Document de Passation : Base de Connaissance RAG (PING 56)

Ce document récapitule les livrables, les instructions d'installation et une analyse critique des tests LLM effectués pour le pipeline RAG (Retrieval-Augmented Generation) du projet PING 56.

---

1. Livrables et Structure du Projet

Le projet est structuré autour de trois composants clés : le script d'ingestion, le script de requêtage, et la base de données vectorielle.

1.1. Structure des Dossiers

Le répertoire racine doit contenir la structure suivante :

Plain Text

```
base_de_reconnaissance_rag/
├── data/                  # Contient tous les documents PDF ingérés
│   ├── reseau/
│   ├── rag_secops_docs/
│   ├── secops_referentiels/
│   └── splunk_docs/
├── venv/                  # Environnement virtuel Python (créé après
installation)
├── vectordb/              # Base de données vectorielle ChromaDB (créée
après ingestion)
├── ingest.py              # Script d'ingestion (Chunking, Vectorisation)
├── query.py               # Script de requêtage (Test RAG)
└── PASSATION.md           # Ce document
```

1.2. Scripts Clés

Fichier	Rôle	Statut
`ingest.py`	Construit la base de connaissance : charge les PDF, effectue le découpage (chunking) et la vectorisation, puis persiste les données dans ChromaDB.	**Finalisé et Optimisé**

`query.py`	Permet de tester la chaîne RAG : charge la base, utilise un LLM (Ollama par défaut) pour répondre aux questions en se basant sur le contexte récupéré.	**Finalisé pour les tests locaux**

---

2. Instructions d'Installation et d'Exécution

2.1. Préparation de l'Environnement

L'environnement a été configuré pour utiliser **Python 3.9** via un environnement virtuel.

Action 1 : Créer et Activer l'Environnement Virtuel

Plain Text

```
# Supprimer l'ancien environnement (si besoin)
Remove-Item -Recurse -Force .\venv

# Créer l'environnement avec Python 3.9 (ajuster le chemin si nécessaire)
& "C:\Users\kouno\anaconda3\envs\architecture-data_env\python.exe" -m venv
venv

# Activer l'environnement
.\venv\Scripts\activate
```

Action 2 : Installer les Dépendances

Plain Text

```
# Installation des librairies RAG, ChromaDB, PDF et le nouveau package
HuggingFace
pip install chromadb langchain-community pypdf cryptography langchain-
huggingface
```

2.2. Ingestion de la Base de Connaissance

Action 3 : Télécharger les Documents

Télécharger tous les documents PDF listés dans le tableau de la section 4 et les placer dans les sous-dossiers de `data/` .

Action 4 : Exécuter le Script d'Ingestion

```
Plain Text
python ingest.py
```

Résultat attendu : Le dossier `vectordb/` est créé, contenant 742 chunks vectorisés.

2.3. Test de la Chaîne RAG (Requêtage)

Action 5 : Préparer le LLM (Choix de l'Ingénieur LLM)

- **Option A (Recommandée pour la Production) :** Utiliser OpenAI. Définir la clé API :

  ```
  Plain Text
  $env:OPENAI_API_KEY="VOTRE_CLE_API"
  ```

- **Option B (Pour les Tests Locaux) :** Utiliser Ollama. Installer Ollama et s'assurer que le modèle `phi3:mini` ou `mistral` est disponible.

Action 6 : Exécuter le Script de Requêtage

```
Plain Text
python query.py
```

Résultat attendu : Le script démarre et affiche l'invite `Analyste SecOps >`.

---

3. Analyse Critique des Tests LLM (Passation)

Cette section est cruciale pour guider l'ingénieur LLM dans le choix du modèle et l'optimisation du prompt.

3.1. Modèles Testés et Résultats

Modèle Testé	Type	RAM Requise	Résultat du Test RAG	Conclusion

`tinyllama`	Local (Ollama)	~1.5 GiB	**Échec.** Hallucinations, non-respect du prompt système, génération de faux code Splunk (API inventée).	**Inutilisable** pour les tâches de raisonnement et de traduction SPL.
`phi3:mini`	Local (Ollama)	3.5 GiB	**Échec.** Bloqué par un manque de RAM (`3.5 GiB requis > 3.0 GiB disponibles`).	**Trop gourmand** en ressources pour la machine de développement.
`mistral:7b-instruct-q4_0`	Local (Ollama)	4.3 GiB	**Échec.** Bloqué par un manque de RAM (`4.3 GiB requis > 3.4 GiB disponibles`).	**Trop gourmand** en ressources pour la machine de développement.

3.2. Recommandation pour l'Ingénieur LLM

Recommandation Principale :

Étant donné la complexité des tâches (génération de requêtes SPL, analyse MITRE ATT&CK), le LLM doit posséder une forte capacité de raisonnement et de suivi d'instructions.

- **Pour la Production :** Utiliser `gpt-4o` ou `gpt-4-turbo` (OpenAI) ou `Claude 3 Sonnet/Opus` (Anthropic). Ces modèles excellent dans la génération de code et le raisonnement technique.
- **Pour les Tests Locaux/Développement :** Si un modèle local est exigé, le modèle `Mixtral 8x7B` (si les ressources le permettent) ou `Llama 3 8B` sont les meilleurs candidats pour la tâche RAG.

Point d'Attention : Le prompt système dans `query.py` est déjà optimisé pour un rôle SecOps. L'ingénieur devra se concentrer sur l'intégration du LLM choisi dans la chaîne RAG.

---

4. Base de Connaissance (Liste des Documents Inclus)

La base de connaissance est maximisée pour couvrir les CU Splunk, Réseau, MITRE ATT&CK et la résolution d'incidents.

Domaine	Titre du Document	URL de Téléchargement Direct	Dossier Recommandé
Splunk/SIEM	The Essential Guide to SIEM	`https://4datasolutions.com/wp-content/uploads/2024/08/the-essential-guide-to-siem.pdf`	`data/splunk_docs/`
Splunk/SIEM	Use Splunk as a SIEM	`https://www.cybermak.net/wp-content/uploads/2019/07/Use-Splunk-as-SIEM.pdf`	`data/splunk_docs/`
Splunk/Réseau	The Essential Guide to Network Data	`https://www.splunk.com/pdfs/ebooks/the-essential-guide-to-network-data.pdf`	`data/splunk_docs/`
Splunk/Architecture	Splunk Validated Architectures	`https://www.splunk.com/en_us/pdfs/white-paper/splunk-validated-architectures.pdf`	`data/splunk_docs/`
Splunk/Pratiques	Best Practices and Better Practices for Users	`https://conf.splunk.com/files/2016/slides/best-practices-and-better-practices-for-users.pdf`	`data/splunk_docs/`
Splunk/CIM	A Look At The Common Information Model	`https://conf.splunk.com/files/2016/slides/the-power-of-data-normalization-a-look-at-cim-under-the-hood.pdf`	`data/splunk_docs/`
Splunk/CIM	The Power of Data Normalization	`https://conf.splunk.com/files/2017/slides/the-power-of-data-normalization-a-look-at-cim-under-the-hood.pdf`	`data/splunk_docs/`
Splunk/CIM	Splunk for Cybersecurity Maturity Model Certification	`https://www.splunk.com/en_us/pdfs/tech-brief/splunk-for-cybersecurity-maturity-`	`data/splunk_docs/`

		model-certification-solution.pdf	
Réseau/Routage	IPv4/IPv6 Addressing and Routing Review	https://netacad.fit.vutbr.cz/wp-content/uploads/ccnp/ce2/ENARSI_Chapter_1.pdf	data/reseau/
Réseau/Protocoles	ARP and DNS – address translation	https://iisi.pcz.pl/Robert_Nowicki/fcn/ARPiDNS_eng.pdf	data/reseau/
RAG & Cybersécurité	RAG for Robust Cyber Defense (PNNL)	https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-36792.pdf	data/rag_secops_docs/
Référentiel SecOps	MITRE ATT&CK®: Design and Philosophy	https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf	data/secops_referentiels/
Référentiel SecOps	MITRE ATT&CK®: Getting Started	https://www.mitre.org/sites/default/files/2021-11/getting-started-with-attack-october-2019.pdf	data/secops_referentiels/
Playbook IR	Federal Government Cybersecurity Incident & Vulnerability Response Playbooks (CISA)	https://www.cisa.gov/sites/default/files/2024-08/Federal_Government_Cybersecurity_Incident_and_Vulnerability_Response_Playbooks_508C.pdf	data/secops_referentiels/
Playbook IR	Public Power Cyber Incident Response Playbook	https://www.publicpower.org/system/files/documents/Public-Power-Cyber-Incident-Response-Playbook.pdf	data/secops_referentiels/
Playbook IR	Cyber Incident Response Plan Guidance and Template	https://database.cyberpolicyportal.org/api/files/1659689651793diio1yngm4e.pdf	data/secops_referentiels/

Spécification	Documentdespécificationv1-00.pdf	(Votre document initial)	`data/` (ou un sous-dossier)

Fin du Document de Passation