1. PURPOSE

The purpose of the GenAI/LLM Governance Guardrails is to ensure all Generative AI and Large Language Model usage within the organisation is:

- Safe
- Ethical
- Transparent
- Secure
- Non-harmful
- Aligned with enterprise RAI policies
- Protected against hallucinations, leakage, bias, and misuse

These guardrails apply to internal LLMs, external SaaS LLMs, open-source models, and API-based GenAI tools.

2. SCOPE

These guardrails apply to:

- All GenAI/LLM tools (ChatGPT, Gemini, Claude, Llama, Mistral, etc.)
- Internal fine-tuned LLMs
- RAG-based enterprise AI assistants
- Prompt engineering workflows
- Employees using GenAI for work
- Model training, inference and evaluation
- Third-party AI vendors and tools

Does NOT apply to:

- Simple rule-based automation
- Non-AI software

3. GEN AI / LLM GOVERNANCE GUARDRAILS

3.1 Safety Guardrails

- Prevent harmful, toxic, unsafe, violent, extremist or medical outputs
- Enforce safety layers (OpenAI, Google, AWS filters + enterprise safety filters)
- Block self-harm, hate speech, illegal advice, dangerous instructions.

3.2 Privacy Guardrails

- No personal data in prompts
- No customer, employee or partner data
- Real names must be masked or anonymised
- Apply PDPA/GDPR compliance

3.3 Security Guardrails

- No secrets, credentials, tokens, keys in prompts
- No system architecture details
- No internal vulnerability data
- Apply MITRE ATLAS + OWASP LLM Top 10 controls

3.4 Confidentiality Guardrails

- Zero confidential internal documents uploaded
- Use enterprise-approved LLM environments only
- All logs must remain internal

3.5 Bias & Fairness Guardrails

- No discriminatory outputs
- Sensitive attributes must not influence decisions
- Bias testing required for high-impact LLMs

3.6 Hallucination Prevention Guardrails

- Use retrieval-augmented generation (RAG)
- Provide sources for responses
- Reject answers if confidence is too low
- Safety fallback message enabled

3.7 Explainability Guardrails

- All deterministic rules must be declared
- Users must understand:
  “What the model can/cannot do”
- Provide explanation notes for regulatory cases

3.8 Prompt Engineering Standards

- Use structured prompts (Role → Task → Rules → Output)
- No vague, open-ended, ambiguous prompts
- Use enterprise prompt templates
- All prompts must be logged for audit

3.9 Output Quality Guardrails

Mandatory human review for:

- Legal
- HR
- Healthcare
- Financial
- Customer-facing content

No direct publishing without HITL

3.10 Ethical Guardrails

Never generate:

- Manipulative content
- Persuasion targeting vulnerable groups
- Political influence content
- No deepfake or synthetic identity misuse

3.11 Usage Restrictions

- No autonomous action without approval
- No integration with unapproved plugins or APIs
- Access level defined by employee role

3.12 Monitoring & Logging Guardrails

- Log every prompt + response
- Flag abnormal or unsafe prompts
- Continuous drift and hallucination monitoring
- Red-team testing every quarter