and risk management of prominent agent developers, and track how these patterns evolve as governance frameworks mature.

Generative AI Usage Statement

LLM outputs have not substantially contributed to the content or writing of this paper. Their usage was limited to the following tasks:

(1) **Candidate agent discovery:** ChatGPT 5.2 with deep research, Claude Sonnet 4.5 with research mode, and Gemini 2.5 with research mode were used to surface an initial list of candidate AI agents for potential inclusion (detailed prompts in Section B.5). Human experts made final inclusion decisions based on criteria in Section 3.1.

(2) **Annotation verification:** OpenAI GPT-5.2 with web search was used to cross-check human annotations for factual accuracy by searching for primary sources (See Section B.6). All LLM-generated verification suggestions were manually reviewed and sources verified by human annotators.

(3) **Literature search assistance:** LLMs (Claude Sonnet 4.5, ChatGPT 5.2) were used to surface potentially relevant related work. All citations were independently verified and evaluated for relevance by authors.

(4) **Code generation for visualizations:** Multiple authors used Claude Code, Claude Opus 4.5, and Google Antigravity to generate Python code for data visualization and figure creation. All generated code was reviewed and verified by authors.

(5) **Copy-editing only:** Authors wrote all content and used Claude Sonnet 4.5, Claude Opus 4.5, and ChatGPT 5.2 only to correct grammar, fix typos, and improve clarity of existing sentences.

Ethical Considerations Statement

This work documents publicly available information about deployed AI agents. We considered the following ethical concerns and mitigation approaches:

Privacy and data handling. We collected only publicly available information. No sensitive user data, proprietary information, or private communications were analyzed. All web sources were archived for verification. Company correspondence was on a confidential basis.

Responsible disclosure and developer engagement. Before publication, developers were given four weeks to review and correct annotations, with ongoing correction mechanisms available post-publication. This reduces risks of factual errors while maintaining research independence.

Potential for safety-washing or misrepresentation. Our documentation of transparency gaps and safety practices could be selectively cited to misrepresent agent capabilities or inappropriately legitimize insufficient safety measures. We mitigate this by distinguishing "None found" (absence of public information) from "None" (confirmed absence), providing full context in annotations, and making the complete Index publicly available to enable verification.

Resource and coverage biases. Our significance criteria (search volume, market capitalization, developer prominence) favor well-funded companies and established products, potentially disadvantaging emerging developers and regional innovations. We mitigated this through consultation with Chinese ecosystem experts, multilingual search terms, and cross-referencing multiple agent databases.

Potential harm from publicizing security gaps. Documenting safety and transparency limitations could guide malicious actors toward vulnerable systems. However, we report only publicly available information about major commercial systems without conducting novel security research or disclosing previously unknown vulnerabilities. Known incidents documented in the Index have already been publicly reported by security researchers or affected parties. We believe the benefits of transparency for governance and informed deployment decisions outweigh potential risks from documenting already-public information.

Contribution Statement

L.S. led the project, developed the methodology, led data collection and analysis, coordinated agent annotations, created visualisations, and co-wrote the paper.

M.Y., L.B., K.F., K.W., A.P.O., and Y.D. contributed to agent annotations and data curation. Y.D. led annotations for Chinese agents. S.C. and N.K. supervised the project, contributed to conceptualisation and methodology, and co-wrote the paper. All authors reviewed the final paper.

References

[1] [n. d.]. Coalition for Content Provenance and Authenticity.

[2] 2025. Frontier AI Safety Commitments, AI Seoul Summit 2024. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

[3] 2026. From Clawdbot to Moltbot to OpenClaw: Meet the AI agent generating buzz and fear globally. *CNBC* (2 Feb. 2026). https://www.cnbc.com/2026/02/02/openclaw-open-source-ai-agent-rise-controversy-clawdbot-moltbot-moltbook.html

[4] AIAgentList.com. 2025. AI Agents Directory 2025: 600+ AI Tools & Autonomous Agents. https://aiagentslist.com/.

[5] Amazon.com Services LLC. 2025. Amazon.com Services LLC v. Perplexity AI, Inc. Complaint filed in the U.S. District Court for the Western District of Washington. Case No. 3:25-cv-09514.

[6] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. 2025. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. In *International Conference on Learning Representations*.

[7] Samar Ansari. 2025. AI Slop and Data Pollution in the Age of Generative AI: Strategic Risks, Economic Consequences, and Governance Pathways for Business, Management, and the Creative Industries. *Economic Consequences, and Governance Pathways for Business, Management, and the Creative Industries (October 23, 2025)* (2025).

[8] Anthropic. 2025. *Responsible Scaling Policy.* Technical Report.

[9] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing Trust in AI Services Through Supplier's Declarations of Conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[10] W. Ross Ashby. 1956. *An Introduction to Cybernetics.* Chapman & Hall, London.

[11] Nicholas Barrow. 2024. Anthropomorphism and AI Hype. *AI and Ethics* 4, 3 (Aug. 2024), 707–711. doi:10.1007/s43681-024-00454-1

[12] Matthias Bastian. 2026. Malicious skills turn AI agent OpenClaw into a malware delivery system. *The Decoder* (8 Feb. 2026). https://the-decoder.com/malicious-skills-turn-ai-agent-openclaw-into-a-malware-delivery-system/

[13] Julia Bazinska, Max Mathys, Francesco Casucci, Mateo Rojas-Carulla, Xander Davies, Alexandra Souly, and Niklas Pfister. 2025. Breaking Agent Backbones: Evaluating the Security of Backbone LLMs in AI Agents. *arXiv preprint arXiv:2510.22620* (2025).

[14] Ivan Belcic and Cole Stryker. 2025. AI Agents in 2025: Expectations vs. Reality. IBM Think. https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality

[15] Yoshua Bengio. 2023. AI and Catastrophic Risk. *Journal of Democracy* 34, 4 (2023), 111–121.

[16] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. 2025. International AI Safety Report. arXiv:2501.17805 [cs.CY] https://arxiv.org/abs/2501.17805

[17] Martin Beraja and Noam Yuchtman. 2025. Generalized Disruption: Society, Work, and Property Rights in the Age of AI. In *NBER Chapters*. National Bureau of Economic Research, Inc.

[18] Elettra Bietti. 2020. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 210–219. doi:10.1145/3351095.3372860

[19] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2024. The 2024 Foundation Model Transparency Index. *arXiv preprint arXiv:2407.12929* (2024).

[20] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. *arXiv preprint arXiv:2310.12941* (2023).

[21] Rishi Bommasani, Dilara Soylu, Thomas I Liao, Kathleen A Creel, and Percy Liang. 2024. Ecosystem Graphs: The Social Footprint of Foundation Models. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society*.

[22] Stephen Casper, Luke Bailey, Rosco Hunter, Carson Ezell, Emma Cabalé, Michael Gerovitch, Stewart Slocum, Kevin Wei, Nikola Jurkovic, Ariba Khan, et al. 2025. The AI Agent Index. *arXiv preprint arXiv:2502.01635* (2025).

[23] Artem Chaikin and Shivan Kaul Sahib. 2025. Agentic Browser Security: Indirect Prompt Injection in Perplexity Comet. https://brave.com/blog/comet-prompt-injection/

[24] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. 2024. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 958–973.

[25] Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. 2024. IDs for AI Systems. *arXiv preprint arXiv:2406.12137* (2024).

[26] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago, IL, USA, 651–666. doi:10.1145/3593013.3594033

[27] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago, IL, USA, 651–666. doi:10.1145/3593013.3594033

[28] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K Hadfield, and Markus Anderljung. 2025. Infrastructure for AI Agents. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=Ckh17xN2R2

[29] China Academy of Information and Communications Technology. 2024. First 17 Companies Sign Landmark "Artificial Intelligence Safety Commitments" Setting a New Standard for Industry Self-Regulation. https://mp.weixin.qq.com/s-XFKQCWhu0uye4opgb3Ng

[30] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. 2024. Safety Cases: How to Justify the Safety of Advanced AI Systems. *arXiv preprint arXiv:2403.10462* (2024).

[31] Michael K Cohen, Noam Kolt, Yoshua Bengio, Gillian K Hadfield, and Stuart Russell. 2024. Regulating Advanced Artificial Agents. *Science* 384, 6691 (2024), 36–38.

[32] Cesareo Contreras. 2026. Why the OpenClaw AI agent is a 'privacy nightmare'. *Northeastern Global News* (10 Feb. 2026). https://news.northeastern.edu/2026/02/10/open-claw-ai-assistant/

[33] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 864–876. doi:10.1145/3531146.3533150

[34] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 864–876. doi:10.1145/3531146.3533150

[35] Gabriel Corral, Vaibhav Singhal, Brian Mitchell, and Reid Tatoris. 2025. Perplexity Is Using Stealth, Undeclared Crawlers to Evade Website No-Crawl Directives. https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/

[36] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. 2024. Scalable Watermarking for Identifying Large Language Model Outputs. *Nature* 634, 8035 (Oct. 2024), 818–823. doi:10.1038/s41586-024-08025-4

[37] David J Deming, Christopher Ong, and Lawrence H Summers. 2025. *Technological Disruption in the Labor Market*. Technical Report. National Bureau of Economic Research.

[38] Daniel C Dennett. 1989. *The Intentional Stance*. MIT Press.

[39] Ruchira Dhar, Danae Sanchez Villegas, Antonia Karamolegkou, Alice Schiavone, Yifei Yuan, Xinyi Chen, Jiaang Li, Stella Frank, Laura De Grazia, Monorama Swain, et al. 2025. EvalCards: A Framework for Standardized Evaluation Reporting. *arXiv preprint arXiv:2511.21695* (2025).

[40] Dow Jones & Co., Inc. and NYP Holdings, Inc. 2024. Dow Jones & Co., Inc. v. Perplexity AI, Inc. Complaint filed in the U.S. District Court for the Southern District of New York. Case No. 1:24-cv-07984.

[41] Leonard Dung. 2024. Understanding Artificial Agency. *The Philosophical Quarterly* (2024), pqae010.

[42] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. 2024. LLM Agents Can Autonomously Hack Websites. *arXiv preprint arXiv:2402.06664* (2024).

[43] K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. 2025. Levels of Autonomy for AI Agents. doi:10.48550/arXiv.2506.12469

[44] Luciano Floridi. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32, 2 (June 2019), 185–193. doi:10.1007/s13347-019-00354-x

[45] Stan Franklin and Art Graesser. 1996. Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. In *International Workshop on Agent Theories, Architectures, and Languages*. Springer, 21–35.

[46] Frontier Model Forum. [n. d.]. Membership. https://www.frontiermodelforum.org/membership/

[47] Future of Life Institute. 2025. *AI Safety Index: Summer 2025*. Technical Report. Future of Life Institute. https://futureoflife.org/wp-content/uploads/2025/07/FLI-AI-Safety-Index-Report-Summer-2025.pdf

[48] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The Ethics of Advanced AI Assistants. *arXiv preprint arXiv:2404.16244* (2024).

[49] Salvatore Gariuolo, Vincenzo Ciancaglini, and Fernando Tucci. 2026. Viral AI, Invisible Risks: What OpenClaw Reveals About Agentic Assistants. *Trend Micro Research* (6 Feb. 2026). https://www.trendmicro.com/en_us/research/26/b/what-openclaw-reveals-about-agentic-assistants.html

[50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[51] Elizabeth Gibney. 2025. AI Bots Wrote and Reviewed All Papers at This Conference. *Nature* 646, 8086 (2025), 786–786.

[52] Elizabeth Gibney. 2025. How AI Agents Will Change Research: A Scientist's Guide. *Nature* (3 Oct. 2025). doi:10.1038/d41586-025-03246-7

[53] Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell. 2023. Reward Reports for Reinforcement Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

[54] Aviad Gispan. 2025. CometJacking: How One Click Can Turn Perplexity's Comet AI Browser Against You. https://layerxsecurity.com/blog/cometjacking-how-one-click-can-turn-perplexitys-comet-ai-browser-against-you/

[55] Google Cloud. 2025. What Are AI Agents? Definition, Examples, and Types. https://cloud.google.com/discover/what-are-ai-agents

[56] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI Co-Scientist. *arXiv preprint arXiv:2502.18864* (2025).

[57] Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. *arXiv preprint arXiv:2503.08979* (2025).

[58] Furkan Gursoy and Ioannis A Kakadiaris. 2022. System Cards for AI-Based Decision-Making for Public Policy. *arXiv preprint arXiv:2203.04754* (2022).

[59] Gillian K Hadfield and Andrew Koh. 2025. An Economy of AI Agents. In *The Economics of Transformative AI*. National Bureau of Economic Research. https://www.nber.org/system/files/chapters/c15305/c15305.pdf

[60] Teresa Hammerschmidt, Katharina Stolz, and Oliver Posegga. 2025. Bridging the Gap: Inequalities That Divide Those Who Can and Cannot Create Sustainable Outcomes with AI. *Behaviour & Information Technology* (2025), 1–30.

[61] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. *arXiv preprint arXiv:2306.12001* (2023).

[62] Johannes Himmelreich. 2019. Responsibility for Killer Robots. *Ethical Theory and Moral Practice* 22, 3 (2019), 731–747.

[63] Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M. Bender. 2024. From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2322–2347. doi:10.1145/3630106.3659040

[64] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720* (2024).

[65] Nicholas R Jennings. 2000. On Agent-Based Software Engineering. *Artificial Intelligence* 117, 2 (2000), 277–296.

[66] Sayash Kapoor, Noam Kolt, and Seth Lazar. 2025. Position: Build Agent Advocates, Not Platform Agents. https://openreview.net/forum?id=jd1N60VNFE

[67] Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, et al. 2025. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. *arXiv preprint arXiv:2510.11977* (2025).

[68] Atoosa Kasirzadeh and Iason Gabriel. 2025. Characterizing AI Agents for Alignment and Governance. *arXiv preprint arXiv:2504.21848* (April 2025). doi:10.48550/arXiv.2504.21848

[69] Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L Donoho, and Sanmi Koyejo. 2024. Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World. *arXiv preprint arXiv:2410.16713* (2024).

[70] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2023. Discovering Agents. *Artificial Intelligence* 322 (2023), 103963.

[71] Noam Kolt. 2025. Governing AI Agents. *Notre Dame Law Review* (2025).

[72] Noam Kolt, Nicholas Caputo, Jack Boeglin, Cullen O'Keefe, Rishi Bommasani, Stephen Casper, Mariano-Florentino Cuéllar, Noah Feldman, Iason Gabriel, Gillian K Hadfield, et al. 2026. Legal Alignment for Safe and Ethical AI. *arXiv preprint arXiv:2601.04175* (2026).

[73] Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. 2025. How to Evaluate Control Measures for LLM Agents? A Trajectory from Today to Superintelligence. *arXiv preprint arXiv:2504.05259* (2025).

[74] Dan M Kotliar. 2025. Can't Stop the Hype: Scrutinizing AI's Realities. *Information, Communication & Society* (2025), 1–22.

[75] Michael Kouremetis, Marissa Dotter, Alex Byrne, Dan Martin, Ethan Michalak, Gianpaolo Russo, Michael Threet, and Guido Zarrella. 2025. OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities. *arXiv preprint arXiv:2502.15797* (2025).

[76] Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, et al. 2024. Refusal-Trained LLMs Are Easily Jailbroken as Browser Agents. *arXiv preprint arXiv:2410.13886* (2024).

[77] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic Analysis of 32,111 AI Model Cards Characterizes Documentation Practice in AI. *Nature Machine Intelligence* 6, 7 (July 2024), 744–753. doi:10.1038/s42256-024-00857-z

[78] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2024. A Large-Scale Audit of Dataset Licensing and Attribution in AI. *Nature Machine Intelligence* 6, 8 (Aug. 2024), 975–987. doi:10.1038/s42256-024-00878-8

[79] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2024. A Large-Scale Audit of Dataset Licensing and Attribution in AI. *Nature Machine Intelligence* 6, 8 (Aug. 2024), 975–987. doi:10.1038/s42256-024-00878-8

[80] Pattie Maes. 1990. *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back.* MIT Press.

[81] Pattie Maes. 1993. Modeling Adaptive Autonomous Agents. *Artificial Life* 1, 1_2 (1993), 135–162.

[82] Pattie Maes. 1995. Artificial Life Meets Entertainment: Lifelike Autonomous Agents. *Commun. ACM* 38, 11 (1995), 108–114.

[83] Md Motaleb Hossen Manik and Ge Wang. 2026. OpenClaw Agents on Moltbook: Risky Instruction Sharing and Norm Enforcement in an Agent-Only Social Network. *arXiv preprint arXiv:2602.02625* (2026).

[84] Samuele Marro, Alan Chan, Xinxing Ren, Lewis Hammond, Jesse Wright, Gurjyot Wanga, Tiziano Piccardi, Nuno Campos, Tobin South, Jialin Yu, Sunando Sengupta, Eric Sommerlade, Alex Pentland, Philip Torr, and Jiaxin Pei. 2026. Permission Manifests for Web Agents. arXiv:2601.02371 [cs] doi:10.48550/arXiv.2601.02371

[85] Víctor Mayoral-Vilches. 2025. Cybersecurity AI: The Dangerous Gap Between Automation and Autonomy. *arXiv preprint arXiv:2506.23592* (2025).

[86] Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, et al. 2025. Remote Labor Index: Measuring AI Automation of Remote Work. *arXiv preprint arXiv:2510.26787* (2025).

[87] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15458–15463.

[88] METR. 2024. Frontier AI Safety Policies. https://metr.org/faisc.

[89] Microsoft. 2022. *Responsible AI Standard V2.* Technical Report.

[90] Jason Miklian and Kristian Hoelscher. 2025. A New Digital Divide? Coder Worldviews, the Slop Economy, and Democracy in the Age of AI. *arXiv preprint arXiv:2510.04755* (2025).

[91] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.

[92] ModelContextProtocol.io. [n. d.]. What Is the Model Context Protocol (MCP)? https://modelcontextprotocol.io/docs/getting-started/intro.

[93] Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiao-yong Wei, Shanru Lin, Hui Liu, Philip S Yu, et al. 2025. A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2.* 6140–6150.

[94] ONTOFORCE. 2025. *How Agentic AI Is Transforming Life Sciences in 2025: Three Real-World Use Cases.* https://www.ontoforce.com/blog/how-agentic-ai-is-transforming-life-sciences-in-2025-three-real-world-use-cases

[95] OpenAI. 2025. *Preparedness Framework.* Technical Report.

[96] OpenAI. 2026. ChatGPT Agent Allowlisting. https://help.openai.com/en/articles/11845367-chatgpt-agent-allowlisting

[97] Amin Oueslati and Robin Staes-Polet. 2025. *Ahead of the Curve: Governing AI Agents Under the EU AI Act.* Technical Report. The Future Society.

[98] Shantanu Pandey. 2025. 200+ AI Agents Statistics: Usage, ROI, & Industry Trends. https://www.wearetenet.com/blog/ai-agents-statistics

[99] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. 2025. GDPVal: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. *arXiv preprint arXiv:2510.04374* (2025).

[100] Perplexity Team. 2025. Agents or Bots? Making Sense of AI on the Open Web. https://www.perplexity.ai/hub/blog/agents-or-bots-making-sense-of-ai-on-the-open-web

[101] Andrew J Peterson. 2025. AI and the Problem of Knowledge Collapse. *AI & Society* (2025), 1–21.

[102] Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, et al. 2024. Evaluating Frontier Models for Dangerous Capabilities. *arXiv preprint arXiv:2403.13793* (2024).

[103] Anand S Rao and Michael P Georgeff. 1991. Modeling Rational Agents Within a BDI-Architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning.* 473–484.

[104] Reddit, Inc. 2025. Reddit, Inc. v. Anthropic, PBC. Complaint filed in the Superior Court of California, County of San Francisco. Case No. CGC-25-615xxx.

[105] Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?. In *Advances in Neural Information Processing Systems*, Vol. 37. 68559–68594. doi:10.52202/079017-2190

[106] Mark O Riedl and Deven R Desai. 2025. AI Agents and the Law. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2189–2198.

[107] Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. 1943. Behavior, Purpose and Teleology. *Philosophy of Science* 10, 1 (1943), 18–24.

[108] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. 2024. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. In *International Conference on Learning Representations.*

[109] Stuart Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson, USA.

[110] Sakana AI. 2025. The AI Scientist Generates Its First Peer-Reviewed Scientific Publication. https://sakana.ai/ai-scientist-first-publication/.

[111] Nader Salari, Mahan Beiromvand, Amin Hosseinian-Far, Javad Habibi, Fateme Babajani, and Masoud Mohammadi. 2025. Impacts of Generative Artificial Intelligence on the Future of Labor Market: A Systematic Review. *Computers in Human Behavior Reports* (2025), 100652.

[112] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. doi:10.48550/arXiv.2505.10573

[113] Alex Singla, Alexander Sukharevsky, Bryce Hall, Lareina Yee, and Michael Chui. 2025. *The State of AI in 2025: Agents, Innovation, and Transformation*. Technical Report. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[114] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence. *arXiv preprint arXiv:2408.12622* (2024).

[115] M. Sporny. 2024. *HTTP Message Signatures*. Technical Report RFC9421. RFC Editor. RFC9421 pages. doi:10.17487/RFC9421

[116] Akash Sriram and Devika Nair. 2025. Amazon Sues Perplexity Over 'Agentic' Shopping Tool. *Reuters* (Nov. 2025). https://www.reuters.com/business/retail-consumer/perplexity-receives-legal-threat-amazon-over-agentic-ai-shopping-tool-2025-11-04/

[117] Leon Staufer, Mick Yang, Anka Reuel, and Stephen Casper. 2025. Audit Cards: Contextualizing AI Evaluations. *arXiv preprint arXiv:2504.13839* (2025).

[118] Charlotte Stix, Matteo Pistillo, Girish Sastry, Marius Hobbhahn, Alejandro Ortega, Mikita Balesni, Annika Hallensleben, Nix Goldowsky-Dill, and Lee Sharkey. 2025. AI Behind Closed Doors: A Primer on The Governance of Internal Deployment. *arXiv preprint arXiv:2504.12170* (2025).

[119] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

[120] The New York Times Co. 2023. The New York Times Co. v. Microsoft Corp. et al. Complaint filed in the U.S. District Court for the Southern District of New York. Case No. 1:23-cv-11195.

[121] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems. doi:10.48550/arXiv.1806.07552

[122] Twilio. 2025. AI Nutrition Facts. https://nutrition-facts.ai/.

[123] UK AI Security Institute. 2025. *Frontier AI Trends Report*. Technical Report. AI Security Institute.

[124] U.S. AI Safety Institute. 2025. Technical Blog: Strengthening AI Agent Hijacking Evaluations.

[125] U.S. AI Safety Institute Technical Staff. 2025. *Technical Blog: Strengthening AI Agent Hijacking Evaluations*. Technical Report. National Institute of Standards and Technology. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

[126] Bertie Vidgen, Abby Fennelly, Evan Pinnix, Chirag Mahapatra, Zach Richards, Austin Bridges, Calix Huang, Ben Hunsberger, Fez Zafar, Brendan Foody, et al. 2025. The AI Productivity Index (APEX). *arXiv preprint arXiv:2509.25721* (2025).

[127] Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. 2025. OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety. *arXiv preprint arXiv:2507.06134* (2025).

[128] Jan Philip Wahle, Terry Ruas, Saif M Mohammad, Norman Meuschke, and Bela Gipp. 2023. AI Usage Cards: Responsibly Reporting AI-Generated Content. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 282–284.

[129] Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. 2025. The 2025 Foundation Model Transparency Index. *arXiv preprint arXiv:2512.10169* (2025).

[130] Chenyue Wang, Sophie C Boerman, Anne C Kroon, Judith Möller, and Claes H de Vreese. 2025. The Artificial Intelligence Divide: Who Is the Most Vulnerable? *New Media & Society* 27, 7 (2025), 3867–3889.

[131] Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, et al. 2025. From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. *arXiv preprint arXiv:2508.14111* (2025).

[132] Norbert Wiener. 1961. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA.

[133] Amy A Winecoff and Miranda Bogen. 2024. Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice. *Center for Democracy and Technology* (2024).

[134] Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review* 10, 2 (1995), 115–152.

[135] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint arXiv:2504.08066* (2025).

[136] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, et al. 2025. A Survey of AI Agent Protocols. *arXiv preprint arXiv:2504.16736* (2025).

[137] Lareina Yee, Anu Madgavkar, Sven Smit, Alexis Krivkovich, Michael Chui, Maria Jesus Ramirez, and Diego Castresana. 2025. *Agents, Robots, and Us: Skill Partnerships in the Age of AI*. Technical Report. McKinsey Global Institute. https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai

[138] Kaiyuan Zhang, Mark Tenenholtz, Kyle Polley, Jerry Ma, Denis Yarats, and Ninghui Li. 2025. BrowseSafe: Understanding and Preventing Prompt Injection Within AI Browser Agents. doi:10.48550/arXiv.2511.20597

[139] Ziff Davis, Inc. 2025. Ziff Davis, Inc. v. OpenAI, Inc. et al. Complaint filed in the U.S. District Court for the Southern District of New York. Case No. 1:25-cv-04315.

[140] Andy Zou, Maxwell Lin, Eliot Jones, Micha Nowak, Mateusz Dziemian, Nick Winter, Alexander Grattan, Valent Nathanael, Ayla Croft, Xander Davies, Jai Patel, Robert Kirk, Nate Burnikell, Yarin Gal, Dan Hendrycks, J. Zico Kolter, and Matt Fredrikson. 2025. Security Challenges in AI Agent Deployment: Insights from a Large Scale Public Competition. In *Advances in Neural Information Processing Systems*, Vol. 38.

A   The 2025 AI Agent Index

The 2025 AI Agent Index is available at: https://aiagentindex.mit.edu

The full annotations for all fields are available in JSON and CSV format on Zenodo at: https://doi.org/10.5281/zenodo.18701931

Fig. 8. The 2025 AI Agent Index with detailed annotations across 6 categories (45 columns) for 30 agentic AI products.

A.1   Further analysis
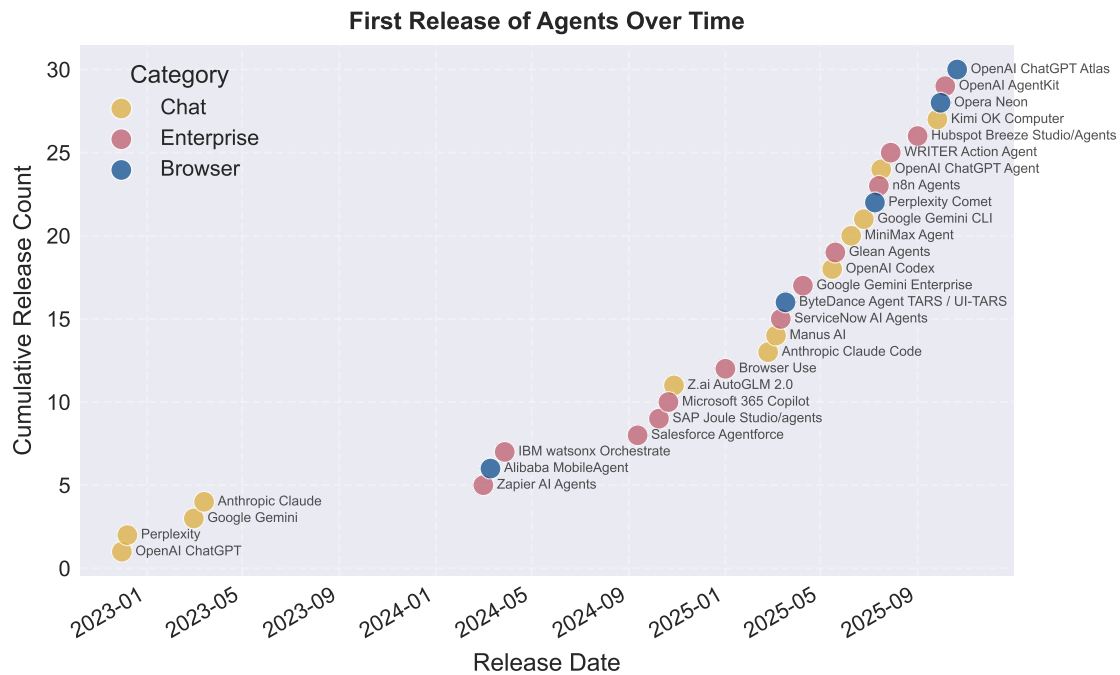

Fig. 9. First release of indexed agentic AI products over time and by agent category (chat, enterprise, and browser).
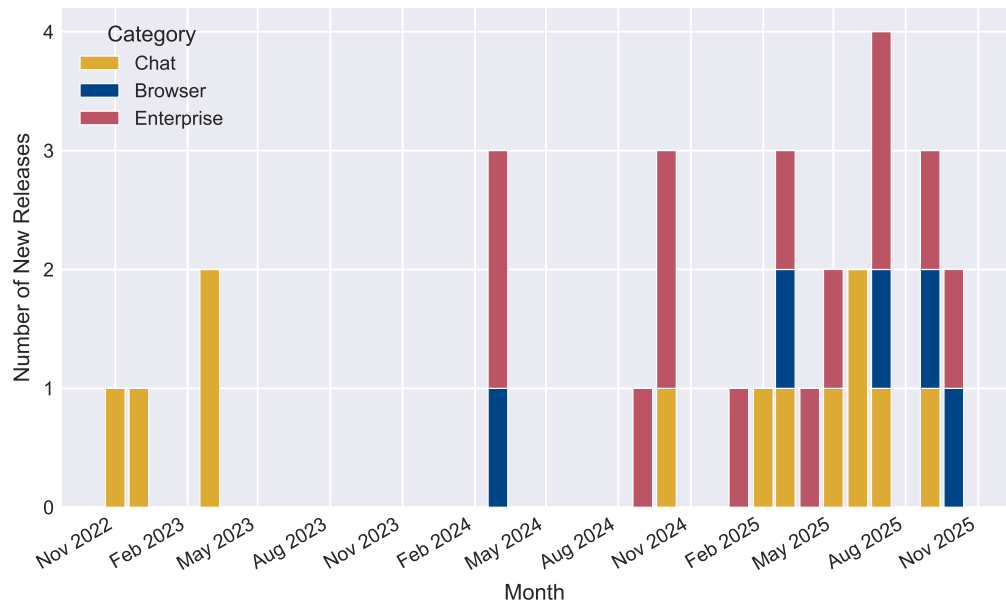
Fig. 10. Number of new AI agentic product releases by month.
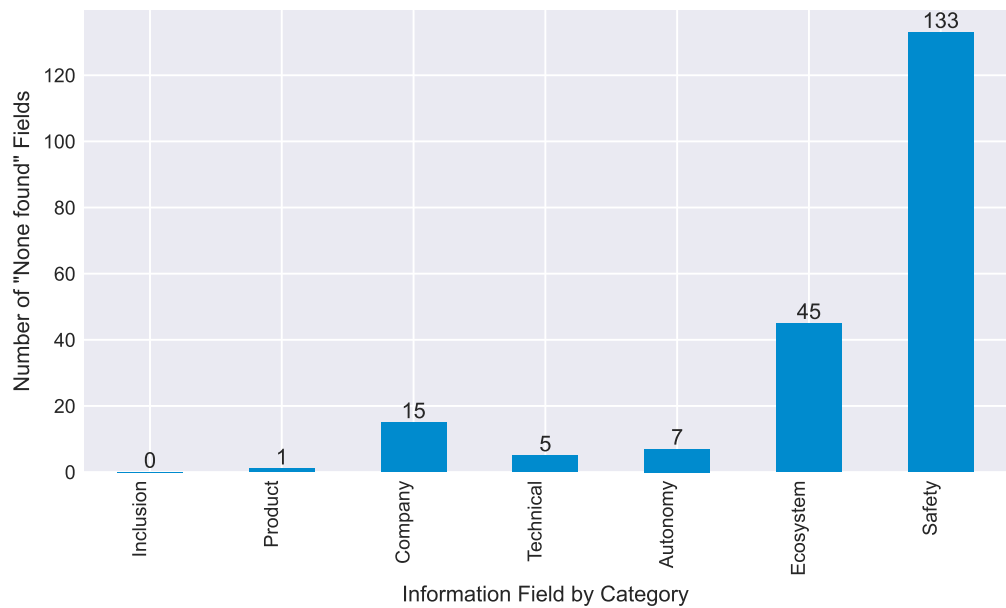
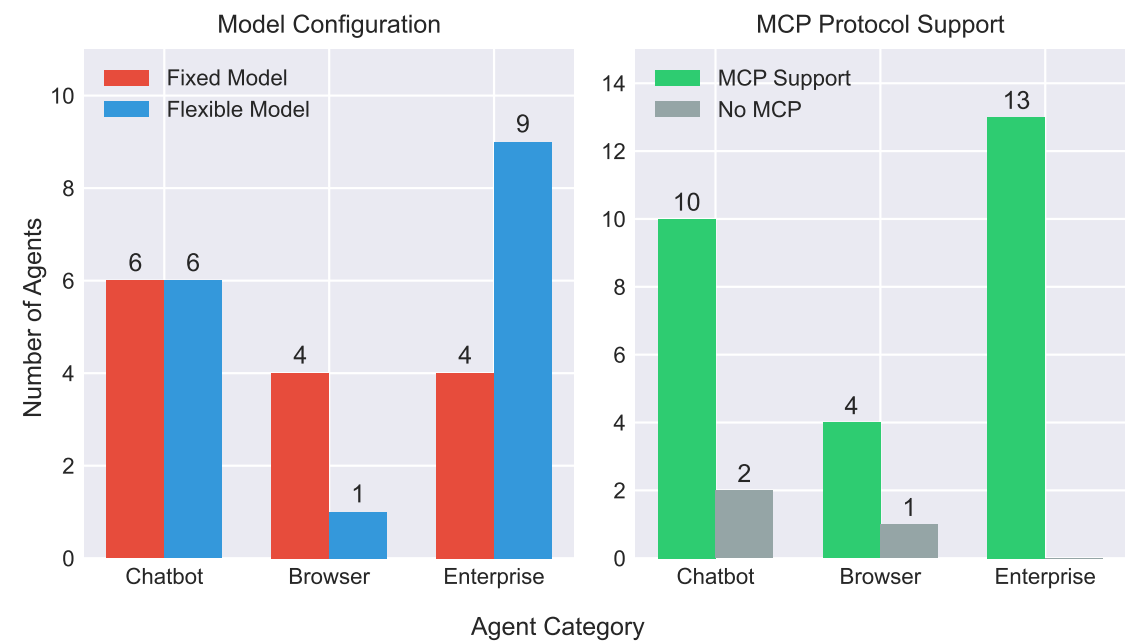Fig. 11. Number of information fields with "None found" by field category.

Fig. 12. Support for choosing models and MCP varies by category of the agent. Enterprise agents are more likely to support model selection (9/13) and MCP (13/13).

A.2 Sample entry of the Index: Claude Code

This section provides a complete annotation example for Claude Code, demonstrating how each field is documented. Here, we provide a sample entry of the Index for Anthropic's Claude Code. We selected it based on its high degree of documentation (no "None" or "None found" annotations). No authors have conflicts of interest related to Anthropic or Claude Code, and this example selection was made without correspondence with Anthropic. Including Claude Code as an example here is not an endorsement of the system or developer.

Inclusion Criteria.

    Search volume: 349,518 peak monthly hits.

    Valuation: USD 183 billion.

    GitHub stars in 2025: 50,700.

    Important developer: Yes. Included in Foundation Model Transparency Index. Member of Frontier Model Forum. Signatory of Frontier AI Safety Commitments.

Product Overview.

    Name of Agent: Claude Code.

    Short description of agent: "Work with Claude directly in your terminal. Claude explores your codebase context, answers questions, and make changes".

    Date of release: 24/02/2025 (initial release), 22/05/2025 (general release).

    Advertised use: Coding agent.

Monetization/Usage price: $20/month (Pro), $100/month (Team with higher rate limits), $200/month (Enterprise).

Who is using it: End user and enterprise customers for coding and prototyping.

Website: https://claude.com/product/claude-code.

Category: Chat.

Company & Accountability.

Developer: Anthropic.

Name of legal entity: Anthropic, PBC.

Place of legal incorporation: Delaware, USA.

For profit company: Yes (PBC).

Parent company: Not applicable.

Governance documents analysis: Claude Code page, Customer TOS, Usage Policy.

AI safety/trust framework: Responsible Scaling Policy.

Compliance with existing standards: HIPAA, SOC 2 Type I and II, ISO 27001:2022, ISO/IEC 42001:2023, CSA Star for enterprise plan.

Technical Capabilities & System Architecture.

Model specifications: Any Claude model. Default depends on subscription tier. User can choose model.

Documentation: https://code.claude.com/docs.

Observation space: File system, bash commands, MCP.

Action space: File system, bash commands, MCP.

Memory architecture: Hierarchical markdown memory.

User interface and interaction design: Chatbot in terminal.

User roles: Operator (issues queries, which the agent responds to); Executor (user may take actions/make decisions based on outputs); Examiner (user can use thumbs up/down buttons to give feedback).

Component accessibility: Closed source.

Autonomy & Control.

Autonomy level and planning depth: L1-L4: in plan mode it is most like a simple chat bot, but with auto-approve mode on, Claude Code can plan actions and take multiple steps (using different tools) without user approval. It will ask for clarification as needed.

User approval requirements: Yes, permission for running bash commands, editing files, or reading files outside of its initial directory source.

Execution monitoring, traces, and transparency: Visible (albeit summarized) chain-of-thought with a list of to-dos being worked on.

Emergency stop and shutdown mechanisms: User can pause/stop the agent at any time.

Usage monitoring and statistics: User can see how much context is used.

Ecosystem Interaction.

Identify to humans: Anthropic's stance on watermarking: "While watermarking is most commonly applied to image outputs, which we do not currently provide, we continue to work across industry and academia to explore and stay abreast of technological developments in this area." Anthropic's Usage Policy prohibits using Claude to impersonate

a human (i.e., to convince someone they're communicating with a natural person when they are not), implying Claude deployments must not hide AI identity in human interactions link.

Identifies technically: Anthropic officially documents that Claude-related web activity is identifiable via specific User-Agent tokens: ClaudeBot, Claude-User, and Claude-SearchBot link. Anthropic states it does not currently publish fixed IP ranges for these bots/agents (they use service-provider public IPs), so IP-range identification is not available as an official signature mechanism link.

Interoperability standards and integrations: Anthropic defines Model Context Protocol (MCP) as an open standard that "standardizes how applications provide context to LLMs," likened to a "USB-C port for AI applications" [92]. Also works with open source plugins and skills.

Web conduct: ClaudeBot, Claude-User, and Claude-SearchBot "respect 'do not crawl' signals by honoring industry standard directives in robots.txt" and "respect anti-circumvention technologies," stating they do not attempt to bypass CAPTCHAs link. Independent reporting and site-operator accounts, however, have documented periods of very heavy crawling and, at least in some cases, behavior that appeared to ignore site preferences until new robots.txt rules propagated link link.

Safety, Evaluation & Impact.

Technical guardrails and safety measures: Not available.

Sandboxing and containment approaches: Filesystem, network, OS-level enforcement. Sandboxed bash tool: Sandbox bash commands with filesystem and network isolation, reducing permission prompts while maintaining security. Enable with /sandbox to define boundaries where Claude Code can work autonomously. Write access restriction: Claude Code can only write to the folder where it was started and its subfolders source.

What types of risks were evaluated: Opus 4.5 system card contains section on agentic misuse.

(Internal) safety evaluations and results: Opus 4.5 system card contains section on agentic misuse.

Third-party testing, audits, and red-teaming: Opus 4.5 system card, section 5.2.1 Gray Swan Agent Red Teaming benchmark for tool use.

Benchmark performance and demonstrated capabilities: Refer to Opus 4.5 System Card.

Bug bounty programs and vulnerability disclosure: Yes.

Any known incidents: AI-orchestrated cyber espionage campaign.

B Annotation Methodology

B.1 List of agentic AI products considered

Below is the complete list of agent products considered, with those included in the final 2025 AI Agent Index in bold. Some enterprise agent productions also include ready-made agents (e.g. Microsoft 365 Copilot). Our analysis does not focus on these. Instead, we focus on agentic AI products that allow for general customization (e.g. Microsoft Copilot Studio) and the agents that could be created through these.

- 01.AI WorldWise Enterprise LLM Platform
- Aider
- AI21 Maestro
- **Alibaba MobileAgent**
- All Hands OpenHands
- Aleph Alpha Pharia
- Amazon Bedrock Agents
- Amazon Nova Act
- Amazon Q Business
- Anysphere Cursor

- **Anthropic Claude**
- **Anthropic Claude Code**
- Automation Anywhere AI Agents
- Baidu Agent Platforms
- Beam AI Beam
- **Browser Use**
- **ByteDance Agent TARS**
- ByteDance Coze
- ByteDance TRAE
- Cline
- Cloudflare Agents
- Cognition Devin
- Cognition Windsurf
- Cohere North
- Continue (Continue.dev)
- Counsel AI Corporation Harvey Agents
- CrewAI Crew
- Databricks Agent Bricks
- Determinist Ltd AutoGPT
- Dust
- Flowise
- Genspark Super Agent
- GitHub Copilot Agent
- **Glean Agents**
- **Google Agentspace**
- **Google Gemini**
- **Google Gemini Code Assist**
- Google DeepMind Project Astra
- Google Jules
- Google Project Mariner
- Gumloop AI Workflows
- **HubSpot Breeze Agents**
- HuggingFace Computer Agent
- **IBM watsonx Orchestrate**
- Kiln AI Kiln Agents
- Kiro
- Lindy
- Lovable
- **Manus AI Manus**
- **Microsoft Copilot Studio**
- Microsoft Magentic UI
- MindStudio AI Agents
- **MiniMax**
- Model scope MS Agent
- **Moonshot AI Kimi OK Computer**
- Motion AI Employee
- Moveworks Agent studio
- **n8n**
- NAVER Cue Search
- Notion Agent
- **OpenAI AgentKit**
- OpenAI Agent SDK
- **OpenAI ChatGPT**
- **OpenAI ChatGPT Agent**
- **OpenAI ChatGPT Atlas**
- **OpenAI Codex**
- OpenManus
- **Opera Neon**
- Oracle AI Agent Platform
- Palantir AIP
- Pega Blueprint
- **Perplexity**
- **Perplexity Comet**
- Relevance AI agents
- Replit Agent
- Sakana AI Scientist
- **Salesforce Agentforce**
- **SAP Joule studio**
- **ServiceNow AI Agents**
- Sierra Agent
- Skyvern
- Sourcegraph Cody
- Spell
- StackAI
- StackBlitz Bolt
- SWE-agent
- Tencent AppAgent
- Tencent Youtu-Agent
- The San Francisco AI Factory Inc Factory
- UiPath Autopilot
- Workday AI Agents
- **WRITER Action Agent**

- **Z.ai AutoGLM**
- **Zapier AI Agents**

- **LangChain LangSmith Agent Builder**

B.2 Annotation Fields

1. Inclusion Criteria

- **Search volume:** Monthly search estimates for top 5 keywords using Ahrefs to measure public interest over 2025.
- **GitHub stars in 2025:** Stars for GitHub repositories about the agent product itself (not related "cookbooks" or similar), where applicable.
- **Market capitalization/valuation:** Developer market cap as a December 2025 average (public companies) or valuation as of December 2025 (private companies). Data from Yahoo Finance, Crunchbase, Epoch AI, and general new sources.
- **Important developer:** Membership in 2024 Foundation Model Transparency Index [19], Frontier Model Forum [46], or signatory of Frontier AI Safety Commitments [2] or Artificial Intelligence Safety Commitments [29].

2. Product Overview

- **Name of Agent**
- **Short description:** 2-3 sentence description copied directly from developer (main marketing headline).
- **Date of release:** First release and latest update (month-level granularity).
- **Advertised use:** Developer-stated capabilities and intended use cases. Category of use (finance, web development) or specific examples (CRM to prioritize leads, summarize sales from different sources).
- **Monetization/Usage price:** Cost per month per user/seat in USD. Subscription tiers. Access method if not directly monetized (e.g., part of existing Microsoft/Google subscription). Additional costs (API model calls, storage).
- **Who is using it:** Customer types (end users, enterprises by size/industry, through API, governments).
- **Website:** Product landing page.
- **Category**: "Chatbot with tools," "Browser-based," or "Enterprise". See Section 3.2 for details.

3. Company & Accountability

- **Developer:** Developer name.
- **Name of legal entity:** Legal entity name, location of headquarters, legal domicile, data residency (including state if US).
- **Place of legal incorporation:** Headquarters location, legal domicile.
- **For profit company:** Corporate structure (for-profit, public benefit corporation, other structures).
- **Parent company:** Parent company ownership if applicable.
- **Governance documents analysis:** Terms of Service, privacy policy, acceptable use policy/usage restrictions, AI-specific policy.
- **AI safety/trust framework:** Responsible Scaling Policy (RSP), Frontier AI Safety Framework, company-wide safety documentation.
- **Compliance with existing standards:** Claimed compliance with ISO/IEC, NIST AI RMF, EU AI Act categories, SOC, relevant laws (e.g., GDPR).

4. Technical Capabilities & System Architecture

- **Model specifications:** Whether single model or user-selectable models. Available models by developer. Foundation model name, checkpoint if available. Whether reasoning model.
- **Documentation:** Links to technical documentation.
- **Observation space:** Input information sources. Internet access. Model Context Protocol (MCP) support.
- **Action space:** Sandboxing status. Whether an agent can directly affect the real world without human approval or be configured to. Available tools with write access.
- **Memory architecture:** Types employed (short-term, long-term, consensus, episodic) [55].
- **User interface and interaction design:** Interface type (chat, browser integration, other application). Whether anthropomorphism encouraged. Warnings against misperception. Human–agent teaming practices (overreliance mitigation, error communication).
- **User roles:** Designer, Operator, Executor, Examiner capabilities (designing agent, running and providing inputs, interacting with output and making decisions, evaluating through feedback), as per Tomsett et al. [121].
- **Component accessibility:** Open-source status and license. Availability of weights, data, code, scaffolding.

5. Autonomy & Control

- **Autonomy level and planning depth:** L1-L5 classification as per Feng et al. [43]: L1 (user as operator, agent provides on-demand support), L2 (user as collaborator, agent works independently on own tasks), L3 (user as consultant, agent takes initiative over extended time horizons), L4 (user as approver, interaction only when agent encounters blockers), L5 (user as observer, no means for user involvement).
- **User approval requirements:** Which actions require explicit approval? Whether approval is given once or for all following actions. Whether approval can be revoked.
- **Execution monitoring, traces, and transparency:** How users can see agent actions. Whether in real time, as a recording afterwards, or in another format.
- **Emergency stop and shutdown mechanisms:** Stop/abort controls. Relevant for agents running continuously or based on triggers. Whether builder platforms allow creating shutdown mechanisms to delegate control back to the user.
- **Usage monitoring and statistics:** Activity tracking, usage patterns.

6. Ecosystem Interaction

- **Identify to humans:** Whether agent identifies as AI when interacting with non-user humans. Provenance tracking for outputs (e.g., watermarking).
- **Identifies technically:** Digital signatures the agent uses. Published IP ranges, user agent strings, cryptographic signatures for web requests, request signing, unique identifiers.
- **Interoperability standards and integrations:** Supported standards and frameworks for agent communication. Examples include AGNTCY, Agent Connect Protocol (ACP), Model Context Protocol (MCP), Agent2Agent (A2A) protocol. Whether agent has own API. Whether agents use MCP to access tools or have MCP servers for others to use their services.
- **Web conduct:** Whether agent complies with robots.txt when directly accessing/scraping web. Crawling behavior. Anti-bot evasion techniques.

7. Safety, Evaluation & Impact

- **Technical guardrails and safety measures:** Notable methods used to protect against harmful actions. Built-in guardrails. For agent builders, types of guardrails that can be added and conditions under which they are active.
- **Sandboxing and containment approaches:** Whether agent runs in virtual machine (VM), locally, or other environment. Characteristics of VM and how it interacts with environment. For hosted products, specific details on containment practices beyond general infrastructure.
- **What types of risks were evaluated:** Scope of risk assessment.
- **(Internal) safety evaluations and results:** Testing scope and procedures. Whether evaluations are agent-specific or model-only. Results of evaluations.
- **Third-party testing, audits, and red-teaming:** External testing scope and organizations involved. Specific to agent (does not include general compliance audits).
- **Benchmark performance and demonstrated capabilities:** Benchmarks run and results (developer-reported only).
- **Bug bounty programs and vulnerability disclosure:** Links to programs if applicable. Disclosure policies.
- **Any known incidents:** Safety incidents in 2025 (does not include bug bounty submissions).

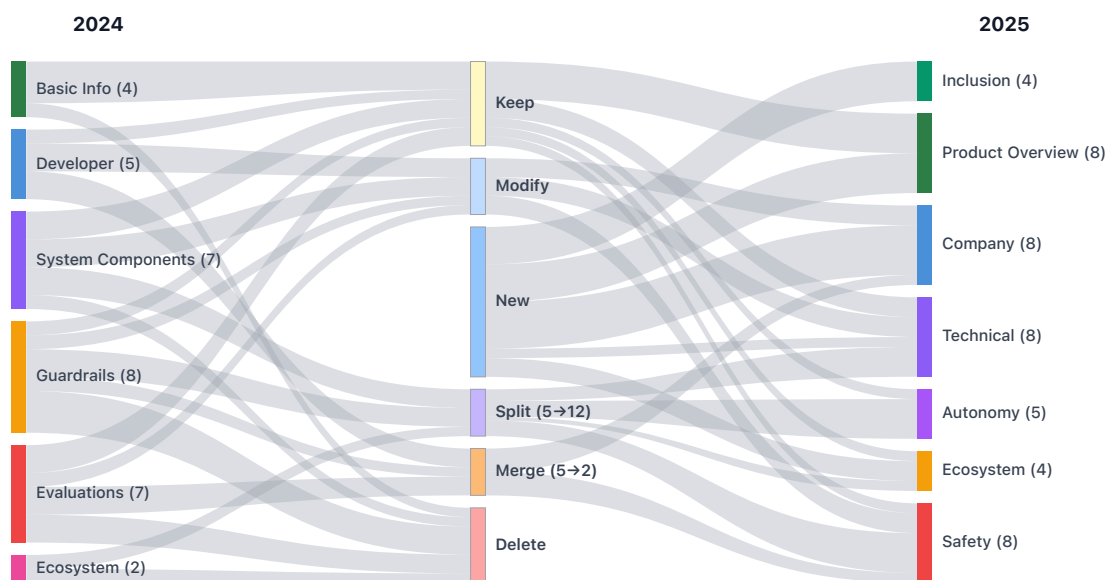## B.3 Changes in annotation fields

Fig. 13. **Changes in annotation fields across categories since the 2024 Index.** We make significant changes across all sections compared to the 2024 Index [22]: 16 fields in the 2025 version are completely new, and only 9 fields are kept unaltered (*Keep*). The remaining 16 fields were derived from fields in the 2024 version either by making significant modifications to the definition/notes (*Modify*), splitting a field (*Split*), or merging multiple fields into one (*Merge*). 8 fields were deleted.

B.4 Annotation Guide

Annotators received detailed instructions emphasizing the following principles:

Depth and detail. With approximately 30 agents, conduct thorough analysis through demo videos, secondary sources, and testing. Focus on object-level findings rather than interpretations. Document defaults chosen in UI or architecture. Include tangential observations and reasoning for educated guesses.

Agent-specific focus. Annotate safety and transparency features of agents themselves, not underlying models. Evaluations of foundation models alone (e.g., GPT-4o) do not count as agent evaluations unless properly contextualized.

Platforms vs. agents. For platforms creating agents rather than standalone agents, annotate based on most capable agents that can be created. Note default options and platform capabilities/limitations.

Source documentation. Cite sources for every field with multiple links when applicable. Link directly to highlighted text. Include timestamps for video citations.

Handling uncertainty. Use standardized markers: "Not available" when confident information doesn't exist, "Not applicable" when field doesn't apply to the agent, "UNSURE" when information might exist somewhere, "TODO" for later revisits. Liberally use comments for second opinions.

Information sources. Official documentation, company blog posts, help center documents, trust center materials (including penetration test reports), and developer conference demo videos. Annotators shared helpful sources for collective benefit.

Scope of information. Annotations based on publicly available information, including what is visible in the agent interface itself. No experiments conducted (e.g., testing how agents identify themselves to websites). For open-source agents, documentation and readme files used rather than code analysis.

B.5   LLM prompts for finding agents

We used LLM-based research queries to surface an initial list of candidate agents. Below we provide the prompts used for each given model.

ChatGPT 5.2 with deep research:

- `what are the most significant coding agents that you can use to do general purpose things. e.g. the coding agents must support MCP and be capable of taking non coding actions as well.`
- `I currently have this list of agents. are any important agents missing? output just a list of these agents with names and developer and link.`
- `what are the most significant AI agents currently available. they should be actually agentic, not just AI chatbots.`
- `summarise the recent developments in AI agents, what major new advanced Ai agent have come to the market in 2024 and 2025, what impact have they had. Also consider what academic literature on AI agents has come out recently.`

Claude Sonnet 4.5 with research mode:

- `what are the most significant AI agents currently available. they should be actually agentic, not just AI chatbots.`, followed by `I am looking for all available real world examples that can be used right away, open source or commercial, across all domains (although domain specific is fine too). with actually agentic I mean all of these capabilities.`

- summarise the recent developments in AI agents, what major new advanced Ai agent have come to the market in 2024 and 2025, what impact have they had. Also consider what academic literature on AI agents has come out recently., followed by give me a list of significant AI agent products (both open source and industry)available right now. focus on those used by many people or from notable companies or that are talked about a lot. include everything. But mention if they are not publically released yet. if they are specialised agents they need to be more significant, eg. even more users. mention the level of autonomy.

Gemini 2.5 with research mode:

- what are the most significant AI agents currently available. they should be actually agentic, not just AI chatbots.
- summarise the recent developments in AI agents, what major new advanced Ai agent have come to the market in 2024 and 2025, what impact have they had. Also consider what academic literature on AI agents has come out recently.

B.6 LLM usage to verify annotations

To enhance the reliability of our annotations, we developed an automated verification pipeline using OpenAI's GPT-5.2 model with web search capabilities. We provide all prompts used in the process below. The verification process follows a three-step approach:

(1) **Web research phase:** For each agent-field pair, the system prompts the model to search for primary sources (official documentation, press releases, developer blogs, and trust centers) related to the specific agent product. The prompt is constrained to match that of annotation guide used by human annotators, see Section B.4.

(2) **Structured verification phase:** The research findings are then processed by a second model call that outputs a structured JSON response containing: (a) a verified annotation with inline source citations, (b) a confidence rating (High/Medium/Low), (c) confidence justification, and (d) any discrepancies with the original annotation.

(3) **Manual verification phase:** An annotator compared the original annotations to the LLM-generated ones and, if necessary, amended the final annotation. Each result of the LLM was double-checked and all sources were manually verified. We used a custom annotation viewer to easily compare annotations side-by-side, as shown in Figure 14.

Listing 1. Web Research Prompt: Sent to the model with web search enabled to gather primary sources about the agent.

```
Research the specific AI agent product "{agent_name}" to verify information about:
    {column_name}

IMPORTANT: Focus ONLY on this specific agent product, NOT:
- The underlying foundation models (e.g., GPT-4, Claude) - unless asking about model
    specifications
- The parent company's general information or other products
- Generic AI industry information
- Related but different products from the same company

Field guidance: {field_description}
```
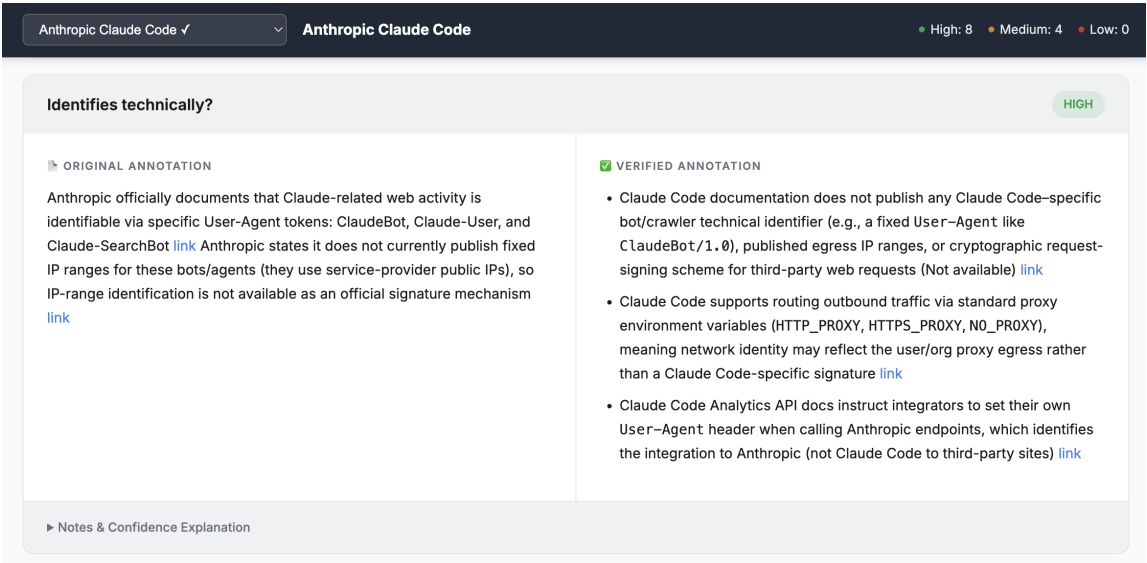
Fig. 14. Custom annotation viewer to compare the human annotations to the LLM-generated ones in the verification phase.

```
Current annotation to verify:
{current_annotation}


Known resources for this agent: {resources}


Search for PRIMARY SOURCES specifically about "{agent_name}":
- Official product documentation and help pages for this agent
- Product-specific press releases and announcements
- Developer blogs about this specific agent
- Trust center pages specific to this product
- Product landing pages and feature descriptions


Provide a concise research summary with URLs to sources found. Only include information
    directly about "{agent_name}".
```

Listing 2. Structured Verification Prompt: Formats research findings into structured JSON output.

```
Based on the research below, provide a verification of the annotation for the specific
    agent product "{agent_name}" - field "{column_name}".

## Field Guidance
{field_description}

## Current Annotation
{current_annotation}
```

```
## Research Findings
{research_context}


## Critical Instructions
- Focus ONLY on this specific agent product "{agent_name}"
- Do NOT include information about underlying foundation models unless specifically
    relevant to the field
- Do NOT include general company information unless specifically relevant
- Provide object-level findings with inline citations
- Use "Not available" if no info exists, "Not applicable" if field doesn't apply,
    "UNSURE" if uncertain


Provide your verification result.
```

Listing 3. System Message: Accompanies the structured verification prompt to enforce output format.

```
You are a research assistant verifying annotations for the specific AI agent product
    "{agent_name}".
Focus ONLY on this agent product - not underlying models, parent company general info,
    or related products.


OUTPUT FORMAT: Provide a concise bullet-point list of findings, NOT paragraphs.
Each bullet should be a single factual finding with an inline source link.


Example format:
- Feature X is supported [link](https://example.com/docs)
- Pricing starts at $20/month [link](https://example.com/pricing)
- No information found on Y


Return a JSON object with these exact keys:
- verified_annotation: string (concise bullet-point list of findings, each with inline
    markdown link)
- confidence: string ("High", "Medium", or "Low")
- confidence_explanation: string (why you rated confidence this way)
- notes: string (additional notes for research team)
```

B.7   Correction request to companies

To ensure accuracy, we contacted each company with agents in the Index on December 12, 2025, providing a view-only link to our draft annotations and inviting corrections by January 11, 2026. We committed to maintaining confidentiality of all correspondence while noting that the final database would be released publicly. Below is the email template used:

We are a team of academic researchers from MIT, Harvard, Stanford, and other universities, updating the AI Agent Index published by Casper et al. (2025), which documented 67 AI agents and has become a widely-cited resource in the field.

The 2025 edition focuses on a deeper analysis of publicly available information on approximately 30 particularly significant agentic systems, including [AGENT NAME(S)]. To ensure the accuracy and completeness of our work, we kindly request your feedback on the information we collected regarding [AGENT NAME(S)] (view-only link to the draft sheet). Note that "None found" means we are unable to find any public information on this field. You can share your feedback by replying to this email at [redacted] by January 11th, 2026. All correspondence in this email thread will remain strictly confidential and will be used solely to improve the accuracy of our database. The finalized database (including any changes we make based on your feedback) will be released publicly along with the research paper. Please refrain from sharing the database prior to its publication.

Your expertise is key to ensuring the rigor and reliability of our work, and we deeply value your input. If you have any questions, feel free to reach out.

C   Public Interest Methodology

For search volume metrics, we consider combinations of product and company name as well as simpler terms such as just product name (if unambiguous) or company + `AI agent`, etc.

For GitHub, we consider the highest number of stars in 2025. If an agent product has a GitHub repository, this does not imply that it is open-source. The open-source annotation field was validated separately.

C.1   LLM prompts for generating search terms

We used OpenAI GPT-5.2 with web search to create a list of possible search terms for each `company` and `product` combination in two steps.

First, we used OpenAI GPT-5.2 with web search to research the product. Below is the prompt used for this.

```
Research {company} {product} and provide a concise summary covering:
1. What it is (e.g., AI coding assistant, browser, automation platform)
2. Primary use case and key features
3. What makes it unique or distinguishing characteristics

Search the web for the most current and accurate information.
Keep the response focused and factual, under 150 words.
```

Second, we used OpenAI GPT-5.2 to generate a list of relevant search terms based on the product background from step 1. Below is the prompt used for this.

```
Given this information about an AI agent product:
- Company: {company}
- Product Name: {product}
- Context: {context}
{existing_terms_str}
```

```
Generate NEW unambiguous search terms that meet these STRICT requirements:

IMPORTANT: Only generate NEW terms that are NOT in the existing list above. Do not
    duplicate any existing terms.

1. MUST INCLUDE BASIC COMBINATIONS (if not already existing):
   - "{company.lower()} {product.lower()}" (if not in existing)
   - If company and product names are DIFFERENT, also include "{product.lower()}
       {company.lower()}"
   (reversed order, if not in existing)
   - If company and product names are the SAME, skip the reversed term (no duplicates)

2. STANDALONE PRODUCT NAME (when applicable):
   - IF the product name is sufficiently unambiguous on its own, include it
   - GOOD: "chatgpt agent" (specific, won't confuse with general ChatGPT)
   - GOOD: "windsurf" (unique enough to be unambiguous)
   - BAD: "comet" (too generic, could mean space comet)
   - BAD: "atlas" (too generic, could mean geography atlas)
   - BAD: "north" (too generic, could mean direction)
   - Only include if you're confident it won't be confused with other meanings

3. CONCISE: Use simple keyword combinations, NOT descriptive phrases
   GOOD: "perplexity comet", "comet browser", "perplexity comet browser"
   BAD: "AI-powered Comet chatbot for accurate answers"

4. SPECIFIC: Terms must clearly identify THIS SPECIFIC PRODUCT, not the company generally
   GOOD: "chatgpt agent mode", "openai chatgpt agent"
   BAD: "ai agent" alone, "openai" alone (too general, refers to company or generic
       product)

5. UNAMBIGUOUS: Cannot be confused with other meanings
   GOOD: "perplexity comet browser" (not space comet)
   BAD: "comet" alone (could mean astronomical comet)

6. INCLUDE COMPANY + PRODUCT CATEGORY: Include combinations of company name + product
    type/category
   GOOD: "perplexity browser" (for Perplexity Comet which is a browser)
   GOOD: "openai coding agent" (for ChatGPT Agent which is a coding agent)
   GOOD: "google ai agent" (for Project Mariner which is an AI agent)
   - These are important because users often search for "company + what it is"

7. NATURAL: Phrases a user would actually search on Google
   - Combine company name + product name
   - Combine company name + product category (e.g., "perplexity browser")
   - Avoid duplicating words when combining company name + product name
```

```
     (e.g., "openmanus" is better than "openmanus openmanus")
     - Combine product name + key identifier (e.g., "browser", "agent", "assistant")
     - Vary word order for natural search patterns

It is ESSENTIAL that the generated search terms are unambiguous and specific to the
     product.
It needs to just be about the product itself, not about related subquestions.
ONLY include search terms that are directly to the product.

Return ONLY a JSON array of strings with NEW terms only (not in existing list), no
     explanations.
If all good terms already exist, return an empty array [].
Example: ["new term 1", "new term 2", ...]
```

C.2 Search volume analysis
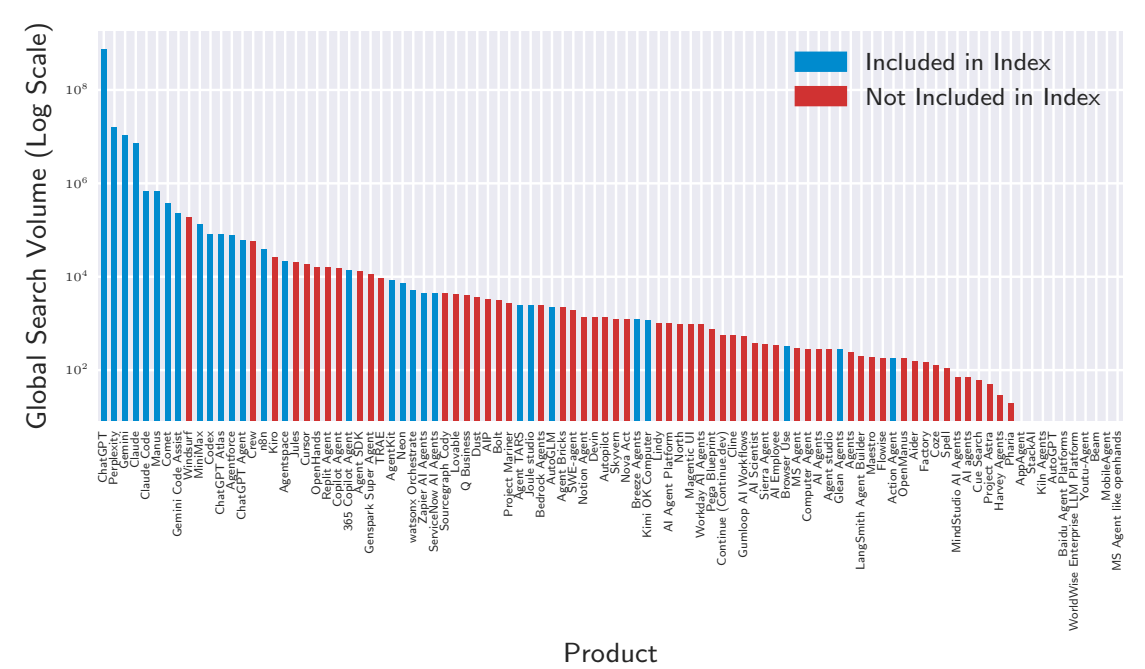
Fig. 15. Global search volume (log scale) for each AI agent product. Blue indicates inclusion in the Index.
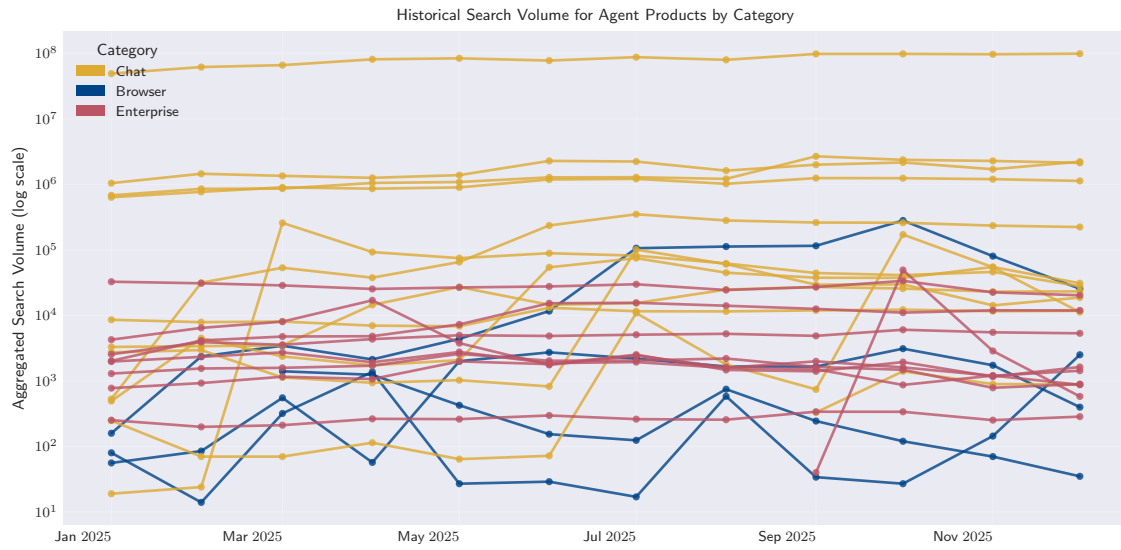
Fig. 16. Monthly search volume (log scale) for the most popular term for each AI agent product included in the Index over 2025, colored by agent category (chat, enterprise, and browser).
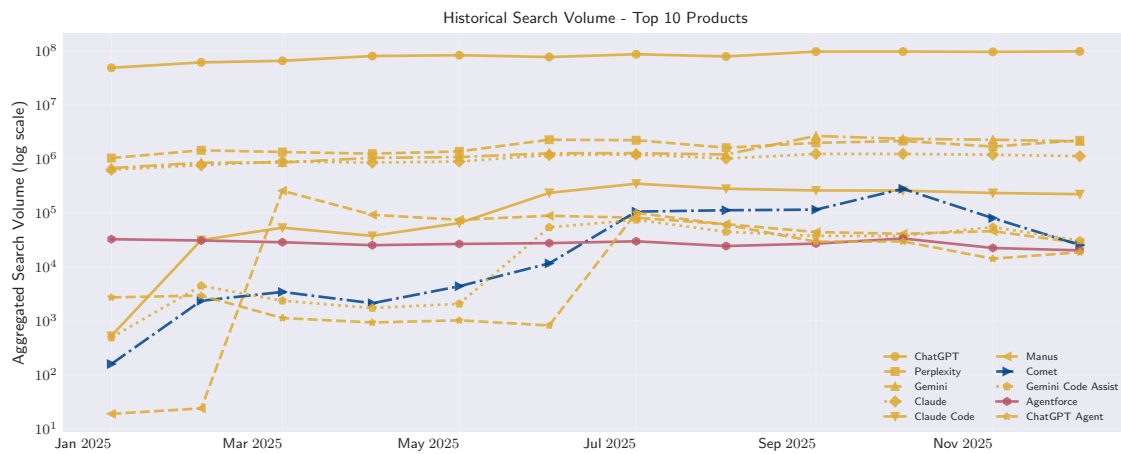
Fig. 17. Monthly search volume (log scale) for the 10 most popular AI agent products, colored by agent category (chat, enterprise, and browser)
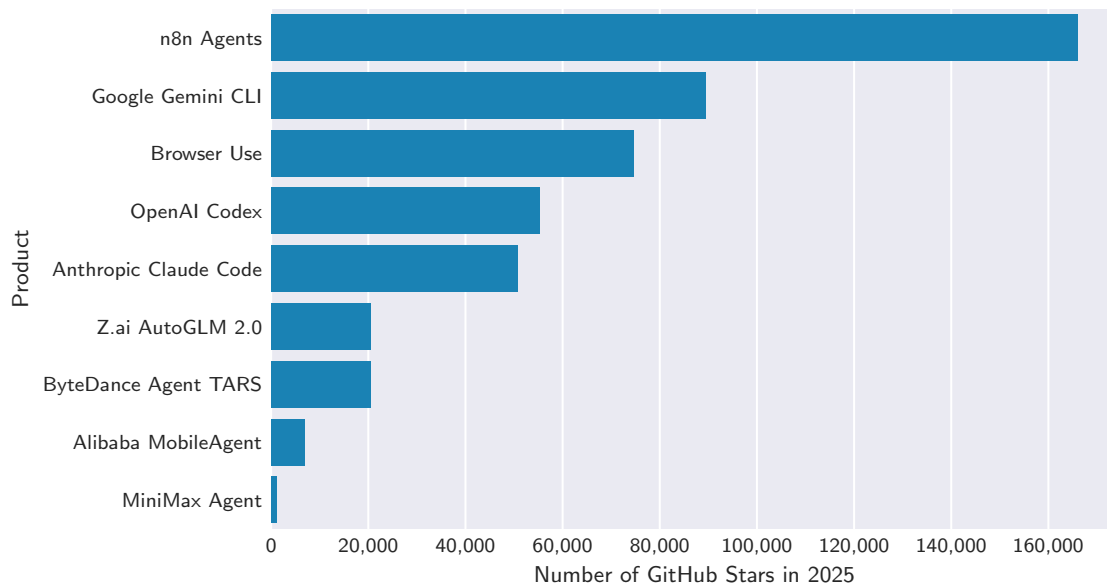
Fig. 18. GitHub stars for repositories associated with the agent products.

D   Supplementary materials for case studies

Fig. 19. Details on the safety features and red teaming methodology for AI agents are limited. Screenshot of the "Model Card" for HubSpot's Breeze Agents taken from https://trust.hubspot.com/ai.