

Public report from the pre-study:

Multiagent systems for improved decision making in industrial value chains

This report has been created by Theander, Sahlgren, Bridges, Munoz and Hoseini at AI Sweden. Sept 2025

About the project

This project was a pre-study with funding from Advanced Digitalisation on System changing initiatives in applied industrial AI, conducted May - October 2025, aiming to establish the foundation for a Swedish ecosystem around industrial generative multi-agent AI systems, supporting the planning of joint efforts and the creation of a strategic roadmap.

Project partners; AI Sweden (Lindholmen Science Park), Astra Zeneca AB, Volvo Lastvagnar, Saab AB, Ericsson AB.

This report summarises the identified industrial needs and challenges, as well as the state of the art for relevant research in those identified areas.

Benchmark of the Swedish industry status, challenges and needs

Stakeholders from Swedish industry, academia and private sector were invited to give input to a survey on their needs and challenges, as well as suggestion for collaborative actions in the area of multiagent systems. The survey was followed up by in-depth interviews and workshops. Around 40 Swedish organisations have contributed.

Needs and benefits identified are:

- Trustworthy Systems, with control functions.
- Efficiency and Automation, including automating administrative tasks, improving workflows, and enabling faster prototyping.
- Information Management, including providing a quick way to find information, summarizing documents, and analysing large datasets.
- Decision Support, including aiding developers and case officers in making better-informed decisions.

- Reusability is very important for frameworks, components, etc.

Identified risks and challenges:

- Reliability and Quality, including concerns about the quality of output, data accuracy, and the risk of hallucinations (fabricating answers).
- Trust and Transparency, acknowledging that there is a lack of insight into how agents are trained, where information originates, and also a general erosion of trust in decisions made by AI.
- Legal and Ethical Issues, including data privacy, IP intrusion, and the potential impact on workforce competence and human relations.
- Understanding the Cost vs Benefits.

In order to meet the needs and challenges, the following areas are suggested for further collaboration:

- Areas for R&D: The most important areas for future research and development are identified as quality and validation, management and orchestration, explainability and trustworthiness, and the ability to integrate AI agents with existing industrial domain knowledge and systems and create sharable resources.
- Needed Support: Companies are looking for shared use cases, best practices, and open, portable solutions that can run in both on-premises and cloud environments.
- Collaboration: Several organizations expressed interest in contributing to a future project, with potential contributions including providing domain knowledge, sharing implementation experience, and building competence cross industries.

Comparison with global surveys

Cap Gemini published a survey during the summer 2025, where 1500 organisations were interviewed, of which 3% were Swedish. [1] The main outcome of that survey, resonates well with the result of this pre-study:

- AI agents drive cost reduction and improve operational outcomes, with 24/7 availability.
- 93% of leaders believe that those who successfully scale AI agents in the next 12 months will gain an edge over industry peers.
- Trust is a major hurdle. Make agent decision-making traceable and auditable.

- Businesses need confidence in AI systems before granting them any level of autonomy.
- Establishing goal alignment between humans and AI agents is a key success factor.

Trust is a major concern that needs to be fulfilled, in order for industry to fully automate the work of agents. This is an important issue that is high on agendas both globally and in our project consortium. Some of these issues will be handled by the market, but it is important that Swedish industry is on top of these questions to keep industrial leadership and to be aware of risks and challenges to take the right measures.

The area of AI multiagent systems evolves rapidly and it is therefore important to be able to work in sprints rather than several years project planned in detail. The main topics will be safety, quality and validation, management and orchestration, continuously developing agents, as well as creation of sharable resources and protocols.

State of the art multi-agent system research, sept 2025

The multiagent system (MAS) area evolves constantly, and it is important to continue to keep track of research evolution. The information below gives a glimpse of the interesting research for identified relevant focus areas, LLMs and multiagent system, architectural design, optimization, and security, safety, security & governance, robust & hybrid MAS by sept 2025.

In 2024–2025, two fast trends shape multi-agent research. Large language models are being used as planners and routers that coordinate specialist agents, while multi-agent reinforcement learning combined with human-in-the-loop training focuses on learning robust, aligned coordination policies [2].

Researchers and engineers are shipping orchestration frameworks and developer kits that let multiple agents communicate, call external tools, and hand off subtasks. A common pattern in deployed stacks is *planner*→*executor*: an LLM or planner generates a high-level plan and specialist executors (API adapters, deterministic services, or learned policies) perform each step; lightweight verifiers are often inserted to block hallucinated or unsafe actions. [3]

1. LLMs and multi-agent systems

One of the more interesting recent papers on the topic of LLMs and agent systems is the paper “Small Language Models are the Future of Agentic AI” [4] that argues for the benefit of using smaller specialized language models in agent workflows. The main argument is that while LLMs are unquestionably more powerful than smaller models, the latter are often sufficient and much more economical for tasks that are “repetitive, scoped, and non-conversational”. In such cases, it is preferable to use models that are “efficient, predictable, and inexpensive.” We will likely see further developments in the direction of using small, more specialized, models in agent workflows.

Another interesting research direction is self-evolving agents, which autonomously and dynamically adjust their profiles, goals and planning strategies to better solve complex tasks. This development is thoroughly covered in the paper “A Comprehensive Survey of Self-Evolving AI Agents: A New Paradigm Bridging Foundation Models and Lifelong Agentic Systems” [5]. Self-evolving agents normally involve updating the LLM parameters, which may risk corrupting the LLM by compromising the knowledge it has acquired during pre-training. An interesting solution to this problem is proposed in the paper “Memento: Fine-tuning LLM Agents without Fine-tuning LLMs” [6], where the agent workflow is augmented by a structured memory system based on case-based reasoning, which retrieves and adapts solutions based on its experience without the need for fine-tuning the LLMs.

In addition to developments in fine-tuning, post-training and specialization of smaller models and self-evolving agent systems, we will likely see further development in long context models that can handle complex tasks and long documents without having to partition the information across several different agents, as well as further developments in reinforcement learning strategies and environments for reasoning models, see e.g. “A Survey of Reinforcement Learning for Large Reasoning Models” by Zhang et al (2025) [7].

Countries to keep an eye on includes the usual suspects, USA and China. A telling example of the current pace of development in the USA is the following statement from one of the state leaders on AI at a seminar at Princeton University this summer on the challenge of regulation of agent systems: “we wish we had more time to think about it, but we don’t”.

2. Architectural Design, Optimization, and Security

Use cases for MAS are rapidly proliferating sophisticated and integrated coding agents (Google, 2025), and cyber threat-intelligence-to-automated-detection pipelines [8]. Major technology companies (e.g., OpenAI, Google, Anthropic) are now actively pursuing MAS services to automate complex tasks such as commerce, agentic coding, office software automation, scheduling, travel booking, social

media video generation (e.g., TikTok style output), and business website development [9,10] - this gives areas of research to avoid.

While MASs composed of smaller, specialized LLMs have been shown to outperform single large LLMs in both quality and efficiency when optimally configured [11], MAS design remains a high-dimensional, combinatorial optimization problem. Current automated topology design efforts often lack guarantees of optimality and result in black-box systems [12]. Most agentic system designs are currently *ad hoc* and unprincipled. A key advancement is that once the graphical structure of an MAS is established, prompt optimization for components across the entire system is achievable using tools like TextGrad [13].

Existing test and evaluation environments for MAS struggle to keep pace with the rapid advancement of foundation models and often lack portability across different applications. Recent advancements offer robust sandboxes that support both unit tests (for component-level accuracy, latency, and security) and full system tests (for identifying systemic problems and evaluating overall costs) [14]. Key issues remain, including the generation of non-public benchmarks to prevent agents from cheating [15]. Some problem-specific studies have generated benchmarks that illuminate MAS issues [16], and LLM judges have rapidly become a standard method of evaluation [17].

Addressing security issues in agentic systems remains a critical need. State-of-the-art contributions currently offer design principles that provide security guarantees specifically against prompt injections [18]. Research is still required to extend these built-in security guidelines to address other forms of vulnerability and provide comprehensive security guarantees.

3. Safety, security & governance

Agentic systems widen the attack surface and increase governance requirements compared with single-model deployments [19]. Practical safety work therefore combines three strands:

- **Engineering controls:** least-privilege connectors, tamper-evident provenance/logging, pre-execution verifiers (small local models or deterministic checkers), and runtime anomaly detection. [3,4.]
- **Adversarial evaluation:** Testing should cover a defined set of threat vectors that are particularly relevant for agents: prompt-injection and tool call leakage, protocol-spoofing and “shadow-agent” abuse, connector compromise (database/API), model extraction/memorization, and training

time poisoning. Each vector is evaluated by both automated scripts and focused human red teaming [2,20, 21]. Adversarial tests are integrated into the development lifecycle: lightweight checks run in pre-merge continuous integration, full adversarial suites run nightly or on feature branches, and a canary/production stage provide continuous live probes and monitors for problems. Test outputs feed measurable metrics (attack success rate, leakage exposure, meantime to detection, false-positive rate, and cost/latency impact) that drive go/no-go decisions and risk prioritization [3,20,22]. Minimum recommended red-team suite for pilot deployments cover

- (1) prompt-injection and tool-call leakage tests,
- (2) protocol-spoofing / agent-to-agent spoof tests,
- (3) connector compromise and least-privilege violation tests,
- (4) membership inference and model-extraction attacks,
- (5) poisoning / data integrity scenarios, and
- (6) end-to-end incident detection / rollback drills.

Run these regularly and record the measured metrics to demonstrate improvement over time [4,19].

- **Governance artifacts:** Human-oversight policies say who must approve agent actions and how approvals are recorded. Change-control requires a shortchange request, tests, and a staged roll out with clear rollback rules before any model or connector goes to production. Post-market monitoring collects a few key metrics (e.g., hallucination, leakage, meantime to detection/remediation (MTTD/MTTR), human overrides), sets alert thresholds and names an on-call responder.

These controls should now be mandatory for industry pilots because they determine whether a system can be deployed safely and legally (impacting cost, privacy and regulatory risk). Hybrid stacks are using small/local LMs (SLMs) on-prem for verification and routine tasks, plus selective cloud calls to large foundation models for heavy planning. Hybrid stacks are a practical way to balance capability with latency, cost and privacy [2,4].

Actors to keep on eye on: Industry labs such as DeepMind, Microsoft Research, Anthropic and NVIDIA drive agent architectures and tooling; academic hubs at UC Berkeley, Stanford, Oxford and ETH lead MARL and safety research; and open communities (AutoGen, Hugging Face) accelerate adoption and reproducibility.

4. Robust & Hybrid Multi-Agent Systems

LLM-based agents are instantiated from powerful, pre-trained foundation models, and are endowed with sophisticated capabilities for natural language understanding, generation, and in-context reasoning, allowing them to tackle complex, knowledge-intensive tasks that are beyond the scope of traditional MARL (Multi-Agent Reinforcement Learning). This shift has introduced a novel set of research challenges (the “hot topics” in current research) that are less about policy optimization and more about system design, orchestration, and the management of information flow. [23].

Context and Memory Management [23] methods address the problem of managing the global task objective, agent-specific instructions, shared knowledge, and the history of interactions in an effective and scalable way (e.g., with hierarchical memory for different levels of abstraction, or a shared "consensus memory" to maintain consistency).

Robust Reasoning [24] methods look at addressing, e.g., the risk of emergent negative phenomena like groupthink, where agents quickly converge on a plausible but incorrect answer, or the "snowballing" of hallucinations, where an initial error from one agent is accepted and amplified by the rest of the group. Ensuring that the collective reasoning process is sound, diverse, and self-correcting remains key.

Looking forward, methods can look to adding explainability & causal grounding to this reasoning process by, e.g., translating reasoning into a sequence of verifiable operations on a knowledge graph or some other world model [25]. This exemplifies a larger push in current research to explore system design of hybrid models, i.e., examining how and where LLM-based agents can interact with traditional MARL specialist agents, assuming more of an “orchestrator” role and providing these with e.g. a reward function informed from the state of a KG. How to dynamically partition and allocate tasks in this sense is a current research question.

Furthermore, looking forward, adaptation to dynamic environments via continuous learning approaches remains an unsolved challenge - how can we ensure optimal, continued collaboration among agents in an environment that is constantly changing (e.g., enterprise intranets) [26]. Is self-organization, both on the individual level but also group level possible in real time, i.e., knowing when to forget or consolidate past knowledge? What are the design parameters of a self-evolving MAS system that generates and autonomously improves a joint world model (e.g. knowledge graph) dynamically as new data/information is learned by agents independently?

Actors to keep an eye on: Industrial use cases in this field will continue to be dominated by players such as DeepMind, Google, Microsoft, and Meta. The opportunity for differentiation will come through research that is focused on increasing efficiency of existing systems and explicitly examine the formal foundations of agent interaction. Some key groups / institutions in this space include the Bernoulli Institute at the University of Groningen, ICL, University of Illinois Urbana-Champaign (ISRL), or the MAS Lab at the University of Tsukuba.

5. Architecture and Human-in-the-Loop

Despite the diverse and high-performing designs available, practitioners often face confusion when selecting the most effective pipeline for their specific task: Which topology is the best choice for my task, avoiding unnecessary communication token overhead while ensuring high quality solution?

G-Designer, an adaptive, efficient, and robust solution for multi-agent deployment has been introduced, which dynamically designs task-aware, customized communication topologies. Specifically, G-Designer models the MAS as a multi-agent network, leveraging a variational graph auto-encoder to encode both the nodes (agents) and a task-specific virtual node, and decodes a task-adaptive and high-performing. [12]

Designing MAS that can effectively interact with humans is gaining importance. This includes developing intuitive communication interfaces and ensuring that agent behaviour is understandable and predictable to human users. [27]

LLM MAS today lack native social behaviour, they should be pre-trained to learn cooperation and competition through multi-agent scenarios and interactive feedback, enabling them to develop socially adaptive behaviours. The authors propose developing MAS LLM frameworks that are natively asynchronous and grounded in standardised, open-source communication protocols, ensuring security, identity, and trust, drawing from established practices in multi-agent distributed systems and communication. [28]

How to engage

We are planning for a 3+3-year project within Advance Digitalisation System Changing Initiatives, building competence aiming to make sure that Swedish industrial companies have prerequisites to implement scalable, secure and sustainable AI multiagent system solutions that clearly strengthen their competitiveness.

Does your organisation want to engage in this collaborative effort coming years? Reach out to Helena Theander: Helena.theander@ai.se and describe how you want to contribute.

References:

1. Weblink: [Capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf](https://capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf)
2. Shuaihang Chen, et al. A survey on LLM-based multi-agent system: Recent advances and new frontiers in application, 2025
3. Qingyun Wu, et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations 2024.,
4. Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, Pavlo Molchanov: arXiv:2506.02153
5. Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, Zaiqiao Meng: arXiv:2508.07407
6. Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, Jun Wang: arXiv:2508.16153
7. "A Survey of Reinforcement Learning for Large Reasoning Models" by Zhang et al (2025): arXiv:2509.08827)
8. Lanka, P., Gupta, K., & Varol, C. (2024). Intelligent Threat Detection—AI-Driven Analysis of Honeypot Data to Counter Cyber Threats. *Electronics*, 13, 2465. <https://doi.org/10.3390/electronics13132465>.
9. OpenAI. (2025). *Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol*. Retrieved from https://openai.com/index/buy-it-in-chatgpt/?utm_source=substack&utm_medium=email.
10. Lambert, J. (2025, September). *ChatGPT: The Agentic App*. Interconnects.
11. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). *Mixture-of-agents enhances large language model capabilities*. arXiv preprint arXiv:2406.04692.
12. Zhang, G., Yue, Y., Sun, X., Wan, G., Yu, M., Fang, J., Wang, K., Chen, T., & Cheng, D. (2025). *G-designer: Architecting multi-agent communication topologies via graph neural networks*. arXiv preprint arXiv:2410.11782.

13. Yuksekgonul, M., et al. (2025). Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055), 609–616.
14. Andrews, P., Pierre, et al. (2025). *ARE: scaling up agent environments and evaluations*. arXiv preprint arXiv:2509.17158.
15. Zhou, K., Zhu, Y., Chen, W., Chen, W. X., Chen, X., Lin, Y., Wen, J. R., & Han, J. (2023). *Don't Make Your LLM an Evaluation Benchmark Cheater*. arXiv preprint arXiv:2311.01964.
16. Li, B., et al. (2024). Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37, 17044–17068.
17. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024). *LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. arXiv preprint arXiv:2412.05579.
18. Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2025). *Defeating prompt injections by design*. arXiv preprint arXiv:2503.18813.
19. Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.
20. OWASP GenAI Security Project. LLM Top 10 for LLMs and Generative AI (v1.1), 2024. Release v1.1, April 11, 2024. OWASP GenAI Security Project (OWASP Foundation). Accessed: 2025-09-29
21. Ziyang Wang, Zhicheng Zhang, Fei Fang, and Yali Du. M3hf: Multi-agent reinforcement learning from multi-phase human feedback of mixed quality, 2025.
22. Langchain: A framework for building LLM applications. <https://python.langchain.com/>, 2024. Accessed: 2025-09-XX
23. LLM Multi-Agent Systems: Challenges and Open Problems. <https://arxiv.org/abs/2402.03578>
24. Improving Factuality and Reasoning in Language Models Through Multiagent Debate. <https://arxiv.org/abs/2305.14325>
25. A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. <https://arxiv.org/abs/2408.05141>
26. Strategy Coopetition Explains the Emergence and Transience of In-Context Learning. <https://arxiv.org/abs/2503.05631>
27. Fully Autonomous AI Agents Should Not be Developed Margaret Mitchell Avijit Ghosh Alexandra Sasha Luccioni Giada Pistilli arXiv:2502.02649v2 [cs.AI] 6 Feb 2025
28. Large Language Models Miss the Multi-agent Mark. Emanuele La Malfa et.al., <https://arxiv.org/abs/2505.21298>