Laporan Kajian Teknis Strategis: Transformasi Digital Operasional Lapangan AP2T Melalui Chatbot Asisten Cerdas Berbasis RAG

1. Pendahuluan: Paradigma Baru Layanan Pelanggan dan Operasional Lapangan

Dalam era digitalisasi utilitas energi, efisiensi operasional petugas lapangan menjadi parameter kritis yang menentukan kualitas layanan pelanggan dan integritas pendapatan perusahaan. PT PLN (Persero), melalui ekosistem Aplikasi Pelayanan Pelanggan Terpusat (AP2T), telah menstandarisasi proses bisnis dari hulu ke hilir. Namun, dinamika di lapangan—khususnya yang dihadapi oleh tim Penertiban Pemakaian Tenaga Listrik (P2TL) dan teknisi pemeliharaan—sering kali melampaui cakupan prosedur statis yang tersedia. Dokumen "Chatbot Asisten Digital Pemandu Petugas AP2T" menguraikan visi strategis untuk menjembatani kesenjangan ini melalui implementasi *Generative Artificial Intelligence* (GenAI) dengan arsitektur *Retrieval-Augmented Generation* (RAG).[1]

Laporan ini menyajikan analisis komprehensif, mencakup desain topologi proses bisnis, arsitektur sistem yang modular, serta spesifikasi infrastruktur berkinerja tinggi. Analisis ini tidak hanya membedah dokumen sumber, tetapi juga mengintegrasikan data pasar terkini mengenai perangkat keras komputasi (GPU NVIDIA A100 vs H100), evolusi basis data vektor, dan standar tata kelola AI global (ISO/IEC 42001:2023) untuk memastikan solusi yang diusulkan bersifat *future-proof*, aman, dan *cost-effective*.

1.1 Latar Belakang Strategis dan Tantangan Operasional

Petugas lapangan AP2T beroperasi dalam lingkungan yang kompleks dan penuh tekanan. Berdasarkan analisis tugas P2TL, petugas harus melakukan pemeriksaan fisik terhadap Alat Pembatas dan Pengukur (APP), mengidentifikasi modus pelanggaran yang semakin canggih, serta menyusun Berita Acara (BA) yang memiliki kekuatan hukum.[2] Tantangan utama yang menghambat efisiensi meliputi:

1. **Fragmentasi Pengetahuan:** Informasi teknis tersebar dalam ribuan halaman manual PDF, SOP, dan edaran direksi yang sulit diakses secara instan di lokasi kerja.
2. **Keterbatasan Kognitif:** Petugas tidak mungkin menghafal seluruh kode kesalahan (*error codes*) dari berbagai merek meteran listrik atau spesifikasi teknis trafo distribusi.
3. **Kendala Lingkungan:** Kondisi lapangan (hujan, terik matahari, tangan kotor) menyulitkan

interaksi dengan perangkat *mobile* konvensional yang berbasis teks.[1]

Implementasi Chatbot Asisten Digital bukan sekadar modernisasi alat bantu, melainkan rekayasa ulang proses bisnis (*Business Process Reengineering*) yang mentransformasi petugas lapangan dari sekadar pelaksana instruksi menjadi "knowledge worker" yang didukung oleh kecerdasan buatan.

---

2. Topologi Proses Bisnis: Rekayasa Ulang Interaksi Lapangan

Topologi proses bisnis dalam konteks ini mendefinisikan bagaimana alur kerja operasional petugas berinteraksi dengan sistem cerdas untuk menyelesaikan tugas-tugas kritis. Transformasi ini mengubah model "pencarian informasi pasif" menjadi "konsultasi aktif kontekstual".

2.1 Peta Alur Kerja Terintegrasi (Integrated Workflow Map)

Sistem dirancang untuk mengintervensi titik-titik friksi dalam siklus kerja petugas P2TL dan Teknik.

2.1.1 Skenario Penanganan Gangguan Teknis (*Troubleshooting*)

Dalam model konvensional, ketika petugas menemukan anomali pada peralatan (misalnya, layar meter padam atau kode error tidak dikenal), mereka harus menghubungi supervisor atau mencari manual fisik di kendaraan operasional. Model *As-Is* ini memiliki latensi tinggi yang berdampak pada *Mean Time to Repair* (MTTR).

Dalam topologi *To-Be* yang diusulkan:

1. **Akuisisi Data Multimodal:** Petugas memotret *nameplate* atau layar meter yang bermasalah melalui aplikasi *mobile*. Jika kondisi tidak memungkinkan mengetik, petugas menggunakan perintah suara (*voice command*) untuk mendeskripsikan gejala.[1]
2. **Analisis Konteks Otomatis:** Sistem backend tidak hanya menerima gambar, tetapi juga memadukannya dengan metadata kontekstual (ID Pelanggan dari Work Order AP2T, lokasi GPS, dan riwayat gangguan sebelumnya).
3. **Retrieval dan Sintesis:** Chatbot mencari basis pengetahuan internal, mencocokkan kode error dari gambar (via OCR) dengan manual teknis spesifik vendor terkait, dan menyajikan solusi langkah-demi-langkah.
4. **Panduan Eksekusi:** Petugas menerima instruksi yang divalidasi. Untuk prosedur kompleks, chatbot menampilkan potongan video tutorial pendek (durasi <30 detik) yang relevan dengan langkah tersebut.[1]

2.1.2 Skenario Penegakan Kepatuhan (P2TL)

Untuk petugas P2TL, akurasi referensi regulasi adalah vital untuk menghindari sengketa hukum dengan pelanggan.

1. **Identifikasi Anomali:** Petugas mengunggah foto instalasi yang dicurigai melanggar (misal: *bypass* pada terminal kWh).
2. **Validasi Regulasi:** Chatbot menganalisis pola visual dan menyarankan pasal pelanggaran yang relevan berdasarkan Direktori Putusan Direksi (DIR) yang berlaku.
3. **Generasi Berita Acara:** Sistem membantu memformulasikan deskripsi temuan teknis untuk dimasukkan ke dalam Berita Acara Hasil Pemeriksaan (BAHP), memastikan bahasa yang digunakan baku dan sesuai standar hukum.[2]

2.2 Matriks Peran dan Akses Informasi (RBAC Business Logic)

Topologi bisnis juga mengatur distribusi informasi berdasarkan peran, menerapkan prinsip *Need-to-Know* yang ketat demi keamanan infrastruktur kritis nasional.

Peran Petugas	Fokus Tugas Utama	Kapabilitas Chatbot yang Diaktifkan	Batasan Akses Data
Petugas P2TL	Pemeriksaan pelanggaran, penyitaan barang bukti, pembuatan BA.[2]	Analisis forensik visual, pencarian pasal regulasi, panduan penyitaan aman.	Akses penuh ke data riwayat pelanggaran; Terbatas pada data teknis gardu induk.
Teknisi APP	Pemeliharaan, penggantian meter, *troubleshooting* konektivitas.	Manual teknis mendalam, skema wiring diagram, video tutorial instalasi.	Akses penuh manual vendor; Tidak ada akses ke data profil finansial pelanggan.
Admin AP2T	Verifikasi laporan, tindak lanjut administrasi.[2]	Dashboard analitik tren pertanyaan lapangan, validasi kualitas jawaban AI.	Akses *read-only* ke log interaksi chatbot untuk audit kualitas.

2.3 Integrasi Multimodal dalam Proses Bisnis

Penggunaan input multimodal (teks, suara, gambar) bukan sekadar fitur tambahan, melainkan

kebutuhan fundamental topologi bisnis lapangan.

- **Speech-to-Text (STT):** Memungkinkan petugas melaporkan kondisi atau bertanya saat sedang memanjat tiang atau bekerja di ruang sempit (*hands-free operation*).[1]
- **Optical Character Recognition (OCR):** Menghilangkan kesalahan manusia (*human error*) dalam pembacaan nomor seri aset atau kode error yang panjang dan kompleks. OCR mengubah data visual menjadi *query* teks yang presisi untuk mesin pencari.[1]

---

3. Topologi Arsitektur Sistem: Desain Modular Enterprise RAG

Arsitektur sistem dibangun di atas prinsip *Microservices* dan *Event-Driven Architecture*, dirancang untuk skalabilitas horizontal guna melayani ribuan petugas secara serentak. Inti dari sistem ini adalah mekanisme *Hybrid Retrieval-Augmented Generation* yang menggabungkan kekuatan pencarian kata kunci deterministik dengan pemahaman semantik probabilistik.

3.1 Arsitektur Tingkat Tinggi (*High-Level Architecture*)

Sistem dibagi menjadi lima lapisan logis (*logical layers*) yang memastikan separasi fungsi (*separation of concerns*) dan kemudahan pemeliharaan.[1]

3.1.1 Layer Interaksi (*User Interaction Layer*)

- **Aplikasi Mobile/Web:** *Frontend* yang dibangun dengan kerangka kerja reaktif (seperti React Native atau Flutter) untuk mendukung *deployment* lintas platform. Fitur utamanya mencakup antarmuka obrolan (*chat interface*), modul kamera terintegrasi untuk pengambilan gambar presisi, dan perekam suara dengan kompresi audio lokal sebelum transmisi.
- **Media Display:** Komponen render yang mampu menampilkan teks terformat (Markdown), gambar diagram, dan *video player* yang mendukung *streaming* adaptif untuk kondisi sinyal rendah.

3.1.2 Layer Gerbang API (*API Gateway & Security*)

Berfungsi sebagai pintu masuk tunggal yang mengelola lalu lintas data.

- **Autentikasi & Otorisasi:** Menggunakan protokol OAuth2/OIDC yang terintegrasi dengan direktori korporat PLN (LDAP/Active Directory). Ini memastikan *Single Sign-On* (SSO) dan penegakan *Role-Based Access Control* (RBAC).[1]
- **Rate Limiting & Throttling:** Mencegah beban berlebih pada *backend* inferensi AI dengan membatasi jumlah *request* per pengguna per menit, melindungi infrastruktur dari serangan DDoS atau penggunaan abnormal.

3.1.3 Layer Orkestrasi Kecerdasan (*Intelligence Orchestration Layer*)

Ini adalah "otak" dari sistem RAG, tempat logika pemrosesan utama terjadi.

- **Query Understanding Module:** Memproses input mentah pengguna.
  - *Rewriting:* Mengubah pertanyaan ambigu ("alat ini rusak") menjadi *query* lengkap ("troubleshooting meteran merk X tipe Y layar blank") menggunakan riwayat percakapan.
  - *Routing:* Menentukan apakah pertanyaan memerlukan pencarian dokumen (RAG), pencarian data pelanggan (Database AP2T), atau sekadar obrolan ringan.
- **Embedding Engine:** Mengonversi teks pertanyaan menjadi vektor representasi numerik. Sistem menggunakan model *dense embedding* berdimensi 768-1024 (misalnya Sentence-BERT atau BGE-M3) yang di-hosting secara lokal untuk keamanan data.[1]

3.1.4 Layer Penyimpanan dan Pengetahuan (*Data & Knowledge Layer*)

Layer ini mengelola persistensi data dalam dua bentuk utama:

- **Vector Database (Milvus):** Menyimpan jutaan vektor dari potongan dokumen teknis. Milvus dipilih karena arsitekturnya yang terdistribusi (*cloud-native*), memisahkan penyimpanan dan komputasi, yang memungkinkan skalabilitas masif dibandingkan FAISS *standalone*.[4] Milvus mendukung indeksasi HNSW (*Hierarchical Navigable Small World*) untuk kecepatan pencarian *approximate nearest neighbor* yang ekstrem (<10ms).
- **Document Store (Object Storage):** Menyimpan file asli (PDF, gambar, video) yang menjadi sumber kebenaran (*ground truth*). MinIO atau solusi S3-compatible *on-premise* direkomendasikan untuk kedaulatan data.

3.1.5 Layer Generasi (*Generation Layer*)

- **LLM Inference Server (vLLM):** Menjalankan *Large Language Model* (seperti Llama 3 atau Qwen 2.5 yang di-*finetune*). Penggunaan vLLM sangat krusial karena fitur *PagedAttention* yang mengoptimalkan manajemen memori GPU, memungkinkan *throughput* 3x lebih tinggi dibandingkan solusi standar seperti HuggingFace Accelerate atau Ollama pada beban konkuren tinggi.[6]

3.2 Strategi Hybrid Retrieval: Mengapa dan Bagaimana?

Dokumen sumber menekankan penggunaan **Hybrid Search** (BM25 + Vector Search). Ini adalah keputusan arsitektural yang vital untuk konteks utilitas teknis.[1]

- **Keterbatasan Vector Search Murni:** Model embedding sering kali gagal menangkap perbedaan presisi pada kode alfanumerik pendek (misalnya, membedakan "MCB-C10" dan "MCB-C16"). Vector search bekerja berdasarkan kedekatan semantik, yang mungkin menganggap kedua kode tersebut "mirip", padahal dalam konteks teknis perbedaannya fatal.
- **Solusi Hybrid:**
  - **BM25 (Sparse Retrieval):** Menggunakan pendekatan statistik (TF-IDF) untuk mencocokkan kata kunci eksak. Jika petugas mencari kode error "Err-04", BM25 akan

menemukan dokumen yang secara eksplisit memuat string tersebut.
    - **Vector Search (Dense Retrieval):** Menangani pertanyaan konseptual seperti "cara mengatasi layar meteran buram", di mana tidak ada kata kunci tunggal yang pasti.
- Fusi Hasil (Reciprocal Rank Fusion - RRF): Algoritma yang menggabungkan peringkat hasil dari kedua metode pencarian tersebut untuk menghasilkan daftar dokumen akhir yang paling relevan. Formula fusi menyeimbangkan skor semantik dan leksikal:

  $$Score_{hybrid} = \alpha \cdot Score_{vector} + (1-\alpha) \cdot Score_{BM25}$$

  Di mana α (alpha) adalah bobot yang dapat disesuaikan (biasanya 0.5).1

3.3 Pipeline Indeksasi dan Chunking Lanjutan

Kualitas jawaban RAG sangat bergantung pada bagaimana data dipotong (*chunking*). Arsitektur ini menerapkan strategi *Content-Aware Chunking*:

- **Manual Teknis:** Menggunakan *Large Chunk* (1000-1500 token) berbasis struktur bab/sub-bab. Hal ini penting untuk menjaga konteks diagram atau tabel spesifikasi agar tidak terpotong di tengah jalan.[1]
- **SOP & Instruksi Kerja:** Menggunakan *Small Chunk* (300-500 token) yang berfokus pada langkah prosedural. Tujuannya adalah agar LLM dapat mengambil satu instruksi spesifik tanpa tercemar informasi tidak relevan dari prosedur lain.
- **Video Tutorial:** Diproses menggunakan *Speech-to-Text* (seperti OpenAI Whisper *self-hosted*) untuk menghasilkan transkrip. Transkrip ini kemudian di-*chunk* dengan menyertakan metadata *timestamp* (misal: "detik 00:30 - 00:45: Cara Membuka Segel"). Ini memungkinkan chatbot memberikan jawaban berupa tautan langsung ke menit video yang relevan.[1]

---

4. Spesifikasi Infrastruktur dan Analisis Teknis

Untuk mendukung arsitektur di atas dengan performa *real-time* (latensi <3 detik) bagi ribuan petugas, diperlukan spesifikasi infrastruktur yang presisi. Pemilihan perangkat keras didasarkan pada analisis *Cost-Performance* dan ketersediaan teknologi di tahun 2025.

4.1 Spesifikasi Server Komputasi (GPU Cluster)

Beban kerja inferensi LLM dan *embedding* membutuhkan akselerator perangkat keras khusus.

4.1.1 Perbandingan Strategis: NVIDIA A100 vs H100

Pemilihan antara arsitektur Ampere (A100) dan Hopper (H100) adalah keputusan CAPEX terbesar.

Parameter Evaluasi	NVIDIA A100 (80GB)	NVIDIA H100 (80GB)	Analisis Dampak untuk AP2T
Arsitektur	Ampere (7nm)	Hopper (4nm)	H100 memperkenalkan *Transformer Engine* yang mempercepat pelatihan/inferensi model berbasis transformer secara drastis.[8]
Memori Bandwidth	~2 TB/s (HBM2e)	~3.35 TB/s (HBM3)	Bandwidth memori adalah *bottleneck* utama dalam inferensi LLM. H100 memberikan respons yang jauh lebih cepat untuk *input context* panjang (RAG dengan banyak dokumen).[9]
Inferensi Throughput	~130 tokens/detik	250-300 tokens/detik	H100 memiliki kapasitas layanan 2.5x lipat, artinya satu server H100 dapat melayani jumlah petugas yang setara dengan 2-3 server A100.[8]
Estimasi Harga (2025)	$10k - $14k (Unit)	$25k - $40k (Unit)	Harga H100 masih premium. Namun, rasio *price-per-token* mulai kompetitif untuk skala besar.[10]
Ketersediaan	Tinggi (Pasar sekunder	Sedang-Tinggi (Lead time	A100 lebih mudah didapat, namun

	melimpah)	membaik)	H100 adalah investasi jangka panjang (5 tahun) yang lebih aman (*future-proof*).

Rekomendasi Infrastruktur:
Mengingat skala PLN dan kebutuhan respons real-time, disarankan menggunakan konfigurasi Hybrid:

- **Production Cluster:** Server dengan **4x NVIDIA H100 80GB**. H100 dipilih karena fitur *MIG (Multi-Instance GPU)* generasi kedua yang memungkinkan satu GPU fisik dipecah menjadi 7 instansi terisolasi secara aman, sangat efisien untuk melayani berbagai model (LLM besar, Embedding kecil, Reranker) dalam satu perangkat keras.[12]
- **Development/Fallback Cluster:** Server dengan **4x NVIDIA A100 80GB** untuk lingkungan pengembangan, pengujian model baru, atau *failover* darurat dengan biaya lebih rendah.

4.2 Spesifikasi Server CPU dan Penyimpanan

Meskipun GPU menangani beban AI, komponen lain krusial untuk orkestrasi dan basis data vektor.

- **CPU:** Prosesor server kelas atas (misal: AMD EPYC Genoa atau Intel Xeon Sapphire Rapids) dengan jumlah *core* tinggi (64+ cores) diperlukan untuk menangani pre-processing data (OCR, ekstraksi teks) yang bersifat CPU-bound.
- **RAM Sistem:** Minimal 1 TB - 2 TB RAM DDR5. Basis data vektor seperti Milvus, meskipun mendukung penyimpanan disk, bekerja optimal jika indeks HNSW dimuat di memori utama untuk latensi pencarian mikrotetik.
- **Storage (NVMe):**
  - **Kecepatan:** Wajib menggunakan NVMe SSD Enterprise Grade (PCIe Gen5) dalam konfigurasi RAID 10. Kecepatan baca acak (*Random Read IOPS*) sangat vital saat sistem menarik *chunk* dokumen asli setelah pencarian vektor selesai.
  - **Kapasitas:** Estimasi awal 50 TB (Usable) untuk menyimpan repositori dokumen digital, video tutorial resolusi tinggi, dan log audit sistem yang tidak boleh dihapus sesuai regulasi.[1]

4.3 Jaringan dan Konektivitas

- **Internal Bandwidth:** Konektivitas antar-node (jika menggunakan klaster multi-server) wajib menggunakan **InfiniBand NDR (400Gb/s)** atau minimal Ethernet 100GbE untuk mencegah *bottleneck* saat transfer tensor paralel antar GPU.[10]
- **Low Latency Edge:** Implementasi CDN (*Content Delivery Network*) internal di intranet PLN untuk mendistribusikan konten statis berat (video/gambar) ke unit-unit wilayah,

mengurangi beban pada pusat data utama.

---

5. Tata Kelola, Keamanan, dan Kepatuhan (ISO 42001)

Implementasi AI di lingkungan infrastruktur kritikal seperti PLN menuntut standar keamanan tertinggi. Desain sistem ini mengadopsi kerangka kerja **ISO/IEC 42001:2023 (Artificial Intelligence Management System)** sebagai landasan tata kelola.[13]

5.1 Manajemen Risiko AI (AI Risk Management)

Sesuai standar ISO 42001, sistem harus memiliki kontrol mitigasi terhadap risiko spesifik AI:

- **Halusinasi & Fabrikasi:** Risiko chatbot memberikan instruksi teknis yang salah.
  - *Mitigasi:* Implementasi mekanisme *Grounding* yang ketat. Sistem diprogram untuk menolak menjawab jika skor relevansi dokumen di bawah ambang batas (misal: < 0.75). Setiap jawaban wajib menyertakan sitasi (halaman/paragraf) ke dokumen sumber.[1]
- **Bias & Toksisitas:** Risiko output yang tidak pantas.
  - *Mitigasi:* Penggunaan *Content Moderation API* (lokal) dan *System Prompt* yang dirancang defensif untuk menjaga nada profesional.[14]

5.2 Keamanan Data dan Privasi (Data Privacy & Security)

- **Enkripsi End-to-End:**
  - *Data-at-Rest:* Seluruh database vektor dan penyimpanan dokumen dienkripsi menggunakan algoritma **AES-256**. Manajemen kunci enkripsi (Key Management) dilakukan terpusat, idealnya menggunakan *Hardware Security Module* (HSM).[1]
  - *Data-in-Transit:* Komunikasi API menggunakan TLS 1.3. Aplikasi mobile menerapkan *Certificate Pinning* untuk mencegah penyadapan data di jaringan publik/seluler.[1]
- **Kontrol Akses Zero Trust:**
  - Sistem tidak mempercayai entitas manapun secara implisit. Setiap permintaan API divalidasi token identitasnya.
  - Penerapan *Attribute-Based Access Control* (ABAC) yang lebih granular dari RBAC. Contoh: Petugas P2TL hanya bisa mengakses dokumen "Area Jawa Bali" jika atribut lokasi mereka sesuai, mencegah kebocoran data lintas wilayah operasional.[1]

5.3 Kepatuhan Regulasi (Regulatory Compliance)

Sistem dirancang untuk mematuhi regulasi perlindungan data yang berlaku (seperti UU PDP di Indonesia dan prinsip GDPR).

- **Audit Trail:** Setiap interaksi, mulai dari *query* petugas hingga jawaban AI, dicatat dalam *immutable log* (log yang tidak bisa diubah). Ini penting untuk audit forensik jika terjadi insiden keselamatan kerja akibat instruksi yang salah.[13]

- **Human-in-the-Loop:** Mekanisme umpan balik di mana petugas dapat memberi *flag* pada jawaban yang salah. Laporan ini diteruskan ke tim ahli manusia (SME) untuk verifikasi dan perbaikan basis pengetahuan, menciptakan siklus perbaikan berkelanjutan.[1]

---

6. Analisis Biaya dan Strategi Pengadaan (Total Cost of Ownership)

Analisis finansial membandingkan dua model pengadaan utama: Membangun infrastruktur sendiri (*On-Premise*) vs Menyewa layanan Cloud (*Cloud Rental*).

6.1 Estimasi Biaya Cloud Rental (OPEX)

Menggunakan asumsi sewa instans GPU H100 dari penyedia layanan *GPU Cloud* khusus (seperti Lambda, GMI Cloud) pada tahun 2025:

- Biaya sewa per GPU H100: ~$2.10 - $4.50 per jam.[10]
- Kebutuhan: 8 GPU (untuk *redundancy* dan *load* puncak) x 24 jam x 365 hari.
- Biaya Tahunan (Optimis @ $2.5/jam): $8 \times 2.5 \times 8760 = \$175,200$ per tahun.
- Biaya 3 Tahun: ~**$525,600** (belum termasuk biaya transfer data, penyimpanan, dan CPU server pendukung).

6.2 Estimasi Biaya On-Premise (CAPEX)

Menggunakan asumsi pembelian perangkat keras:

- Server 8x H100 (HGX/DGX System): ~**$350,000 - $450,000** (One-time cost).[10]
- Biaya Listrik & Cooling (3 tahun): Estimasi 10-15% dari harga hardware per tahun.
- Total Biaya 3 Tahun: ~**$500,000 - $600,000**.

6.3 Kesimpulan Finansial

Meskipun biaya awal *On-Premise* terlihat tinggi, *Break Even Point* (BEP) tercapai di sekitar tahun ke-2.5 dibandingkan sewa cloud. Namun, faktor penentu utama untuk AP2T/PLN bukanlah biaya semata, melainkan **Kedaulatan Data** dan **Keamanan**. Infrastruktur *On-Premise* memberikan kontrol mutlak terhadap data sensitif kelistrikan nasional, menghilangkan risiko data terekspos ke penyedia cloud publik pihak ketiga. Oleh karena itu, **model On-Premise (atau Private Cloud di Data Center PLN)** adalah rekomendasi strategis yang mutlak.

---

7. Roadmap Implementasi dan Manajemen Risiko

Implementasi proyek skala besar ini dibagi menjadi tiga fase strategis untuk meminimalkan

risiko gangguan operasional.[1]

Fase 1: Fondasi dan Kapabilitas Inti (Bulan 1-3)

- **Fokus:** Membangun infrastruktur *backend* dan digitalisasi dokumen prioritas.
- **Aktivitas Kunci:**
  - Pengadaan dan instalasi server GPU (A100/H100).
  - Setup *Environment* Kubernetes untuk orkestrasi kontainer (vLLM, Milvus, API Gateway).
  - Ingest manual teknis kritis (Meter, Trafo, APP) dengan strategi *chunking* berbasis bab.
- **Deliverable:** Chatbot versi Alpha (Text-only) yang dapat diakses terbatas oleh tim *Lead Field Officer*.

Fase 2: Ekspansi Multimodal dan Integrasi (Bulan 4-5)

- **Fokus:** Meningkatkan pengalaman pengguna dan akurasi pencarian.
- **Aktivitas Kunci:**
  - Implementasi *Hybrid Search* (Integrasi Elasticsearch/BM25 dengan Milvus).
  - Deployment modul OCR dan Speech-to-Text.
  - Integrasi API dengan sistem Manajemen Aset PLN untuk verifikasi data aset secara *real-time*.
- **Deliverable:** Chatbot versi Beta dengan kemampuan analisis foto dan suara, diuji coba di satu Unit Induk Wilayah (UIW).

Fase 3: Skalabilitas Penuh dan Optimasi (Bulan 6-7)

- **Fokus:** Kesiapan produksi masal dan kepatuhan standar.
- **Aktivitas Kunci:**
  - Optimasi performa: Implementasi *Semantic Caching* (Redis) untuk jawaban umum guna mengurangi beban GPU.
  - Audit keamanan dan kepatuhan ISO 42001.
  - Pelatihan masif petugas lapangan dan sosialisasi SOP penggunaan asisten digital.
- **Deliverable:** *Go-Live* nasional dengan dasbor analitik performa penuh.

7.1 Manajemen Risiko

- **Risiko Adopsi:** Petugas senior resisten terhadap teknologi baru.
  - *Mitigasi:* Desain UI/UX yang sangat sederhana (mirip WhatsApp) dan program gamifikasi insentif bagi pengguna aktif.
- **Risiko Teknis:** Kegagalan sistem saat beban puncak (misal: paska pemadaman massal).
  - *Mitigasi:* Desain *Failover* ke mode "Lite" (pencarian kata kunci sederhana tanpa LLM) jika beban GPU mencapai 95%.

Penutup

Laporan ini menegaskan bahwa pengembangan Chatbot Asisten Digital AP2T adalah langkah strategis yang layak dan krusial. Dengan memadukan arsitektur RAG canggih, infrastruktur komputasi *state-of-the-art*, dan tata kelola keamanan yang ketat, PLN dapat memberdayakan ribuan petugas lapangannya dengan kecerdasan kolektif perusahaan, mendorong efisiensi, akurasi, dan keselamatan kerja ke tingkat yang belum pernah tercapai sebelumnya.

Karya yang dikutip

1. Chatbot Asisten Digital Pemandu Petugas AP2T.pdf
2. Tugas Dan Kewenangan Petugas P2TL | PDF - Scribd, diakses Januari 4, 2026, https://id.scribd.com/presentation/516316709/TUGAS-DAN-KEWENANGAN-PETUGAS-P2TL
3. The Best Open-Source Embedding Models in 2026 - BentoML, diakses Januari 4, 2026, https://www.bentoml.com/blog/a-guide-to-open-source-embedding-models
4. Top 5 Open Source Vector Databases for 2025 (Milvus vs. Qdrant. vs Weaviate vs Faiss. etc.) | by Fendy Feng | Medium, diakses Januari 4, 2026, https://medium.com/@fendylike/top-5-open-source-vector-search-engines-a-comprehensive-comparison-guide-for-2025-e10110b47aa3
5. Best Vector Databases in 2025: A Complete Comparison Guide - Firecrawl, diakses Januari 4, 2026, https://www.firecrawl.dev/blog/best-vector-databases-2025
6. Ollama vs. vLLM: A deep dive into performance benchmarking | Red Hat Developer, diakses Januari 4, 2026, https://developers.redhat.com/articles/2025/08/08/ollama-vs-vllm-deep-dive-performance-benchmarking
7. Performance vs Practicality: A Comparison of vLLM and Ollama | by Robert McDermott, diakses Januari 4, 2026, https://robert-mcdermott.medium.com/performance-vs-practicality-a-comparison-of-vllm-and-ollama-104acad250fd
8. Comparing NVIDIA H100 vs A100 GPUs for AI Workloads | OpenMetal IaaS, diakses Januari 4, 2026, https://openmetal.io/resources/blog/nvidia-h100-vs-a100-gpu-comparison/
9. NVIDIA A100 vs. H100: Choosing the Right GPU for Your AI Workloads - Clarifai, diakses Januari 4, 2026, https://www.clarifai.com/blog/nvidia-a100-vs.-h100-choosing-the-right-gpu-for-your-ai-workloads
10. How Much Does the NVIDIA H100 GPU Cost in 2025? Buy vs. Rent Analysis - GMI Cloud, diakses Januari 4, 2026, https://www.gmicloud.ai/blog/how-much-does-the-nvidia-h100-gpu-cost-in-2025-buy-vs-rent-analysis
11. NVIDIA A100 80GB Price in 2025 - DGX A100 vs. H100 vs. RTX 4090 - Direct Macro, diakses Januari 4, 2026, https://directmacro.com/blog/post/nvidia-a100-in-2025
12. NVIDIA A100 vs H100 vs H200: GPU Comparison for AI | E2E Networks, diakses

Januari 4, 2026, https://www.e2enetworks.com/blog/nvidia-a100-vs-h100-vs-h200-gpu-comparison
13. AI lifecycle risk management: ISO/IEC 42001:2023 for AI governance | AWS Security Blog, diakses Januari 4, 2026, https://aws.amazon.com/blogs/security/ai-lifecycle-risk-management-iso-iec-42001-2023-for-ai-governance/
14. ISO/IEC 42001:2023 Guide to AI Management & IT Security - Linford & Company LLP, diakses Januari 4, 2026, https://linfordco.com/blog/iso-42001-it-security/
15. Embedding models comparison | SoftwareMill, diakses Januari 4, 2026, https://softwaremill.com/embedding-models-comparison/
16. NVIDIA H100 Price - Is It Worth the Investment? - TRG Datacenters, diakses Januari 4, 2026, https://www.trgdatacenters.com/resource/nvidia-h100-price/