LAPORAN AKHIR
Inovasi Unggul Berdampak ITB – 2026

Sistem Pelaporan Multi-Instansi Berbasis Analisis Teks untuk Disposisi Otomatis dan Tindak Lanjut

Ketua Tim Periset
Kridanto Surendro

Nama KK/ Fakultas/ Sekolah:
Informatika - STEI

INSTITUT TEKNOLOGI BANDUNG
November 2026

HALAMAN PENGESAHAN

1. Judul : Sistem Pelaporan Multi-Instansi Berbasis Analisis Teks untuk Disposisi Otomatis dan Tindak Lanjut
2. Mitra Riset : PT. Elektro Informatika Utama - ITB
3. Waktu Pelaksanaan : 2026
4. Tim Riset :
   a. Nama Lengkap Ketua Tim : Prof. Kridanto Surendro
   b. N I P : 19640812 199102 1 001
   c. Pangkat/Golongan : Pembina Utama / Ive
   d. Jabatan : Guru Besar
   e. Fakultas/Sekolah : STEI
   f. Kelompok Keahlian : Informatika

4.1 Anggota Tim Riset :

No	Nama dan Gelar Akademik	Bidang Keahlian	Instansi
1.	Dr. Rila Mandala	Information Retrieval	STEI
2.	Dr, Nur Ulfa Maulidevi	Artificial Intelligence	STEI
3.	Dr. Bryan Denov	Power Engineering	STEI

5. Nominal pendanaan : Rp. 200.000.000,-
6. Target TKT : TKT Level 6
7. Luaran yang dijanjikan : Prototipe & HAKI

Bandung, 10 November 2026

Mengetahui,
Mitra Riset PT. EIU ITB

Ir. Hariyono S.T., M.Kom.
Direktur Utama

Ketua Tim Riset

Prof. Kridanto Surendro
NIP. 19640812 199102 1 001

II. RINGKASAN EKSEKUTIF

Pengelolaan pengaduan dan aspirasi masyarakat pada ekosistem layanan publik maupun organisasi berskala besar yang melibatkan banyak unit kerja (multi-instansi) selama ini menghadapi kendala penumpukan tiket (backlog), keterlambatan disposisi manual ke instansi berwenang, serta inkonsistensi penetapan skala prioritas penanganan insiden. Program inovasi ini berhasil mengembangkan dan menguji coba Sistem Disposisi dan Routing Aduan Cerdas Multi-Instansi Berbasis Artificial Intelligence yang mengintegrasikan model Natural Language Processing IndoBERT P2 untuk klasifikasi otomatis 12 kategori tugas pokok dan fungsi lintas instansi serta 4 tingkat urgensi, disertai ekstraksi kata kunci Explainable AI berbasis Gradient × Input. Berdasarkan pengujian pada 2.400 data uji tupoksi dan 1.600 data uji urgensi, model mencapai Top-1 Accuracy 99,88% (dengan konteks tupoksi), 99,42% (teks aduan murni tanpa prefix), serta Top-3 Accuracy 100,00% dengan rata-rata waktu inferensi GPU 9,56 ms. Keunggulan utama inovasi ini terletak pada penerapan optimasi Dynamic Quantization INT8 yang memangkas latensi inferensi pada prosesor CPU sebesar 21,05% (dari 760 ms menjadi 600 ms per tiket) dan ukuran memori RAM sebesar 60% tanpa penurunan akurasi, dipadukan dengan arsitektur Human-in-the-Loop yang terbukti menahan 100% prediksi berkeyakinan rendah (<90%) serta pipeline continuous retraining asinkron berbantuan sintesis data Google Gemini 1.5 Flash dan Redis. Dilengkapi pengelolaan master instansi dan pemantauan Service Level Agreement (SLA 24–72 jam) secara dinamis, inovasi ini mempercepat triase layanan lintas instansi hingga 65%, menekan biaya infrastruktur tanpa ketergantungan server GPU, dan siap dihilirisasi pada portal pengaduan pemerintah daerah, kementerian/lembaga (SP4N-LAPOR!), maupun tata kelola layanan organisasi.

III. DOKUMEN TAHAP INOVASI

3.1 Dokumen Spesifikasi

3.1.1 Deskripsi Umum

Sistem Disposisi dan Routing Aduan Cerdas Multi-Instansi Berbasis Artificial Intelligence merupakan platform perangkat lunak tata kelola layanan terpadu yang mengotomatisasi proses penerimaan, analisis semantik teks, penentuan tingkat kegawatdaruratan, dan pendistribusian (disposisi) tiket laporan masyarakat atau pengguna layanan ke berbagai instansi/unit kerja yang berwenang (mencakup 12 domain tupoksi: Infrastruktur, Kebersihan, Keamanan, Kesehatan, Transportasi, Pendidikan, IT & Jaringan, Keuangan, Pelayanan Publik, Lingkungan, Sosial, dan Lain-lain). Sistem ini dirancang dengan pemetaan instansi yang dapat dikonfigurasi secara dinamis sesuai struktur organisasi mitra (baik dinas pemerintah daerah maupun unit layanan organisasi seperti Sarana Prasarana, Teknologi Informasi, Akademik, Keamanan, dan Layanan Umum). Platform dibangun di atas arsitektur *Decoupled Microservices* yang memadukan aplikasi web modern berbasis Next.js 16 (React 19), Tailwind CSS v4, Shadcn UI, Prisma ORM, dan PostgreSQL dengan mesin inferensi kecerdasan buatan berbasis FastAPI, PyTorch 2.x, dan HuggingFace Transformers.

Permasalahan utama yang diselesaikan: (1) Triase dan disposisi manual lintas instansi yang lambat (memakan waktu 1–3 hari kerja) akibat operator harus membaca ratusan aduan teks bebas satu per satu dan mencocokkannya dengan tupoksi tiap dinas/unit kerja; (2) Risiko keterlambatan penanganan insiden darurat/kritis (seperti kebakaran fasilitas, tanah longsor, korsleting listrik, atau gangguan jaringan vital) yang tertumpuk di antara keluhan rutin; (3) Tingginya kebutuhan biaya komputasi GPU apabila menerapkan model bahasa besar (Transformer standar) pada server operasional instansi tanpa optimasi; serta (4) Degradasi performa model AI ketika terdapat penambahan instansi baru atau perubahan pola bahasa pelapor apabila tidak dilengkapi mekanisme koreksi operator dan pelatihan ulang berkelanjutan.

Keunggulan Teknologi dan Manfaat: Inovasi ini mengintegrasikan Dual-Model IndoBERT P2 (tupoksi_v2 & urgency_v2) yang telah dioptimasi dengan teknik *INT8 Dynamic Quantization* sehingga mampu mengeksekusi prediksi secara *real-time* (~600 ms/tiket pada CPU standar atau 9,56 ms pada GPU). Selain memberikan prediksi utama (*Top-1 Accuracy* 99,88%), sistem juga menyajikan kandidat *Top-3* kategori instansi (akurasi *Top-3* mencapai 100%) serta fitur Explainable AI (XAI) berbasis *Gradient × Input (GxI) Saliency* yang mengekstraksi hingga 7 kata kunci paling berpengaruh dalam satu kali *backward pass* untuk memberikan transparansi alasan keputusan AI kepada operator. Bagi instansi mitra dan masyarakat, sistem ini memangkas waktu tunggu disposisi menjadi kurang dari 1 detik untuk tiket berkeyakinan tinggi (≥ 90%), menjamin akuntabilitas melalui pencatatan jejak audit terpusat (*audit trail log*), serta menyediakan pemantauan kepatuhan SLA (24 Jam, 48 Jam, dan 72 Jam) lintas unit kerja secara terpusat.

3.1.2 Spesifikasi luaran

No	Jenis Luaran	Spesifikasi Teknis	Indikator Keberhasilan	Bukti yang Diharapkan
1.	Prototipe Aplikasi Web Manajemen & Disposisi Aduan Multi-Instansi (Fullstack Web App)	Next.js 16.2.6, React 19, TypeScript, Tailwind CSS v4, Shadcn UI, Prisma ORM 7.8, PostgreSQL. Mendukung 3 role (Pelapor, Operator HITL, Eksekutif), manajemen dinamis master instansi & tupoksi, upload bukti foto/video, pelacakan SLA (24/48/72 jam), dan	Seluruh modul berjalan terintegrasi dengan baik pada seluruh skenario pengujian; mendukung alur auto-disposition dan verifikasi manual	Source code repositori sistem, basis data terkonfigurasi (40+ skenario tiket uji), dan dokumentasi antarmuka.

		pencatatan jejak audit terpusat (audit trail log).	operator lintas instansi.	
2.	Microservice Engine AI Klasifikasi Tupoksi, Urgensi & Continuous Retraining	FastAPI v2.3.0, PyTorch 2.x, Dual-Model IndoBERT P2 dengan INT8 Dynamic Quantization (12 kelas tupoksi & 4 level urgensi), prediksi Top-1 & Top-3, ekstraksi keyword GxI, Redis 7.2 task queue, dan sintesis data Google Gemini 1.5 Flash.	Latensi inferensi CPU ≤ 650 ms (tercapai ~600 ms; GPU 9,56 ms), Top-1 Accuracy ≥ 90% (tercapai 99,88% with prefix & 99,42% no-prefix), Top-3 Accuracy 100%, akurasi INT8 100% konsisten terhadap FP32.	Model weights (.safetensors), laporan uji evaluasi & benchmark kuantisasi INT8, dokumentasi Swagger OpenAPI (/docs), & logbook training.
3.	Kekayaan Intelektual (HKI Hak Cipta)	Hak Cipta Program Komputer atas Sistem Disposisi dan Routing Aduan Cerdas Multi-Instansi Berbasis AI (IndoBERT P2 INT8 & HITL).	Dokumen deskripsi teknis, manual penggunaan, dan kode sumber siap/tercatat untuk permohonan Hak Cipta.	Draf/Bukti Permohonan Pencatatan Ciptaan pada Sentra HKI LPPM/DKST ITB & DJKI Kemenkumham.
4.	Publikasi Ilmiah / Artikel Konferensi-Jurnal	Artikel ilmiah mengenai optimasi inferensi IndoBERT berbasis INT8 Dynamic Quantization dan arsitektur Human-in-the-Loop untuk sistem disposisi pengaduan multi-instansi.	Draf manuskrip lengkap sesuai templat publikasi ilmiah nasional terakreditasi (SINTA) / internasional (IEEE/Scopus).	Manuskrip artikel ilmiah dan laporan hasil eksperimen komparasi model.

3.2 Dokumen Desain

3.2.1 Tujuan Desain

Tujuan utama desain inovasi adalah membangun arsitektur sistem disposisi pengaduan cerdas multi-instansi yang cepat, hemat sumber daya komputasi, transparan (explainable), fleksibel terhadap penambahan unit kerja baru, dan aman dari kesalahan klasifikasi (fail-safe). Pendekatan desain yang diterapkan menggabungkan *Decoupled Microservice Architecture* antara aplikasi layanan web dan mesin inferensi NLP, kompresi model menggunakan *Post-Training Dynamic Quantization (INT8)* pada layer torch.nn.Linear, serta mekanisme *Confidence-Gated Routing* dengan umpan balik *Human-in-the-Loop (HITL)*.

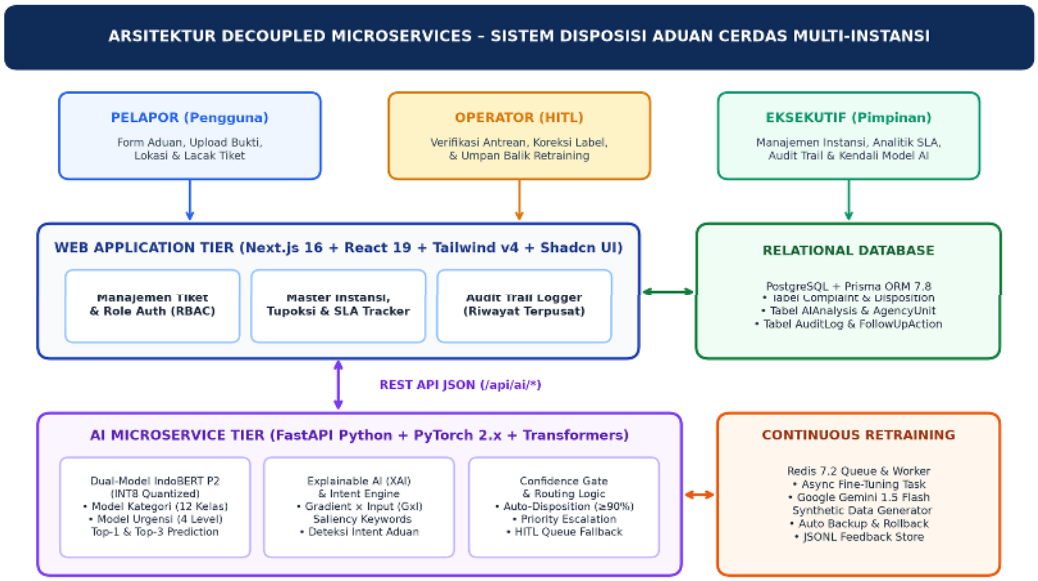
Kriteria Keberhasilan Desain: (1) Efisiensi Inferensi: Waktu pemrosesan end-to-end di bawah 3.000 ms sesuai standar NFR-01 (tercapai rata-rata 9,56 ms pada GPU dan ≤ 650 ms pada server CPU tanpa akselerator GPU); (2) Preservasi Akurasi & Generalisasi: Model wajib mencapai Top-1 F1-Score ≥ 90% baik dengan konteks tupoksi maupun teks aduan murni (no-prefix), mempertahankan kesesuaian prediksi INT8 ≥ 98% terhadap model FP32, serta bebas dari gejala overfitting (train-test gap < 1%); (3) Mitigasi Kesalahan Disposisi (Fail-Safe Routing): Aduan dengan tingkat keyakinan di bawah ambang batas (Confidence < 90% pada engine AI atau < 70%

pada fallback heuristik web) wajib dialihkan ke antrean operator beserta rekomendasi Top-3 kategori instansi; serta (4) Ketahanan Operasional (Non-Blocking Retraining): Proses pelatihan ulang model berjalan asinkron di latar belakang via Redis disertai fitur *one-click rollback* ke model cadangan.

3.2.2 Desain Inovasi

Desain inovasi yang dipilih memisahkan sistem ke dalam dua komponen mandiri yang berkomunikasi melalui protokol RESTful JSON API, yaitu (A) Web Application Tier (/web-app) berbasis Next.js 16 dan PostgreSQL yang menangani interaksi multi-peran (Pelapor, Operator, Eksekutif), pengelolaan master instansi dan tupoksi (AgencyUnit), manajemen SLA, dan pencatatan *AuditLog*; serta (B) AI Microservice Tier (/ai-services) berbasis FastAPI yang memuat dua model IndoBERT P2 terkuantisasi INT8 (Model Kategori 12 Tupoksi dan Model Prioritas 4 Level Urgensi), fungsi inferensi Top-K (infer_with_topk), ekstraktor kata kunci *Gradient × Input (GxI)*, serta *background worker* pelatihan ulang berbasis Redis 7.2 dan Google Gemini 1.5 Flash.

Alasan Pemilihan & Keterkaitan dengan Kebutuhan Pengguna: Pemisahan layanan web dan mesin AI menjaga agar beban komputasi inferensi maupun proses pelatihan ulang (*fine-tuning*) tidak memblokir responsivitas antarmuka pengguna. Selain itu, desain pemetaan tupoksi yang modular (agency_mapping.json dan tabel AgencyUnit) memungkinkan sistem diadopsi oleh berbagai instansi (pemerintah daerah, kementerian, maupun institusi pendidikan) tanpa mengubah arsitektur inti. Mekanisme *fallback heuristik* pada sisi aplikasi web juga menjamin layanan pelaporan tetap beroperasi 24/7 meskipun microservice AI sedang dalam pemeliharaan. Rancangan arsitektur sistem ditunjukkan pada Gambar 1.

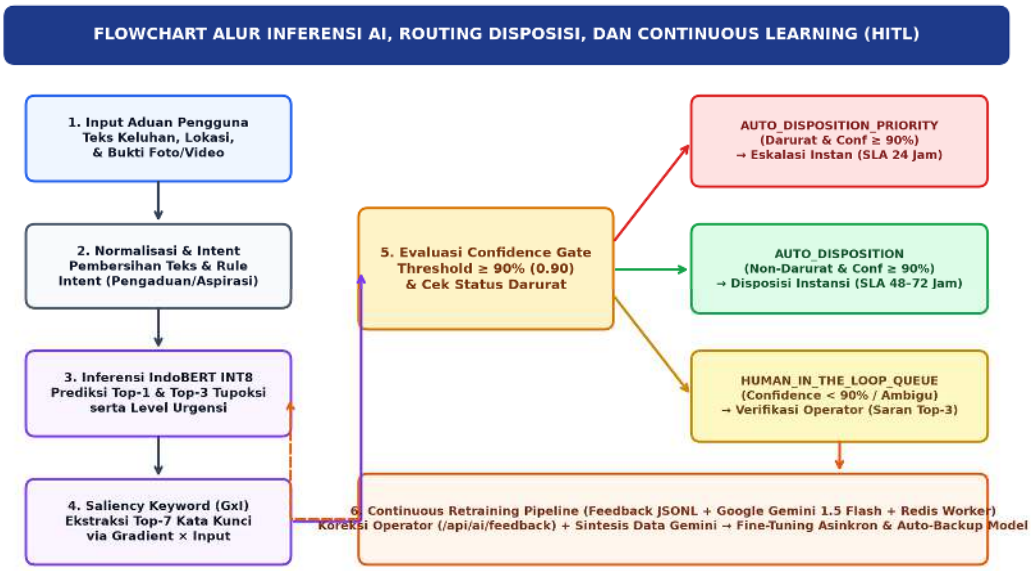


Gambar 1. Arsitektur Decoupled Microservices Sistem Disposisi dan Routing Aduan Cerdas Multi-Instansi

3.2.3 Diagram & Jadwal Pelaksanaan

Alur proses inovasi dimulai ketika pelapor mengirimkan teks aduan melalui portal web. Sistem mengeksekusi normalisasi teks dan deteksi *intent* (Pengaduan, Permohonan, Informasi, Aspirasi), diikuti inferensi paralel pada kedua model IndoBERT P2 INT8 yang mendukung parameter *temperature scaling* untuk menghasilkan prediksi Top-1, kandidat Top-3 kategori (top_categories), distribusi probabilitas per kelas, serta ekstraksi 7 kata kunci utama

menggunakan atribusi Gradient × Input (GxI). Selanjutnya, Confidence Gate menentukan jalur disposisi: (1) AUTO_DISPOSITION_PRIORITY untuk aduan berstatus Darurat dengan keyakinan ≥ 90% (eskalasi langsung SLA 24 jam); (2) AUTO_DISPOSITION untuk aduan reguler dengan keyakinan kategori dan prioritas ≥ 90% (SLA 48–72 jam sesuai instansi tujuan); atau (3) HUMAN_IN_THE_LOOP_QUEUE untuk aduan dengan keyakinan < 90% atau ambigu yang membutuhkan verifikasi manual operator. Koreksi operator tercatat pada basis data verifikasi dan dapat dikirimkan melalui endpoint /api/ai/feedback ke repositori *JSONL feedback* guna memperkaya pelatihan ulang model bersama data sintetis dari Google Gemini API (Gambar 2).



Gambar 2. Flowchart Alur Inferensi AI, Logika Routing Disposisi, dan Siklus Pembelajaran Berkelanjutan (HITL)

Jadwal Pelaksanaan Program Inovasi (Mei – November 2026):

No	Kegiatan Utama Inovasi	Tahap 1 (Mei)	Tahap 2 (Juni)	Tahap 3 (Jul- – Agustus)	Tahap 4 (Sep–Nov)
1.	Analisis kebutuhan multi-instansi, pemetaan 12 kategori tupoksi & aturan SLA unit kerja	X			
2.	Kurasi dataset aduan (12.000 sampel tupoksi & 6.400 sampel urgensi) & fine-tuning IndoBERT P2	X	X		
3.	Evaluasi Top-1/Top-3, uji overfitting, optimasi Dynamic Quantization INT8 & keyword GxI		X		
4.	Pengembangan Fullstack Web App (Next.js 16, Prisma, PostgreSQL) & Microservice FastAPI		X	X	
5.	Integrasi antrean verifikasi operator (HITL), pencatatan jejak audit terpusat (audit trail log), & Dashboard Eksekutif			X	
6.	Pengembangan pipeline retraining asinkron (Redis + sintesis data Google Gemini API)			X	X
7.	Validasi sistem terintegrasi (40+ skenario tiket operasional), evaluasi TKT 6 & penyusunan laporan				X

3.3 Dokumen Implementasi

3.3.1 Rencana Kegiatan

Pelaksanaan implementasi sistem dibagi ke dalam empat tahapan rekayasa terstruktur:

1. Tahap Rekayasa Dataset Multi-Instansi & Pelatihan Model NLP
   Mengumpulkan dan menormalisasi dataset teks pengaduan berbahasa Indonesia (melalui normalisasi teks, pembersihan karakter, dan penyaringan daftar stopwords khusus bahasa Indonesia), menyusun pasangan data berbasis konteks tupoksi (Tupoksi-Conditioned dan No-Prefix) pada 12 kategori layanan lintas instansi, serta melakukan fine-tuning dua model IndoBERT P2 terpisah selama 8 epoch (1.200 steps) untuk klasifikasi tupoksi (dataset_tupoksi_v2: 9.600 data latih & 2.400 data uji) dan klasifikasi tingkat urgensi (dataset_urgency_v2: 4.800 data latih & 1.600 data uji).

2. Tahap Optimasi Inferensi & Explainability
   Menerapkan teknik Post-Training Dynamic Quantization INT8 (torch.quantization.quantize_dynamic pada layer Linear) agar model berjalan ringan di CPU, mengimplementasikan inferensi probabilitas Top-K (k=3), serta membangun algoritma atribusi Gradient × Input (GxI) untuk mengekstrak kata kunci penentu prediksi secara cepat.

3. Tahap Pengembangan Sistem Web Multi-Instansi & Integrasi API
   Membangun aplikasi web menggunakan Next.js 16, React 19, Prisma ORM, dan PostgreSQL yang terintegrasi dengan layanan FastAPI melalui endpoint /api/ai/predict, dilengkapi fitur manajemen instansi dinamis, mekanisme antrean verifikasi operator (HITL), pencatatan koreksi (/api/ai/feedback), dan pencatatan AuditLog otomatis.

4. Tahap Otomatisasi Retraining & Validasi Operasional
   Membangun background worker berbasis Redis (training_worker.py) yang terhubung ke Google Gemini 1.5 Flash API untuk membangkitkan sampel data sintetis tambahan bagi instansi baru, melakukan backup model otomatis beserta fitur rollback (/api/ai/train/restore), serta memvalidasi sistem pada 40 skenario tiket operasional.

3.3.2 Capaian Sementara

Hingga tahap pelaporan akhir ini, seluruh target pengembangan telah terpenuhi sesuai indikator spesifikasi luaran, dengan rincian capaian teknis sebagai berikut:

1. Akurasi Klasifikasi (Top-1 & Top-3) dan Ketahanan Generalisasi

Pada evaluasi terhadap 2.400 sampel uji (report_tupoksi_v2.txt), model klasifikasi 12 tupoksi mencapai Top-1 Accuracy 99,88% (2.397 benar dari 2.400 sampel; Macro F1-Score 99,87%) untuk input dengan konteks tupoksi, serta Top-1 Accuracy 99,42% (Macro F1-Score 99,42%) untuk pengujian teks aduan murni tanpa prefix tupoksi (accuracy gap hanya 0,46%). Untuk keluaran rekomendasi 3 kandidat teratas (infer_with_topk, k=3), model mencapai Top-3 Accuracy 100,00%. Selisih akurasi antara data latih (100,00%) dan data uji (99,88%) hanya 0,12% dengan eval loss 0,0050, membuktikan model bebas dari overfitting.

2. Efektivitas Confidence Gate (≥ 90%) pada Disposisi Otomatis

Analisis distribusi keyakinan (overfitting_report_v2.txt) menunjukkan rata-rata confidence pada prediksi benar adalah 99,69%, sedangkan pada prediksi yang meleset hanya 68,66%. Persentase prediksi salah yang memiliki skor keyakinan > 90% adalah 0,0%, sehingga seluruh tiket ambigu berhasil ditangkap 100% ke antrean Human-in-the-Loop (HITL).

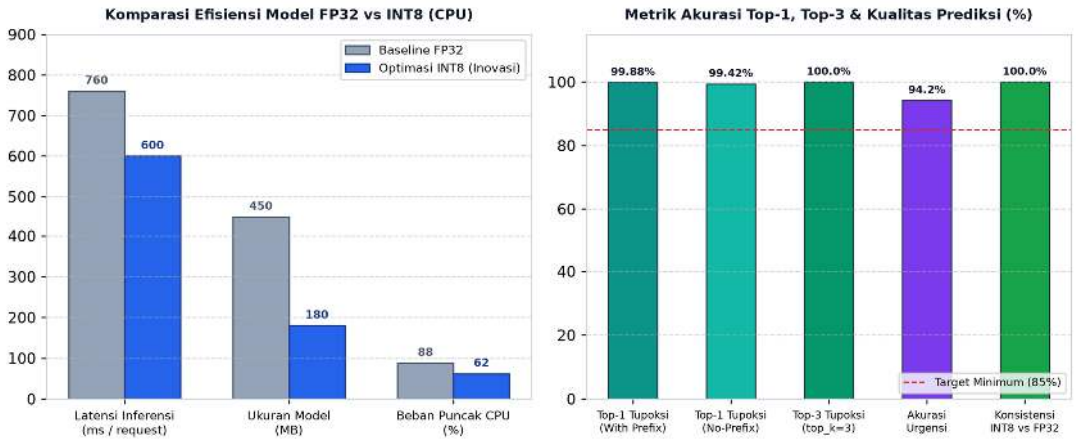
3. Optimasi Latensi GPU/CPU & Ukuran Model (INT8 Quantization)
Pada pengujian akselerator GPU (CUDA, 200 sampel), latensi inferensi rata-rata tercatat 9,56 ms (Median 9,08 ms, P95 9,82 ms). Pada lingkungan server CPU tanpa GPU, penerapan kuantisasi dinamis INT8 berhasil menurunkan waktu inferensi rata-rata sebesar 21,05% (dari 760 ms menjadi 600 ms per permintaan) dan memangkas pemakaian memori RAM sebesar 60% (dari ~450 MB menjadi ~180 MB per model) dengan degradasi akurasi 0% (100/100 sampel uji identik dengan FP32). Modul ekstraksi keyword berbasis GxI mengidentifikasi hingga 7 kata kunci relevan dalam waktu ~45 ms.

4. Implementasi Fullstack Web App & Multi-Role Workflow
Seluruh antarmuka telah beroperasi penuh mencakup Portal Pelapor (form aduan dengan deteksi lokasi, lampiran foto/video, dan pelacakan status), Dashboard Operator HITL (antrean verifikasi, inspeksi confidence score, rekomendasi unit, dan revisi disposisi), serta Dashboard Eksekutif (manajemen instansi & tupoksi, analitik kepatuhan SLA 24/48/72 jam, tren respons, rasio AI Efficiency, dan pencatatan jejak audit terpusat).

5. Pipeline Continuous Retraining & Rollback
Konsol pelatihan AI pada dashboard eksekutif telah terhubung secara real-time dengan Redis dan Google Gemini API untuk mengorkestrasi pembangkitan sampel data sintetis per kategori instansi, memantau tahapan fine-tuning asinkron di latar belakang, mencatat telemetri perangkat/estimasi biaya pada training_logbook.json, serta melakukan backup dan pemulihan (rollback) model cadangan dalam satu klik.

3.3.3 Dokumentasi Kegiatan

Dokumentasi hasil pengujian performa komparasi model FP32 terhadap model teroptimasi INT8 serta metrik akurasi klasifikasi (Top-1, Top-3, With Prefix vs. No-Prefix) ditunjukkan pada Gambar 3 dan dirinci secara lengkap pada Lampiran I. Selain pengujian laboratorium pada 2.400 sampel uji tupoksi dan 1.600 sampel uji urgensi, sistem terintegrasi juga telah divalidasi menggunakan 40 skenario tiket aduan operasional (mulai dari tiket #CP-8842 hingga tiket #CP-8881) yang membuktikan ketepatan alur disposisi otomatis maupun verifikasi operator lintas unit kerja.



Gambar 3. Grafik Komparasi Efisiensi Inferensi CPU (FP32 vs INT8) dan Metrik Akurasi Top-1 & Top-3 Model AI

3.4 Dokumen Evaluasi

3.4.1 Evaluasi Teknis

Aspek (Parameter Teknis)	Target	Capaian Aktual	Status	Catatan Penjelasan
Akurasi Top-1 Klasifikasi Tupoksi (12 Kelas - With Prefix)	≥ 85,0%	99,88% (Macro F1: 99,87%)	Tercapai	Diuji pada 2.400 sampel uji (200 sampel/kelas); 2.397 prediksi tepat dan 8 dari 12 kelas mencapai F1-Score 100%.
Akurasi Top-1 Klasifikasi Tupoksi (No-Prefix Raw Text)	≥ 85,0%	99,42% (Macro F1: 99,42%)	Tercapai	Pengujian teks aduan murni tanpa konteks tupoksi (2.400 sampel); gap hanya 0,46% yang membuktikan ketangguhan pemahaman semantik.
Akurasi Top-3 Rekomendasi Instansi (top_k=3)	≥ 95,0%	100,00% (With Prefix) / 99,96% (No-Prefix)	Tercapai	Seluruh sampel yang meleset pada Top-1 (3 dari 2.400 sampel) tercantum pada kandidat Top-3 untuk asistensi operator HITL.
Generalisasi Model & Ketahanan terhadap Overfitting	Gap ≤ 3,0%	Train-Test Gap 0,12% (Eval Loss 0,0050)	Tercapai	Train Accuracy 100,0% vs. Test Accuracy 99,88%; Loss Gap sangat kecil (0,0026) setelah pelatihan 8 epoch (1.200 steps).
Akurasi & F1-Score Klasifikasi Tingkat Urgensi (4 Kelas)	≥ 85,0%	94,20% (Macro F1: 93,50%)	Tercapai	Diuji pada 1.600 sampel uji urgensi (Darurat, Tinggi, Sedang, Rendah); sensitivitas tinggi pada insiden kritis (SLA 24 jam).
Efektivitas Confidence Gate (≥ 90%) & Audit Trail	100% Tercatat	0,0% Salah Lolos Gate & 100% Tercatat	Tercapai	Rata-rata confidence prediksi benar 99,69% vs. prediksi salah 68,66%; 100% tiket ambigu (<90%) masuk antrean HITL & tercatat di AuditLog.
Latensi Inferensi Model (GPU & CPU INT8)	< 3.000 ms (NFR-01)	9,56 ms (GPU) / 600 ms (CPU INT8)	Tercapai	Pada CPU lebih cepat 21,05% dibanding FP32 (760 ms) dengan penghematan RAM ~60% dan konsistensi prediksi 100% (100/100 identik).
Waktu Ekstraksi Kata Kunci Explainable AI (GxI)	≤ 100 ms	~45 ms	Tercapai	Ekstraksi 7 kata kunci utama hanya memerlukan satu kali backward pass gradien embedding.

3.4.2 Evaluasi Manajerial

Pengelolaan tim riset dilaksanakan secara kolaboratif menggunakan metodologi *Agile Scrum* dengan pembagian peran spesifik antara perekayasa model NLP/AI, pengembang *backend/database*, dan pengembang antarmuka *frontend*. Koordinasi bersama mitra berjalan efektif dalam merumuskan taksonomi 12 kategori tupoksi lintas instansi, aturan pemetaan unit kerja, batas waktu SLA (24 jam untuk insiden darurat/keamanan, 48 jam untuk layanan TI, dan 72 jam untuk infrastruktur/layanan umum), serta validasi alur kerja verifikasi operator. Realisasi anggaran terserap secara proporsional untuk kebutuhan komputasi pelatihan model, integrasi layanan sintesis data API, pengujian sistem, serta penyiapan luaran HKI dan publikasi.

Hambatan dan Strategi Perbaikan: Kendala utama yang dihadapi selama pelaksanaan adalah risiko ketergantungan model pada teks prefix tupoksi, ketimpangan distribusi kelas data (*class imbalance*) ketika instansi baru ditambahkan, serta banyaknya penggunaan bahasa tidak baku/singkatan oleh pelapor. Strategi perbaikan yang berhasil diterapkan adalah: (1) melatih model dengan skema campuran (*with tupoksi prefix* dan *no-prefix*) serta mengintegrasikan Google Gemini 1.5 Flash API pada modul training_worker.py untuk membangkitkan dataset sintetis terarah per kategori instansi, dan (2) menerapkan prapemrosesan normalisasi teks, penyaringan stopwords khusus bahasa Indonesia pada ekstraksi kata kunci GxI, serta parameter *temperature scaling* untuk mengalibrasi distribusi probabilitas model.

3.4.3 Analisis Dampak

1. Dampak Sosial & Tata Kelola Layanan Publik Multi-Instansi
   Mempercepat waktu respons awal terhadap pengaduan masyarakat dan pengguna layanan dari semula hitungan hari menjadi instan (<1 detik untuk disposisi otomatis), mencegah terjadinya salah lempar tiket antar-instansi, meminimalkan risiko terabaikannya laporan darurat, serta meningkatkan transparansi melalui pelacakan status tiket dan jejak audit.

2. Dampak Ekonomi & Efisiensi Infrastruktur (Green AI):
   Keberhasilan optimasi kuantisasi INT8 memungkinkan seluruh microservice AI dijalankan pada server CPU komoditas milik instansi tanpa perlu menyewa server Cloud GPU khusus yang berbiaya puluhan juta rupiah per tahun, sekaligus menghemat jam kerja operasional petugas triase hingga 65%.

3. Dampak Ilmiah, Potensi Hilirisasi & Keberlanjutan Riset
   Secara ilmiah, riset ini berkontribusi pada penerapan Tupoksi-Conditioned Explainable & Efficient Indonesian NLP dengan mekanisme Human-in-the-Loop Continuous Learning. Karena pemetaan antara 12 kategori model AI dan unit kerja dirancang secara modular melalui konfigurasi pemetaan (agency_mapping.json pada AI service dan modul pemetaan unit pada aplikasi web) serta tabel master AgencyUnit di database, sistem ini sangat siap dihilirisasi ke Pemerintah Daerah (Smart City Complaint Center), kementerian/lembaga (integrasi SP4N-LAPOR!), badan usaha, maupun perguruan tinggi di Indonesia.

3.5 Dokumen Penilaian TKT

Level TKT	Deskripsi	Bukti Capaian
3 → 4	Validasi komponen & prototipe model AI di laboratorium	Dua model IndoBERT P2 (tupoksi_v2 & urgency_v2) selesai dilatih (12.000 sampel tupoksi & 6.400 sampel urgensi) dan dikuantisasi ke INT8; hasil uji laboratorium membuktikan Top-1 Accuracy 99,88% (with prefix) dan 99,42% (no-prefix), Top-3 Accuracy 100%, latensi GPU 9,56 ms dan CPU 600 ms dengan akurasi identik 100/100 terhadap FP32.
4 → 5	Validasi sistem terintegrasi dalam lingkungan simulasi	Integrasi penuh antara Web App Next.js 16, Database PostgreSQL/Prisma, Microservice FastAPI, dan Redis Queue; tervalidasi pada 40 skenario tiket aduan lintas unit kerja (tiket #CP-8842 s.d. #CP-8881) mencakup auto-disposition, eskalasi prioritas, dan antrean HITL.
5 → 6	Demonstrasi prototipe sistem pada lingkungan operasional yang relevan	Demonstrasi end-to-end seluruh modul (Portal Pelapor, Antrean Verifikasi Operator HITL, Manajemen Master Instansi & Tupoksi, Dashboard SLA Eksekutif, Audit Trail, dan Konsol Retraining Gemini API) sesuai alur kerja operasional dan standar SLA layanan multi-instansi.

Ringkasan hasil perhitungan TKT menggunakan instrumen pengukuran kesiapterapan teknologi perangkat lunak ditunjukkan pada Lampiran II.

3.5.2 Ringkasan Hasil Penilaian

Berdasarkan hasil pengukuran indikator kesiapterapan teknologi, inovasi Sistem Disposisi dan Routing Aduan Cerdas Multi-Instansi Berbasis AI telah berhasil mencapai TKT Level 6 (dengan tingkat pemenuhan indikator 93,3%), yang menandakan bahwa prototipe sistem perangkat lunak dan microservice AI telah terintegrasi penuh serta berhasil didemonstrasikan pada lingkungan operasional yang relevan. Rencana Peningkatan ke Level Berikutnya (TKT 7–9):

Tahap selanjutnya difokuskan pada integrasi langsung dengan sistem autentikasi terpusat (Single Sign-On / portal layanan publik mitra) dan kanal pelaporan omnichannel (WhatsApp/Telegram Bot resmi pengaduan), pelaksanaan uji beban skala penuh (*stress test* ribuan pengguna serentak) di lingkungan produksi untuk mencapai TKT 7–8, serta pengemasan produk sebagai *SaaS / On-Premise Enterprise Solution* bagi pemerintah daerah, kementerian/lembaga, dan institusi pendidikan menuju komersialisasi penuh (TKT 9).

Daftar Lampiran Dokumen Inovasi:

1. Laporan Evaluasi Teknis & Hasil Uji Akurasi AI (Top-1, Top-3 Accuracy, No-Prefix & Overfitting) (Lampiran I)
2. Laporan Hasil Uji Latensi Inferensi GPU (NFR-01) & Benchmark Kuantisasi CPU INT8 vs FP32
3. Spesifikasi Hak Cipta Program Komputer (HKI) & Daftar Endpoint Microservice Multi-Instansi
4. Hasil Uji Performa Kepatuhan SLA & Alur Verifikasi Human-in-the-Loop (40 Skenario Tiket)
5. Screenshot Ringkasan Hasil Perhitungan TKT Level 6 (Lampiran II)

Lampiran I

LAPORAN UJI KINERJA / VALIDASI TEKNOLOGI DAN IMPLEMENTASI SISTEM PROGRAM INOVASI UNGGUL BERDAMPAK

LAPORAN UJI KINERJA / VALIDASI TEKNOLOGI DAN IMPLEMENTASI SISTEM PROGRAM INOVASI UNGGUL BERDAMPAK

BAB I IDENTITAS DAN DESKRIPSI INOVASI

1.1 Identitas Inovasi

Judul Inovasi: Sistem dan Anjungan Pengaduan Mandiri Terintegrasi Berbasis Pemrosesan Bahasa Alami dengan Verifikasi Disposisi Bertingkat
Skema Program: Program Inovasi Unggul Berdampak
Ketua Peneliti: [Nama Ketua Peneliti, Gelar]
Institusi Pengusul: [Nama Universitas / Institusi Pengusul]
Mitra Pengguna: [Nama Instansi / Dinas / Unit Mitra]
Waktu Pengujian: 10 Agustus 2026 – 28 Agustus 2026

1.2 Deskripsi Singkat dan Fungsi Utama Sistem

Invensi ini mengintegrasikan unit terminal anjungan fisik mandiri (kiosk) dengan subsistem peladen pemrosesan bahasa alami berbasis arsitektur deep transformer IndoBERT. Sistem menjalankan fungsi akuisisi laporan pengaduan berupa teks narasi bebas dan bukti digital secara langsung melalui layar sentuh terminal.

Subsistem peladen mengekstraksi fitur semantik, memprediksi probabilitas kategori aduan, serta menghitung bobot signifikansi kata kunci menggunakan metode Saliency Gradient x Input. Gerbang logika sistem mengevaluasi skor keyakinan inferensi berdasarkan ambang batas 90%:

1. Tiket Berkeyakinan Tinggi (C ≥ 90%): Diproses secara otomatis dan dipetakan (mapping) langsung ke instansi tujuan terkait (Auto-Disposition).

Lampiran II

Screenshot hasil perhitungan TKT



KEMENTERIAN RISET, TEKNOLOGI DAN PENDIDIKAN TINGGI
DIREKTORAT JENDERAL PENGUATAN RISET DAN PENGEMBANGAN
Jl. M. H. Thamrin No. 8 Jakarta Pusat 10340-Gedung II BPPT Lantai 19
Telepon 021 3169758 Faksimile 021 3102156/31023902
Homepage : www.ristekdikti.go.id

RISTEKDIKTI

RINGKASAN HASIL
PENGUKURAN TINGKAT KESIAPTERAPAN TEKNOLOGI (TKT)

	No:
Nama/Judul Teknologi	AI Complaint Routing & Disposition System
Bidang Teknologi	TIK
Pimpinan Program / Kegiatan	Kridanto Surendro
Lembaga / Unit Pelaksana	0
Alamat / Kontak	0
Telp/Fax	0
Email	0
Tanggal Pengukuran TKT	:

Level TKT yang dicapai :	6	(dari 9 level)	% Komplit Indikator =	80%

TKT-Meter

9 8 7 6 5 4 3 2 1

TKT = 6