**Abstract**

The thesis investigates the development of evolvable medical agents through a hybrid simulation and planning framework. Building on the Agent Hospital simulacrum — a virtual environment where autonomous agents powered by large language models simulate full clinical workflows — we explore novel directions for memory-based evolution, GUI interaction, and complex medical reasoning. Our aim is to address three key limitations of existing agent systems: static knowledge representations, lack of dynamic GUI capabilities, and underdeveloped reasoning in interdisciplinary cases.

The research comprises three original investigations, each extending the capabilities of doctor agents beyond those introduced in Agent Hospital.

**Experiment 1: Memory-Augmented Agent Evolution.** We embed episodic and narrative memory structures into doctor agents, enabling them to learn from past successes and failures. Using scaling-law-style evaluation, we show that memory significantly improves diagnostic accuracy as the number of patient cases increases.

**Experiment 2: GUI-Based Workflow Simulation.** We integrate the Agent S framework to enable agents to perform interface-level tasks — such as registration, triage, and prescription — in a simulated hospital GUI. The agent’s performance is evaluated through task success rates, latency, and recovery under time constraints, reflecting real-world operational challenges.

**Experiment 3: Complex Reasoning via Prompt Architectures.** We benchmark agents’ reasoning capabilities on complicated clinical cases involving comorbidities and cross-department logic. Prompt engineering, tool-use routines, and biomedical reasoning modules are compared to determine their effectiveness in improving clinical decision quality.

The thesis presents the following contributions to agentic medical AI:

1. A unified framework combining simulation (Agent Hospital) and interface planning (Agent S) for end-to-end task learning.
2. A memory evolution mechanism for agents that yields measurable accuracy gains across diverse disease categories.
3. A methodology for benchmarking agent reasoning on complex medical cases, advancing toward testable and certifiable agentic AI.

We demonstrate that evolving LLM-based agents in synthetic, feedback-rich settings can produce scalable and safe clinical reasoning tools with potential for broader domains such as public health, finance, and legal assistance.