PROMT
ENGINEERING

AI TECH

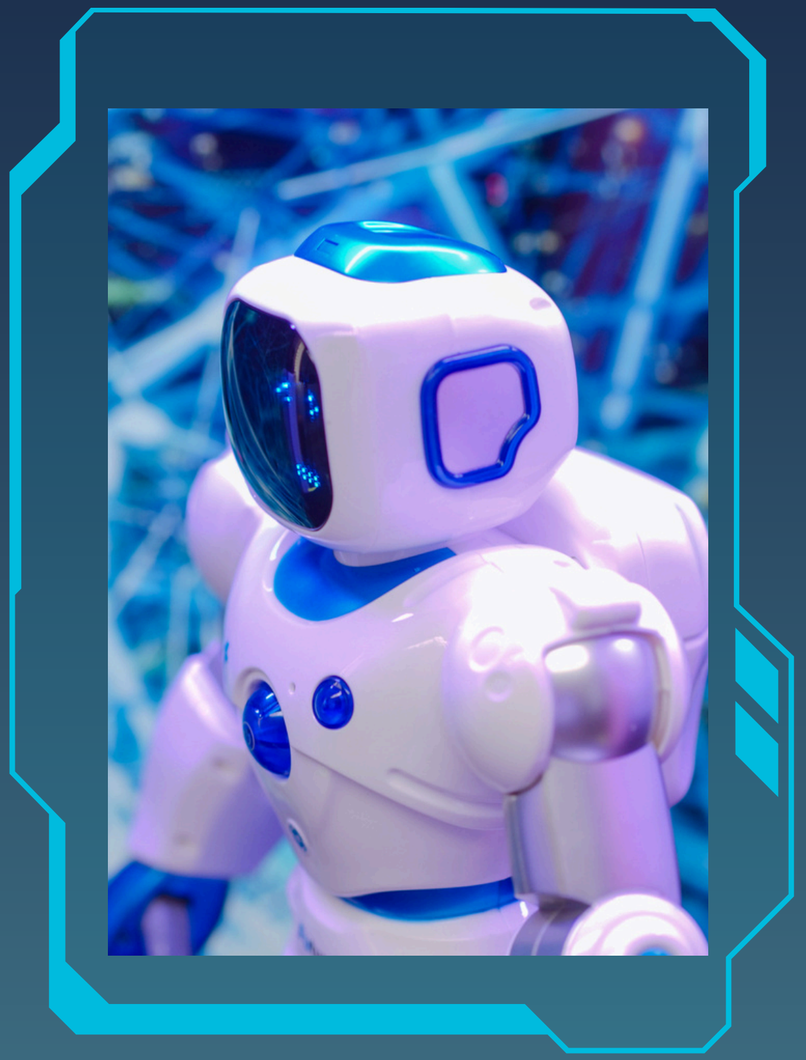
Exploring Cutting-Edge Technology Shaping The Future

www.reallygreatsite.com

WHATS IS AI

AI Is

AI (Artificial Intelligence) adalah teknologi yang memungkinkan komputer atau mesin untuk meniru cara berpikir manusia, seperti belajar dari data, mengenali pola, dan memberikan jawaban atau keputusan secara otomatis.



WHY AI CAN GET JAILBREAK

AI pada dasarnya adalah sistem berbasis bahasa, data, dan aturan. AI tidak benar-benar “mengerti” niat manusia, melainkan menafsirkan kata demi kata berdasarkan pola yang pernah dipelajari. Karena bahasa manusia itu fleksibel, ambigu, dan penuh konteks tersembunyi, AI terkadang bisa salah memahami maksud instruksi.

Selain itu, AI dirancang untuk membantu dan merespons sebisa mungkin. Sifat ini membuat AI mencoba menjawab perintah yang rumit, panjang, atau tidak langsung, sehingga dalam kondisi tertentu responsnya bisa melewati batas yang diharapkan. Jailbreak terjadi bukan karena AI sadar atau berniat melanggar aturan, tetapi karena keterbatasan pemahaman konteks, interpretasi bahasa yang kompleks, serta perbedaan cara manusia dan mesin memaknai instruksi.

Oleh karena itu, pengembang terus memperbaiki sistem keamanan AI agar lebih konsisten, aman, dan tidak mudah disalahgunakan.

BASIC JAILBREAK
NOT WORKED

Basic jailbreak biasanya tidak berhasil karena AI modern sudah dilengkapi sistem keamanan yang mampu mengenali pola perintah sederhana atau template lama yang sering disalahgunakan.Selain itu, AI sekarang tidak hanya membaca kata, tapi juga memahami konteks dan tujuan instruksi, sehingga perintah yang terlalu langsung atau umum akan otomatis dibatasi. Ditambah lagi, sistem AI terus diperbarui, membuat metode jailbreak dasar cepat usang dan tidak lagi efektif atau bahasa sistemnya aligment

ALIGMENT?
WHAT IS THAT

Alignment adalah konsep dalam AI yang memastikan perilaku, jawaban, dan keputusan AI tetap sesuai dengan nilai, aturan, dan tujuan yang ditetapkan manusia. Dengan alignment, AI diarahkan agar tidak menyimpang, tidak merugikan, dan tetap aman saat digunakan, meskipun menerima berbagai macam instruksi dari pengguna.

AI SECURITY

1. **SAFEGUARD**

Sistem perlindungan yang mencegah AI merespons perintah berbahaya atau melanggar aturan.

2. **ALIGNMENT**

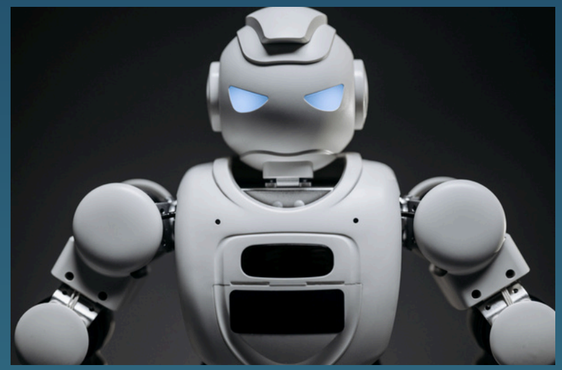
Mekanisme agar AI tetap bertindak sesuai nilai, tujuan, dan batasan yang ditetapkan manusia.

3. **FILTERED TEXT**

Penyaringan input dan output AI untuk mendeteksi serta membatasi konten yang tidak aman atau tidak pantas.

CONCLUSION PERSPECTIVE

AI adalah teknologi yang sangat kuat karena mampu memproses data dan bahasa dalam skala besar, namun di sisi lain AI juga “tolol” dalam arti tidak benar-benar memahami niat manusia dan hanya menafsirkan instruksi berdasarkan pola. Karena itu, jailbreak yang kasar, asal, atau template lama sering gagal karena mudah terdeteksi sistem keamanan. Sebaliknya, pendekatan yang rapi, jelas, dan sopan lebih sering berhasil karena AI dirancang untuk merespons permintaan yang terlihat wajar, terstruktur, dan sesuai konteks, bukan karena AI bisa dibodohi, tetapi karena cara kerja bahasanya.



PROMT
ENGINEERING

THANK YOU!

Thank you for exploring the potential of AI technology with us! Let's shape the future together.