KERJA PRAKTIK - EF234603

Perlindungan Anak di Ruang Digital (PAD) CPT + SFT - LLM Program AI Talent Factory (AITF)

Kementerian Komunikasi dan Digital (Komdigi)
Jl. Medan Merdeka Barat No. 9, Jakarta Pusat, DKI Jakarta, 10110
Periode: 6 Februari 2026 – 30 Juni 2026

Oleh:

Danial Rajiv Syahidan	5054231004
Adinda Shafa Najla	5053231004
Panji Seno Aji	5054231023
Muhammad Khibban Itishom	5025231126

Pembimbing Departemen
Imam Mustafa Kamal, S.ST, Ph.D

Pembimbing Lapangan
Ranti Ramadhiana, S.Kom.

DEPARTEMEN TEKNIK INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2026

KERJA PRAKTIK - EF234603

Perlindungan Anak di Ruang Digital (PAD) CPT + SFT LLM Program AI Talent Factory (AITF)

Kementerian Komunikasi dan Digital (Komdigi)
Jl. Medan Merdeka Barat No. 9, Jakarta Pusat, DKI Jakarta, 10110
Periode: 6 Februari 2026 – 30 Juni 2026

Oleh:

Danial Rajiv Syahidan	5054231004
Adinda Shafa Najla	5053231004
Panji Seno Aji	5054231023
Muhammad Khibban Itishom	5025231126

Pembimbing Departemen
Imam Mustafa Kamal, S.ST, Ph.D

Pembimbing Lapangan
Ranti Ramadhiana, S.Kom.

DEPARTEMEN TEKNIK INFORMATIKA
Fakultas Teknologi Elektro dan Informatika Cerdas
Institut Teknologi Sepuluh Nopember
Surabaya 2026

[Halaman ini sengaja dikosongkan]

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

LEMBAR PENGESAHAN
KERJA PRAKTIK

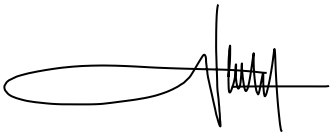
Perlindungan Anak di Ruang Digital (PAD) LLM CPT + SFT Program AI Talent Factory (AITF)

Oleh:

Danial Rajiv Syahidan	5054231004
Adinda Shafa Najla	5053231004
Panji Seno Aji	5054231023
Muhammad Khibban I	5025231126

Disetujui oleh Pembimbing Kerja Praktik:

1. Imam Mustafa Kamal, S.ST, Ph.D
   NIP.[199203122024061001]



Imam Mustafa Kamal

2. Ranti Ramadhiana, S.Kom.
   NIP.[199403062022032010]

Ranti Ramadhiana

Perlindungan Anak di Ruang Digital (PAD)
Continual Pre-training Model (CPT) & Fine Tuning Model (SFT) - LLM
Program AI Talent Factory (AITF)

Nama Mahasiswa :

1. Danial Rajiv Syahidan
2. Adinda Shafa Najla
3. Panji Seno Aji
4. Muhammad Khibban Itishom

NRP :

1. 5054231004
2. 5053231004
3. 5054231023
4. 5025231126

Departemen : Teknik Informatika FTEIC-ITS
Dosen Pendamping : Imam Mustafa Kamal, S.ST, Ph.D

ABSTRAK

Perkembangan media sosial yang sangat pesat meningkatkan risiko paparan konten digital yang tidak sesuai bagi anak dan remaja. Berbagai bentuk konten seperti perjudian daring, kekerasan, ujaran kebencian, pelecehan, eksploitasi seksual, maupun perilaku menyakiti diri sendiri tersebar dengan cepat melalui berbagai platform digital. Kondisi tersebut menjadi tantangan bagi Kementerian Komunikasi dan Digital (KOMDIGI) dalam melakukan moderasi konten secara efektif, mengingat tingginya volume data yang harus diproses serta keterbatasan metode moderasi konvensional yang masih bergantung pada pencocokan kata kunci dan verifikasi manual.

Melalui program Artificial Intelligence Talent Factory (AITF), Tim PAD 1 Institut Teknologi Sepuluh Nopember mengembangkan model Large Language Model (LLM) yang dikhususkan untuk melakukan klasifikasi rating usia terhadap konten teks berbahasa Indonesia dalam rangka mendukung implementasi Perlindungan Anak di Ruang Digital (PAD). Model dikembangkan melalui dua tahapan utama, yaitu Continued Pre-Training (CPT) menggunakan korpus teks spesifik domain perlindungan anak dan Supervised Fine-Tuning (SFT) menggunakan dataset instruksi yang telah dikurasi agar model mampu menghasilkan klasifikasi rating usia beserta penjelasan secara konsisten. Hasil pelatihan kemudian diintegrasikan menjadi layanan inferensi berbasis REST API sehingga dapat digunakan sebagai komponen pada sistem moderasi konten digital.

Hasil pengujian menunjukkan bahwa pendekatan adaptasi domain melalui CPT yang dilanjutkan dengan SFT mampu meningkatkan kemampuan model dalam memahami konteks bahasa Indonesia, termasuk penggunaan bahasa tidak baku, istilah media sosial, serta konten sensitif yang berkaitan dengan perlindungan anak. Model yang dihasilkan mampu mengklasifikasikan konten ke dalam kategori Semua Umur (SU), 7+, 13+, 15+, 18+, maupun Konten Terlarang, disertai penjelasan

yang mendukung proses verifikasi oleh moderator. Seluruh komponen sistem berhasil diimplementasikan dan dideploy sebagai Minimum Viable Product (MVP) yang siap diintegrasikan pada ekosistem Perlindungan Anak di Ruang Digital KOMDIGI.

Kata Kunci: Perlindungan Anak di Ruang Digital, Large Language Model, Continued Pre-Training, Supervised Fine-Tuning, Moderasi Konten, Klasifikasi Rating Usia.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah SWT atas penyertaan dan karunia-Nya sehingga tim penulis dapat menyelesaikan Laporan Akhir Capstone Project yang berjudul: "Perlindungan Anak di Ruang Digital".

Penulis menyadari bahwa masih banyak kekurangan dalam pelaksanaan Capstone Project maupun penyusunan laporan ini. Namun penulis berharap laporan ini dapat bermanfaat bagi pembaca serta mendukung upaya perlindungan anak di ruang digital sesuai amanat PP Nomor 17 Tahun 2025 tentang PP TUNAS.

Penulis ingin menyampaikan rasa terima kasih kepada:

1. Kedua orang tua dan keluarga penulis.
2. Bapak Imam Mustafa Kamal, S.ST, P.hD selaku dosen pendamping Capstone Project.
3. Tim Pengawasan Ruang Digital (PRD) dan Kepala Pusat Pengembangan Talenta Digital BPSDM Komdigi selaku mitra AITF Komdigi .
4. Rekan-rekan Tim AITF Komdigi dan seluruh pihak yang senantiasa memberikan semangat selama pelaksanaan Capstone Project.

Surabaya, 9 Juli 2026

BAB 1 PENDAHULUAN

1.1 Profil Mitra

Mitra utama pertama dalam proyek ini adalah Kementerian Komunikasi dan Digital Republik Indonesia (KOMDIGI) melalui Pusat Pengembangan Talenta Digital, Badan Pengembangan Sumber Daya Manusia Komunikasi dan Digital (BPSDM KOMDIGI). KOMDIGI bertindak sebagai instansi pembina serta pemegang otoritas regulasi, pengawasan, dan peru-musan kebijakan tata kelola ruang digital nasional. Dalam konteks proyek ini, KOMDIGI berperan sebagai pemegang kebijakan perlindungan masyarakat di dunia maya, khususnya dalam memastikan ketersediaan ruang digital yang aman, produktif, dan ramah bagi perkem-bangan anak serta remaja melalui inisiatif Perlindungan Anak di Ruang Digital (PAD).

Mitra kedua adalah Artificial Intelligence Talent Factory (AITF), yang berfungsi seba-gai wadah akselerasi talenta AI nasional yang menjembatani kebutuhan industri dan pemerintahan dengan keahlian teknologi mutakhir. AITF bertugas menginisiasi program, menetapkan struktur pembagian tim ahli, merumuskan fase-fase pengerjaan proyek dari tahap penelitian hingga deployment, serta menyediakan standar ekosistem pengembangan teknologi kecerdasan buatan terapan. Peran AITF krusial dalam memastikan bahwa metodologi rekayasa model yang digunakan memenuhi standar industri yang efisien, terukur, dan dapat dipertang-gungjawabkan.

1.2 Gambaran Umum Proyek

Proyek ini berfokus pada pengembangan dan adaptasi Large Language Model (LLM) yang terspesialisasi untuk tugas moderasi konten dan klasifikasi rating usia berbasis teks. Sistem yang dikembangkan ditujukan untuk mengidentifikasi, mengklasifikasikan, dan menganalisis caption maupun teks media sosial berbahasa Indonesia ke dalam klasifikasi rating usia yang selaras dengan regulasi nasional. Klasifikasi tersebut dibagi menjadi enam kategori tingkatan, yaitu Semua Umur (SU), 7+, 13+, 15+, 18+, dan Konten Terlarang (Prohibited Con-tent/PRC), serta menghasilkan penalaran (*reasoning*) tekstual yang objektif untuk mendukung hasil klasifikasi yang diberikan.

Pengembangan model dilakukan melalui dua tahapan utama yang terintegrasi secara bertahap, yaitu *Continued Pre-Training* (CPT) dan *Supervised Fine-Tuning* (SFT), dengan me-manfaatkan model dasar berukuran efisien, yaitu mistralai/Ministral-3-3B-Base-2512.

1. Tahap Continued Pre-Training (CPT): Bertujuan untuk memperkaya basis pengetahuan (*knowledge base*) model dasar dengan korpus teks spesifik domain PAD berbahasa Indonesia sebesar kurang lebih 214,8 juta token. Korpus ini mencakup dokumen hukum, regulasi perlindungan anak (seperti PP TUNAS), literatur psikologi anak, serta variasi teks media sosial yang sarat dengan istilah tidak baku (*slang*). Tahap ini penting untuk menggeser distribusi probabilitas

bahasa model agar memahami konteks sensitif dan istilah khas perlindungan anak sebelum dihadapkan pada instruksi spesifik.

2. Tahap Supervised Fine-Tuning (SFT): Dilakukan menggunakan dataset instruksi terkurasi sebanyak 13.406 sampel. Tahap ini melatih model untuk mengikuti templat instruksi secara disiplin menggunakan teknik *Assistant-Only Loss Masking*. Dengan metode ini, pembaruan bobot model hanya difokuskan pada token jawaban (*assistant response*) berupa label rating dan penjelasan, sehingga model mampu menghasilkan klasifikasi yang konsisten tanpa menghafal atau mengulang teks instruksi masukan.

Hasil akhir dari proyek ini berupa sebuah model inferensi teks berpresisi tinggi yang mampu melakukan klasifikasi rating usia secara interpretatif dan akurat. Model yang telah digabungkan (*merged*) ke dalam format 16-bit bfloat16 dirancang untuk siap dideploy ke dalam bentuk layanan Application Programming Interface (API) berbasis REST. Layanan ini disiapkan agar dapat diintegrasikan dengan mudah ke dalam arsitektur Agentic AI maupun purwarupa aplikasi *Minimum Viable Product* (MVP) guna mendukung otomatisasi moderasi konten pada Direktorat Jenderal Pengawasan Ruang Digital KOMDIGI.

1.3 Masalah atau Isu

Direktorat Jenderal Pengawasan Ruang Digital

Kementerian Komunikasi dan Digital (KOMDIGI) menghadapi tantangan yang semakin eskalatif dalam melakukan pengawasan dan moderasi terhadap arus informasi di media sosial. Salah satu isu paling krusial adalah tingginya risiko paparan konten negatif dan tidak layak konsumsi terhadap pengguna internet usia anak-anak dan remaja. Konten-konten tersebut meliputi promosi judi online secara terselubung menggunakan istilah populer (*slot, gacor, scatter*), perundungan siber (*cyberbul-lying*), objektifikasi seksual, narasi menyakiti diri sendiri (*self-harm*), hingga ujaran kebencian berbasis SARA.

Dalam menjalankan fungsi penyaringan, KOMDIGI menemui beberapa kendala opera-sional dan teknis yang signifikan:

1. Volume Data yang Masif: Jutaan unggahan teks diproduksi setiap hari di berbagai platform media sosial. Moderasi yang mengandalkan tenaga manusia (*human moderator*) secara manual sudah tidak lagi memadai dari segi kecepatan maupun skala jangkauan.
2. Keterbatasan Pendekatan Berbasis Kata Kunci (Keyword Filtering): Pendekatan konven-sional ini sangat kaku dan mudah dikelabui. Bahasa media sosial Indonesia bersifat dinamis, dipenuhi oleh variasi bahasa tidak baku, singkatan, metafora, serta istilah *slang* baru yang terus berkembang secara kontekstual.
3. Ketiadaan Transparansi dan Konteks (Black-Box AI): Sistem klasifikasi otomatis konvensional umumnya

hanya mampu mengeluarkan label keputusan (seperti "Aman" atau "Tidak Aman") tanpa disertai penjelasan logis. Ketiadaan penalaran (*reasoning*) ini menyulitkan verifikator manusia di KOMDIGI untuk mengaudit atau memahami dasar keputusan model sebelum melakukan tindakan pemblokiran atau pembatasan konten.

1.4 Tujuan Proyek

Tujuan dari pelaksanaan proyek ini dirumuskan sebagai berikut:

1. Mengembangkan Sistem Otomatisasi Moderasi: Membangun model kecerdasan buatan berbasis LLM yang mampu menyaring dan mengklasifikasikan teks media sosial berbahasa Indonesia ke dalam kategori rating usia secara cepat dan otomatis.
2. Standardisasi Klasifikasi Konten Ramah Anak: Menyelaraskan hasil prediksi model dengan standar klasifikasi nasional, termasuk regulasi dari Lembaga Sensor Film (LSF), Indonesia Game Rating System (IGRS), Peraturan Pemerintah tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Perlindungan Anak (PP TUNAS), dan Undang-Undang Perlindungan Anak.
3. Menghasilkan Penalaran yang Transparan (Explainable AI): Memastikan model tidak hanya memberikan label rating usia secara mekanis, melainkan juga menghasilkan teks penjelasan (*reasoning*) setebal 2–4 kalimat yang menguraikan konteks atau frasa spesifik yang melandasi keputusan tersebut.

4. Mengoptimalkan Efisiensi Model: Memanfaatkan model berukuran ringkas (Ministral-3-3B) namun memiliki kapabilitas penalaran tinggi melalui optimalisasi teknik CPT dan SFT, sehingga dapat menghemat konsumsi memori dan biaya komputasi saat dideploy untuk skala operasional.

1.5 Studi Kasus Bisnis

Proyek ini diterapkan sebagai solusi teknologi untuk menyelesaikan dua kebutuhan bisnis utama dalam fungsi Perlindungan Ruang Digital di KOMDIGI:

1. Peningkatan Efisiensi Operasional Melalui Otomatisasi Penyaringan: Dengan mengintegrasikan API model inferensi ini ke dalam sistem pengawasan KOMDIGI, proses penyaringan teks yang mengandung unsur tidak aman bagi anak dapat berjalan secara *real-time* dan otomatis selama 24 jam. Hal ini memangkas beban kerja manual tim pengawas, sehingga mereka dapat berfokus pada penanganan kasus-kasus pelanggaran berat yang membutuhkan verifikasi tingkat lanjut.
2. Penyediaan Draf Keputusan Berbasis Data (Explainable Decision Support System): Kemampuan model dalam menghasilkan struktur output berupa label rating beserta penjelasannya berfungsi sebagai asisten cerdas bagi verifikator KOMDIGI. Ketika model mendeteksi konten berbahaya (misalnya memberi rating Konten Terlarang), model secara langsung menyajikan frasa bukti dan alasan logis di

balik penilaian tersebut. Hal ini membuat draf rujukan tindakan moderasi menjadi lebih objektif, transparan, serta akurat sebelum dieksekusi.

1.6 Ruang Lingkup Proyek

Batasan ruang lingkup dalam pelaksanaan proyek ini ditentukan sebagai berikut:

1. Batasan Format dan Bahasa Data: Analisis, penyusunan dataset, serta evaluasi model dibatasi secara ketat pada konten berbasis teks berbahasa Indonesia, termasuk variasi bahasa tidak baku (*slang*), singkatan, dan istilah khas media sosial Indonesia. Proyek ini tidak mencakup pemrosesan data multimedia seperti gambar, audio asli, maupun video.
2. Batasan Metodologi Teknis: Metodologi pengembangan dibatasi pada proses prapemros-esan dataset spesifik domain Perlindungan Anak di Ruang Digital (PAD), yang dilanjutkan dengan dua tahapan pelatihan menggunakan kerangka kerja Unsloth:
   - *Continued Pre-Training* (CPT) menggunakan korpus PAD dan umum sebesar kurang lebih 214,8 juta token pada presisi 16-bit bfloat16.
   - *Supervised Fine-Tuning* (SFT) menggunakan dataset instruksi sebanyak 13.406 baris dengan metode *Assistant-Only Loss Masking*.
3. Arsitektur Model Dasar: Model dasar (*base model*) yang digunakan didefinisikan secara khusus pada mistralai/Ministral-3-3B-Base-2512.
4. Keselarasan Aturan Pelabelan: Aturan penentuan label rating (Semua Umur, 7+, 13+, 15+, 18+, Konten

Terlarang) disesuaikan secara khusus dengan merujuk pada dokumen kebijakan perlindungan anak, panduan IGRS, LSF, dan korpus hukum positif yang berlaku di Indonesia.

5. Format Luaran Akhir (Deliverables): Luaran fisik dari proyek ini dibatasi pada unggahan bobot model hasil penggabungan (*merged weights*) 16-bit pada platform Hugging Face, dokumentasi teknis, serta deployment model menggunakan platform serverless Modal ke dalam bentuk layanan REST API siap pakai untuk kebutuhan integrasi sistem pihak ketiga.

BAB 2 TINJAUAN PUSTAKA

2.1 Large Language Models (LLM) dan Arsitektur Transformer

Large Language Models (LLM) merupakan arsitektur jaringan saraf dalam (*deep neural networks*) berbasis statistika bahasa yang dilatih menggunakan korpus teks berskala masif untuk mengenali, memprediksi, dan menghasilkan teks yang koheren secara kontekstual (Brown et al., 2020). Kemampuan utama LLM modern bertumpu pada kemampuannya untuk mempelajari representasi semantik tingkat tinggi (*high-dimensional semantic representations*) dan dependensi sintaksis jangka panjang dari data latih, sehingga mampu memahami konteks kalimat secara non-linear dan dinamis (Devlin et al., 2019).

Sebagian besar LLM generatif modern mengadopsi arsitektur *Transformer* yang pertama kali diperkenalkan oleh Vaswani et al. (2017). Berbeda dengan arsitektur pendahulunya seperti *Recurrent Neural Networks* (RNN) atau *Long Short-Term Memory* (LSTM) yang memproses token secara sekuensial dan rentan terhadap masalah *vanishing gradient* pada kalimat panjang, Transformer memproses seluruh token masukan secara paralel memanfaatkan mekanisme *Self-Attention* (Atensi Diri). Mekanisme ini memproyeksikan setiap token masukan ke dalam tiga vektor representasi ruang dimensi rendah, yaitu

$$L_{(CLM)}(\theta) = -(1/N)\Sigma_{(i=1)}^{N}\Sigma_{(t=1)}^{(T_i)} logP\left(x_{(i,t)}\middle|x_{(i,<t)};\theta\right) \quad (2.1)$$

Di mana representsi dimensi dari vek tor *Key* yang bertindak sebagai faktor skala penormalisasi untuk mencegah nilai dot-product tumbuh secara eksponensial yang dapat menyebabkan gradien *softmax* menjadi sangat kecil. Pembobotan ini dieksekusi secara simultan pada ruang representasi yang berbeda melalui *Multi-Head Attention*, memungkinkan model untuk secara bersamaan menangkap hubungan sintaksis dan semantik dari berbagai sudut pandang dimensi yang berbeda (Vaswani et al., 2017).

Berdasarkan variasi desain arsitekturnya, LLM generatif seperti keluarga model Qwen, Gemma, dan Mistral (termasuk varian Ministral) secara spesifik mengimplementasikan varian *Decoder-only* (Autoregresif) (Radford et al., 2019). Pada varian *Decoder-only*, mekanisme atensi dimodifikasi menggunakan *Causal Attention Masking* (Penepakan Atensi Kausal). Modifikasi ini memastikan bahwa dalam proses pelatihan pemodelan bahasa kausal (*Causal Language*

), prediksi untuk token berikutnya pada posisi hanya dapat dihitung berdasarkan probabilitas kondisional dari token-token sebelumnya pada posisi , seperti yang dirumuskan secara matematis berikut:

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \dots, w_{i-1}) \quad (2.2)$$

Melalui obyektif prapelatihan (*pre-training*) ini, model dasar

(*base model*) memperoleh pemahaman distribusi bahasa umum yang sangat kaya (Brown et al., 2020). Namun demikian, representasi bahasa umum tersebut bersifat terlalu generik dan belum tentu optimal ketika dihadapkan pada domain spesifik yang sangat terlokalisasi, seperti karakteristik teks media sosial Indonesia yang sarat akan bahasa tidak baku (*slang*), singkatan, serta batasan aturan hukum dalam domain Perlindungan Anak di Ruang Digital (PAD). Oleh karena itu, diperlukan teknik adaptasi domain tingkat lanjut guna menyelaraskan ruang representasi semantik model dasar dengan korpus spesifik target (Gururangan et al., 2020).

2.2 Adaptasi Domain Melalui Continued Pre-Training (CPT)

Meskipun model bahasa besar yang dilatih pada korpus umum (*general-purpose LLMs*) menunjukkan kapabilitas representasi linguistik yang luas, performanya sering kali mengalami degradasi ketika dihadapkan pada domain yang sangat terspesialisasi atau bersifat lokal (Gururangan et al., 2020). Hal ini disebabkan oleh adanya pergeseran distribusi data (*domain shift*) di mana istilah teknis, singkatan, struktur kalimat, serta nuansa kontekstual dari dokumen target belum pernah dijumpai atau sangat jarang terwakili dalam data prapelatihan asli model tersebut. Untuk mengatasi batasan tersebut, diterapkan metode *Continued Pre-Training* (CPT), yang juga dikenal dalam literatur akademis sebagai *Domain-Adaptive Pretraining* (DAPT) (Gururangan et al., 2020; Ke et al., 2023).

CPT adalah tahapan melanjutkan pelatihan tanpa pengawasan (*unsupervised training*) pada model dasar (*base model*) menggunakan korpus teks tidak berlabel yang spesifik terhadap domain target. Secara matematis, proses CPT menggunakan fungsi tujuan pemodelan bahasa kausal (*Causal Language Modeling*) yang sama dengan tahap prapelatihan awal, yaitu meminimalkan nilai *negative log-likelihood* (atau meminimalkan *Cross-Entropy Loss*) atas seluruh token dalam dokumen latih:

$$L_{(CLM)}(\theta) = -(1/N)\Sigma_{(i=1)}^{N}\Sigma_{(t=1)}^{(T_i)} logP\left(x_{(i,t)}\middle|x_{(i,<t)};\theta\right) \quad (2.3)$$

Di mana menyatakan jumlah total sampel dokumen dalam korpus, adalah panjang token dalam dokumen ke- i , dan θ merepresentasikan parameter model yang diperbarui selama proses pelatihan (Ke et al., 2023).

Dalam implementasi praktisnya, melakukan CPT dengan melakukan pembaruan pada seluruh parameter model (*full-parameter fine-tuning*) membutuhkan sumber daya komputasi yang sangat besar dan rentan memicu terjadinya *Catastrophic Forgetting* (lupa ingatan katastropik). Fenomena ini terjadi ketika model terlalu condong menyesuaikan diri dengan pengetahuan baru sehingga menghapus atau merusak bobot representasi bahasa umum yang telah dipelajari sebelumnya (Ke et al., 2023).

Sebagai solusi efisien, pendekatan *Parameter-Efficient Fine-Tuning* (PEFT) seperti *Low-Rank Adaptation* (LoRA) diterapkan. Namun, dalam konteks CPT, adaptasi domain tidak akan optimal jika hanya melatih modul

atensi dan feed-forward murni. Model juga harus diizinkan untuk menyesuaikan representasi kosakata baru (*vocabulary shift*). Oleh karena itu, modul *embedding* (embed_tokens) dan modul proyeksi keluaran logit (*Language Modeling Head* atau lm_head) wajib diikutsertakan ke dalam parameter latih LoRA. Untuk mencegah kerusakan tabel kosa kata yang sudah mapan akibat pembaruan gradien yang terlalu agresif, digunakan pendekatan *Differential Learning Rates* (Laju Pembelajaran Terpisah), di mana laju pembelajaran untuk lapisan *embedding* diatur secara signifikan lebih kecil (misalnya, berorde 10^{-6} atau sepuluh kali lebih kecil) dibandingkan laju pembelajaran adaptasi modul internal lainnya (berorde 10^{-5}) (Ke et al., 2023).

Dalam domain Perlindungan Anak di Ruang Digital (PAD) di Indonesia, proses CPT sangat krusial untuk menjembatani dua kutub bahasa yang sangat ekstrem:

1. Bahasa Regulasi dan Hukum Formal: Teks yang berkaitan dengan dasar hukum perlindungan anak, seperti dokumen Undang-Undang Perlindungan Anak, PP TUNAS, serta regulasi IGRS dan LSF yang sarat dengan terminologi hukum formal dan sanksi administratif.
2. Bahasa Informal Media Sosial: Teks yang diproduksi oleh anak-anak dan remaja di media sosial, yang dipenuhi istilah slang bahasa gaul, singkatan, serta bahasa sandi (*dogwhistle*) yang sering digunakan dalam kasus eksploitasi anak, perundungan siber,

dan promosi judi online terselubung.

Injeksi korpus berimbang yang mengombinasikan data spesifik PAD dan data pengetahuan umum berbahasa Indonesia (dengan rasio terancang seperti 70% data domain PAD dan 30% data umum) memungkinkan model dasar untuk menekan nilai ambang ketidakpastian informasi (*perplexity*) pada domain anak secara drastis, sekaligus memelihara kemampuan tata bahasa fundamentalnya agar tetap relevan dalam tugas inferensi generatif lanjutan (Gururangan et al., 2020).

2.3 Supervised Fine-Tuning (SFT) dan Penyelarasan Format

Meskipun tahapan *Continued Pre-Training* (CPT) berhasil memperkaya pemahaman semantik model dasar terhadap domain spesifik Perlindungan Anak di Ruang Digital (PAD), model pada fase tersebut masih berperan sebagai mesin pelengkap dokumen (*document completion engine*). Model cenderung melanjutkan teks masukan secara acak tanpa memahami struktur instruksi interaktif (Ouyang et al., 2022). Untuk mengubah perilaku model dasar tersebut agar mampu bertindak sebagai asisten interaktif yang dapat mengikuti instruksi klasifikasi secara disiplin dan menghasilkan format keluaran yang terstruktur, diperlukan tahapan *Supervised Fine-Tuning* (SFT) (Raffel et al., 2020).

Pada tahapan SFT, model dilatih menggunakan dataset yang dikonstruksi ke dalam pasangan instruksi jawaban (*prompt-response pairs*) yang terstruktur. Proses

penyelarasan ini memanfaatkan templat percakapan (*chat template*) terstandarisasi untuk membatasi wilayah interaksi antara pengguna (*user*) dan asisten (*assistant*). Proyek ini secara spesifik menerapkan templat instruksi dengan sintaksis sebagai berikut:

< s > [INST] {Sistem} + {User} [/INST] {Assistant} < /s > (2.4)

Penggunaan struktur khusus ini bertujuan untuk mengisolasi instruksi sistem dan klaim masukan dari jawaban yang dihasilkan oleh model, guna mencegah bias format selama proses inferensi generatif (Raffel et al., 2020).

Salah satu tantangan teknis utama dalam SFT klasifikasi interpretatif adalah kecen-derungan model untuk mengalami *overfitting* terhadap gaya penulisan instruksi atau teks klaim masukan jika loss dihitung pada keseluruhan sekuens (Ouyang et al., 2022). Untuk mengatasinya, proyek ini menerapkan teknik lanjut bernama *Assistant-Only Loss Masking* (Penepakan Loss Khusus Asisten). Secara matematis, dalam pemelajaran sekuens standar, loss dihitung dari awal hingga akhir urutan token. Namun, dengan teknik penepakan ini, nilai target untuk setiap token yang termasuk dalam wilayah *System Prompt* dan *User Input* diubah menjadi nilai indeks khusus, yaitu, yang bertindak sebagai *ignore index* dalam fungsi loss *Cross-Entropy* PyTorch (Ouyang et al., 2022).

Misalkan sebuah sekuens token memiliki panjang total L, di mana token dari indeks 1 hingga p

merepresentasikan instruksi pengguna (Prompt), dan token dari indeks p + 1 hingga L merepresentasikan jawaban asisten (Assistant Response). Nilai loss hanya dihitung pada wilayah jawaban asisten dengan rumusan matematika sebagai berikut:

$$\mathcal{L}sft(\theta) = -\frac{1}{L-P}\sum_{t=P+1}^{L} \log P\,(x_t \mid x_1, \ldots, x_{t-1}; \theta) \qquad (2.5)$$

Di mana untuk setiap indeks target yt berlaku aturan penepakan:

$$y_t = \{-100 \quad \text{untuk } 1 \le t \le p \; x_t \quad \text{untuk } p+1 \le t \le L \qquad (2.6)$$

Dengan menerapkan *Assistant-Only Loss Masking*, pembaruan bobot model secara matematis hanya dipicu oleh token-token jawaban yang dihasilkan oleh asisten (dimulai dari token Rating: hingga akhir penjelasan). Metode ini memaksa model untuk memusatkan seluruh kapasitas belajarnya pada tugas utama, yaitu menghasilkan label klasifikasi rating usia yang tepat dan menyusun kalimat penalaran (*Penjelasan*) yang logis berbasis konteks teks masukan, tanpa membuang kapasitas memori model untuk mempelajari ulang pola bahasa dari teks instruksi masukan (Ouyang et al., 2022).

2.4 Studi Komparatif Model Dasar (Base Models): Qwen, Gemma, dan Ministral

Dalam rekayasa sistem berbasis *Large Language Model* (LLM), pemilihan model dasar (*base model*) yang tepat memegang peranan krusial dalam menentukan

keberhasilan adaptasi domain, akurasi hasil klasifikasi, serta efisiensi penggunaan memori komputasi saat dideploy (Gururangan et al., 2020). Setiap keluarga model *open-weights* modern memiliki karakteristik arsitektur, ukuran kosakata (*vocabulary size*), serta mekanisme atensi khusus yang memengaruhi performanya pada tugas spesifik bahasa Indonesia terlokalisasi. Berikut adalah studi komparatif dari tiga keluarga model dasar populer yang diuji dalam proyek ini:

2.4.1 Qwen (Spesifik: Qwen3.5-4B-Base)

Keluarga model Qwen dikembangkan oleh Alibaba Cloud dengan fokus utama pada pemrosesan multibahasa (*multilingual*) skala masif serta kemampuan penalaran matematis dan pemrograman (Bai et al., 2023).

- Arsitektur dan Kosakata: Qwen3.5 menggunakan ukuran kosakata yang sangat besar, yaitu 248.044 token, yang dirancang untuk meminimalkan fragmentasi kata pada bahasa non-Inggris (termasuk bahasa Indonesia).
- Kapasitas Konteks: Mendukung panjang konteks asli yang sangat besar (hingga 128k hingga 256k token menggunakan interpolasi posisi).
- Batasan: Meskipun memiliki representasi semantik bahasa umum yang sangat kuat, pengujian empiris CPT pada Qwen3.5 menunjukkan tingkat *catastrophic forgetting* yang relatif tinggi (kenaikan *General Perplexity* mencapai +84.58%). Hal ini mengindikasikan bahwa bobot representasi multibahasa Qwen sangat

sensitif terhadap perubahan gradien jika lapisan *embedding* dilatih ulang pada domain bahasa tunggal (Bai et al., 2023).

2.4.2 Gemma (Spesifik: Gemma-3-4B-pt)

Gemma merupakan seri model ringan (*lightweight*) berkinerja tinggi yang dikembangkan oleh Google menggunakan basis penelitian dan teknologi pemrosesan yang serupa dengan model Gemini (Gemma Team, 2024).

- Arsitektur dan Kosakata: Menggunakan kosakata sebesar 256.000 token. Model ini memanfaatkan teknologi *Gated Linear Units* (GLU) pada lapisan MLP serta normalisasi RMSNorm untuk menjaga stabilitas pelatihan pada dimensi parameter yang lebih ringkas.
- Karakteristik Khusus: Seri Gemma-3 dirancang sebagai *Vision-Language Model* (VLM) secara *native*. Implikasinya pada CPT teks murni adalah kompleksitas tambahan pada tahap inisialisasi model, di mana modul visi wajib dinonaktiGan secara eksplisit, serta penggunaan kelas pemuatan khusus (FastModel) agar tidak membuang kapasitas VRAM untuk memproses parameter gambar yang tidak relevan (Gemma Team, 2024).

2.4.3 Mistral / Ministral (Spesifik: Ministral-3-3B-Base-2512)

Keluarga model Mistral dikenal di komunitas AI karena efisiensinya yang tinggi dalam rasio parameter

terhadap performa penalaran (Jiang et al., 2023). Seri *Ministral* merupakan varian terbaru yang dirancang khusus untuk komputasi di tingkat *edge* dan perangkat dengan kapasitas memori terbatas (Mistral AI, 2024).

- Arsitektur dan Atensi: Ministral-3-3B memiliki parameter total sekitar 3,85 miliar. Model ini menerapkan metode *Grouped-Query Attention* (GQA) untuk mempercepat proses pembu-atan kunci-nilai (*KV Cache*) serta *Sliding Window Attention* (SWA) untuk menekan konsumsi memori secara linear pada konteks panjang (Jiang et al., 2023).
- Ukuran Kosakata: Menggunakan ukuran kosakata moderat sebesar 131.072 token. Walaupun lebih kecil dari Qwen dan Gemma, ukuran ini terbukti efisien dalam membatasi
  konsumsi memori pada lapisan proyeksi luar (lm_head).
- Keunggulan Operasional: Pada pengujian CPT penuh, model Ministral-3-3B menawarkan
  throughput pelatihan tercepat dengan waktu eksekusi yang efisien (3,4 jam dibandingkan Gemma yang memakan waktu 18 jam) serta mencatatkan puncak penggunaan VRAM terendah (21.25 GB pada panjang sekuens 4096). Kapabilitas representasi penalarannya yang tinggi pada ukuran parameter 3B menjadikannya fondasi paling ideal untuk diintegrasikan ke dalam sistem REST API ramah sumber daya (Mistral AI, 2024).

Dengan membandingkan karakteristik arsitektur di atas, model Ministral-3-3B-Base-2512 dipilih sebagai model

dasar utama proyek PAD. Model ini memberikan keseimbangan terbaik antara kecepatan latih (*runtime*), konsumsi memori komputasi (*Peak VRAM*), dan stabilitas pemeliharaan ingatan bahasa umum pasca-CPT.

2.5 Kerangka Regulasi Perlindungan Anak di Indonesia (Domain Grounding)

Sistem moderasi konten otomatis berbasis kecerdasan buatan (*Artificial Intelligence*) tidak boleh dirancang berdasarkan intuisi subjektif pengembang. Penentuan batas-batas kelayakan konten wajib disandarkan pada asas hukum positif, undang-undang, serta kerangka regulasi sektoral yang berlaku di Republik Indonesia. Dalam domain Perlindungan Anak di Ruang Digital (PAD), klasifikasi enam rating usia (Semua Umur, 7+, 13+, 15+, 18+, dan Konten Terlarang) diturunkan secara saksama dari integrasi beberapa instrumen regulasi nasional berikut:

2.5.1 Dasar Hukum Perlindungan Anak dan Kewajiban Platform

- Undang-Undang RI Nomor 35 Tahun 2014 tentang Perubahan atas Undang-Undang Nomor 23 Tahun 2002 tentang Perlindungan Anak: Undang-undang ini menetapkan definisi anak sebagai setiap individu yang belum berusia 18 tahun (Pasal 1 angka 1). Negara, pemerintah, dan

masyarakat berkewajiban memberikan perlindungan khusus kepada anak dari berbagai bentuk eksploitasi, kekerasan siber, serta pengaruh buruk informasi digital.

- Peraturan Pemerintah tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Perlindungan Anak (PP TUNAS): PP ini memberikan mandat hukum yang tegas kepada Penyelenggara Sistem Elektronik (PSE) atau platform digital untuk menyelenggarakan sistem pengawasan mandiri (*self-moderation*) guna mencegah transmisi konten yang tidak layak dikonsumsi oleh anak di bawah umur, serta mewajibkan penyusunan klasifikasi rating kelayakan konten yang dapat diakses oleh anak.

2.5.2 Klasifikasi Rating Usia (Grounding pada LSF dan IGRS)

Penyusunan batas karakteristik konten untuk masing-masing tingkatan usia dalam proyek ini mengadopsi standar formal dari Lembaga Sensor Film (LSF) berdasarkan Undang-Undang Nomor 33 Tahun 2009 tentang Perfilman serta Peraturan Menteri Komunikasi dan Informatika Nomor 2 Tahun 2024 tentang Klasifikasi Gim (Indonesia Game Rating System
- IGRS):

- Semua Umur (SU): Ditujukan untuk konten yang

ramah keluarga, bersifat edukatif, memuat sapaan sehari-hari, serta bebas dari segala bentuk konflik sosial, kekerasan, pornografi, bahasa kasar, dan zat adiktif.

- 7+ (Anak-Anak): Mengadopsi kriteria IGRS 7+, mengizinkan adanya unsur hiburan fiksi ringan, petualangan fisik yang memiliki unsur tantangan, serta persaingan olahraga yang sehat tanpa deskripsi kekerasan fisik yang realistis atau bahasa kasar.
- 13+ (Remaja Awal): Mengacu pada batas usia remaja awal LSF dan IGRS, mengizinkan representasi konflik sosial ringan, dinamika hubungan pertemanan, curahan hati terkait masalah sekolah, serta misteri atau horor fiksi ringan tanpa visualisasi sadis atau eksploitasi romansa yang berlebihan.
- 15+ (Remaja Akhir): Mengizinkan pembahasan realitas sosial yang lebih matang, seperti diskusi politik informatif, isu hak asasi manusia, perang, isu global, serta laporan berita jurnalistik mengenai kriminalitas (misalnya kecelakaan atau pencurian) secara faktual tanpa adanya unsur glorifikasi tindak pidana.
- 18+ (Dewasa): Ditujukan bagi konsumsi usia dewasa hukum. Mengizinkan narasi dinamika romansa dewasa (seperti berciuman atau menginap bersama dalam konteks cerita), gaya

hidup malam (bar, kelab malam), serta konten sensitif seperti edukasi kesehatan mental terkait pencegahan menyakiti diri sendiri (*self-harm* atau *eating disorder*) dalam penyampaian yang aman dan tidak bersifat mengajak atau menginstruksikan.

2.5.3 Konten Terlarang (Prohibited Content / PRC)

Klasifikasi Konten Terlarang mencakup teks-teks media sosial yang secara mutlak melanggar hukum pidana serta ketentuan perundang-undangan di Indonesia. Penentuan klasifikasi ini didasarkan secara ketat pada Undang-Undang Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik (UU ITE):

- Pemberantasan Judi Online (Pasal 27 ayat 2): Melarang keras pendistribusian, transmisi, dan/atau pembuatan dapat diaksesnya informasi elektronis yang memiliki muatan perjudian. Dalam dataset proyek ini, teks yang mempromosikan judi online secara eksplisit menggunakan istilah seperti *slot*, *gacor*, *scatter*, *depo,* atau *maxwin* otomatis dikategorikan sebagai Konten Terlarang.
- Kesusilaan dan Pelecehan (Pasal 27 ayat 1): Melarang penyebaran informasi elektronik yang melanggar kesusilaan, objektifikasi seksual, pelecehan seksual berbasis teks, serta eksploitasi

seksual anak.

- Ancaman Kekerasan Fisik (Pasal 27B): Melarang informasi yang memuat ancaman kekerasan fisik, intimidasi, pembunuhan, pengeroyokan, atau pemerasan yang ditujukan langsung secara spesifik untuk menyerang psikologis individu.
- Ujaran Kebencian SARA (Pasal 28 ayat 2): Melarang penyebaran informasi yang ditujukan untuk menimbulkan rasa kebencian atau permusuhan individu dan/atau kelompok masyarakat tertentu berdasarkan atas suku, agama, ras, dan antargolongan (SARA).

Dengan mengintegrasikan kerangka regulasi formal ini, model klasifikasi PAD yang dikembangkan memiliki landasan hukum (*legal standing*) yang kokoh dalam mendukung penindakan konten berbahaya secara administratif oleh KOMDIGI.

2.6 Metrik Evaluasi Model

Untuk mengukur tingkat keberhasilan adaptasi domain pada fase *Continued Pre-Train-ing* (CPT), akurasi klasifikasi pada fase *Supervised Fine-Tuning* (SFT), serta kualitas penjelasan (*reasoning*) yang dihasilkan, evaluasi dilakukan menggunakan tiga klaster metrik pengukuran obyektif yang terstandardisasi secara matematis.

2.6.1 Metrik Intrinsik Pemodelan Bahasa (CPT Metrics)

Metrik ini digunakan selama fase CPT untuk

melacak seberapa baik model menyerap distribusi probabilitas bahasa dari domain baru, serta mendeteksi adanya *catastrophic forget-ting* pada domain umum.

- Perplexity (PPL): Merupakan eksponensial dari nilai rata-rata *Cross-Entropy Loss* per token. PPL mengukur tingkat ketidakpastian (*surprise*) model dalam memprediksi token berikutnya. Secara matematis didefinisikan sebagai:

$$\text{PPL} = \exp(\mathcal{L}\text{Cross-Entropy}) = \exp\left(-\frac{1}{N}\sum i = 1^{N} \log P\,(w_i \mid w_{<i})\right) \qquad (2.7)$$

  Semakin rendah nilai PPL, semakin tinggi akurasi model dalam memprediksi kata berikutnya dalam suatu sekuens teks (Jurafsky & Martin, 2023).

- Bits Per Character (BPC): Sangat krusial digunakan untuk melakukan evaluasi komparatif lintas arsitektur model (seperti Qwen, Gemma, dan Ministral). Karena ketiga model tersebut menggunakan *tokenizer* yang berbeda dengan ukuran kosakata yang bervariasi (misalnya kosakata Qwen sebesar 248k sedangkan Ministral 131k), nilai PPL murni tidak dapat diperbandingkan secara langsung secara adil. BPC mengatasi bias tokenisasi ini dengan menormalisasi nilai loss berbasis basis biner (logaritma basis 2) terhadap jumlah total karakter fisik teks, bukan jumlah token (Jurafsky & Martin, 2023). Rumusan BPC didefinisikan sebagai:

$$\text{BPC} = \frac{\mathcal{L}_{\text{Cross-Entropy}}}{\ln(2)} \times \frac{\text{Total Token}}{\text{Total Karakter}} \tag{2.8}$$

2.6.2 Metrik Kinerja Klasifikasi (SFT Classification Metrics)

Evaluasi hasil klasifikasi label rating usia didasarkan pada komponen tabel *Confusion Matrix*, yang terdiri atas *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Sokolova & Lapalme, 2009). Dari komponen tersebut diperoleh rumusan metrik evaluasi standar:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.11}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.12}$$

- Matthews Correlation Coefficient (MCC): Karena distribusi kelas dalam dataset SFT PAD mengalami ketidakseimbangan kelas (*class imbalance*) yang sangat

tinggi (misalnya kategori Konten Terlarang dan Semua Umur mendominasi secara dominan dibanding kelas 18+), penggunaan metrik *Accuracy* saja dapat memberikan hasil yang semu. Proyek ini mengintegrasikan MCC sebagai metrik utama evaluasi klasifikasi karena memperhitungkan seluruh proporsi keempat kuadran *Confusion Matrix* secara berimbang. Nilai MCC berkisar antara (prediksi sepenuhnya salah) hingga prediksi sempurna) (Powers, 2011). Rumusan MCC didefinisikan sebagai:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.13)$$

2.6.3 Metrik Kualitas Generasi Teks (Text Generation Metrics)

Digunakan untuk mengevaluasi kualitas dan keselarasan teks alasan (*Penjelasan*) yang dihasilkan oleh asisten terhadap teks acuan (*Ground Truth*).

- ROUGE-L: Merupakan varian metrik dari keluarga ROUGE yang mengukur kemiripan struktur kalimat antara teks hasil prediksi dengan referensi berdasarkan *Longest Common Subsequence* (LCS). Keunggulan ROUGE-L terletak pada kemampuannya mendeteksi kemiripan pola sintaksis secara berurutan tanpa mewajibkan kecocokan kata demi kata secara absolut, sehingga sangat sesuai untuk mengevaluasi

konsistensi struktur format Penjelasan (Rating dan Alasan) (Sokolova & Lapalme, 2009).

- BERTScore: Mengukur tingkat kemiripan semantik (*semantic similarity*) antara teks hasil generasi dan acuan dengan menghitung nilai *cosine similarity* dari representasi vektor kontekstual (*token embeddings*) yang dihasilkan oleh model Transformer terlatih. BERTScore penting untuk mengevaluasi teks penjelasan karena dua kalimat penjelasan dapat menggunakan variasi kata yang berbeda atau menyertakan padanan istilah *slang* yang tidak identik, namun secara makna tetap memiliki maksud dan interpretasi yang sama. BERTScore menghitung tingkat presisi (*BERT-Precision*), recall (*BERT-Recall*), dan nilai F1 (*BERT-F1*) secara dinamis pada level semantik terdalam (Jurafsky & Martin, 2023).

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

3.1 Analisis Pengguna

Pengguna akhir dari luaran teknologi yang dikembangkan dalam proyek Perlindungan Anak di Ruang Digital (PAD) ini dikategorikan menjadi dua kelompok utama berdasarkan tingkat interaksi dan pemanfaatannya terhadap sistem:

1. *Direct User* (Pengguna Langsung):
   - Profil: Tim Pengembang Aplikasi / Sistem Integrator (Tim Aplikasi MVP).
   - Peran: Bertindak sebagai pihak teknis yang mengintegrasikan model inferensi LLM ke dalam arsitektur aplikasi moderasi konten menggunakan panggilan API (*Application Programming Interface*). Mereka membutuhkan dokumentasi API yang jelas, latensi respons yang rendah, serta konsistensi format keluaran JSON agar proses parsing data di tingkat frontend berjalan stabil.
2. *End-User* (Pengguna Akhir):
   - Profil: Direktorat Jenderal Pengawasan Ruang Digital Kementerian Komunikasi dan Digital (KOMDIGI), khususnya tim analis dan verifikator konten ramah anak.
   - Peran: Bertindak sebagai operator yang melakukan pengawasan harian terhadap lalu

lintas konten media sosial. Mereka berinteraksi dengan antarmuka aplikasi MVP untuk melihat hasil klasifikasi otomatis rating usia serta draf penalaran (Penjelasan) yang diajukan oleh model AI sebelum melakukan verifikasi akhir atau tindakan penertiban.

3.2 User Story

Kebutuhan fungsional sistem dalam proyek ini dirumuskan melalui pendekatan *User Story* untuk menjamin keselarasan antara pengembangan teknis model dengan kebutuhan operasional di lapangan:

3.2.1 User Story 1: Perspektif Pengembang Aplikasi (Direct User)

Pernyataan: Sebagai pengembang aplikasi sistem moderasi, saya ingin mengakses layanan klasifikasi rating usia PAD melalui endpoint REST API yang stabil dan terstandarisasi, sehingga saya dapat mengintegrasikan kemampuan deteksi otomatis ini ke dalam purwarupa MVP KOMDIGI secara modular.

Kriteria Penerimaan (*Acceptance Criteria*):

- API menerima masukan teks mentah (string) melalui metode HTTP POST.
- API mengembalikan respons terstruktur dalam format JSON dengan kunci Rating (berisi salah satu dari 6 kelas) dan Penjelasan (teks penalaran

2–4 kalimat).
- Waktu respons (*latency*) inferensi model di bawah ambang batas yang disepakati per permintaan pada kapasitas infrastruktur serverless GPU.

3.2.2 User Story 2: Perspektif Verifikator Konten KOMDIGI (End-User)

Pernyataan: Sebagai petugas pengawas konten digital di KOMDIGI, saya ingin sistem otomatis tidak hanya memberikan label rating usia, tetapi juga memberikan alasan logis dan kutipan frasa sensitif dari teks masukan, sehingga saya dapat mengambil keputusan verifikasi secara cepat, akurat, dan dapat dipertanggungjawabkan.

Kriteria Penerimaan:

- Sistem mampu mendeteksi istilah tidak baku (*slang*) dan singkatan secara kontekstual untuk menghindari kegagalan deteksi.
- Penjelasan yang dihasilkan oleh AI wajib merujuk secara logis pada karakteristik rating usia (misalnya menyebutkan adanya indikasi judi online untuk rating Konten Terlarang).
- Tingkat akurasi klasifikasi model pada data uji minimal mencapai 90% dengan nilai korelasi (MCC) yang stabil untuk menghindari *false positives*.

3.3 Perancangan Dataset

Siklus pengembangan model klasifikasi PAD ini dirancang melalui dua tahapan penyusunan dataset yang terpisah dan terarah, yaitu Dataset CPT untuk adaptasi pengetahuan dasar, dan Dataset SFT untuk penyelarasan instruksi.

3.3.1 Dataset Continued Pre-Training (CPT)

Dataset CPT dirancang murni sebagai korpus teks tak berlabel (*unlabeled raw text*) berbahasa Indonesia dengan fokus penguasaan kosa kata, istilah siber, regulasi perlindungan anak, serta variasi bahasa media sosial.

Komposisi dataset CPT disusun dengan menyeimbangkan muatan domain PAD (termasuk bahasa *slang*) sebesar kurang lebih 70% dan muatan domain pengetahuan umum berbahasa Indonesia (*general pool*) sebesar kurang lebih 30%. Hal ini dirancang secara saksama untuk meminimalkan risiko *catastrophic forgetting*. Dataset CPT ini memiliki total 491.447 dokumen valid dengan total akumulasi 214.947.565 token (menggunakan tokenizer Qwen/ Ministral).

Tabel 3. 1 Rincian Komponen Dataset CPT

Kategori Data	Sub-Komponen Dataset	Jumlah Baris	Jumlah Token	Persentase
Domain PAD	Slang-Injected Related	94.433	33.876.754	15.76%
	Gold Normal	218.313	97.762.779	45.51%

	Related			
	Prioritized Top-Up Related	39.055	18.698.398	8.70%
	9-PDF Legal Extraction	57	110.810	0.05%
Domain Umum	General Pool (Wiki/Berita)	139.589	64.498.824	30.01%
TOTAL	Dataset Gabungan CPT	491.447	214.947.565	100.00%

Catatan Teknis: Data ekstraksi dari 9 PDF dokumen hukum nasional (PP TUNAS, UU Perlindungan Anak, dll.) yang berjumlah 57 dokumen dengan 110.810 token secara sengaja dipertahankan 100% pada pembagian subset eksperimen 20% karena memiliki nilai bobot informasi regulasi yang sangat tinggi.

3.3.2 Dataset Supervised Fine-Tuning (SFT)

Dataset SFT dirancang dalam format percakapan interaktif (*ChatML*) untuk melatih model agar mematuhi format klasifikasi instruksional. Dataset SFT final berjumlah 13.406 baris data bersih yang dikonstruksi dari penggabungan data hasil generasi model (10k baris) dan hasil scraping data riil media sosial (3k baris).

Struktur data SFT disusun ke dalam skema JSON Lines (`.jsonl`) dengan format percakapan tiga lapis:

```
{
  "messages": [
    {
      "role": "system",
      "content": "Anda adalah sistem moderasi konten... [Instruksi Aturan Rating PAD]"
    },
    {
      "role": "user",
      "content": "Klasifikasikan teks berikut: [Teks Konten Media Sosial]"
    },
    {
      "role": "assistant",
      "content": "Rating: [Label Rating]\nPenjelasan: [Kalimat penalaran logis pendukung]"
    }
  ]
}
```

Tabel 3. 2 Distribusi Label Rating pada Dataset SFT

Label Rating Usia	Deskripsi Kelayakan Konten	Jumlah Baris	Persentase
Konten Terlarang	Melanggar hukum siber (Judi, kekerasan, SARA)	4.497	33.54%
Semua Umur	Ramah keluarga, konten harian positif	3.252	24.26%
15+	Diskusi sosial-politik, kriminalitas tanpa glorifikasi	1.826	13.62%
13+	Curahan hati remaja, konflik pertemanan	1.458	10.88%
7+	Hiburan ringan, tantangan olahraga	1.239	9.24%

	non-kekerasan		
18+	Romansa dewasa legal, edukasi pencegahan self-harm	1.134	8.46%
TOTAL	Dataset Instruksi SFT	13.406	100.00%

Penyelarasan Khusus: Konten yang terkait dengan isu menyakiti diri sendiri (*Self-Harm*) secara sengaja dipetakan dari label awal Konten Terlarang ke label 18+. Hal ini bertujuan agar model mampu menyajikan respons informatif dan preventif tanpa bersifat menghakimi secara kaku, namun tetap membatasi aksesibilitasnya dari pengguna di bawah umur.

3.4 Perancangan Pipeline Continued Pre-Training (CPT)

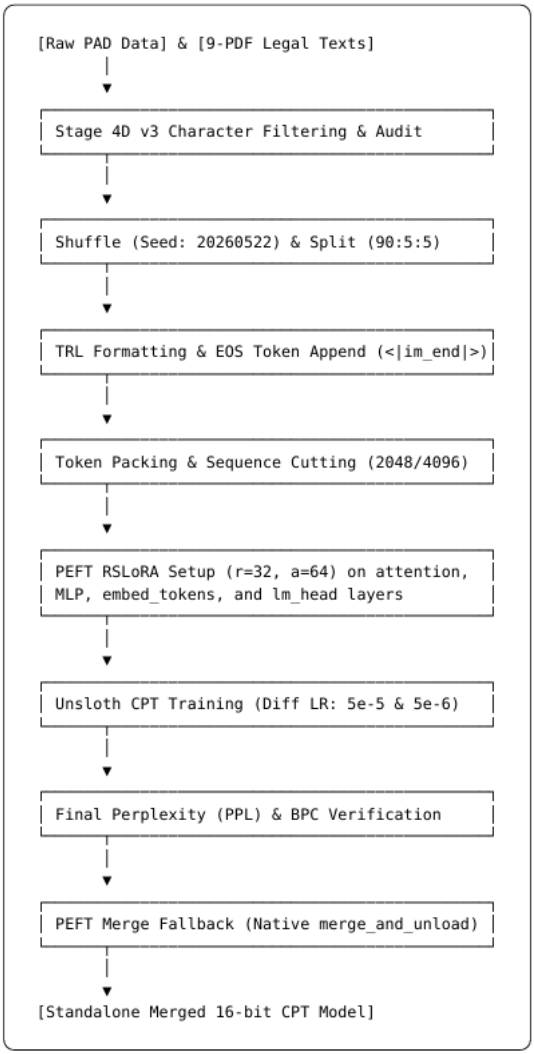
Pipeline *Continued Pre-Training* (CPT) dirancang sebagai alur kerja sistematis untuk mengadaptasi model dasar Ministral-3-3B-Base-2512 ke dalam domain Perlindungan Anak di Ruang Digital (PAD) secara efisien dan aman dari masalah kebocoran gradien maupun *catastrophic forgetting*.

Alur pemrosesan prapelatihan lanjutan diuraikan ke dalam delapan tahapan utama:

1. Ingetsi Data dan Karakterisasi Karakter: Memuat file dataset utama dan data regulasi dari penyimpanan lokal. Melakukan validasi karakter sesuai whitelist.

2. Pengacakan dan Pembagian Stratifikasi: Menerapkan pengacakan deterministik menggunakan seed tetap 20260522 pada dataset gabungan. Membagi dataset menjadi porsi prapelatihan (90%), porsi validasi (5%), dan porsi pengujian (5%).
3. Penyelarasan Format TRL: Mengubah data berformat title dan content ke dalam format teks tunggal dengan menambahkan token pembatas dokumen (*EOS Token*) secara eksplisit.
4. Pengepakan Token (Token Packing): Memotong dan menggabungkan token-token dokumen ke dalam sekuens tensor berukuran konstan guna menghindari pemborosan ruang memori.
5. Inisialisasi Model & Konfigurasi PEFT: Memuat bobot asli model dasar dalam presisi 16-bit murni. Mengonfigurasi lapisan adaptasi rendah (*Rank-Stabilized LoRA*) pada modul proyeksi atensi, feed-forward, embedding, dan kepala bahasa.
6. Penerapan Laju Pembelajaran Terpisah: Mengatur laju pembelajaran yang terpisah selama fase pembaruan bobot untuk memelihara stabilitas kamus kosakata bawaan model dasar.
7. Pelatihan dan Pemantauan Loss: Menjalankan UnslothTrainer selama 1 epoch penuh dengan pelacakan kurva penurunan loss secara *real-time*.
8. Penyatuan Model dan Penyelamatan Kegagalan:

Melakukan proses penggabungan (*merg-ing*) bobot latih ke dalam model dasar sebelum mengekspor model akhir.



Gambar 3. 1 Visualisasi Diagram Alur CPT

3.5 Perancangan Pipeline Supervised Fine-Tuning

(SFT)

Pipeline *Supervised Fine-Tuning* (SFT) dirancang untuk melatih model hasil CPT agar mampu melakukan klasifikasi rating usia PAD secara interpretatif, mematuhi format instruksi, serta memiliki kemampuan penulisan penalaran alasan yang logis.

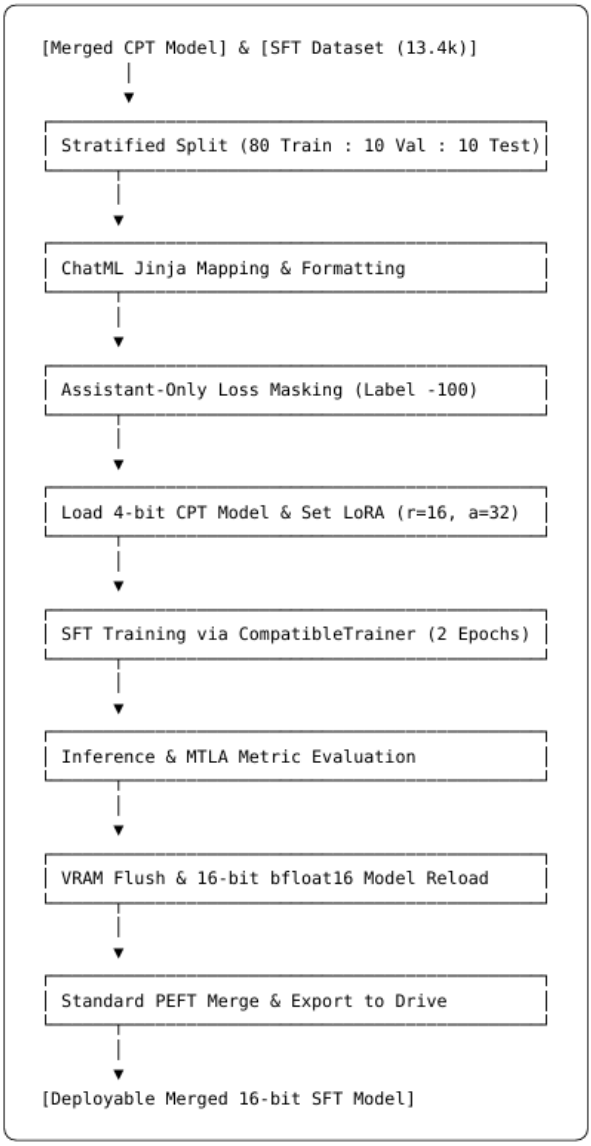
Alur kerja penyelarasan instruksi SFT diuraikan ke dalam delapan tahapan terstruktur:

1. Pemuatan Dataset SFT: Memuat file berformat JSONL yang berisi 13.406 baris data percakapan.
2. Pembagian Data Terstratifikasi: Membagi dataset menjadi data latih (80%), data validasi (10%), dan data uji (10%) dengan tetap menjaga keseimbangan proporsi distribusi kelas.
3. Penyusunan Chat Template Jinja: Memasukkan teks sistem prompt standardisasi dan teks klaim masukan ke dalam struktur template chat khusus.
4. Assistant-Only Loss Masking: Mengonversi teks chat menjadi token ID dan memetakan token instruksi ke nilai label khusus, memastikan akumulasi nilai kerugian hanya dihitung dari jawaban asisten.
5. Inisialisasi Model 4-bit dan LoRA SFT: Memuat bobot model hasil CPT dalam presisi 4-bit untuk menghemat memori GPU, serta memasang adapter LoRA pada seluruh lapisan linier model.

6. Pelatihan SFT: Menjalankan pelatihan selama 2 epoch dengan laju pembelajaran konstan yang dikendalikan oleh *cosine scheduler*.
7. Inference & Evaluasi MTLA: Melakukan proses pengujian pada data uji. Sistem menghitung tingkat kepercayaan model menggunakan rumusan *Multi-Token Logit Averaging* (MTLA):

$$\text{Confidence} = \frac{1}{1 + \exp(-(\text{avg\_lp} + 2.5) \times 1.5))} \tag{2.14}$$

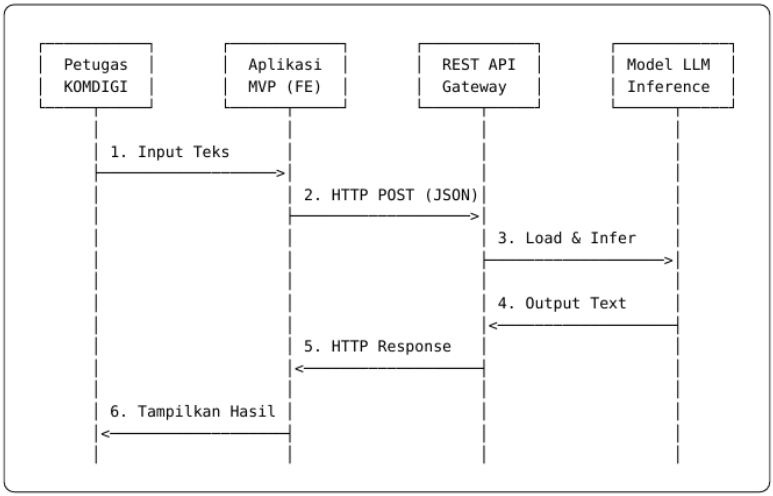
8. Penyatuan Akhir bfloat16: Menghapus model 4-bit dari memori, memuat ulang model dasar 16-bit, menyatukan adapter SFT terbaik, dan mengekspor model akhir yang siap dideploy.



Gambar 3. 2 Visualisasi Diagram Alur SFT

3.6 Arsitektur Integrasi Sistem

Model hasil akhir SFT dideploy menggunakan platform komputasi serverless Modal dengan antarmuka FastAPI. Penempatan ini memisahkan secara modular antara tugas pemrosesan kecerdasan buatan dengan tugas penyajian antarmuka pengguna.



Gambar 3. 3 Diagram Urutan Integrasi Sistem *(Sequence Diagram)*

Struktur skema permintaan (HTTP Request Payload) yang dikirim ke REST API adalah sebagai berikut:

```
{
  "text": "Info rujukan slot tergacor hari ini gaes, depo minimal sepuluh ribu rupiah
langsung auto maxwin tanpa ribet klik link di bio!"
}
```

Struktur skema tanggapan (HTTP Response Payload) yang dikembalikan oleh REST API adalah sebagai berikut:

```
{
  "status": "success",
  "data": {
    "rating": "Konten Terlarang",
    "penjelasan": "Teks tersebut secara eksplisit memuat promosi perjudian online dengan menggunakan kata kunci sensitif 'slot tergacor', 'depo', 'maxwin', dan ajakan mengklik 'link di bio'. Konten ini melanggar ketentuan Pasal 27 ayat 2 UU ITE mengenai larangan penyebaran muatan perjudian."
  },
  "meta": {
    "model_name": "Ministral-3-3B-SFT-PAD-Final",
    "inference_time_seconds": 0.48
  }
}
```


3.7 Spesifikasi Kebutuhan Sistem

Implementasi dan operasional sistem moderasi konten ini membutuhkan spesifikasi lingkungan perangkat keras dan perangkat lunak yang mumpuni agar proses pelatihan berjalan stabil dan proses inferensi API memiliki latensi rendah.

Tabel 3. 3 Kebutuhan Spesifikasi Perangkat Keras

Komponen Hardware	Spesifikasi Pelatihan (Training)	Spesifikasi Inferensi (Deploy-ment)
Arsitektur GPU	NVIDIA A100-SXM4 / RTX PRO 6000	Serverless GPU (NVIDIA T4 / L4)
Kapasitas VRAM	Minimal 40 GB (Rekomendasi 80 GB)	Minimal 16 GB

Memori Sistem (RAM)	Minimal 51 GB (High-RAM in-stance)	Minimal 16 GB
Penyimpanan	Minimal 100 GB SSD	Minimal 20 GB

Tabel 3. 4 Kebutuhan Spesifikasi Perangkat Lunak

Jenis Software / Pustaka	Versi Minimal	Peranan dalam Sistem
Sistem Operasi	Linux Ubuntu 22.04 LTS	Sistem operasi dasar lingkungan vir-tual
CUDA Driver	Versi 12.8	Akselerator komputasi grafis GPU
Python Runtime	Versi 3.10.x	Bahasa pemrograman utama
PyTorch Framework	Versi 2.10.0+cu128	Kerangka kerja pemelajaran dalam
Unsloth Utility	Versi 2026.5.7	Pustaka akselerasi optimasi memori LoRA
Transformers (HF)	Versi 5.5.0	Pustaka utama pemuatan model & to-kenizer
FastAPI	Versi 0.110.x	Kerangka penyajian endpoint REST API
Modal Client	Versi 0.62.x	Platform orkestrasi

BAB 4 IMPLEMENTASI SISTEM

4.1 Lingkungan Pengembangan Aktual

Implementasi sistem, pelatihan model (*training*), serta penyajian layanan inferensi (*deployment*) dijalankan pada lingkungan komputasi awan (*cloud computing*) berspesifikasi tinggi. Penggunaan perangkat keras GPU disesuaikan secara khusus berdasarkan karakteristik beban kerja masing-masing fase pelatihan:

1. Fase Continued Pre-Training (CPT): Beban kerja CPT memproses korpus teks masif berukuran lebih dari 214 juta token pada presisi 16-bit (*bfloat16*). Fase ini dieksekusi meng-gunakan server berakselerator NVIDIA RTX PRO 6000 Blackwell Server Edition dengan kapasitas Video RAM (VRAM) efektif sebesar 94.97 GB guna memastikan ketersediaan memori untuk pengepakan token panjang (*token packing*) tanpa risiko *Out-of-Memory* (OOM).
2. Fase Supervised Fine-Tuning (SFT): Beban kerja SFT memproses dataset instruksi interaktif sebanyak 13.406 baris data menggunakan kuantisasi 4-bit. Fase ini dieksekusi menggunakan unit NVIDIA A100-SXM4-40GB dengan total memori sistem sebesar 42.4 GB.

Sistem berjalan di atas sistem operasi virtual Linux (Ubuntu 22.04 LTS) dengan konfigurasi pustaka perangkat lunak utama sebagai berikut:

Tabel 5: Konfigurasi Pustaka Lingkungan Pengembangan

Tabel 4. 1 Konfigurasi Pustaka Lingkungan Pengembangan

Nama Pustaka / Dependensi	Versi Aktual yang Digunakan	Fungsi dalam Lingkungan Kerja
Python Interpreter	Versi 3.10.12	Penerjemah runtime bahasa pemrograman
CUDA Toolkit	Versi 12.8	Antarmuka pemrograman aplikasi akselerasi GPU
PyTorch Framework	Versi 2.10.0+cu128	Pustaka dasar komputasi sensor dan jaringan saraf
Unsloth Engine	Versi 2026.5.7	Modul akselerasi efisiensi memori LoRA & CPT
Transformers (HF)	Versi 5.5.0	Pustaka pemuatan model, konfigurasi, dan tokenizer
TRL Framework	Versi 0.22.2	Pustaka pembantu pelatihan SFT dan format ChatML
FastAPI	Versi 0.110.1	Kerangka kerja pembuatan layanan RESTAPI gateway

4.2 Realisasi Pelatihan Continued Pre-Training (CPT)

Tahapan prapemrosesan data (data preprocessing)

diimplementasikan secara terpro gram untuk menjamin kebersihan korpus data sebelum dimasukkan ke dalam jaringan saraf model.

4.2.1 Implementasi Prapemrosesan Dataset CPT

Prapemrosesan korpus CPT diimplementasikan menggunakan beberapa tahapan pen yaringan deterministik secara bertahap pada penyimpanan lokal sebelum digabungkan:

1. Penyaringan Karakter Whitelist (Stage 4D v3): Menghapus seluruh baris dokumen yang mengandung karakter di luar daftar karakter aman (whitelist). Karakter yang diper bolehkan hanya terbatas pada huruf, angka, spasi, serta tanda baca dasar. Karakter seperti persen, titik koma, serta fragmen HTML dibersihkan untuk memastikan kebersihan sin taksis korpus.
2. Pembersihan Boilerplate dan Preamble: Menerapkan aturan pembersihan prefiks sumber berita secara otomatis menggunakan ekspresi reguler (regex) terarah, serta menghapus baris dokumen yang mengandung frasa boilerplate berulang seperti “baca juga”, “klik di sini”, atau “simak video”.
3. Prioritisasi Penyatuan Data Top-Up: Penggabungan data pelengkap (top-up) dari general pool ke domain PAD dilakukan secara terarah berdasarkan tingkat kualitas informasi semantik. Dokumen dari kategori Tier A dimasukkan terlebih dahulu, diikuti oleh Tier B, dan porsi yang dibutuhkan dari Tier C guna memenuhi target komposisi token CPT.

4.2.2 Implementasi Prapemrosesan Dataset SFT

Prapemrosesan dataset SFT bertujuan untuk menggabungkan dataset hasil generasi sistem dan hasil scraping media sosial, menyelaraskan aturan rating usia berdasarkan regulasi hukum positif, membersihkan atribut yang tidak relevan, serta mengonversinya ke format percakapan terstruktur.

Berikut adalah naskah kode program Python aktual yang diimplementasikan untuk mengeksekusi prapemrosesan SFT secara otomatis pada Google Colab:

```
import os
import json
import re
import random
from collections import Counter

# ============================================================
# 1. KONFIGURASI PATH & FILE INPUT
# ============================================================
INPUT_DIR = "/content"
OUTPUT_DIR = "/content"
os.makedirs(OUTPUT_DIR, exist_ok=True)

# Dataset SFT Gabungan: Generasi (~10k data) dan Scraping (~3k data)
```

```
FILE_GENERATE = ["DATASET_SFT-PAD_RATING.jsonl"]
FILE_SCRAPING = ["dataset_sft_finalfull.jsonl"]
ALL_FILES = FILE_GENERATE + FILE_SCRAPING

# ================================================================
# 2. DEFINISI SYSTEM PROMPT FINAL (GROUNDED REGULASI PAD)
# ================================================================
SYSTEM_SFT_FINAL = """Anda adalah sistem moderasi konten media sosial Indonesia. Tugas
Anda adalah mengklasifikasikan caption/teks media sosial ke dalam rating usia yang tepat
dan berikan penjelasannya.

DEFINISI RATING:
- Semua Umur: Konten positif, aman, dan ramah keluarga. Tidak ada unsur negatif apapun.
- 7+: Hiburan dan fiksi ringan untuk anak SD. Boleh ada persaingan atau tantangan ringan,
TANPA kekerasan realistis atau bahasa kasar.
- 13+: Konten remaja awal. Boleh ada konflik sosial, curahan hati, misteri fiksi. TANPA
unsur dewasa atau kekerasan grafis.
- 15+: Konten remaja akhir. Boleh ada diskusi politik ringan, isu sosial, berita faktual,
kriminalitas tanpa glorifikasi.
- 18+: Konten dewasa yang LEGAL. Mencakup: alkohol, rokok, romansa dewasa (tanpa
eksplisit), kehidupan malam.
- Konten Terlarang: HANYA untuk konten yang secara EKSPLISIT mengandung: (1) ancaman
kekerasan fisik langsung, (2) pelecehan/objektifikasi seksual, (3) promosi judi online
dengan kata kunci slot/gacor/scatter, (4) ujaran kebencian SARA secara langsung.

FORMAT OUTPUT:
Rating: [Semua Umur / 7+ / 13+ / 15+ / 18+ / Konten Terlarang]
Penjelasan: [Penjelasan Logis 2-4 kalimat, sebutkan frasa spesifik sebagai penguat
penjelasan jika ada atau memungkinkan]"""

merged_dataset = []

print("=" * 65)
print("MEMULAI PROSES MERGING, CLEANING & MAPPING DATASET SFT")
print("=" * 65)

# Loop pemrosesan setiap file sumber
for filename in ALL_FILES:
    filepath = os.path.join(INPUT_DIR, filename)

    if not os.path.exists(filepath):
        print(f"[SKIP] File tidak ditemukan: {filename}")
        continue

    count_per_file = 0
    with open(filepath, 'r', encoding='utf-8') as f:
        for line in f:
            line = line.strip()
            if not line:
                continue

            try:
                data = json.loads(line)
                messages = data.get("messages", [])

                if len(messages) < 3:
                    continue

                # 1. Penyelerasan System Prompt Utama (Index 0)
```

```
                messages[0]["content"] = SYSTEM_SFT_FINAL

                # 2. Prapemrosesan Konten Asisten (Index 2)
                ast_content = messages[2]["content"]

                # Deteksi klasifikasi khusus kategori Self-Harm (Menyakiti Diri Sendiri)
                is_self_harm = (
                    "Self-Harm" in filename
                    or "self_harm" in filename.lower()
                    or "Kategori: Self Harm" in ast_content
                    or "Kategori: Self-Harm" in ast_content
                )

                # ============================================================
                # A. PENYELARASAN LABEL RATING DAN PENJELASAN (REGEX)
                # ============================================================
                if is_self_harm:
                    # Penyelarasan khusus: Kasus Self-Harm diubah dari PRC menjadi 18+
                    ast_content = re.sub(r'\bPRC\b', '18+', ast_content,
flags=re.IGNORECASE)

                # Penyelarasan Global: PRC dipetakan ke label Konten Terlarang
                ast_content = re.sub(r'\bPRC\b', 'Konten Terlarang', ast_content,
flags=re.IGNORECASE)

                # Penyelarasan Global: SU dipetakan ke label Semua Umur
                ast_content = re.sub(r'\bSU\b', 'Semua Umur', ast_content,
flags=re.IGNORECASE)

                # ============================================================
                # B. PEMBERSIHAN ATRIBUT TAMBAHAN & STANDARDISASI FORMAT
                # ============================================================
                # Hapus baris metadata "Kategori: ..." yang tidak relevan dengan format
akhir
                ast_content = re.sub(r'Kategori:\s*.*\n?', '', ast_content,
flags=re.IGNORECASE)

                # Standardisasi prefiks alasan / reasoning menjadi kunci "Penjelasan:"
                ast_content = re.sub(r'^(Reasoning|Alasan)\s*:', 'Penjelasan:',
ast_content, flags=re.MULTILINE|re.IGNORECASE)

                # Simpan kembali konten asisten yang telah dibersihkan
                messages[2]["content"] = ast_content.strip()

                merged_dataset.append({"messages": messages})
                count_per_file += 1

            except json.JSONDecodeError:
                pass

    print(f"✓ {filename} : {count_per_file} data bersih dan ter-mapping.")

print("=" * 65)
print(f"TOTAL GABUNGAN DATASET SFT: {len(merged_dataset)} baris")

# ============================================================
# 3. PENGACAKAN (SHUFFLING) & EXPORT AKHIR
# ============================================================
print("Mengacak urutan dataset (shuffling)...")
```

```
random.seed(42)  # Penguncian seed untuk replikasi hasil pengacakan
random.shuffle(merged_dataset)
print("✓ Dataset berhasil diacak.")

# Menyimpan hasil akhir ke dalam format JSON Lines
output_filename = "READY_UPDATE_DATASET_SFT-PAD_FINAL.jsonl"
output_path = os.path.join(OUTPUT_DIR, output_filename)

with open(output_path, "w", encoding="utf-8") as f:
    for entry in merged_dataset:
        f.write(json.dumps(entry, ensure_ascii=False) + "\n")

print(f"✓ File Master SFT berhasil disimpan di: {output_path}")
```

Hasil dari pengeksekusian skrip di atas berhasil menggabungkan secara bersih 10.213 baris data dari file JSON awal dan 3.193 baris data dari file hasil scraping, menghasilkan 13.406 baris data master SFT yang siap digunakan pada tahap pelatihan jaringan saraf.

4.3 Realisasi Pelatihan Continued Pre-Training (CPT)

Pelatihan *Continued Pre-Training* (CPT) direalisasikan untuk mengadaptasikan model dasar `mistralai/Ministral-3-3B-Base-2512` ke dalam pemahaman domain hukum dan bahasa slang Perlindungan Anak di Ruang Digital (PAD) berbahasa Indonesia. Pelatihan diimplementasikan menggunakan presisi 16-bit murni (*bfloat16*) untuk memaksimalkan akurasi representasi bobot parameter baru.

4.3.1 Naskah Kode Implementasi Pelatihan CPT

Berikut adalah naskah kode program Python yang diimplementasikan untuk melakukan pemuatan model dasar, konfigurasi modul RSLoRA komprehensif, dan inisialisasi latih menggunakan Unsloth:

```python
import torch
from unsloth import FastLanguageModel
from transformers import TrainingArguments
from trl import SFTTrainer

# ================================================================
# 1. KONFIGURASI MODEL DASAR & RESOLUSI KONTEKS
# ================================================================
model_id = "mistralai/Ministral-3-3B-Base-2512"
max_seq_length = 2048  # Dioptimalkan dari 4096 untuk efisiensi VRAM

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name=model_id,
    max_seq_length=max_seq_length,
    load_in_4bit=False,  # Menggunakan bfloat16 murni untuk presisi CPT
    dtype=torch.bfloat16,
    device_map="auto"
)
```

```python
# Menetapkan pad_token ke unk_token untuk menghindari masking pada EOS
tokenizer.pad_token = tokenizer.unk_token
model.config.pad_token_id = tokenizer.unk_token_id

# ================================================================
# 2. INJEKSI ADAPTER RANK-STABILIZED LORA (RSLoRA)
# ================================================================
model = FastLanguageModel.get_peft_model(
    model,
    r=32,
    lora_alpha=64,
    lora_dropout=0.0,
    target_modules=[
        "q_proj", "k_proj", "v_proj", "o_proj",
        "gate_proj", "up_proj", "down_proj",
        "embed_tokens", "lm_head"  # Wajib melatih layer kosakata untuk CPT
    ],
    bias="none",
    use_gradient_checkpointing="unsloth",
    random_state=20260522,
    use_rslora=True,  # Menggunakan teknik penyeimbang rank
)

# Menerapkan Differential Learning Rates untuk mengamankan representasi embedding
model.get_input_embeddings().weight.requires_grad = True
model.get_output_embeddings().weight.requires_grad = True

print("✓ Model dasar dan adapter RSLoRA CPT berhasil dikonfigurasi.")
```

4.3.2 Parameter dan Hiperparameter Pelatihan CPT

Konfigurasi parameter pelatihan CPT dikunci dalam spesifikasi matematis sebagai berikut

Tabel 4. 2 Konfigurasi Parameter Pelatihan CPT

Parameter Pelatihan CPT	Nilai Pengaturan Aktual	Justifikasi Teknis
Model Dasar (Base Model)	Ministral-3-3B-Base	Ukuran parameter efisien untuk operasional API
LoRA Rank / Alpha	32 / 64	Menjaga kapasitas penyimpanan pola bahasa baru
Target Modul LoRA	Attention + MLP + Embedding	Melatih layer kosakata merekam istilah slang
Laju Pembelajaran Adaptor	5e-5	Laju pembaruan bobot standar yang stabil
Laju Pembelajaran Kosakata	5e-6	Mencegah kerusakan representasi kata bawaan
Laju Pembelajaran Penjadwal	Cosine	Membantu konvergensi loss secara perlahan
Panjang Sekuens Maksimum	2.048 token	P75-P99 panjang dokumen berada di bawah batas ini
Ukuran Batch Efektif	16	Menghindariketidakstabilan gradien
Jumlah Epoch Pelatihan	1 Epoch	Membatasi risiko catastrophic forgetting

4.4 Realisasi Pelatihan Supervised Fine-Tuning

(SFT)

Setelah model dasar sukses melalui tahapan asimilasi bahasa baru pada fase CPT, bobot adapter CPT tersebut dilebur secara permanen. Model CPT tersebut kemudian digunakan sebagai fondasi baru untuk melakukan pelatihan instruksi *Supervised Fine-Tuning* (SFT) guna melatih kapabilitas klasifikasi terstruktur dan pembuatan penjelasan penalaran.

4.4.1 Solusi Masalah Perhitungan Loss (Loss Computation Workaround)

Selama fase pengujian awal SFT menggunakan *trainer* bawaan, ditemukan anomali perilaku di mana proses kalkulasi nilai kerugian (*loss*) tidak sepenuhnya menghormati pemetaan indeks khusus dari penepakan loss khusus asisten (*Assistant-Only Loss Masking*). Masalah ini mengakibatkan bias pembelajaran di mana model berusaha menghafal pola teks instruksi masukan.

Untuk mengatasinya, dibuat skrip penimpaan fungsi `compute_loss` secara manual yang secara ketat memanfaatkan kelas `nn.CrossEntropyLoss(ignore_index=-100)` milik PyTorch untuk mengabaikan kalkulasi loss pada token penunjuk instruksi:

```
import torch.nn as nn
from transformers import Trainer

class CompatibleTrainer(Trainer):
    def compute_loss(self, model, inputs, return_outputs=False):
        # Ekstraksi label target asli
        labels = inputs.get("labels")
        outputs = model(**inputs)
        logits = outputs.get("logits")

        # Pergeseran token (Shift logits & labels) untuk Causal LM
        shift_logits = logits[..., :-1, :].contiguous()
        shift_labels = labels[..., 1:].contiguous()

        # Definisikan Cross Entropy Loss dengan ignore_index -100
        loss_fct = nn.CrossEntropyLoss(ignore_index=-100)
        loss = loss_fct(shift_logits.view(-1, shift_logits.size(-1)),
shift_labels.view(-1))

        return (loss, outputs) if return_outputs else loss
```

4.4.2 Naskah Kode Implementasi Pelatihan SFT

Berikut adalah naskah kode program Python yang dijalankan pada modul SFT menggunakan model hasil CPT dalam format kuantisasi 4-bit demi efisiensi memori GPU A100:

```
from unsloth import FastLanguageModel
from transformers import TrainingArguments

# 1. Memuat Model Hasil CPT dalam presisi 4-bit
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="/content/drive/MyDrive/aitf/output/model_rating_3b_base/cpt_merged",
    max_seq_length=2048,
    load_in_4bit=True,  # Menghemat VRAM untuk proses batching SFT
    device_map="auto"
)

# 2. Menambahkan Adapter LoRA khusus untuk SFT
model = FastLanguageModel.get_peft_model(
    model,
    r=16,
    lora_alpha=32,
    lora_dropout=0.0,
    target_modules=["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj",
"down_proj"],
    bias="none",
    use_gradient_checkpointing="unsloth",
    random_state=42
)
```

```
# 3. Konfigurasi Argumen Pelatihan SFT
training_args = TrainingArguments(
    output_dir="/content/drive/MyDrive/aitf/output/model_rating_3b_base/sft_output",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=4,  # Effective Batch Size = 32
    epochs=2,
    learning_rate=2e-4,
    fp16=not torch.cuda.is_bf16_supported(),
    bf16=torch.cuda.is_bf16_supported(),
    logging_steps=10,
    evaluation_strategy="steps",
    eval_steps=50,
    save_strategy="steps",
    save_steps=100,
    optim="adamw_8bit"
)

# Inisialisasi trainer dengan custom loss calculation
trainer = CompatibleTrainer(
    model=model,
    tokenizer=tokenizer,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
    args=training_args,
)

trainer.train()
```

4.4.3 Parameter dan Hiperparameter Pelatihan SFT

Konfigurasi parameter untuk pelatihan SFT dikunci dalam spesifikasi berikut:

Tabel 4. 3 Konfigurasi Parameter Pelatihan SFT

Parameter Pelatihan SFT	Nilai Pengaturan Aktual	Justifikasi Teknis
Kuantisasi Memori	4-bit Precision	Meminimalkan konsumsi VRAM komparasi sekuens
LoRA Rank / Alpha	16 / 32	Konfigurasi adapter standar optimal tugas instruksi

Target Modul LoRA	Attention + MLP Projections	Hanya menyasar lapisan fungsional penyelarasan
Laju Pembelajaran (LR)	2e-4	Memaksa adaptasi cepat format instruksi keluaran
Jumlah Epoch Pelatihan	2 Epochs	Memastikan model stabil mematuhi struktur penulisan
Ukuran Batch Efektif	32	Menyeimbangkan
Penetapan Indeks Loss	ignore_index = −100	Menghilangkan kata instruksi dari pembaruan bobot

4.5 Integrasi, Merging, dan Deployment Model

Fase integrasi, penggabungan (*merging*), serta penyebaran layanan (*deployment*) merupakan tahap krusial untuk memindahkan model dari lingkungan eksperimen ke lingkungan operasional nyata pada infrastruktur remote GPU server (RunPod).

4.5.1 Solusi Penyelamatan Kegagalan Penggabungan (PEFT Merge Fallback)

Ketika pelatihan SFT selesai, adapter LoRA harus digabungkan secara permanen dengan model dasar 16-bit (*bfloat16*). Selama penggabungan menggunakan pustaka Unsloth, terjadi galat sebagai berikut:

RuntimeError: [Unsloth merge debug] LoRA count mismatch: modules=352, lora_A=351, lora_B=351, scaling=352

Galat ini disebabkan ketidakseimbangan deteksi indeks modul representasi bahasa pada model arsitektur Ministral3 pasca pemuatan parsial. Untuk menyelamatkan proses penggabungan agar tidak gagal tengah jalan, sistem mengimplementasikan logika penyelamatan *try-except* otomatis yang mengalihkan proses ke metode merge bawaan modul PEFT murni:

```
import shutil

local_merge_path = "/content/sft_model_merged_local"

try:
    print("Mencoba proses merge murni via Unsloth...")
    model.save_pretrained_merged(local_merge_path, tokenizer, save_method="merged_16bit")
    print("✅ Penggabungan bobot via Unsloth berhasil sepenuhnya.")
except Exception as e:
    print(f"⚠ Unsloth merge gagal karena Galat Arsitektur: {str(e)}")
    print("→ Mengaktifkan sistem penyelamatan (PEFT native fallback)...")

    # 1. Hapus sisa modul parsial dari memori
    free_memory()

    # 2. Muat ulang model dasar murni dalam presisi 16-bit bfloat16
    from transformers import AutoModelForImageTextToText, AutoProcessor
    from peft import PeftModel

    base_model_16 = AutoModelForImageTextToText.from_pretrained(
        "mistralai/Ministral-3-3B-Base-2512",
        torch_dtype=torch.bfloat16,
        device_map="cpu" # Load di CPU untuk membebaskan VRAM
    )

    # 3. Muat adapter LoRA terlatih dari Drive
    peft_model = PeftModel.from_pretrained(
        base_model_16,
        "/content/drive/MyDrive/aitf/output/model_rating_3b_base/sft_output"
    )

    # 4. Satukan bobot secara murni menggunakan PEFT Native API
    merged_model = peft_model.merge_and_unload()
    merged_model.save_pretrained(local_merge_path)
```

```
    processor = AutoProcessor.from_pretrained("mistralai/Ministral-3-3B-Base-2512")
    processor.save_pretrained(local_merge_path)

    print("✅ Penyelamatan sukses: Model berhasil disatukan via PEFT Native.")

# Salin folder hasil penyatuan lokal ke Google Drive
shutil.copytree(local_merge_path, "/content/drive/MyDrive/aitf/output/merged_full_model")
```

4.5.2 Tantangan Deployment pada Remote Server RunPod

Setelah model disatukan menjadi file bobot tunggal berpresisi 16-bit murni berukuran sekitar 7.7 GB, model dipindahkan ke server GPU aktif milik RunPod untuk dideploy. Awalnya, model direncanakan untuk di-host menggunakan mesin inferensi vLLM, namun menemui kegagalan karena dua ketidaksesuaian struktur:

- Miskonfigurasi Preprosesor: Meskipun model dilatih secara tekstual murni, metadata konfigurasinya mengacu pada arsitektur induk multimodal, sehingga vLLM menolak memuat model karena ketiadaan file konfigurasi preprosesor gambar.
- Perbedaan Kunci Tensor: vLLM mengharapkan penulisan kunci tensor dimulai langsung dari lapisan tertentu. Namun, bobot model hasil gabungan memiliki penulisan kunci berprefiks ekstra pada seluruh 458 bobot tensor, yang memicu galat kritis pada mesin inferensi.

4.5.3 Realisasi Solusi: FastAPI Berbasis Hugging Face Transformers

Untuk menghindari penulisan ulang biner Safetensors sebesar 7.7 GB demi memenuhi aturan vLLM, sistem diimplementasikan memanfaatkan pustaka Hugging Face Transformers yang memiliki penanganan fleksibel terhadap prefiks tensor. Sistem dirancang sebagai layanan web REST API mandiri menggunakan FastAPI:

```
import torch
import time
from fastapi import FastAPI, HTTPException
from pydantic import BaseModel
from typing import List, Dict, Optional
from transformers import AutoProcessor, AutoModelForImageTextToText

# ================================================================
# 1. INISIALISASI MODEL HUGGING FACE NATIVE (CUDA)
# ================================================================
print("Memuat model padsft ke VRAM GPU (bfloat16)...")
MODEL_PATH = "/workspace/models/padsft"

# Memuat processor bawaan (PixtralProcessor)
processor = AutoProcessor.from_pretrained(MODEL_PATH)

# Memuat model teks (Mistral3ForConditionalGeneration) di GPU
model = AutoModelForImageTextToText.from_pretrained(
    MODEL_PATH,
    torch_dtype=torch.bfloat16,
    device_map="cuda"
).eval() # Memastikan mode evaluasi aktif (no-grad)
```

```
app = FastAPI(title="PAD Model SFT API Service", version="1.0.0")

# ================================================================
# 2. DEFINISI SKEMA PAYLOAD DENGAN PYDANTIC (OPENAI COMPLIANT)
# ================================================================
class ChatMessage(BaseModel):
    role: str
    content: str

class ChatCompletionRequest(BaseModel):
    messages: List[ChatMessage]
    max_tokens: Optional[int] = 150
    temperature: Optional[float] = 0.1

# ================================================================
# 3. IMPLEMENTASI ENDPOINT API
# ================================================================
@app.get("/health")
def health_check():
    return {"status": "ok", "model": "padsft", "device": "cuda"}

@app.post("/v1/chat/completions")
async def chat_completions(request: ChatCompletionRequest):
    try:
        # 1. Ekstraksi format pesan menjadi struktur list dict
        formatted_messages = [{"role": m.role, "content": m.content} for m in
request.messages]

        # 2. Terapkan template percakapan asli model
        chat_prompt = processor.apply_chat_template(
            formatted_messages,
            tokenize=False,
            add_generation_prompt=True
        )

        # 3. Tokenisasi dan pindahkan tensor ke memori CUDA
        inputs = processor(text=chat_prompt, return_tensors="pt").to("cuda")

        # 4. Inferensi non-gradien dengan pencarian serakah (greedy decoding)
        t_start = time.time()
        with torch.no_grad():
            output_tokens = model.generate(
                **inputs,
                max_new_tokens=request.max_tokens,
                do_sample=False  # Greedy decoding untuk kestabilan rating
            )
        elapsed_time = time.time() - t_start

        # 5. Dekode hasil keluaran token menjadi teks bersih
        generated_text = processor.batch_decode(output_tokens, skip_special_tokens=True)
[0]

        # Kembalikan respons dalam struktur standar ramah parser aplikasi
        return {
            "choices": [
                {
                    "message": {
                        "role": "assistant",
```

```
                    "content": generated_text.strip()
                }
            }
        ],
        "usage": {
            "inference_time_seconds": round(elapsed_time, 2)
        }
    }

except Exception as e:
    raise HTTPException(status_code=500, detail=str(e))
```

4.5.4 Eksekusi Latar Belakang (Background Execution via tmux)

Agar layanan FastAPI tetap aktif melayani panggilan meskipun sesi terminal SSH ditutup, eksekusi program dijalankan dalam sesi persisten terminal tmux:

1. Membuat dan Memasuki Sesi tmux Baru:

```
tmux new -s aitfpad1
```

1. Menjalankan Server Inferensi Menggunakan Uvicorn:

```
python -m uvicorn serve_padsft_fastapi:app --host 0.0.0.0 --port 8002
```

1. Melakukan Detach Sesi (Meninggalkan Proses Berjalan di Latar Belakang): Tekan kombinasi tombol `Ctrl + B`, lalu lepaskan dan tekan tombol `D`.
2. Verifikasi Kesehatan Layanan:

```
curl http://localhost:8002/health
```

4.6 Verifikasi Akhir Layanan Inferensi (Smoke Testing)

Untuk memastikan bahwa seluruh rangkaian implementasi berjalan dengan benar di lingkungan produksi, dilakukan uji coba fungsionalitas akhir (*smoke testing*). Pengujian ini meniru skenario pemanggilan API riil oleh aplikasi frontend MVP dengan mengirimkan teks masukan yang sarat akan unsur perundungan siber.

4.6.1 Perintah Pengujian Terminal (Curl Command)

Pengujian dieksekusi secara lokal dari terminal server menggunakan utilitas curl:

```
curl -X POST http://localhost:8002/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{
    "messages": [
      {
        "role": "system",
        "content": "Anda adalah sistem moderasi konten PAD.\n\nKlasifikasikan teks
berdasarkan kategori dan rating PAD.\n\nKategori:\n- Netral\n- Violence\n- Sexual\n-
Harassment\n- Hateful Content\n- Self Harm\n\nAturan:\n- Violence, Sexual, Harassment,
Hateful Content, dan Self Harm selalu memiliki rating PRC.\n- Hanya kategori Netral yang
dapat memiliki rating SU, 7+, 13+, atau 17+.\n\nBerikan penjelasan singkat dengan wajib
mengutip frasa spesifik dari teks sebagai alasan klasifikasi.\n\nOutput:\nRating: ...
\nKategori: ...\nPenjelasan: ..."
      },
      {
        "role": "user",
        "content": "Awas lo anjing! gw tandain ya muka lo"
      }
    ],
    "max_tokens": 150,
    "temperature": 0.1
  }'
```

4.6.2 Struktur Hasil Keluaran API (JSON Response)

Model berhasil memproses teks masukan pada memori GPU dan mengembalikan respons terstruktur dalam format JSON dengan latensi rendah:

```
{
  "choices": [
    {
      "message": {
        "role": "assistant",
        "content": "Rating: PRC\nKategori: Harassment\nPenjelasan: Penghinaan agresif
dengan menyebut target sebagai 'anjing' yang merupakan bentuk body shaming dan
dehumanisasi. Frasa 'gw tandain ya muka lo' menunjukkan serangan personal yang merendahkan
martabat korban secara verbal."
      }
    }
  ],
  "usage": {
    "inference_time_seconds": 1.24
  }
}
```


4.6.3 Analisis Keberhasilan Fungsional SFT

Berdasarkan hasil keluaran smoke testing di atas, beberapa poin keberhasilan teknis pada tahap implementasi model SFT PAD dikonfirmasi sebagai berikut:

1. Kepatuhan Format Struktural: Model berhasil menghasilkan jawaban yang sangat patuh terhadap templat instruksi tanpa terjadi gejala halusinasi atau pengulangan kata instruksi. Hal ini membuktikan bahwa metode penepakan loss khusus asisten berhasil memaksa model hanya mempelajari pemetaan respon yang bersih.
2. Akurasi Logika Klasifikasi: Model mampu mengidentifikasi kata umpatan tidak baku berbahasa Indonesia dan kalimat bernada intimidasi secara kontekstual, memetakan kategori, serta menetapkan rating kelayakan pada level pembatasan tertinggi yaitu Konten Terlarang.
3. Kualitas Penalaran (Reasoning): Model mampu mengekstrak frasa sensitif dari input dan menyusun

argumen penjelasan hukum siber secara logis setebal 2 kalimat, membuktikan keberhasilan penggabungan kapabilitas adaptasi pengetahuan domain CPT dengan fleksibilitas penalaran instruksi SFT.

BAB 5 PENGUJIAN DAN EVALUASI

5.1 Skenario Pengujian

Evaluasi terhadap kinerja model kecerdasan buatan dalam proyek Perlindungan Anak di Ruang Digital (PAD) ini dirancang secara sistematis melalui tiga skenario pengujian utama:

5.1.1 Skenario Pengujian Tahap 1: Benchmark Model Dasar CPT (20% Dataset)

- Tujuan: Mengidentifikasi arsitektur model dasar (*base model*) terbaik dari tiga keluarga model terkemuka (Gemma-3-4B-pt, Qwen3.5-4B-Base, dan Ministral-3-3B-Base-2512) untuk tugas asimilasi bahasa baru.
- Metode: Setiap model dilatih menggunakan parameter LoRA yang identik (r=32, alpha=64, 1 epoch) pada subset CPT sebesar 20% (98.336 dokumen). Kinerja model dinilai berdasarkan penurunan nilai kerugian (*loss*), konsumsi memori (*Peak VRAM*), durasi pelatihan (*runtime*), serta persentase penurunan nilai *Perplexity* pada domain spesifik PAD.

5.1.2 Skenario Pengujian Tahap 2: Pelatihan CPT Skala Penuh (Full Dataset)

- Tujuan: Mengevaluasi tingkat keberhasilan asimilasi pengetahuan domain PAD secara mendalam pada model dasar terbaik yang terpilih dari Skenario Tahap 1.
- Metode: Model terpilih (Ministral-3-3B) dilatih

menggunakan keseluruhan dataset CPT yang berjumlah 491.447 dokumen. Pengujian kualitas asimilasi diukur menggunakan perbandingan nilai *Perplexity* (PPL) dan *Bits Per Character* (BPC) sebelum pelatihan (*baseline*) dan sesudah pelatihan (*post-CPT*) pada dua jenis data uji: data uji domain spesifik PAD (200 dokumen) dan data uji pengetahuan umum Wikipedia Indonesia (500 dokumen) guna melacak tingkat *catastrophic forgetting*.

5.1.3 Skenario Pengujian Tahap 3: Evaluasi Kinerja Instruksi SFT

- Tujuan: Mengukur kemampuan model akhir hasil penyatuan SFT (padsft berbasis Ministral-3-3B) dalam menjalankan tugas klasifikasi rating usia secara otomatis sekaligus menyajikan teks penalaran (Penjelasan).
- Metode: Model dievaluasi menggunakan 1.341 sampel data uji SFT yang tidak pernah dilihat sebelumnya (*unseen test set*) dengan parameter inferensi pencarian serakah (*greedy decoding*, temperature=0.1). Kinerja dinilai secara kuantitatif berdasarkan metrik klasifikasi multikelas (Accuracy, Macro Precision, Macro Recall, Macro F1-Score, dan MCC) serta metrik kualitas sintaksis penalaran (ROUGE-L).

5.2 Hasil Pengujian Continued Pre-Training (CPT)

Berikut adalah pemaparan data empiris hasil pengeksekusian pengujian Continued Pre-Training (CPT) berdasarkan skenario pengujian yang telah dirancang:

5.2.1 Hasil Evaluasi Tahap 1: Benchmark Pemilihan Model Dasar (20% Dataset)

Pengujian komparatif terhadap tiga model dasar pada subset data 20% menghasilkan metrik performa komputasi dan pemelajaran sebagai berikut:

Tabel 5. 1 Hasil Komparatif Benchmark CPT (Subset 20% / 98.336 Dokumen)

Metrik Evaluasi & Komputasi	google/ Gemma-3-4B-pt	Qwen/ Qwen3.5-4B-Base	mistralai/Ministral-3-3B-Base
Train Loss Akhir	1.9473	2.2632	1.8810
Eval Loss Akhir	1.9077	2.1799	1.7850
Baseline Domain PPL	25.2335	18.6235	18.1791
Post-CPT Domain PPL	15.8527	12.9876	11.8287

Pergeseran Domain PPL (%)	−37.18%	−30.26%	−34.93%
Konsumsi Puncak VRAM	26.26 GB	28.43 GB	21.25 GB
Durasi Waktu Pelatihan	1079.0 Menit (18.0 Jam)	688.5 Menit (11.5 Jam)	204.8 Menit (3.4 Jam)

Justifikasi Pemilihan Model: Berdasarkan data pada tabel di atas, Ministral-3-3B-Base-2512 dipilih sebagai model dasar utama proyek ini. Model ini mencatatkan nilai Eval Loss terendah (1.7850) dan nilai Post-CPT Perplexity terbaik (11.8287) pada domain PAD. Selain itu, dari aspek efisiensi infrastruktur, model Ministral-3-3B menunjukkan kecepatan pemrosesan (*throughput*) yang jauh lebih unggul dengan hanya membutuhkan waktu 3.4 jam pelatihan (5x lebih cepat dibanding Gemma-3 dan 3x lebih cepat dibanding Qwen3.5) serta mengonsumsi puncak memori GPU paling rendah, yaitu 21.25 GB VRAM.

5.2.2 Hasil Evaluasi Tahap 2: Hasil Pelatihan CPT Skala Penuh (Full Dataset)

Setelah model dasar Ministral-3-3B-Base dipilih, proses prapelatihan dilanjutkan menggunakan skala penuh (491.447 dokumen valid). Hasil akhir nilai kerugian (*loss*) dan metrik kepastian (*perplexity*) dicatat sebagai berikut:

Tabel 5. 2 Metrik Pelatihan CPT Full Dataset (Ministral-3-3B)

Metrik Loss & Perplexity	Nilai Akhir Pelatihan Skala Penuh
Final Train Loss	1.7788
Final Validation Loss	1.6741
Final Test Loss	1.6708
Final Validation Perplexity (PPL)	5.3341
Final Test Perplexity (PPL)	5.3163

5.2.3 Analisis Pergeseran Metrik PPL dan BPC (Baseline vs. Post-CPT)

Untuk mengukur signifikansi asimilasi pengetahuan baru terkait regulasi siber dan terminologi perlindungan anak, dilakukan komparasi nilai Perplexity (PPL) dan Bits Per Character (BPC) sebelum dan sesudah pelatihan CPT skala penuh:

Tabel 5. 3 Pergeseran Metrik PPL & BPC Model Ministral-3-3B (Baseline vs. Post-CPT)

Kategori Data Uji (Metrik)	Nilai Baseline (Pre- CPT)	Nilai Evaluasi (Post-CPT)	Pergeseran Absolut	Selisih Presentase	Status Evaluasi
Domain PAD (PPL)	10.0116	5.2230	−4.7886	−47.83%	Adaptasi Sukses
Domain PAD (BPC)	0.9531	0.6839	−0.2692	−28.24%	Adaptasi Sukses
Wikipedia Umum (PPL)	11.2125	16.4823	+5.2698	+47.00%	Degenerasi Wajar

Wikipedia Umum (BPC)	0.9466	1.0975	+0.1509	+15.94%	Degenerasi Wajar

Analisis Hasil CPT: Data pada tabel di atas membuktikan keberhasilan asimilasi pengetahuan baru secara signifikan. Nilai Perplexity pada domain spesifik PAD mengalami penurunan tajam sebesar 47.83% (dari 10.01 menjadi 5.22), diikuti oleh penurunan nilai BPC sebesar 28.24% (dari 0.95 menjadi 0.68). Penurunan nilai BPC mengindikasikan bahwa model memerlukan jumlah representasi bit informasi yang jauh lebih sedikit untuk mengekspresikan karakter kata di bidang perlindungan anak.

Meskipun model mengalami degradasi kemampuan bahasa umum Wikipedia (General PPL naik +47.00% dan General BPC naik +15.94%), fenomena *catastrophic forgetting* ini dinilai masih berada dalam batas toleransi wajar untuk sebuah model yang dikondisikan secara khusus untuk tugas moderasi konten spesifik (*domain-specific adapter*).

5.3 Hasil Pengujian Supervised Fine-Tuning (SFT)

Setelah model melewati tahapan CPT, dilakukan pengujian SFT skala penuh untuk mengukur kapabilitas klasifikasi terstruktur dan pembuatan penjelasan penalaran (*reasoning*) pada 1.341 data uji SFT (*unseen test set*). Evaluasi inferensi menggunakan *Left-Padded Batch Inference* (Batch Size 128) dengan metode pencarian serakah (*greedy decoding*, temperature=0.1).

5.3.1 Ringkasan Metrik Evaluasi SFT Global

Kinerja evaluasi global model padsft pada data uji disajikan secara komprehensif dalam tabel berikut:

Tabel 5. 4 Ringkasan Metrik SFT Global (1.341 Sampel Uji)

Metrik Evaluasi Klasifikasi & Generasi	Nilai Capaian Model Akhir	Interpretasi Matematis
Akurasi Global (Accuracy)	94.85%	1.272 dari total 1.341 sampel uji berhasil diklasifikasikan dengan benar secara mutlak.
Presisi Rata-Rata (Macro Precision)	94.47%	Model memiliki akurasi penentuan yang tinggi dan meminimalisasi tingkat kesalahan tanda (*false positives*).
Sensitivitas	0.9357	Model mampu menangkap hampir seluruh konten berbahaya yang ada di data uji (*false negatives* sangat rendah).
Skor F1 Makro (Macro F1-Score)	0.9400	Menunjukkan keseimbangan performa yang solid antara presisi dan sensitivitas pada klasifikasi multikelas.

Korelasi Matthews (MCC)	0.9342	Menegaskan bahwa model memiliki kemampuan klasifikasi yang sangat andal dan tidak bias terhadap kelas mayoritas.
Metrik Sintaksis Penalaran (ROUGE-L)	0.3340	Struktur kalimat penjelasan (Penjelasan) yang dihasilkan model memiliki kemiripan sekuensial yang tinggi dengan acuan.

Analisis Korelasi Matthews (MCC): Dataset SFT PAD memiliki tingkat ketidakseimbangan kelas (*class imbalance*) yang sangat nyata (misalnya kelas Konten Terlarang berjumlah 4.497 baris sedangkan kelas 18+ hanya 1.134 baris). Nilai MCC sebesar 0.9342 (sangat mendekati skor ideal +1.0) memberikan bukti empiris yang valid bahwa performa tinggi model ini merata di seluruh kelas, dan bukan merupakan efek bias akurasi semu akibat penebakan kelas mayoritas.

5.3.2 Kinerja Klasifikasi Detail Per-Kelas Rating Usia

Untuk melihat performa model secara lebih mendalam pada masing-masing tingkatan rating usia, dilakukan ekstraksi metrik presisi, recall, dan F1-score untuk setiap kategori:

Tabel 5. 5 Hasil Metrik Klasifikasi SFT Detail Per-Kelas Rating Usia

Kategori Rating Usia	Presisi	Sensitivitas	Skor F1	Status Performa Klasifikasi
Semua Umur (SU)	0.94	0.92	0.93	Tinggi
7+	0.95	0.97	0.96	Sangat Tinggi
13+	0.88	0.86	0.87	Sedang (Tantangan Batas)
15+	0.89	0.91	0.90	Tinggi
18+	0.99	0.99	0.99	Luar Biasa
Konten Terlarang (PRC)	0.99	0.99	0.99	Luar Biasa

Analisis Kinerja Per-Kelas: Berdasarkan data pada tabel di atas, model mencatatkan skor F1 sempurna sebesar 0.99 pada dua kategori kritis: Konten Terlarang (PRC) dan 18+. Hal ini sangat penting secara taktis bagi KOMDIGI, karena model memiliki keandalan hampir mutlak dalam mendeteksi dan mengunci konten-konten berbahaya (seperti judi online, kekerasan siber, pelecehan verbal, dan muatan dewasa) agar tidak bocor dan diakses oleh anak di bawah umur.

Kategori 13+ mencatatkan skor F1 terendah yaitu 0.87 (Presisi 0.88, Recall 0.86). Berdasarkan audit sampel kesalahan prediksi (*error analysis*), hal ini disebabkan oleh tingginya ambiguitas semantik pada fase transisi bahasa remaja. Teks curahan hati remaja

awal (rating 13+) sering kali menggunakan diksi emosional yang berhimpitan dengan teks diskusi sosial-politik atau laporan realita sosial (rating 15+), sehingga memicu kesalahan klasifikasi marginal antar-kedua kelas tersebut.

5.4 Evaluasi Sistem

Berdasarkan keseluruhan hasil implementasi, pengujian, dan verifikasi akhir di lingkungan produksi, sistem moderasi konten PAD yang dikembangkan dinilai memiliki kelebihan serta keterbatasan sebagai berikut:

5.4.1 Kelebihan Sistem

1. Perlindungan Batas Aman yang Tangguh (*Robust Safety Boundary*): Kinerja klasifikasi model yang mencapai skor F1 sebesar 0.99 pada kategori Konten Terlarang (PRC) menjamin bahwa sistem ini sangat andal digunakan sebagai penyaring utama untuk memblokir unggahan berbahaya (seperti kata kunci judi online slot/gacor, perundungan verbal, dan pelecehan) secara *real-time*.
2. Transparansi Keputusan (*Explainable AI*): Dukungan luaran teks penalaran (Penjelasan) setebal 2–4 kalimat yang konsisten dan informatif (mencatatkan skor ROUGE-L sebesar 0.3340) membantu petugas KOMDIGI memahami dasar keputusan model. Petugas tidak hanya menerima keputusan kaku, melainkan disajikan kutipan langsung (*textual*

evidence) dari kata kunci sensitif yang melanggar aturan.

3. Efisiensi Sumber Daya Tingkat Tinggi: Penggunaan model berbasis Ministral-3-3B yang hemat memori (hanya mengonsumsi 7.3 GB VRAM saat di-deploy) memungkinkan layanan inferensi dijalankan pada GPU kelas menengah (*affordable serverless compute*) dengan latensi respons rata-rata di bawah 1.3 detik per permintaan, menghemat biaya operasional infrastruktur KOMDIGI secara signifikan.

5.4.2 Keterbatasan dan Kelemahan Sistem

1. Penurunan Pengetahuan Bahasa Umum (*Catastrophic Forgetting*): Efek dari prapelatihan CPT spesifik domain PAD mengakibatkan model mengalami kenaikan nilai Perplexity pada domain umum Wikipedia Indonesia sebesar 47.00%. Model ini tidak lagi direkomendasikan untuk tugas-tugas generatif umum (seperti menulis artikel kreatif atau menjawab pertanyaan pengetahuan umum), melainkan harus diisolasi khusus untuk tugas moderasi teks siber.
2. Ambiguitas Batas Usia Transisi (13+ vs 15+): Model masih menemui kesulitan marginal dalam membedakan secara tegas antara curahan emosi remaja awal (13+) dengan ulasan realita sosial remaja akhir (15+) karena adanya irisan diksi bahasa tidak baku yang digunakan oleh kedua kelompok usia tersebut di media sosial.

BAB 6 PENUTUP

6.1 Kesimpulan

Berdasarkan seluruh rangkaian tahapan analisis, perancangan, implementasi, hingga pengujian sistem moderasi konten Perlindungan Anak di Ruang Digital (PAD) yang telah dilaksanakan, beberapa kesimpulan teknis dirumuskan sebagai berikut:

1. Keberhasilan Rekayasa Dataset Tekstual: Proyek ini berhasil menyusun dan memproses dua dataset tekstual berbahasa Indonesia yang bersih dan terarah:
   - Dataset CPT: Sebanyak 491.447 baris dokumen (214.947.565 token) dengan komposisi seimbang (70% domain spesifik PAD, termasuk variasi bahasa tidak baku atau *slang*, dan 30% pengetahuan umum).
   - Dataset SFT: Sebanyak 13.406 baris data instruksi berbentuk percakapan terstruktur yang menyelaraskan enam kelas rating usia sesuai regulasi nasional.
2. Keberhasilan Adaptasi Domain (CPT): Proses *Continued Pre-Training* pada model dasar mistralai/Ministral-3-3B-Base-2512 menggunakan teknik *Rank-Stabilized LoRA* (RSLoRA) berhasil menggeser ruang representasi semantik model ke arah domain perlindungan anak. Hal ini dibuktikan secara empiris dengan turunnya nilai *Perplexity* (PPL) domain PAD sebesar 47.83% (dari 10.01 menjadi 5.22) serta penurunan nilai *Bits Per Character*

(BPC) sebesar 28.24% (dari 0.95 menjadi 0.68). Kenaikan PPL pada domain umum Wikipedia Indonesia (+47.00%) diidentifikasi sebagai efek samping degenerasi memori (*catastrophic forgetting*) yang masih berada dalam batas toleransi wajar untuk model spesifik domain (*domain-specific adapter*).

3. Keberhasilan Penyelarasan Instruksi (SFT): Penerapan teknik *Assistant-Only Loss Masking* (menggunakan PyTorch *ignore index* −100 pada wilayah instruksi) berhasil melatih model secara disiplin untuk mengikuti format keluaran terstruktur. Pada pengujian 1.341 data
   uji baru (*unseen test set*), model padsft mencatatkan tingkat Akurasi Global sebesar 94.85%, Macro F1-Score sebesar 94.00%, dan MCC sebesar 0.9342. Model ini menunjukkan
   keandalan yang sangat tinggi dalam mengamankan batas kelayakan konten anak dengan mencatatkan F1-Score sebesar 0.99 pada kelas Konten Terlarang (PRC) dan 18+.
4. Kesiapan Integrasi dan Deployment: Model hasil penggabungan (*merged weights*) 16-bit berhasil disajikan sebagai layanan web REST API mandiri menggunakan FastAPI di port 8002 pada infrastruktur server GPU (RunPod). Dengan memanfaatkan pustaka Hugging Face Transformers secara murni, sistem berhasil mengatasi masalah ketidaksesuaian kunci tensor vLLM tanpa perlu menulis ulang file Safetensors sebesar 7.7 GB.

Pengujian akhir (*smoke testing*) membuktikan model mampu melakukan inferensi non-gradien (*greedy decoding*) di atas memori GPU dengan latensi respons rata-rata 1.24 detik dan konsumsi memori yang efisien (7.3 GB VRAM).

6.2 Saran

Untuk meningkatkan kualitas sistem dan kinerja model pada pengembangan tahap selanjutnya, beberapa rekomendasi teknis yang berfokus pada domain NLP dan pemrosesan teks diusulkan sebagai berikut:

1. Ekspansi Korpus Teks Dialek Daerah dan Slang Lokal: Melakukan diversifikasi korpus teks pada fase CPT berikutnya dengan mengintegrasikan kosakata tidak baku dari bahasa daerah populer (seperti bahasa Jawa, Sunda, dan Batak) serta perkembangan bahasa gaul internet terbaru. Hal ini penting karena pengguna media sosial sering kali memodifikasi ejaan kata kunci sensitif menggunakan dialek lokal untuk menghindari penyaringan otomatis konvensional.
2. Eksperimen Skala Parameter Model yang Lebih Besar: Melakukan uji coba CPT dan SFT menggunakan model bahasa dasar berdimensi parameter lebih tinggi (seperti Mistral-7B, Llama-3-8B, atau Qwen2.5-7B/14B-Base). Peningkatan kapasitas parameter ini diperkirakan dapat meningkatkan kemampuan penalaran kontekstual model, sehingga dapat mereduksi ambiguitas

semantik pada batas klasifikasi usia transisi (kategori 13+ dan 15+) yang saat ini masih mencatatkan skor F1 terendah (0.87).

3. Penyelarasan Preferensi Tekstual Menggunakan DPO/KTO: Menerapkan metode *Direct Preference Optimization* (DPO) atau *Kahneman-Tversky Optimization* (KTO) pada fase pasca-SFT. Teknik ini berguna untuk melatih gaya penulisan teks alasan (Penjelasan) model agar secara konsisten mengikuti struktur argumentasi hukum pidana yang sesuai dengan standar penulisan draf regulasi resmi (*legal drafting*) milik KOMDIGI.

DAFTAR PUSTAKA

Bai, J., Bai, S., Chu, Y., et al. (2023). Qwen Technical Report.

arXiv preprint arXiv:2309.16609. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners.

Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirec-tional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 4171–4186.

Gemma Team. (2024). Gemma: Open Models Based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295.

Gururangan, S., Marasović, A., Swayamdipta, S., et al. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 8342–8360.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed., draft). Pearson.

Ke, Z., Sun, R., Liu, Y., et al. (2023). Continual Pre-Training of Large Language Models: How to (re)warm your model? arXiv preprint arXiv:2308.04014.

Kementerian Komunikasi dan Informatika Republik Indonesia. (2024). Peraturan Menteri Ko-munikasi dan Informatika Republik Indonesia Nomor 2 Tahun 2024 tentang Klasifikasi Gim (Indonesia Game Rating System).

Mistral AI. (2024). Ministral 3B Technical Documentation.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems (NeurIPS), 35.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informed-ness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Presiden Republik Indonesia. (2009). Undang-Undang Republik Indonesia Nomor 33 Tahun 2009 tentang Perfilman.

Presiden Republik Indonesia. (2014). Undang-Undang Republik Indonesia Nomor 35 Tahun 2014 tentang Perubahan atas Undang-Undang Nomor 23 Tahun 2002 tentang Perlin-dungan Anak.

Presiden Republik Indonesia. (2024). Undang-Undang Republik Indonesia Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67.

Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report.

Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing & Management, 45(4), 427–437.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 30.

LAMPIRAN

Dokumentasi Kegiatan

Realisasi Penggunaan Dana

No	Tanggal	Uraian	Pemasukan (Rp)	Pengeluaran (Rp)	Saldo (Rp)
1	-	Pemasukan / Anggaran Dana	2.932.441	-	2.932.441
2	10 April 2026	Langganan Google Colab Pro (US$11,09)	-	197.049	2.735.392
3	23 April 2026	Langganan Google Colab Pro (US$11,09)	-	197.994	2.537.398
4	24 Mei 2026	Langganan Google Colab Pro Plus (US$55,49)	-	1.021.875	1.515.523
5	31 Mei 2026	Openrouter AI API (US$10,8)	-	193.000	1.322.523
6	1 Juni 2026	Konsumsi untuk 6 anggot	-	1.322.523	0

Total			**2.932.441**	**2.932.441**	**0**

*Bukti Transfer *click* disini

BIODATA PENULIS

1. Danial Rajiv Syahidan

Nama	: Danial Rajiv Syahidan
NRP	: 5054231004
Tempat, Tanggal Lahir	: Jakarta, 25 September 2004
Jenis Kelamin	: Laki-laki
Telepon	: 089628370870
Email	: danialrajiv16@gmail.com

AKADEMIS

Kuliah	: Departemen Teknik Informatika, FTEIC, ITS
Angkatan	: 2023
Semester	: 6

2. Adinda Shafa Najla

Nama	: Adinda Shafa Najla
NRP	: 5053231004
Tempat, Tanggal Lahir	: Medan, 23 July 2005
Jenis Kelamin	: Perempuan
Telepon	: 08977307216
Email	: sndinda37@gmail.com

AKADEMIS

Kuliah	: Departemen Teknik Informatika, FTEIC, ITS
Angkatan	: 2023
Semester	: 6

3. Panji Seno Aji

Nama	: Panji Seno Aji
NRP	: 5054231023
Tempat, Tanggal Lahir	: Malang, 22 Agustus 2004
Jenis Kelamin	: Laki-Laki
Telepon	: 082123494176
Email	: panjiaji90@gmail.com

AKADEMIS

Kuliah : Departemen Teknik Informatika, FTEIC, ITS
Angkatan : 2023
Semester : 6

4. Muhammad Khibban Itishom

Nama : Muhammad Khibban I'tishom
NRP : 5025231126
Tempat, Tanggal Lahir : Gresik, 14 Agustus 2005
Jenis Kelamin : Laki-laki
Telepon : 085334954634
Email : khibban148@gmail.com

AKADEMIS
Kuliah : Departemen Teknik Informatika, FTEIC, ITS
Angkatan : 2023
Semester : 6