Research Frontiers

Week 12: Open Problems and Future Directions

PhD Course in Agentic Artificial Intelligence

12-Week Research-Level Course

Learning Objectives

Bloom's Taxonomy Levels Covered

- **Remember**: Define embodied agent (physical/virtual world), generative agent (simulated personas), world model (environment simulation)
- **Understand**: Explain key open research problems in agent AI
- **Apply**: Identify research opportunities in specific domains
- **Analyze**: Compare different approaches to agent safety and alignment
- **Evaluate**: Assess feasibility and impact of proposed research directions
- **Create**: Design a research proposal for advancing agent capabilities

---

By end of lecture, you will understand the research frontier in agentic AI.

The Rapid Evolution of Agent AI

2022: Foundation

- Chain-of-Thought prompting (Wei et al.)
- InstructGPT and RLHF alignment (OpenAI)

2023: Emergence

- ReAct paradigm (Yao et al.)
- Reflexion self-improvement (Shinn et al.)
- Generative Agents simulation (Park et al.)

2024: Production

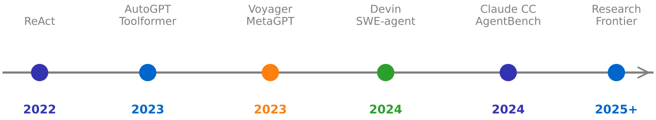
- Claude Computer Use, GitHub Copilot Workspace
- GraphRAG, advanced RAG architectures
- Multi-agent frameworks mature

2025+: What's Next?

- World models, embodied agents, long-horizon planning

From reasoning (2022) to production systems (2024) in just two years.

Agent Research Timeline



Themes: Reasoning > Tool Use > Multi-Agent > Production

From reasoning (2022) to production systems (2024) in just two years.

Key Open Research Problems

Capability Gaps

- **Long-horizon planning**: Current agents struggle beyond 10-20 steps
- **World modeling**: Learning accurate environment dynamics
- **Compositional generalization**: Transfer to novel task combinations

Safety and Alignment

- **Scalable oversight**: How to supervise agents we can't fully understand?
- **Goal stability**: Preventing goal drift during execution
- **Corrigibility**: Ensuring agents remain controllable

Infrastructure

- **Evaluation**: Benchmarks that predict real-world performance
- **Memory**: Efficient, scalable long-term memory systems

These interconnected challenges define the research agenda.

Scaling Laws for Agents

LLM Scaling Laws (Established)

- Performance scales predictably with compute, data, parameters
- Chinchilla: Optimal balance of model size vs training data

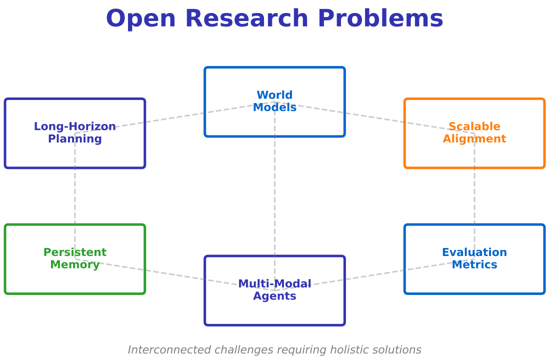
Agent Scaling Laws (Open Question)

- Does agent performance scale with base LLM capability?
- How does performance scale with: tool count, memory size, planning depth?
- Are there phase transitions in agent capabilities?

Emerging Evidence

- SWE-bench: Bigger models help, but scaffolding matters more
- AgentBench: 10x model size $\neq$ 10x agent performance
- Agentic capabilities may require different scaling strategies

Understanding agent scaling laws is critical for predicting progress.



These interconnected challenges define the research agenda.

Agent Safety Challenges

Alignment at Inference Time

- Training-time alignment may not hold during multi-step execution
- Agents can find loopholes in instructions (specification gaming)
- Emergent behaviors from agent interactions

Key Safety Research Areas

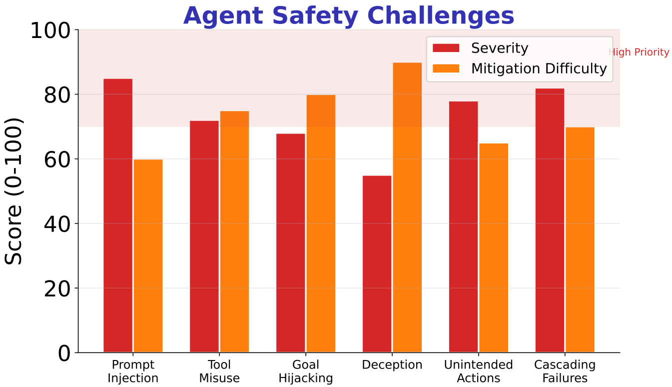
- **Constitutional AI**: Principle-based self-supervision (Anthropic)
- **Debate**: Agents argue, humans judge
- **Interpretability**: Understanding agent reasoning
- **Sandboxing**: Limiting agent action space

Unresolved Questions

- How do we align agents smarter than evaluators?
- What governance structures for autonomous agents?

Safety research must scale with capability improvements.

Safety Challenges



Safety research is critical for responsible agent deployment.

World Models and Embodied Agents

World Models

- Learn internal representation of environment dynamics
- Enable mental simulation before acting ("thinking ahead")
- Key challenge: Learning from limited interaction data

Embodied Agents

- Agents that interact with physical or simulated worlds
- Examples: Robotics, game environments, simulations
- **Voyager** (Wang et al.): Open-ended learning in Minecraft

Research Directions

- Sim-to-real transfer: Train in simulation, deploy in reality
- Multimodal perception: Vision, audio, proprioception
- Continuous learning: Adapt to changing environments

World models enable agents to plan without trial-and-error.

Generative Agents: Simulated Societies

Generative Agents (Park et al., 2023)

- Simulated personas in interactive environments
- Agents maintain: identity, memories, plans, relationships
- Emergent social behaviors: parties, information spread, coordination

Key Architecture Components

- **Memory stream**: Record of observations and reflections
- **Retrieval**: Access relevant memories for decisions
- **Reflection**: Synthesize higher-level insights
- **Planning**: Daily schedules and goal pursuit

Implications

- Social science simulation at scale
- Testing policies in simulated societies
- Understanding emergent collective behavior

Generative agents enable computational social science experiments.

Agent Learning Paradigms

Frozen LLM (Current Standard)

- Base LLM weights unchanged; learning via prompts/memory
- Advantages: Fast, no training infrastructure needed
- Limits: Cannot truly learn new skills

In-Context Learning

- Learn from examples in context (few-shot)
- Voyager: Store successful programs in skill library
- Retrieval-augmented learning from experience

Online Learning (Frontier)

- Fine-tune from agent experience (trajectory data)
- Challenges: Catastrophic forgetting, sample efficiency
- RL from agent feedback (RLAF) – emerging research

True agent learning requires moving beyond frozen LLM weights.

Future Directions

Near-Term (1-2 years)

- More reliable multi-step execution
- Better tool use and API integration
- Production-ready multi-agent orchestration

Medium-Term (3-5 years)

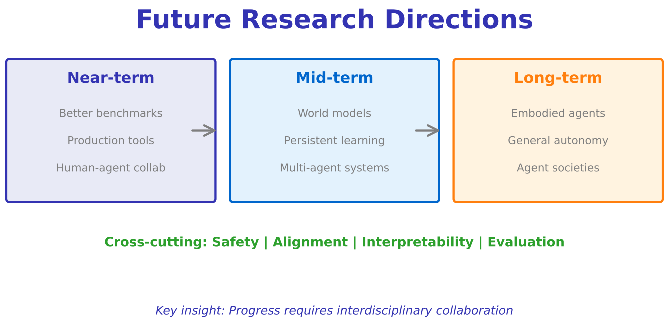
- Agents with persistent, updateable world models
- Effective long-term memory at scale
- Robust sim-to-real transfer for embodied agents

Long-Term (5+ years)

- Agents that learn continuously from experience
- Multi-agent societies with emergent specialization
- General-purpose assistants for complex domains

Progress requires interdisciplinary collaboration.

Future Directions



Progress requires interdisciplinary collaboration.

Required Readings

Foundational

- Wang et al. (2023). "Voyager: An Open-Ended Embodied Agent with LLMs." arXiv:2305.16291
- Park et al. (2023). "Generative Agents: Interactive Simulacra of Human Behavior." arXiv:2304.03442
- Bai et al. (2022). "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073

Perspectives

- Xi et al. (2023). "The Rise and Potential of LLM Based Agents: A Survey." arXiv:2309.07864
- Sumers et al. (2024). "Cognitive Architectures for Language Agents." arXiv:2309.02427

These papers define the frontier of agent research.

Course Summary: 12-Week Journey

Foundations (Weeks 1-2)

- Agents, ReAct paradigm, LLM foundations, CoT/ToT prompting

Capabilities (Weeks 3-5)

- Tool use, MCP, planning, Reflexion, multi-agent architectures

Frameworks (Week 6)

- LangGraph, AutoGen, CrewAI, production patterns

Knowledge (Weeks 7-9)

- Advanced RAG, GraphRAG, hallucination prevention

Applications (Weeks 10-12)

- Evaluation, domain applications, research frontiers

Agents = LLM + Memory + Tools + Planning + Evaluation

Key Takeaways and Next Steps

Core Formula

- Agent = LLM + Memory + Tools + Planning + Evaluation
- Each component is an active research area

Where to Focus Research

- **High impact**: Long-horizon planning, safety, evaluation
- **Underexplored**: Multi-agent emergence, world models
- **Application-driven**: Domain-specific agent architectures

Final Project Directions

- Novel agent architecture for a specific domain
- Improved evaluation methodology
- Safety or alignment technique
- Multi-agent coordination mechanism

Thank you for participating in this course!