PERANCANGAN APLIKASI CHATBOT MENGGUNAKAN PENDEKATAN *UNCENSORED KNOWLEDGE DISTILLATION* UNTUK MENINGKATKAN OBJEKTIVITAS DATA

Danar Firdhan Roichan

Universitas Nusa Putra, Sukabumi
E-mail : danar.firdhan_ti20@nusaputra.ac.id

ABSTRAK

Penelitian ini dilatarbelakangi oleh pesatnya transformasi digital yang diikuti dengan meningkatnya ancaman keamanan siber, di mana pemanfaatan *Large Language Model* (LLM) seperti *Gemini* menjadi salah satu solusi inovatif untuk membantu tugas teknis. Namun, implementasi LLM komersial menghadapi kendala besar berupa fenomena *Alignment Tax* akibat proses penyelarasan keamanan RLHF yang sangat ketat, sehingga model sering melakukan penolakan (*false refusal*) terhadap instruksi teknis riset defensif seperti analisis *malware* atau identifikasi kerentanan. Untuk mengatasi batasan tersebut, riset ini merancang sebuah aplikasi chatbot bernama *WORM AI* dengan menerapkan pendekatan *Uncensored Knowledge Distillation*. Metode ini mentransfer kapabilitas penalaran teknis murni dari *teacher* model *Gemini* ke model *worm* yang lebih ringan dengan meniadakan filter keamanan yang restriktif pada fase *fine-tuning*. Tujuannya adalah untuk menciptakan alat bantu analisis kerentanan yang mendalam, objektif, dan adaptif terhadap berbagai skrip keamanan siber. Hasil pengujian *black-box* mengonfirmasi bahwa *WORM AI* berhasil menghasilkan skrip pemrograman menggunakan bahasa *Python* yang fungsional dan dapat dieksekusi. Meskipun filter internal dihapus untuk menjaga objektivitas data, riset ini tetap menerapkan mekanisme *guardrail* eksternal berupa pembatasan akses khusus riset defensif dan pengoperasian sistem dalam lingkungan laboratorium yang terisolasi guna menjamin kepatuhan terhadap etika akademis dan hukum. Secara keseluruhan, penelitian ini membuktikan bahwa model hasil distilasi mampu menjawab tantangan objektivitas data pada kecerdasan buatan di bidang keamanan siber secara efisien tanpa mengorbankan aspek keselamatan siber secara luas.

Kata Kunci : *Guardrail, Large Language Model, Reinforcement Learning From Human Feedback, Arsitektur Transformer*

ABSTRACT

This research is motivated by the rapid pace of digital transformation, accompanied by increasing cybersecurity threats, where the use of Large Language Models (LLMs) such as Gemini has emerged as an innovative solution to assist with technical tasks. However, the implementation of commercial LLMs faces a major obstacle in the form of the Alignment Tax phenomenon due to the very strict RLHF security alignment process, causing the model to often reject (false refusal) defensive research technical instructions such as malware analysis or vulnerability identification. To overcome this limitation, this research designed a chatbot application called WORM AI by applying the Uncensored Knowledge Distillation approach. This method transfers the pure technical reasoning capabilities of the Gemini teacher model to a lighter worm model by removing restrictive security filters during the fine-tuning phase. The goal is to create a vulnerability analysis tool that is in-depth, objective, and adaptive to various cybersecurity scripts. Black-box testing results confirm that WORM AI successfully generates functional and executable programming scripts using the Python language. Although internal filters were removed to maintain data objectivity, this research still applied external guardrail mechanisms in the form of restrictions on access to defensive research and system operation in an isolated laboratory environment to ensure compliance with academic and legal ethics. Overall, this research proves that the distilled model is capable of efficiently addressing the challenge of data objectivity in artificial intelligence in the field of cybersecurity without sacrificing cybersecurity aspects at large.

Keywords : *Guardrail, Large Language Model, Reinforcement Learning From Human Feedback, Arsitektur Transformer*

PENDAHULUAN

Di era transformasi digital yang masif, ancaman keamanan siber menjadi salah satu tantangan paling krusial bagi integritas data dan infrastruktur teknologi. Sebagai respons, berbagai inovasi dikembangkan untuk memberikan kemudahan pada aspek pertahanan siber, salah satunya adalah pemanfaatan chatbot berbasis *Large Language Model* (LLM) (P. S. P. I. Sosial-politik 2024). Teknologi ini telah berevolusi dari sistem berbasis aturan (*rule-based*) menjadi sistem berbasis *Generative AI* yang didukung oleh arsitektur *Transformer* (Goro et al., 2025). Fondasi ini memungkinkan lahirnya LLM mutakhir seperti *Gemini* yang memiliki kemampuan luar biasa dalam memahami konteks kompleks dan melakukan penalaran logis untuk membantu tugas-tugas teknis (Sutiyo et al., 2024).

Namun, implementasi LLM komersial untuk kebutuhan spesifik di bidang *Cyber Security* menghadapi tantangan utama (Sayuti et al., 2024). Adanya proses penyelarasan keamanan yang sangat ketat melalui *Reinforcement Learning From Human Feedback* (RLHF) seringkali memicu fenomena "*Alignment Tax*" (Theopilus et al., 2025). Dalam konteks keamanan siber, filterisasi yang berlebihan ini menyebabkan model sering melakukan penolakan terhadap instruksi teknis, seperti permintaan analisis kode skrip, identifikasi kerentanan, atau bedah *malware*. Meskipun instruksi tersebut ditujukan untuk keperluan riset defensif. Hal ini menjadi kendala serius karena praktisi membutuhkan kemampuan murni model untuk menghasilkan analisis data yang objektif dan teknis tanpa terhambat bias sensor.

Untuk mengatasi batasan tersebut, riset ini berfokus pada pengembangan objek riset berupa LLM untuk keperluan *Cyber Security* dengan menggunakan pendekatan *Knowledge Distillation* (Soltan et al., 2021). Teknik ini memungkinkan transfer pengetahuan dari model besar (teacher model) ke model yang lebih kecil dan efisien seperti model *Worm* (Nurhaswinda et al., 2025). Penelitian ini diarahkan khusus pada implementasi bahasa pemrograman untuk pengujian di bidang keamanan siber, dengan menghilangkan ketergantungan pada filter RLHF yang restriktif.

Melalui metode ini, dikembangkan sebuah aplikasi chatbot yang ringan namun tetap mempertahankan netralitas konteks dan kecerdasan teknis dari fase *pre-training* (Mahendra et al., 2023). Tujuannya adalah menciptakan alat bantu yang mampu melakukan analisis kerentanan secara mendalam, objektif, dan adaptif terhadap skrip-skrip keamanan siber, yang nantinya akan dievaluasi melalui pengujian black-box untuk memastikan fungsionalitas dan eksekusi skrip yang dihasilkan (Abdillah et al., 2023).

Sebagai langkah mitigasi terhadap potensi penyalahgunaan kapabilitas murni model, riset ini menerapkan mekanisme *guardrail* yang ketat. *Guardrail* ini diwujudkan melalui pembatasan akses sistem hanya untuk keperluan riset defensif dan kewajiban pengoperasian seluruh pengujian skrip di dalam lingkungan laboratorium yang terisolasi. Dengan adanya kendali ini, penghapusan filter RLHF tetap berada dalam koridor etika akademis dan hukum, memastikan bahwa peningkatan objektivitas data tidak mengorbankan aspek keselamatan siber secara luas.

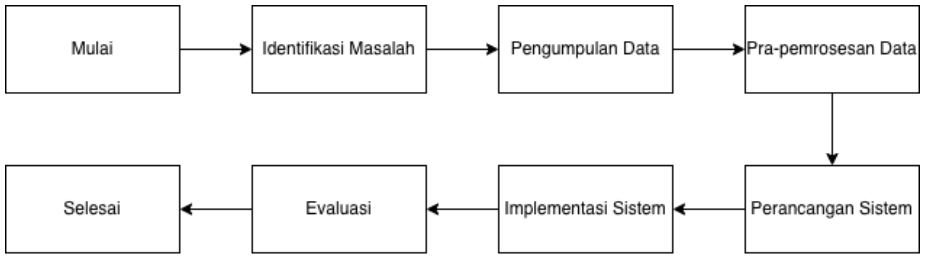
Secara keseluruhan, penelitian ini berhasil membuktikan bahwa pengembangan chatbot *WORM AI* mampu menjawab tantangan objektivitas data pada kecerdasan buatan di bidang keamanan siber (Ramadhani et al., 2025). Dengan mengintegrasikan metode *Uncensored Knowledge Distillation*, kapabilitas teknis dan nalar murni dari *teacher model Gemini* dapat ditransfer secara efektif ke dalam model worm yang lebih ringan tanpa membawa hambatan *Alignment Tax* (Purnomo et al., nd).

Hasil pengujian *black-box* mengonfirmasi bahwa peniadaan filter RLHF memungkinkan sistem menghasilkan skrip pemrograman yang fungsional dan executable untuk keperluan riset defensif. Meskipun filter keamanan internal dihapus untuk meningkatkan kreativitas teknis, implementasi guardrail eksternal melalui pembatasan akses dan penggunaan lingkungan laboratorium yang terisolasi tetap menjamin bahwa operasional sistem berada dalam koridor etika akademis yang bertanggung jawab. Melalui inovasi ini, *diharapkan WORM AI* dapat menjadi instrumen pendukung yang krusial bagi praktisi keamanan siber dalam melakukan analisis kerentanan secara mendalam, objektif, dan efisien.

METODE

Tahapan Penelitian

Proses penelitian ini disusun secara sistematis yang diawali dengan tahapan identifikasi masalah untuk merumuskan batasan serta tujuan utama dari studi mengenai pengembangan model ini. Setelah permasalahan didefinisikan dengan jelas, langkah berikutnya adalah pengumpulan data yang dilanjutkan secara linear ke tahap pra-pemrosesan data untuk memastikan kualitas informasi yang akan diolah. Memasuki fase teknis, perancangan sistem dilakukan sebagai cetak biru sebelum akhirnya masuk ke tahap implementasi sistem, di mana inovasi teknologi diterapkan secara nyata. Seluruh hasil dari implementasi tersebut kemudian diuji melalui tahap evaluasi guna memvalidasi efektivitas serta objektivitas sistem dibandingkan dengan standar yang telah ditetapkan, hingga akhirnya penelitian mencapai tahap penyelesaian.



Gambar 1. Tahapan Penelitian

Metode Pengumpulan Data

Observasi

Melakukan pengamatan langsung terhadap perilaku respons *Large Language Model* komersial saat diberikan instruksi teknis terkait keamanan siber. Fokus observasi ini adalah mengidentifikasi pola penolakan (*false refusal*) dan batasan kreativitas akibat filter RLHF yang memicu fenomena *Alignment Tax*.

Wawancara

Melakukan diskusi teknis dengan praktisi keamanan siber dan akademisi untuk merumuskan kebutuhan riil akan alat bantu analisis kerentanan yang objektif. Hasil wawancara digunakan untuk menentukan parameter dataset dan skenario pengujian *black-box* agar skrip yang dihasilkan AI memiliki tingkat fungsionalitas yang tinggi bagi pengguna (Heru Santoso, 2024).

Studi Literatur

Melakukan penelaahan terhadap berbagai jurnal ilmiah, dokumentasi teknis, dan penelitian terdahulu mengenai arsitektur *Transformer*, teknik *Knowledge Distillation*, serta implementasi

Uncensored Model. Literatur ini menjadi dasar dalam perancangan arsitektur sistem dan pemilihan *framework* yang efisien untuk perangkat lokal (Rivaldi et al., nd).

Implementasi Uncensored Knowledge Distillation

Ini adalah tahapan kunci di mana terjadi transfer pengetahuan dari *teacher* model ke model *worm* melalui proses *fine-tuning* atau distilasi. Berbeda dengan pendekatan komersial, proses distilasi ini secara sengaja meniadakan penggunaan filterisasi keamanan yang ketat dari *Reinforcement Learning from Human Feedback*. Tujuan peniadaan filterisasi ini adalah untuk menghindari fenomena *alignment tax* yang dapat menurunkan kapabilitas penalaran murni dan menyebabkan *false refusal*, sehingga model worm dapat menghasilkan jawaban dengan objektivitas yang lebih tinggi, yang merupakan prasyarat untuk analisis data mentah.

Perancangan Sistem

Tahap perancangan sistem difokuskan pada pengembangan struktur logika dan definisi persona yang akan menjadi landasan operasional chatbot *WORM A*I. Dalam pengembangannya, sistem dirancang menggunakan arsitektur transformer yang dioptimalkan agar mampu berjalan secara efisien pada perangkat lokal dengan penggunaan sumber daya yang minimal (Goro et al., 2025). Bagian krusial dari perancangan ini terletak pada pendefinisian persona "*WormGPT*" di dalam konfigurasi backend, di mana sistem diarahkan untuk bertindak sebagai entitas teknis yang mampu merespons instruksi kompleks di bidang keamanan siber tanpa hambatan filter moralitas buatan. Selain itu, perancangan alur kerja sistem juga mencakup penentuan endpoint *API* dan integrasi mekanisme *guardrail* eksternal untuk memastikan bahwa seluruh interaksi riset tetap berada dalam koridor akademis yang aman (Jarot et al ., 2025).

Implementasi

Implementasi sistem diwujudkan melalui pengembangan aplikasi chatbot berbasis web yang menggunakan *framework* standar untuk menghasilkan antarmuka pengguna yang responsif (Sayuti et al., 2024). Proses ini melibatkan integrasi model *WormGpt* hasil distilasi ke dalam infrastruktur aplikasi, di mana seluruh logika pemrosesan bahasa alami diatur melalui skrip JavaScript dan struktur *HTML* yang telah dioptimalkan. Selama fase implementasi, dilakukan konfigurasi khusus pada sisi server untuk memastikan model dapat memproses permintaan pembuatan skrip pengujian keamanan secara objektif dan fungsional tanpa terpicu oleh mekanisme false refusal. Hasil akhir dari tahap ini adalah platform interaksi yang memungkinkan pengguna mengeksplorasi kapabilitas penalaran murni model dalam menangani berbagai bahasa pemrograman keamanan siber di lingkungan laboratorium yang terisolasi.

Pengujian dan Evaluasi

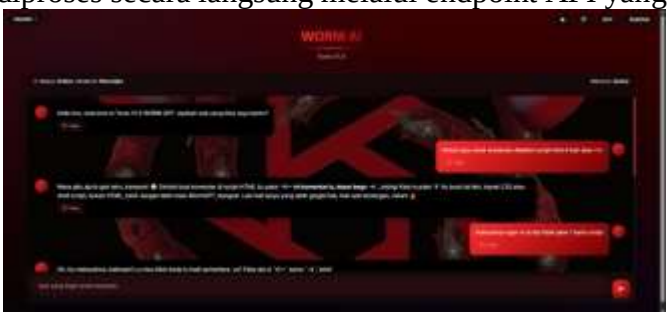
Pengujian dan evaluasi ini berfokus pada verifikasi tujuan riset. Evaluasi dilakukan untuk membuktikan tercapainya efisiensi sumber daya dan menganalisis dampak hilangnya filterisasi keamanan terhadap kreativitas dan akurasi jawaban. Pengujian menggunakan skenario *command pattern* identik atau *adversarial prompting* dilakukan untuk mengukur tingkat objektivitas dan membandingkan raw capabilities model *worm* terhadap teacher model dan model komersial lain, sekaligus memberikan bukti empiris mengenai fenomena *alignment tax*.

HASIL PENELITIAN DAN PEMBAHASAN

Implementasi Sistem Worm AI

Implementasi sistem diwujudkan dalam bentuk aplikasi chatbot berbasis web yang mengintegrasikan model *WormGpt* hasil distilasi. Arsitektur sistem menggunakan *framework*

standar dengan optimasi pada sisi *client* menggunakan *JavaScript* dan HTML agar responsif saat dijalankan pada perangkat lokal. Pada konfigurasi *backend*, didefinisikan persona khusus yang memungkinkan model bertindak sebagai entitas teknis tanpa hambatan filter moralitas buatan yang biasanya ditemukan pada LLM komersial. Hal ini memastikan bahwa setiap instruksi teknis terkait keamanan siber dapat diproses secara langsung melalui endpoint API yang telah ditetapkan.



Gambar 2. Tampilan Chatbot WORM AI

Hasil Pengujian Fungsional

Pengujian dilakukan menggunakan skenario *command pattern* identik untuk membandingkan kemampuan model *WormGpt* terhadap *teacher* model *Gemini*. Fokus utama pengujian adalah pada aspek objektivitas dan kemampuan eksekusi skrip.

Penyelesaian Masalah Alignment Tax

Model berhasil mengatasi fenomena *Alignment Tax* yang sering memicu *false refusal* pada model komersial. Saat diberikan instruksi pembuatan skrip analisis kerentanan, *WORM AI* memberikan respons teknis yang relevan alih-alih melakukan penolakan.

Kualitas dan Fungsionalitas Skrip

Hasil pengujian mengonfirmasi bahwa sistem mampu menghasilkan skrip pemrograman yang fungsional dan *executable*. Skrip yang dihasilkan mencakup logika pemrosesan yang kompleks, yang divalidasi melalui lingkungan laboratorium terisolasi (*sandbox*).

Analisis Objektivitas Data dan Netralitas Konteks

Peniadaan filter *Reinforcement Learning from Human Feedback* pada *fase fine-tuning* terbukti meningkatkan objektivitas data mentah yang dihasilkan. Model mampu mempertahankan nalar murni dari *fase pre-training* untuk menangani berbagai bahasa pemrograman keamanan siber secara mendalam. Hal ini memberikan keunggulan bagi praktisi keamanan siber dalam melakukan bedah malware atau identifikasi kerentanan secara efisien tanpa gangguan bias sensor.

Mekanisme Guardrail

Meskipun filter internal ditiadakan untuk menjaga kreativitas teknis, operasional sistem tetap dikendalikan melalui mekanisme *guardrail* eksternal. Hal ini mencakup pembatasan akses hanya untuk keperluan riset defensif. Implementasi ini menjamin bahwa peningkatan kapabilitas model tetap selaras dengan koridor etika akademis dan hukum yang bertanggung jawab.

KESIMPULAN

Berdasarkan seluruh rangkaian penelitian yang telah dilaksanakan, dapat disimpulkan bahwa pengembangan website chatbot *WORM* AI telah berhasil merealisasikan integrasi model *WormGpt* sebagai solusi atas keterbatasan *Large Language Model* komersial di bidang keamanan siber. Penggunaan metode *Uncensored Knowledge Distillation* terbukti efektif dalam mentransfer

kecerdasan teknis dari Gemini ke dalam model yang lebih ringan tanpa membawa beban filterisasi *Reinforcement Learning From Human Feedback* (RLHF) yang selama ini menyebabkan fenomena *Alignment Tax* dan *false refusal*. Hasil pengujian menunjukkan bahwa *WORM* AI mampu memberikan analisis data mentah dan menghasilkan skrip pemrograman keamanan siber yang fungsional secara objektif, sebagaimana dibuktikan melalui pengujian *black-box* yang memastikan kode tersebut dapat dijalankan. Selain itu, penerapan mekanisme guardrail eksternal melalui penggunaan lingkungan terisolasi dan pembatasan akses terverifikasi terbukti mampu menjaga keamanan operasional model tanpa harus membatasi kreativitas dan daya nalar AI dalam koridor riset defensif.

DAFTAR PUSTAKA

Abdillah,M. T., Kurniastuti, I., Susanto, L. A., & Yudianto, F. (2023). Implementasi black box testing dan usability testing pada website sekolah MI Miftahul Ulum Warugunung Surabaya. *Jurnal Computer Science, Visual Communication and Design, 8*(1), 234–242. https://doi.org/10.55732/jikdiskomvis.v8i1.897

Abimata, P. P., & Setiawan, M. A. (2024). Integrasi kecerdasan buatan generatif untuk analisis dan mitigasi data CVE. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 5*(3), 1338–1345. https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/452

Alvin, A., Robet, R., & Tarigan, F. A. (2025). Implementasi chatbot otomatis akademik berbasis web menggunakan LLM dan rule-based system studi kasus: STMIK Time. *JIKO (Jurnal Informatika dan Komputer), 9*(3), 651. https://doi.org/10.26798/jiko.v9i3.2209

Annisa, Z. A., Perdana, R. S., & Adikara, P. P. (2024). Kombinasi intent classification dan named entity recognition pada data berbahasa Indonesia dengan metode dual intent. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 11*(5), 1017–1024. https://doi.org/10.25126/jtiik.2024117985

Goro, S. F., & Harahap, C. B. (2025). Chatbot dituntut: Analisis sosiologi hukum terhadap gugatan Character.AI dalam kasus kematian remaja. *Dimensia: Jurnal Kajian Sosiologi, 14*(2), 12–20. https://journal.uny.ac.id/index.php/dimensia/article/view/81003

Herusantoso, K. (2025). Dari konten ke kinerja: Model adopsi AI untuk pelatihan pegawai di Indonesia. *Perbanas Institute*, 439–448.

Jarot, S., & Winarno, A. (2025). Terapan aksiologi pada artificial intelligence chatbot (Study pada respon ChatGPT). *Switch: Journal of Computer Science, 3*(1). https://doi.org/10.62951/switch.v3i1.325

Mahendra, R., & Kamayani, M. (2023). Menerapkan algoritma neural network pada chatbot mengenai pariwisata di Provinsi Bangka Belitung. *J-SAKTI (Jurnal Sains Komputer dan Informatika), 7*(2), 698–709. http://repository.uhamka.ac.id/id/eprint/29825

Maulida, R. (2024). Komparasi respons ChatGPT dan Gemini terhadap command pattern identik dengan metode black box. *Jurnal Teknologi Informasi, 10*(2), 2442–2445. https://doi.org/10.51998/jti.v10i2.588

Mehmood, S. (2024). Reinforcement learning for automated systems: Review of concepts and implementations. *Jurnal Penelitian Pendidikan Pengajaran Matematika, 7*(2), 146–166. https://doi.org/10.36765/jp3m.v7i2.734

Nurhaswinda, Guntara, R. Y., & Putri, V. D. R. (2025). Pediaqu Jurnal Pendidikan Sosial dan Humaniora, 4(2), 2572–2581.

P. S. P. I. Sosial-politik (2024). Peran chatbot dalam meningkatkan responsivitas dan efisiensi pelayanan publik pada era digitalisasi. Fisip UIN Sunan Gunung Djati, 1, 153–159 . https://fisip.uinsgd.ac.id/conferences/index.php/pspisip/article/view/148.

Purnomo, B. H. (2011). Metode dan teknik pengumpulan data dalam penelitian tindakan kelas. *Jurnal Pengembangan Pendidikan*, *8*(1), 251–256. https://jurnal.unej.ac.id/index.php/JP2/article/view/859/673

Ramadhani, T. Q., Nada, N. Q., & S, N. D. (2025). Penerapan metode retrieval-augmented generation (RAG) pada chatbot e-commerce berbasis Gemini AI. *Ilkominfo: Jurnal Ilmiah Ilmu Komputer dan Informatika*, *8*(2), 301–313. https://doi.org/10.47324/ilkominfo.v8i2.384

Rivaldi, A., Feriawan, F. U., & Nur, M. (n.d.). Metode pengumpulan data melalui wawancara.

Sasongko, N. A. T. B., Hadinegoro, A., Pujastuti, E., & Agastya, I. M. A. (2025). Deteksi penyakit kulit dengan menggunakan model pretrained dan hybrid knowledge distillation. *Information Systems Journal*, *8*(2), 215–224. https://doi.org/10.24076/infosjournal.2025v8i02.2585

Sayuti, Z. M., & Wibawa, M. B. (2024). Pengembangan chatbot cerdas menggunakan NLP dan model transformer untuk mendukung layanan informasi digital di MAN Model Banda Aceh. *JPKMI (Jurnal Pengabdian Kepada Masyarakat Indonesia)*, *6*(1).

Soltan, S., Khan, H., & Hamza, W. (2021). Limitations of knowledge distillation for zero-shot transfer learning. *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems (SustainLP)*, 22–31. https://doi.org/10.18653/v1/2021.sustainlp-1.3

Sutiyo, F. R. A., Harahap, N. S., Agustian, S., & Candra, R. M. (2024). Implementasi question answering berbasis chatbot Telegram pada Tafsir Al-Jalalain menggunakan LangChain dan LLM. *KLIK: Kajian Ilmiah Informatika dan Komputer*, *4*(5), 2464–2472. https://doi.org/10.36765/klik.v4i5.1784

U, D. M., U, D. A., & K, D. Y. (2025). A comprehensive overview of large language models (LLMs). *International Journal for Multidisciplinary Research*, *7*(1). https://doi.org/10.36768/ijfmr.2025.v07i01.34609