RINGKASAN

Pengembangan Sistem *Document Enhancement* Berbasis Ai Dan Analisis Kontekstual Untuk *Optimasi Retrieval-Augmented Generation* (Rag) Melaluipengayaan Konteks, Ekstraksi Teks Tabel, Dan Peningkatan Akurasi OCR (Studi Kasus: PT Inspigo Inovasi Indonesia), Kimi Dandy Yudanarko, NIM E41220493, Tahun 2025, Teknik Informatika,Politeknik Negeri Jember, Dr. Denny Trias Utomo, S.Si., M.T. (Dosen Pembimbing), Bapak Setyo Guritno, (CEO PT Inspigo Inovasi Indonesia), Bapak Charisma Pribadi (Pembimbing Lapangan).

PT Inspigo Inovasi Indonesia (Inspigo) merupakan perusahaan rintisan teknologi edukasi (EdTech) berbasis di Jakarta yang berfokus pada pengembangan solusi pembelajaran berbasis kecerdasan buatan untuk klien korporasi. Inspigo mengembangkan berbagai produk AI, seperti AI Mentor dan Inspigo AI *Roleplay*, yang memanfaatkan dokumen internal perusahaan sebagai sumber utama pengetahuan dalam mendukung proses onboarding, pelatihan berkelanjutan, dan pengembangan kompetensi. Tantangan utama dalam pengembangan layanan tersebut terletak pada kemampuan AI untuk menghasilkan jawaban yang akurat, relevan, dan kontekstual, mengingat dokumen internal klien umumnya bersifat panjang, tidak terstruktur, dan mengandung informasi tersirat. Untuk menjawab tantangan tersebut, Inspigo mengembangkan sistem *Document Enhancement* AI yang bertujuan meningkatkan kualitas dokumen melalui pengayaan konteks, eksplisitasi informasi tersirat, serta peningkatan akurasi ekstraksi teks dan tabel dari dokumen PDF. Sistem ini dikembangkan menggunakan pendekatan *Retrieval-Augmented Generation* (RAG) yang mengombinasikan pencarian semantik dengan *Large Language Model* (LLM) guna meningkatkan performa layanan AI Mentor secara signifikan.

Pipeline pemrosesan dokumen pada sistem Document Enhancement AI terdiri atas lima tahapan utama yang saling terintegrasi dan dijalankan secara berurutan, dimulai dari tahap ekstraksi hingga vektorisasi. Tahap ekstraksi berfungsi mengonversi dokumen PDF ke format Markdown terstruktur dengan

memanfaatkan PyMuPDF, pdfplumber, dan Tesseract OCR, sehingga struktur dan isi dokumen asli tetap terjaga dan menghasilkan dokumen versi awal (v1). Tahap berikutnya adalah enhancement, yang bertujuan memperkaya dokumen secara kontekstual menggunakan model AI generatif melalui LangChain dengan pendekatan structured output, di mana dokumen diproses dalam beberapa window secara paralel dan menghasilkan keluaran JSON terstruktur. Selanjutnya, tahap persetujuan diterapkan sebagai mekanisme pengendalian biaya dan kualitas melalui dual approval flow, yang mencakup persetujuan estimasi biaya sebelum pemrosesan AI serta persetujuan hasil enhancement sebelum dokumen dilanjutkan ke tahap berikutnya. Tahap sintesis kemudian menggabungkan dokumen versi awal dengan hasil enhancement yang telah disetujui melalui penyisipan catatan kaki berformat Markdown secara token-aware, sehingga dihasilkan dokumen versi enhanced (v2) yang utuh dan konsisten. Tahap akhir adalah vektorisasi, yang mempersiapkan dokumen untuk pencarian semantik dan implementasi RAG dengan memecah dokumen menjadi beberapa chunk, mengonversinya ke vektor menggunakan model embedding, serta menyimpannya ke dalam basis data vektor Pinecone beserta metadata pendukung untuk menjaga keterlacakan dan akurasi pencarian.

Pengembangan sistem *Document Enhancement* AI meningkatkan kualitas dokumen klien dan memperkuat kapabilitas AI Mentor dalam menyediakan layanan pembelajaran yang akurat dan kontekstual. Sistem ini mengintegrasikan pendekatan teknis, pengendalian kualitas, dan pertimbangan biaya dalam arsitektur yang siap produksi. Kegiatan magang ini memberikan pengalaman langsung dalam penerapan teknologi AI modern dan praktik pengembangan perangkat lunak profesional. Proyek ini mencerminkan sinergi antara dunia pendidikan dan kebutuhan industri dalam mendorong transformasi digital edukasi korporasi.