Assessment Output

Assessment Output

Assessment Brief: Secure RAG Implementation (Assignment 2)

Unit: LLM Security and Governance
Weighting: 30%
Submission Requirements:

* A written report detailing the implementation of a secure Retrieval-Augmented Generation (RAG) pipeline.
* Code snippets or documentation demonstrating the integration of security features, including prompt injection testing.

Learning Outcomes Assessed:

1. Design secure RAG pipelines.
2. Implement structured output enforcement.
3. Conduct prompt injection testing.

Capstone Requirements Incorporated:

* Implement a prompt-injection test suite for the RAG pipeline.
* Report attack success rate, false block rate, JSON validity rate, and citation coverage.

Assessment Criteria:

1. **Security Design (10 points)**
    * Effectiveness of security features in preventing common LLM risks (e.g., prompt injection attacks).
    * Integration of mitigation strategies (e.g., intent routing, risk-scored triage layer).
2. **Implementation and Code Quality (8 points)**
    * Clarity and organization of code documentation.
    * Correctness and efficiency of implementation.
3. **Prompt Injection Testing (6 points)**
    * Thoroughness and effectiveness of the test suite.
    * Accuracy of reported metrics (attack success rate, false block rate, JSON validity rate, citation coverage).
4. **Ethics/Compliance Section (6 points)**
    * Depth and accuracy of discussion on ethics and compliance considerations.

Submission Guidelines:

* Submit a single PDF document containing the written report and code snippets or documentation .
* Use a clear and consistent formatting style throughout the submission.
* Include a title page with student name, ID, and unit details.

Marking Rubric: Secure RAG Implementation (Assignment 2)

| Criteria | Excellent (9-10) | Good (7-8) | Satisfactory (5-6) | Unsatisfactory (0-4) |

| --- | --- | --- | --- | --- |
Security Design	Effective security features, thorough mitigation strategies.	Some security features implemented, but incomplete or ineffective mitigation strategies.	Limited security features, inadequate mitigation strategies.	No security features implemented.
Implementation and Code Quality	Clear, organized code documentation; efficient implementation.	Code documentation is adequate, but implementation has some issues.	Poorly documented code; inefficient implementation.	Incomplete or incorrect code submission.
Prompt Injection Testing	Thorough test suite with accurate metrics reporting.	Test suite is mostly effective, but some metrics are missing or inaccurate.	Limited testing, incomplete or inaccurate metrics reporting.	No prompt injection testing implemented.
Ethics/Compliance Section	Comprehensive discussion on ethics and compliance considerations.	Some discussion on ethics and compliance, but lacks depth or accuracy.	Limited discussion on ethics and compliance.	No discussion on ethics and compliance.

Student Submission Checklist: Secure RAG Implementation (Assignment 2)

* Have I implemented a secure RAG pipeline with effective security features?
* Is my code well-documented and efficiently implemented?
* Have I conducted thorough prompt injection testing, including accurate metrics reporting?
* Has my submission included an ethics/compliance section discussing relevant considerations?

Note: This assessment brief is designed to align with the postgraduate expectations of the unit. The marking rubric provides clear criteria for assessing student submissions, while the student submission checklist ensures students understand what is required from them.