INTISARI

Institusi Pendidikan Tinggi menghadapi peningkatan permintaan dukungan kesejahteraan mahasiswa namun masih bergantung pada kerangka kerja konseling reaktif yang seringkali gagal menjangkau mahasiswa sebelum krisis memuncak. Skripsi ini mengusulkan dan mengevaluasi kerangka kerja AI agentic proaktif yang dirancang untuk menjembatani kesenjangan *insight-to-action* dengan memungkinkan intervensi dini dan manajemen sumber daya berbasis data. Kami memperkenalkan *Safety Agent Suite*, arsitektur multi-agen terpisah yang mendistribusikan tanggung jawab klinis dan operasional kepada agen khusus di bawah pengawasan manusia. Sistem ini mencakup: (i) **Aika**, orkestrator Meta-Agent yang menyediakan antarmuka pengguna terpadu dan melakukan penyaringan risiko Tingkat 1 segera; (ii) **Safety Triage Agent (STA)** untuk analisis risiko percakapan Tingkat 2 yang komprehensif; (iii) **Therapeutic Coach Agent (TCA)** yang memberikan intervensi mikro terapeutik berbasis Cognitive Behavioral Therapy (CBT); (iv) **Case Management Agent (CMA)** untuk koordinasi operasional; dan (v) **Insights Agent (IA)** untuk analitik manajemen sumber daya yang menjaga privasi. Untuk menyeimbangkan responsivitas dengan kedalaman analisis, kami menggunakan arsitektur pemantauan risiko dua tingkat yang menggabungkan penyaringan segera dengan analisis percakapan mendalam untuk memungkinkan intervensi dini. Sistem multi-agen dibangun dengan LangGraph dan mencakup perlindungan untuk penggunaan alat, redaksi, dan kemampuan audit.

Kami membangun prototipe fungsional dalam platform UGM-AICare dan melakukan evaluasi berbasis skenario yang menitikberatkan secara eksklusif pada kinerja arsitektur agen: sensitivitas dan *False Negative Rate* (FNR) triase pada skenario krisis sintetis; keandalan orkestrasi melalui tingkat keberhasilan pemanggilan fungsi dan transisi state; latensi ujung-ke-ujung; verifikasi kepatuhan privasi; serta kualitas coaching melalui rubrik kepatuhan CBT dengan penilaian ahli dan validasi LLM. **Skripsi ini berfokus secara spesifik pada desain dan evaluasi kerangka multi-agen itu sendiri**—agen spesialis berbasis BDI, lapisan orkestrasi Aika, dan perilaku kolektif mereka dalam konteks percakapan kritis keselamatan. Desain basis data, komponen antarmuka pengguna, dan infrastruktur deployment didokumentasikan sebagai konteks implementasi namun bukan subjek evaluasi formal. Temuan utama meliputi: (1) arsitektur keselamatan dua tingkat mencapai **False Negative Rate 0%** melalui deteksi komplementer Tingkat 1 (sensitivitas 72%) dan Tingkat 2 (sensitivitas 100%); (2) akurasi orkestrasi sebesar **64,71%**, dengan kegagalan yang mayoritas bersifat konservatif (eskalasi berlebihan), mengindikasikan kebutuhan penyempurnaan prompt; dan (3) kualitas respons terapeutik sebesar **4,08/5,0**, melampaui ambang batas target 3,5. Hasil-hasil ini menunjukkan kelayakan teknis keselamatan proaktif, orkestrasi agen yang fungsional dengan area yang teridentifikasi untuk penyempurnaan, dan dukungan yang menjaga privasi, mengonfirmasi kapasitas sistem untuk menutup kesenjangan *insight-to-action* di bawah pengawasan manusia. Kami membahas pertimbangan etis, prinsip *privacy by design*, keterbatasan penelitian, dan kebutuhan studi klinis lapangan di masa depan dengan pengguna riil.

Kata kunci: Sistem Multi-Agen; Arsitektur BDI; Orkestrasi Agen; Triase Keselamatan; LangGraph; Human-in-the-Loop; Kesejahteraan Mahasiswa; Evaluasi Berbasis Skenario

ABSTRACT

Higher Education Institutions face rising demand for student well-being support while relying on reactive counseling frameworks that often fail to reach students before crises escalate. This thesis proposes and evaluates a proactive, agentic AI framework designed to bridge the critical 'insight-to-action' gap by enabling early intervention and data-driven resource management. We introduce the *Safety Agent Suite*, a decoupled multi-agent architecture that distributes clinical and operational responsibilities to specialized agents under human oversight. The system features: (i) **Aika**, a Meta-Agent orchestrator that provides a unified user interface and performs immediate Tier 1 risk screening; (ii) a **Safety Triage Agent (STA)** for comprehensive Tier 2 conversational risk analysis; (iii) a **Therapeutic Coach Agent (TCA)** delivering Cognitive Behavioral Therapy (CBT)-based micro-interventions; (iv) a **Case Management Agent (CMA)** for operational coordination; and (v) an **Insights Agent (IA)** for privacy-preserving resource management analytics. To balance responsiveness with depth, we employ a two-tier risk monitoring architecture that combines immediate screening with deep conversational analysis to enable early intervention. The multi-agent system is built with LangGraph and includes guardrails for tool use, redaction, and auditability.

We implement a functional prototype within the UGM-AICare platform and conduct scenario-based evaluations focused exclusively on agent architecture performance: triage sensitivity and False Negative Rate (FNR) on synthetic crisis scenarios; orchestration reliability via tool-call success and state transition behavior; end-to-end latency; privacy compliance verification; and coaching quality via CBT adherence rubrics with expert assessment and LLM validation. **This thesis focuses specifically on the design and evaluation of the multi-agent framework itself**—the BDI-based specialist agents, Aika orchestration layer, and their collective behavior in safety-critical conversational contexts. Database design, user interface components, and deployment infrastructure are documented as implementation context but are not subjects of formal evaluation. Key findings include: (1) the two-tier safety architecture achieves a **0% False Negative Rate** through complementary Tier 1 (72% sensitivity) and Tier 2 (100% sensitivity) detection; (2) orchestration accuracy of **64.71%**, with failures predominantly conservative (over-escalation), indicating need for prompt refinement; and (3) therapeutic response quality of **4.08/5.0**, exceeding the 3.5 target threshold. These results demonstrate the technical feasibility of proactive safety, functional agent orchestration with identified areas for refinement, and privacy-preserving support, confirming the system's capacity to close the insight-to-action gap under human-in-the-loop supervision. We discuss ethical considerations, privacy by design principles, research limitations, and outline requirements for future clinical field studies with real users.

Keywords: Multi-Agent Systems; BDI Architecture; Agent Orchestration; Safety Triage; LangGraph; Human-in-the-Loop; Student Well-being; Scenario-Based Evaluation