Devoir 1 – Conception et Itération d'un Prompt avec promptFoo

Thème : OpenClaw – Framework d'agents LLM

1. Introduction

L'essor des Large Language Models (LLM) a favorisé l'émergence d'architectures multi-agents capables d'orchestrer des tâches complexes en combinant génération linguistique, planification et interaction avec des outils externes.

Dans ce contexte, OpenClaw constitue un exemple de framework open source dédié à l'orchestration d'agents LLM mais il pose également un problème de sécurité quand à léguer le controle de nos systèmes à des agents IA.

L'objectif de ce travail est de concevoir un prompt structuré, testé avec promptFoo, permettant de générer une note de synthèse académique complète sur OpenClaw.

Le processus inclut une conception itérative, l'intégration de sources externes via un fichier TXT (simulation RAG statique), ainsi que la mise en place de mécanismes de validation automatique.

2. Chronologie de conception du prompt

Version 1 – Prompt simple :
Une première version demandait une note structurée. Les résultats étaient trop généraux, manquant de profondeur technique et présentant un risque d'hallucination.

Version 2 – Ajout du rôle système :
Un rôle expert en architectures LLM a été introduit afin de fixer un registre technique de niveau master/ingénieur.

Version 3 – Structure académique obligatoire :
Imposition d'une structure numérotée (I à VI) pour encadrer la production et permettre une validation automatique.

Version 4 – Contraintes quantitatives :
Ajout d'un minimum de 1200 mots, deux scénarios détaillés et un schéma conceptuel textuel.

Version 5 – Réduction des hallucinations :
Température fixée à 0.2, interdiction explicite d'inventer des fonctionnalités et obligation de mentionner les incertitudes.

Version 6 – Intégration d'un corpus externe (RAG statique) :
Création d'un fichier sources_openclaw.txt contenant des articles, liens et résumés.
Ce fichier est injecté via {{file:...}} dans promptFoo afin de contraindre la génération à un corpus fermé.

3. Intégration d'un corpus documentaire externe (Simulation RAG)

Afin d'améliorer la fiabilité du contenu, les sources ont été externalisées dans un fichier texte distinct.

Avantages :
- Séparation des données et des instructions.
- Réduction des hallucinations.
- Reproductibilité des résultats.
- Simulation d'un pipeline Retrieval-Augmented Generation (RAG) statique.

Le modèle reçoit explicitement les documents en contexte et doit s'appuyer exclusivement sur ceux-ci.
Il lui est interdit d'introduire des informations non présentes dans les sources.

4. Choix pédagogiques et contraintes techniques

Plusieurs décisions structurantes ont été adoptées :

- Niveau technique imposé : Master / Ingénieur.
- Style académique neutre sans opinion personnelle.
- Comparaison obligatoire avec une architecture LLM simple.
- Deux scénarios pas-à-pas pour illustrer les mécanismes.
- Description textuelle d'un schéma d'architecture.
- Obligation de citer les sources internes sous la forme [SOURCE 1].

Ces choix visent à garantir rigueur, profondeur analytique et traçabilité des informations.

5. Tests automatiques avec promptFoo

Des assertions automatiques ont été intégrées :

- Vérification de la présence des sections (I. Contexte, II. Architecture, etc.).
- Vérification du nombre minimal de mots (>= 1200).
- Vérification de la présence de citations internes ([SOURCE 1], [SOURCE 2]).

Ces tests transforment le prompt en artefact mesurable et permettent une amélioration continue.

6. Analyse des risques et mesures de mitigation

Risque 1 – Hallucination technique :
Mitigation : corpus fermé + température basse + interdiction explicite d'invention.

Risque 2 – Confusion avec d'autres frameworks :
Mitigation : centrage strict sur OpenClaw et comparaison contrôlée.

Risque 3 – Dérive narrative :
Mitigation : structure académique numérotée et validation automatique.

Risque 4 – Surinterprétation des sources :
Mitigation : obligation de mentionner les incertitudes.

7. Conclusion

Ce travail démontre qu'un prompt efficace relève d'une véritable démarche d'ingénierie. L'intégration d'un corpus documentaire externe, la structuration stricte, les contraintes quantitatives et la validation automatique via promptFoo permettent de transformer un simple prompt en dispositif contrôlé et optimisé.

L'approche adoptée illustre la convergence entre prompt engineering et principes RAG, dans une logique de réduction des hallucinations et d'amélioration de la fiabilité scientifique.