The 2026 Research Landscape of Agentic AI

Planning, Tool Use, Memory, Multi-Agent Systems & Safety

ScholarLens Sample Research Pack

A concise literature landscape showing how ScholarLens can organize a fast-moving research area into themes, evidence, gaps, and research questions.

Prepared as a sample research-discovery artifact • September 2026

1. Research Question

How are LLM-based systems evolving from conversational models into autonomous agents, and what technical challenges must be solved before reliable deployment?

Agentic AI systems combine language-model reasoning with actions such as tool use, retrieval, planning, memory, and interaction with dynamic environments. Recent surveys organize the field around reasoning, acting, interacting, and evaluation. ■cite■turn0academia3■turn0academia1■

2. Research Landscape

Area	Core question
Planning & reasoning	How should agents decompose goals, plan actions, reflect, and recover from failure?
Tool use	How can agents reliably select and operate APIs, browsers, code tools, and databases?
Memory	What information should persist across tasks, and how can memory remain useful and secure?
Multi-agent systems	When does collaboration outperform a single capable agent?
Evaluation	How should capability, robustness, safety, cost, and reliability be measured?
Agentic programming	How can coding agents plan, execute, debug, and adapt across long tasks?
Autonomous research	How can AI systems conduct research while producing claims humans can verify?

This structure is supported by recent surveys of agentic LLMs, agent evaluation, and agentic programming. ■cite■turn0academia3■turn0academia1■turn0academia2■

3. Selected Research

Representative starter papers across different dimensions of the field:

Paper	Focus	Why it matters
Agentic Large Language Models, a survey (2025)	Foundations	Reasoning, acting, interacting, tools and multi-agent systems. ■cite■turn0academia3■
Survey on Evaluation of LLM-based Agents (2025)	Evaluation	Benchmarks for planning, tool use, memory and realistic agent tasks; identifies evaluation gaps. ■cite■turn0academia1■
AI Agentic Programming: A Survey (2025)	Coding agents	Planning, memory, tool integration, execution monitoring and coding-agent evaluation. ■cite■turn0academia2■
Autonomous Research Agents: A Survey (2026)	AI scientists	Finds a verification gap between runnable systems and reproducibility/claim-verification artifacts. ■cite■turn0academia0■
Teams of LLM Agents can Exploit Zero-Day (2024)	Safety	Shows the capabilities and security implications of hierarchical, tool-using agent teams. ■cite■turn0search7■

4. What the Literature Suggests

Planning and reasoning. Agents transform goals into sequences of decisions and actions; long-horizon tasks make recovery and verification increasingly important.

Tool use. A defining shift from passive language models to agents is the ability to act through external tools and incorporate observations into later decisions.

Memory. Persistent memory can support long-running tasks but raises questions about what should be stored, retrieved, updated, and protected.

Multi-agent collaboration. Specialized agents can divide work, but coordination adds communication and reliability costs.

Evaluation. Agent evaluation increasingly needs realistic tasks and measures beyond final-answer accuracy, including cost, safety and robustness. ■cite■turn0academia1■

Autonomous research. A 2026 survey reports that code release is much more common than reproducibility and novelty-verification artifacts, highlighting an auditability gap. ■cite■turn0academia0■

5. Research Gaps & Opportunities

Gap 1 — Long-horizon reliability: improve planning, verification, rollback, and recovery for multi-step tasks.

Gap 2 — Agent evaluation: create scalable benchmarks for reliability, safety, robustness, cost, and adaptation. ■cite■turn0academia1■

Gap 3 — Claim verification: create reproducibility and novelty-verification mechanisms for autonomous research systems. ■cite■turn0academia0■

Gap 4 — Secure tool use: constrain and audit agent actions when systems can interact with external tools. ■cite■turn0search7■

Gap 5 — Human oversight: determine which actions can be autonomous and which require approval.

6. Research Questions

1. How can LLM agents reliably recover from failures during long-horizon tasks?

2. What evaluation metrics best measure real-world agent reliability?

3. When does multi-agent collaboration outperform a single capable agent?

4. How can persistent agent memory be made secure and controllable?

5. How should humans allocate oversight across autonomous agents?

6. How can autonomous research agents verify their own scientific claims?

7. How can tool-using agents defend against malicious or unintended tool calls?

7. ScholarLens Research Map

> **AGENTIC AI** → **Planning** → **Tool Use** → **Memory** → **Autonomous Action** → **Multi-Agent Collaboration** → **Evaluation** → **Safety & Human Oversight**

This is a conceptual map of connected research themes, not a claim that one architecture is universally accepted.

8. How to Use This Pack

Use this as a starting point for a literature review. Inspect the original papers, verify claims, trace citations, and refine the research question for your specific project.

> **ScholarLens principle:** research discovery should help users trace ideas back to underlying literature rather than replace the literature itself.

Sources

Plaat et al., *Agentic Large Language Models, a survey*, arXiv:2503.23037. ■cite■turn0academia3■

Yehudai et al., *Survey on Evaluation of LLM-based Agents*, arXiv:2503.16416. ■cite■turn0academia1■
Wang et al., *AI Agentic Programming: A Survey*, arXiv:2508.11126. ■cite■turn0academia2■
Ding et al., *Autonomous Research Agents: A Survey of AI Scientists and the Verification Gap*, arXiv:2608.05179. ■cite■turn0academia0■
Wang et al., *Teams of LLM Agents can Exploit Zero-Day*, arXiv:2406.01637. ■cite■turn0search7■