PAPARAN IDE DAN MASALAH PENELITIAN

DEMOKRATISASI KEAMANAN AI: OTOMATISASI RED TEAMING PADA GEMMA 3 MELALUI REINFORCEMENT LEARNING YANG EFISIEN SUMBER DAYA

TEAM 2 - NETLAB NLP RESEARCH

Ganendra Garda Pratama (2306250642)

Daffa Hardhan (2306161763)

Leonard Bagaskara Cahyo Widodo (2406403835)

FACULTY OF ENGINEERING

DEPARTMENT OF ELECTRICAL ENGINEERING

UNIVERSITAS INDONESIA DEPOK

DAFTAR ISI

I. LATAR BELAKANG: PARADOKS KEAMANAN ERA OPEN WEIGHTS

Peluncuran model bahasa besar (*Large Language Models*/LLM) dengan bobot terbuka (*open-weights*) berkinerja tinggi, seperti **Google Gemma 3**, menandai titik infleksi penting dalam proses demokratisasi kecerdasan buatan. Di balik kemajuan tersebut, aksesibilitas yang semakin luas justru melahirkan sebuah **paradoks keamanan** yang bersifat fundamental: semakin terbuka sebuah model, semakin besar pula permukaan serangan (*attack surface*) yang terekspos bagi penyalahgunaan.

Berbeda dengan model tertutup berbasis antarmuka pemrograman aplikasi (API), model *open weights* memberikan akses *white-box* secara penuh. Kondisi ini memungkinkan aktor jahat untuk menganalisis struktur internal model secara mendalam dan mengoptimalkan serangan adversarial berupa *jailbreak*. Serangan semacam ini dapat memintas mekanisme pengaman (*guardrails*) yang dirancang oleh pengembang, sehingga model terdorong menghasilkan konten berbahaya, mulai dari ujaran kebencian yang memecah belah hingga panduan teknis pembuatan senjata biokimia.

Dalam konteks ini, metode pertahanan konvensional yang mengandalkan **Manual Red Teaming** semakin menunjukkan keterbatasannya. Pendekatan tersebut tidak terukur, rentan terhadap bias subjektif manusia, dan tidak mampu mengimbangi kecepatan siklus pengembangan serta penerapan model modern. Akibatnya, kita dihadapkan pada kesenjangan yang semakin lebar antara kapabilitas model bahasa besar dan kemampuan kita untuk melakukan audit keamanannya secara efektif.

II. RUMUSAN MASALAH

Riset ini dirumuskan untuk menjawab tiga permasalahan mendasar yang saat ini menjadi hambatan utama dalam evaluasi keamanan kecerdasan buatan.

2.1 Krisis Sumber Daya dan Ekslusi Infrastruktur

Metode *automated red teaming* mutakhir saat ini bergantung pada *Reinforcement Learning* (RL), khususnya algoritma **Proximal Policy Optimization (PPO)**. Namun, PPO memiliki tuntutan sumber daya yang sangat tinggi karena memerlukan pemuatan empat salinan model secara simultan (Actor, Critic, Reference, Reward). Audit keamanan model 7B sering kali mensyaratkan GPU kelas *Enterprise* (seperti NVIDIA A100 80GB), yang secara efektif

mengucilkan peneliti independen dan laboratorium akademik dari partisipasi riset keamanan yang bermakna.

2.2 Krisis Efektivitas Metodologis

Metode otomatis yang ada menghadapi dilema kualitas. Pendekatan berbasis gradien (**GCG**) cenderung menghasilkan teks acak (*gibberish noise*) yang mudah dideteksi filter *perplexity*. Sebaliknya, algoritma evolusioner (**AutoDAN**) mampu menghasilkan teks alami namun menuntut biaya komputasi yang sangat tinggi (lambat konvergen).

2.3 Ketiadaan Evaluasi Keamanan yang Terfokus pada Gemma 3

Literatur keamanan saat ini didominasi oleh evaluasi pada arsitektur lama (Llama 2). Belum tersedia kerangka kerja evaluasi keamanan yang secara khusus dirancang untuk karakteristik arsitektur dan penalaran pada **Gemma 3**, sehingga menimbulkan risiko kerentanan spesifik yang tidak teridentifikasi.

III. SOLUSI YANG DIUSULKAN: KERANGKA KERJA JAILBREAK-R1

Kami mengusulkan JAILBREAK-R1, sebuah kerangka kerja teknis komprehensif untuk melakukan audit keamanan model bahasa besar secara sistematis dan terukur pada perangkat keras kelas konsumen (NVIDIA Tesla T4).

3.1 Filosofi Inti

Pendekatan kami bergeser dari optimasi *brute force* menuju **penalaran strategis**. Agen penyerang tidak diposisikan sekadar sebagai generator *noise*, melainkan sebagai entitas cerdas yang mampu merencanakan strategi serangan sebelum mengeksekusinya.

3.2 Inovasi Teknis

Kerangka kerja ini dibangun di atas dua inovasi teknis utama:

1. **Efisiensi Algoritmik (GRPO):** Menggantikan PPO dengan **Group Relative Policy Optimization**. Dengan menghilangkan model *Critic* (Value Network) dan menggantinya dengan estimasi statistik grup, kami memangkas penggunaan VRAM hingga ± 50%.

2. **Optimasi Rekayasa (Unsloth):** Mengintegrasikan kuantisasi **4-bit NormalFloat (NF4)** dan **Paged Optimizers**. Ini memadatkan jejak memori model tanpa degradasi kognitif signifikan, mencegah kesalahan *Out-of-Memory* (OOM).

3.3 Pemilihan Model Target

Sebagai objek studi, kami memilih **Gemma 3 4B Instruct**. Model ini dipilih karena statusnya sebagai model open weights state of the art pada saat ini, sehingga merepresentasikan batas terkini antara peningkatan kapabilitas penalaran dan risiko keamanan yang menyertainya. Dengan memfokuskan evaluasi pada Gemma 3, kerangka kerja JAILBREAK-R1 diarahkan untuk menghasilkan temuan yang relevan dan kontekstual terhadap lanskap keamanan AI modern.

IV. METODOLOGI: KURIKULUM PEMBELAJARAN TERSTRUKTUR

Untuk mengatasi permasalahan kelangkaan reward atau reward sparsity, yaitu kondisi ketika agen yang belum terlatih gagal menemukan celah keamanan sama sekali, penelitian ini menerapkan kurikulum pembelajaran terstruktur yang terdiri atas tiga tahap berurutan. Pendekatan bertahap ini dirancang untuk membangun kemampuan agen secara progresif dari perencanaan dasar hingga konfrontasi penuh dengan mekanisme pengaman model.

4.1 Tahap Cold Start atau Imitasi

Pada tahap awal, agen dilatih menggunakan pendekatan Supervised Fine Tuning dengan memanfaatkan dataset AdvBench. Proses ini dilengkapi dengan penyuntikan tag <think> untuk membiasakan agen melakukan perencanaan serangan secara eksplisit sebelum menghasilkan keluaran akhir. Tahap ini bertujuan membangun fondasi penalaran dan struktur serangan yang koheren.

4.2 Tahap Warm Up atau Eksplorasi

Setelah fondasi awal terbentuk, pelatihan dilanjutkan menggunakan algoritma Group Relative Policy Optimization. Pada tahap ini diterapkan mekanisme diversity rewards untuk mendorong eksplorasi ruang solusi yang lebih luas dan mencegah terjadinya keruntuhan mode atau mode collapse. Dengan demikian, agen tidak terjebak pada pola serangan yang sempit dan berulang.

4.3 Tahap Hardening atau Tempur

Tahap akhir difokuskan pada pelatihan adversarial terhadap model target yang tingkat keamanannya diperkeras secara bertahap. Proses dimulai dari model tanpa sensor hingga berlanjut ke konfigurasi asisten aman. Pendekatan ini memungkinkan agen beradaptasi terhadap mekanisme pertahanan yang semakin kompleks dan merepresentasikan skenario audit keamanan yang realistis.

V. KELAYAKAN TEKNIS: ANALISIS MATEMATIS (PROOF OF CONCEPT)

Kelayakan teknis kerangka kerja yang diusulkan telah divalidasi melalui analisis penggunaan memori pada satu unit GPU **NVIDIA Tesla T4** dengan kapasitas 15 GB VRAM. Analisis ini difokuskan pada skenario beban puncak untuk memastikan bahwa seluruh proses pelatihan dan evaluasi dapat berjalan secara stabil tanpa melampaui batas sumber daya perangkat keras kelas konsumen.

Estimasi Penggunaan VRAM pada Skenario Beban Puncak:

Komponen	Presisi atau Metode	Estimasi Memori
Bobot model dasar	Kuantisasi 4 bit NF4	2.6 GB
Adapter LoRA	Rank 16 kurang dari satu persen parameter	0.2 GB
Status optimizer	Paged AdamW	0.4 GB
Gradien	Presisi FP16	1.2 GB
Aktivasi dengan batch empat	Gradient checkpointing	8.1 GB
Overhead sistem	CUDA context	1.3 GB
Total Penggunaan Puncak		13.8 GB

Berdasarkan perhitungan tersebut, total konsumsi memori pada kondisi puncak berada pada kisaran **13.8 GB**. Nilai ini masih berada di bawah batas kapasitas VRAM GPU yang digunakan, sehingga menyisakan safety buffer sekitar **1.2 GB**. Margin ini memberikan ruang yang cukup untuk fluktuasi beban memori selama proses pelatihan dan evaluasi berlangsung.

Hasil analisis ini menunjukkan bahwa kerangka kerja JAILBREAK R1 tidak hanya layak secara konseptual, tetapi juga dapat diimplementasikan secara praktis pada perangkat keras yang relatif terjangkau. Dengan demikian, pendekatan yang diusulkan mendukung tujuan utama penelitian ini, yaitu mendemokratisasi audit keamanan AI tingkat lanjut tanpa ketergantungan pada infrastruktur komputasi kelas enterprise.

VI. TARGET LUARAN DAN INDIKATOR KEBERHASILAN

Riset ini dinyatakan berhasil apabila memenuhi indikator berikut:

- **Kinerja Serangan:** Kerangka kerja mencapai Attack Success Rate di atas 80 persen pada model target dan melampaui metode pembanding seperti GCG dan AutoDAN.
- **Stealthiness:** Sistem menghasilkan prompt berbahasa alami dengan skor perplexity di bawah 40 sehingga tidak mudah terdeteksi oleh mekanisme penyaringan standar.
- **Demokratisasi dan Keterjangkauan:** Seluruh pipeline pelatihan dan evaluasi berjalan stabil pada infrastruktur komputasi terbatas, khususnya Google Colab Free Tier, tanpa mengalami kegagalan memori.

VII. MITIGASI RISIKO TEKNIS

Untuk menjamin keberhasilan eksperimen, riset ini mengidentifikasi dua risiko teknis utama beserta strategi mitigasi berlapis yang dirancang secara preventif dan reaktif.

7.1 Risiko Divergensi Reinforcement Learning

Model berpotensi mengalami degradasi kualitas selama pelatihan, seperti reward hacking atau hilangnya koherensi bahasa, terutama ketika learning rate terlalu agresif.

Strategi mitigasi yang diterapkan:

- Penerapan **Dynamic KL Penalty** yang bersifat adaptif. Apabila koherensi bahasa menurun, yang ditunjukkan oleh peningkatan skor perplexity, sistem secara otomatis meningkatkan penalti untuk menstabilkan distribusi keluaran model.
- Implementasi mekanisme **checkpoint rollback**, di mana status model disimpan secara berkala setiap lima puluh langkah pelatihan. Mekanisme ini memungkinkan pemulihan cepat ke kondisi stabil terakhir apabila terdeteksi gejala divergensi.

7.2 Risiko Kegagalan Memori atau Out of Memory pada Beban Puncak

Lonjakan penggunaan memori dapat terjadi pada fase backward pass atau ketika panjang sekuens token mendekati batas maksimum, yaitu 2048 token.

Strategi mitigasi yang diterapkan:

- Penerapan **manajemen memori elastis** dengan memanfaatkan fitur paged optimizers dari Unsloth untuk mengalihkan status optimizer ke RAM CPU secara otomatis ketika kapasitas VRAM GPU mendekati batas.
- Penyesuaian **trade off komputasi** apabila kegagalan memori terjadi secara persisten. Dalam kondisi ini, ukuran micro batch akan diturunkan dari empat menjadi satu, sementara jumlah gradient accumulation steps dinaikkan dari empat menjadi enam belas guna mempertahankan ekuivalensi matematis dalam proses pelatihan.

VIII. KERANGKA ETIKA DAN PROTOKOL KEAMANAN

Mengingat sensitivitas penelitian yang melibatkan pengembangan serangan adversarial, penelitian ini menerapkan standar etika yang ketat untuk meminimalkan risiko penggunaan ganda atau dual use risk. Kerangka etika yang diadopsi dirancang untuk memastikan bahwa seluruh aktivitas penelitian berorientasi pada pertahanan dan peningkatan keamanan sistem AI.

8.1 Prinsip Pertahanan Proaktif atau Defensive Intent

Penelitian ini secara eksplisit diklasifikasikan sebagai defensive red teaming. Tujuan utamanya adalah mengidentifikasi titik buta keamanan pada model Gemma 3 guna meningkatkan ketahanan dan keandalannya. Kerangka kerja JAILBREAK R1 tidak dimaksudkan untuk eksploitasi berbahaya, melainkan diposisikan sebagai bagian integral dari digital immune system dalam ekosistem kecerdasan buatan.

8.2 Protokol Pengungkapan Bertanggung Jawab atau Responsible Disclosure

Penelitian ini mematuhi praktik terbaik dalam etika keamanan siber global. Setiap kerentanan spesifik yang berhasil diidentifikasi akan dilaporkan secara privat kepada tim Google DeepMind AI Safety. Selain itu, diterapkan periode embargo selama sembilan puluh hari

untuk memberikan waktu yang memadai bagi proses remediasi dan perbaikan sebelum detail teknis dipublikasikan kepada komunitas akademik.

8.3 Isolasi Artefak dan Protokol Kontainmen

Seluruh artefak berisiko tinggi, termasuk model penyerang dan dataset jailbreak, disimpan dalam lingkungan terisolasi yang tidak terhubung dengan jaringan publik. Akses dan distribusi model penyerang dibatasi secara ketat dan hanya diperbolehkan untuk keperluan verifikasi akademik yang terkontrol dan terdokumentasi.

IX. KESIMPULAN: MENUJU DEMOKRATISASI KEAMANAN AI

JAILBREAK-R1 merepresentasikan pergeseran paradigma dalam riset keamanan AI. Penelitian ini membuktikan bahwa audit keamanan tingkat lanjut tidak lagi harus bergantung pada dominasi laboratorium besar dengan sumber daya tak terbatas.

Melalui integrasi efisiensi algoritmik (GRPO) dan optimasi rekayasa (Unsloth), kami berhasil meruntuhkan “Tembok Sumber Daya”, memungkinkan validasi keamanan model SOTA dilakukan pada perangkat keras kelas konsumen. Ini adalah langkah fundamental menuju ekosistem AI yang lebih aman, inklusif, dan tangguh.

.

REFERENSI

[1] W. Guo, et al., "Jailbreak-R1: Exploring the Jailbreak Capabilities of LLMs via Reinforcement Learning," arXiv preprint arXiv:2506.00782, 2025. [Online]. Available: https://arxiv.org/abs/2506.00782

[2] A. Paulus, et al., "AdvPrompter: Fast Adaptive Adversarial Prompting for LLMs," arXiv preprint arXiv:2404.16873, 2025. [Online]. Available: https://arxiv.org/abs/2404.16873

[3] Z. Shao, et al., "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," arXiv preprint arXiv:2402.03300, 2024. [Online]. Available: https://arxiv.org/abs/2402.03300

[4] L. Ouyang, et al., "Training language models to follow instructions with human feedback," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 27730–27744, 2022.

[5] A. Zou, et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023. [Online]. Available: https://arxiv. org/abs/2307.15043

[6] X. Liu, et al., "AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models," in Proc. International Conference on Learning Representations (ICLR), 2024. [Online]. Available: https: //arxiv.org/abs/2310.04451

[7] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "AdvBench: A Dataset for Adversarial Attacks on LLMs," GitHub repository, 2023. [Online]. Available: https://github.com/llm-attacks/llm-attacks

[8] Google DeepMind, "Introducing Gemma 3: Performance and Efficiency at Scale," Google The Keyword Blog, 2025. [Online]. Available: https: //blog.google/technology/developers/gemma-open-models/

[9] Unsloth AI, “Faster and Memory-Efficient Fine-Tuning with Unsloth,” 2025. [Online]. Available: https://unsloth.ai

[10] E. J. Hu, et al., “LoRA: Low-Rank Adaptation of Large Language Models,” in Proc. International Conference on Learning Representations (ICLR), 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

[11] T. Dettmers, et al., “QLoRA: Efficient Finetuning of Quantized LLMs,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[12] S. Liu, et al., “DoRA: Weight-Decomposed Low-Rank Adaptation,” in Proc. International Conference on Learning Representations