

Un RAG sur les conditions d'assurance habitation

Un assistant qui répond à « suis-je couvert ? » en citant les articles du contrat. Voici la méthode, les résultats mesurés, les échecs et ce qui reste à faire.

Statut : en ligne sur rag.aksel-sorensen.fr (HTTPS), hébergé sur mon VPS. Les articles s'affichent immédiatement ; la réponse rédigée par le modèle 4B peut prendre plusieurs minutes et échouer.

Le problème

Savoir si un sinistre est couvert demande de lire le contrat, de repérer les clauses pertinentes et de vérifier leurs conditions. Un modèle qui répond seul, sans source, affirme souvent avec assurance. Le risque à maîtriser est une réponse fausse présentée comme certaine.

Ce que j'ai construit

Le code est organisé en cinq étapes, chacune dans son module :

1. **Découpage**. Les conditions générales Swiss Life habitation (édition 01.2017) sont découpées en 135 clauses, chacune avec son numéro, son titre et sa page.
2. **Indexation**. Les clauses sont coupées en 405 passages d'environ 1 000 caractères, avec chevauchement. Chaque passage est vectorisé avec multilingual-e5-small (384 dimensions).
3. **Recherche hybride**. La recherche combine le sens (vecteurs) et les mots-clés (BM25). Une petite table de synonymes relie le langage courant au vocabulaire du contrat : « volé » vers « vol », « brûler » vers « incendie ».

4. **Génération.** Les passages les plus pertinents, dans la limite de 4 000 caractères, sont envoyés à un modèle Qwen avec trois consignes : citer chaque article, dire « incertain » quand un fait manque, ne jamais citer un article absent.
5. **Évaluation.** Un script rejoue les cas de test et vérifie que l'article attendu est bien envoyé au modèle, puis que le verdict est correct.

Résultats mesurés

Cinq cas de test, avec pour chacun l'article attendu d'après le texte du contrat.

La recherche

L'article attendu est envoyé au modèle dans **4 cas sur 5**. Le cas raté est la bougie : l'article 9.1 n'entre pas dans les passages retenus, qui portent sur le jardin et la tempête.

La réponse, sur mon PC (Qwen 9B)

Ces essais tournent sur mon ordinateur, pas sur le serveur. Le temps de réponse est de 5 à 40 secondes une fois le modèle chargé.

CAS	ATTENDU	OBTENU
Vélo volé dans le local commun, porte non fermée	Non couvert (9.23)	Non couvert
Vélo attaché à un antivol en U, volé dans la rue	Couvert (9.23)	Couvert
Vélo volé dans une cave, porte forcée	Dépend des conditions de 9.12	« Couvert » : faux, les conditions de construction ne sont pas vérifiées
Bougie qui fond sur la table du jardin	Non couvert sans flamme (9.1)	« Non couvert », mais en citant 9.5 au lieu de 9.1

CAS	ATTENDU	OBTENU
Foyer qui met le feu à la maison	Non couvert : incendie « en dehors d'un foyer normal » (9.1)	Non couvert

Sur ces cinq cas, le verdict est correct dans quatre. Les citations sont incomplètes ou fausses dans deux cas, dont la cave.

La réponse, sur le serveur (Qwen 4B)

Le serveur est un VPS à 4 processeurs et 7,8 Go de RAM, partagé avec d'autres services. Le même modèle y donne des résultats nettement moins bons :

- Vélo, local commun : « incertain » après 333 secondes.
- Cave forcée : « couvert » après 196 secondes, ce qui est faux.
- Bougie : erreur 500 d'Ollama, sans réponse.
- L'évaluation complète s'est arrêtée au premier cas, sur la même erreur 500. Je n'ai donc pas de résultat complet pour le serveur.

Benchmark

Même jeu de cinq cas, même découpage et même recherche, mais deux machines différentes. Le graphique montre le temps de réponse ; les barres partagent une échelle commune (340 secondes = 100 %).

■ PC, Qwen 9B ■ VPS, Qwen 4B

Vélo, local commun	<div> <div>40 s (1er appel, chargement du modèle) · verdict correct</div> <div>333 s · « incertain » (attendu : non couvert)</div> </div>
Cave, porte forcée	<div> <div>6 s · « couvert » (faux)</div> <div>196 s · « couvert » (faux)</div> </div>
Bougie sur la table	<div> <div>7 s · « non couvert », article cité erroné (9.5 au lieu de 9.1)</div> </div>

VPS : erreur 500, pas de réponse

Foyer, feu dans la maison 5 s · « non couvert » (correct)
VPS : non mesuré

Vélo, antivol en U 18 s · « couvert » (correct)
VPS : non mesuré

MESURE	PC, QWEN 9B	VPS, QWEN 4B
Recherche : article attendu envoyé au modèle	4 cas sur 5	4 cas sur 5 (même recherche)
Verdict correct	4 cas sur 5	0 cas sur 2 réponses obtenues
Temps de réponse médian	7 s	264 s (2 cas, 196 et 333 s)
Erreurs	aucune	1 erreur 500 sur 3 cas tentés

Limites de cette mesure : cinq cas seulement, une seule CGV, et des verdicts jugés à la main. Le 1er appel du PC inclut le chargement du modèle. Le VPS n'a pas été mesuré sur les cas foyer et antivol, après l'erreur 500.

Ce qui a échoué, et pourquoi

- **Le rappel de la recherche.** Pour la bougie, l'article 9.1 n'est pas retenu. Le mot « bougie » n'existe pas dans le contrat, et la recherche dépend de la table de synonymes.
- **Le raisonnement sur les conditions.** Pour la cave, le modèle lit l'exclusion de 9.10 et ne vérifie pas les conditions de l'article 9.12 avant de conclure.
- **La taille du modèle.** Le 4B fait plus d'erreurs que le 9B, et il est aussi lent sur ce serveur. Le 9B ne tient pas dans la mémoire disponible du VPS sans toucher aux autres services.
- **Les erreurs 500 sur le serveur.** Je n'ai pas établi la cause avec certitude. Les logs ne montrent pas clairement un manque de mémoire, donc je ne l'affirme

pas.

- **Une erreur de ma part au départ.** Mes premières clauses étaient inventées. Le test sur le texte réel a montré qu'elles étaient différentes. Règle retenue : ne jamais citer un article qu'on n'a pas lu dans la source.

Les chiffres ci-dessus ne sont pas une mesure de qualité générale. Ce sont cinq cas, sur une seule CGV, sans lecture des conditions particulières.

Ce que je ferais ensuite

- Séparer la recherche de la génération : la recherche reste sur le serveur, la génération tourne là où le modèle est assez rapide.
- Ajouter un reclassement des passages, pour mieux trier ceux qui décident réellement.
- Écrire au moins trente cas de test, avec la réponse attendue rédigée à partir du contrat, et mesurer le taux de bonnes réponses.
- Ajouter plusieurs contrats, avec leur assureur, leur produit et leur version, pour ne jamais mélanger les conditions.
- Vérifier, sur le serveur, si plus de mémoire permet de faire tourner le 9B dans de bonnes conditions. C'est une hypothèse, pas une mesure.

Limites

Une seule CGV est indexée. Les conditions particulières du contrat, qui peuvent modifier la réponse, ne sont pas lues. Le projet n'est pas un avis juridique.

Démonstration en ligne : rag.aksel-sorensen.fr.