References

Foundational Works

1. Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems*, 4(3), 382-401.

2. Dwork, C., Lynch, N., & Stockmeyer, L. (1988). Consensus in the Presence of Partial Synchrony. *Journal of the ACM*, 35(2), 288-323.

3. Jøsang, A., Ismail, R., & Boyd, C. (2007). A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43(2), 618-644.

Prompt Injection and LLM Security

1. Qi, X., et al. (2024). Visual Adversarial Examples Jailbreak Aligned Large Language Models. *AAAI 2024*, 38(19), 21527-21536.

2. Perez, F., & Ribeiro, I. (2023). Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs. *EMNLP 2023*.

3. Greshake, K., et al. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *ACM AISec 2023*, 79-90.

4. Liu, Y., et al. (2023). Prompt Injection Attack Against LLM-Integrated Applications. *arXiv:2306.05499*.

5. Zou, A., et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*.

6. Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *NeurIPS 2023*.

Constitutional AI and Alignment

1. Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.
2. Askell, A., et al. (2021). A General Language Assistant as a Laboratory for Alignment. *arXiv:2112.00861*.

Multiagent Systems

1. Wooldridge, M. (2009). *An Introduction to Multiagent Systems* (2nd ed.). John Wiley & Sons.
2. Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.
3. Hong, S., et al. (2023). MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. *arXiv:2308.00352*.
4. Wu, Q., et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv:2308.08155*.

Trust in Distributed Systems

1. Marsh, S. P. (1994). Formalising Trust as a Computational Concept. *PhD Thesis, University of Stirling*.
2. Gambetta, D. (1988). Can We Trust Trust? In *Trust: Making and Breaking Cooperative Relations*, 213-237.
3. Sabater, J., & Sierra, C. (2005). Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, 24(1), 33-60.

Adversarial ML

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ICLR 2015*.
2. Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. *IEEE S&P 2017*, 39-57.

Formal Verification

1. Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model Checking*. MIT Press.
2. Alur, R. (2015). *Principles of Cyber-Physical Systems*. MIT Press.

Cognitive Security

1. Waltzman, R. (2017). The Weaponization of Information: The Need for Cognitive Security. *RAND Corporation*.
2. Beskow, D. M., & Carley, K. M. (2019). Social Cybersecurity: An Emerging National Security Requirement. *Military Review*, 99(2), 117.

Agent Frameworks

1. LangChain. (2023). LangGraph: Build Stateful Multi-Actor Applications. *Documentation*.
2. CrewAI. (2024). Framework for Orchestrating Role-Playing, Autonomous AI Agents.
3. Anthropic. (2024). Claude Code: AI-Powered Software Engineering.

2025 Agentic AI Security

1. OWASP Foundation. (2025). OWASP Top 10 for LLM Applications 2025.
2. OWASP GenAI Security Project. (2025). OWASP Top 10 for Agentic Applications 2026.
3. Chen, W., Zhang, Y., & Liu, J. (2025). A Multi-Agent LLM Defense Pipeline Against Prompt Injection Attacks. *arXiv:2509.14285*.
4. Jo, Y., Kim, S., & Park, J. (2025). Byzantine-Robust Decentralized Coordination of LLM Agents. *arXiv:2507.14928*.
5. Wang, H., Li, X., & Chen, Y. (2025). Rethinking the Reliability of Multi-agent System: A Perspective from Byzantine Fault Tolerance. *arXiv:2511.10400*.
6. Debenedetti, E., Zhang, J., & Carlini, N. (2025). Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents. *NAACL 2025 Findings*.

Red Teaming and Benchmarks

1. Perez, E., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. *ACL 2023 Findings*.

2. Mazeika, M., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *ICML 2024*.

3. Chao, P., et al. (2024). JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv:2404.01318*.

4. Sun, L., et al. (2024). TrustLLM: Trustworthiness in Large Language Models. *arXiv:2401.05561*.

5. Liu, X., et al. (2023). AgentBench: Evaluating LLMs as Agents. *arXiv:2308.03688*.

6. Mialon, G., et al. (2023). GAIA: A Benchmark for General AI Assistants. *arXiv:2311.12983*.

Eusocial Intelligence and Swarm Systems

1. Wilson, E. O. (1971). *The Insect Societies*. Belknap Press of Harvard University Press.

2. Grassé, P.-P. (1959). La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. La théorie de la stigmergie. *Insectes Sociaux*, 6(1), 41-80.

3. Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press.

4. Lenoir, A., D'Ettorre, P., Errard, C., & Hefetz, A. (2001). Chemical Ecology and Social Parasitism in Ants. *Annual Review of Entomology*, 46, 573-599.

5. Seeley, T. D. (2010). *Honeybee Democracy*. Princeton University Press.

6. Kilner, R. M., & Langmore, N. E. (2011). Cuckoos Versus Hosts in Insects and Birds: Adaptations, Counter-adaptations