

Toward Trustworthy Agent Societies in Complex Environments

Tianyu Shi

McGill University
tianyu.shi3@mcgill.ca

LLM agents are moving from chatbots toward systems that must remember, select specialized roles, construct workflows, and act in consequential environments. My talk surveys a systems agenda for *trustworthy agent societies*: building the infrastructure through which agents accumulate experience, coordinate heterogeneous capabilities, improve their workflows, and verify the consequences of their actions.

Infrastructure for persistent and efficient agents. SAGE combines feedback, reflection, and memory optimization among user, assistant, and checker agents (Liang et al. 2025). SEDM extends this idea to a scalable distributed memory substrate: replay-based admission determines what enters memory, while scheduling, consolidation, and knowledge diffusion determine how experience is reused across tasks (Xu et al. 2025). At the inference layer, MARS introduces margin-aware verification for speculative decoding, improving speed while preserving generation quality (Song et al. 2026). Together, these works ask how agent systems should allocate limited memory and inference resources without losing the evidence needed for reliable improvement.

Role-aware orchestration and coordination. CCMA connects cascading LLM-agent cooperation with cooperative decision-making in intelligent mobility (Zhang et al. 2025). DAAO adapts workflow depth, operators, and model assignment to query difficulty (Su et al. 2026b); Role-RL assigns distinct LLMs to empirically optimal roles in long-context processing (He, Shi et al. 2026); and ComfySearch constructs functional creative workflows through validation-guided exploration (Su et al. 2026a). These systems frame coordination as explicit resource allocation rather than merely adding more agents.

Closed-loop improvement and verifiable execution. DebFlow uses debate and reflection to iteratively create and improve agent workflows (Su et al. 2025). SECRM-2D, EVM-QuestBench, and PeopleSearchBench test whether agent outputs satisfy analytic constraints, execute intended state changes, and meet criteria-grounded requirements (Shi et al. 2025; Yang et al. 2026; Shi et al. 2026). Hide-and-Seek AML further illustrates how detector and transaction agents can co-evolve in a domain simulation, turning deployment gaps into data and evaluation problems (Wang et al. 2025).

Future work. I will build agent infrastructure for self-

evolving intelligence: composable experience and memory, role-aware orchestration, and closed-loop training and evaluation. I aim to develop trustworthy agent societies in which heterogeneous agents can specialize, coordinate, debate, compete, and collaborate under verifiable protocols and human governance. The objective is a practical path toward increasingly general-purpose, self-evolving, and accountable agentic intelligence.

References

- He, L.; Shi, T.; et al. 2026. Role-RL: Online Long-Context Processing with Role Reinforcement Learning for Multiple LLMs in Their Optimal Roles. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- Liang, X.; et al. 2025. Self-evolving Agents with Reflective and Memory-augmented Abilities. *Neurocomputing*.
- Shi, T.; et al. 2025. SECRM-2D: Safe, Efficient, and Comfortable Reinforcement-Learning-Based Autonomous Driving. *IET Intelligent Transport Systems*.
- Shi, T.; et al. 2026. PeopleSearchBench: Evaluating AI-Powered People Search Platforms with Criteria-Grounded Verification. In *EMNLP Industry Track*. Accepted.
- Song, J.; et al. 2026. MARS: Unleashing the Power of Speculative Decoding via Margin-Aware Verification. In *Findings of the Association for Computational Linguistics: ACL*, 8652–8663.
- Su, J.; et al. 2025. DebFlow: Automating Agent Creation via Agent Debate. In *CFAgentic Workshop at ICML*.
- Su, J.; et al. 2026a. ComfySearch: Autonomous Exploration and Reasoning for ComfyUI Workflows. *EMNLP findings*.
- Su, J.; et al. 2026b. Difficulty-Aware Agentic Orchestration for Query-Specific Multi-Agent Workflows. In *The Web Conference*.
- Wang, Q.; Tsai, W. T.; Shi, T.; Tang, W.; and Du, B. 2025. Hide and Seek in Transaction Networks: A Multi-Agent Framework for Simulating and Detecting Money Laundering Activities. *Complex and Intelligent Systems*, 11(6): 271.
- Xu, H.; Hu, J.; Song, X.; Zhang, K.; Yu, L.; Tang, Y.; Duan, Y.; Ai, L.; and Shi, B. 2025. SEDM: Scalable Self-Evolving Distributed Memory for Agents. In *NeurIPS 2025 Workshop on Scaling Environments for Agents (SEA)*. Accepted.
- Yang, P.; et al. 2026. EVM-QuestBench: An Execution-Grounded Benchmark for Natural-Language Transaction Code Generation. In *ACL*.
- Zhang, M.; et al. 2025. CCMA: Cascading Cooperative Multi-Agent Autonomous Driving with Large Language Models. *Expert Systems with Applications*.