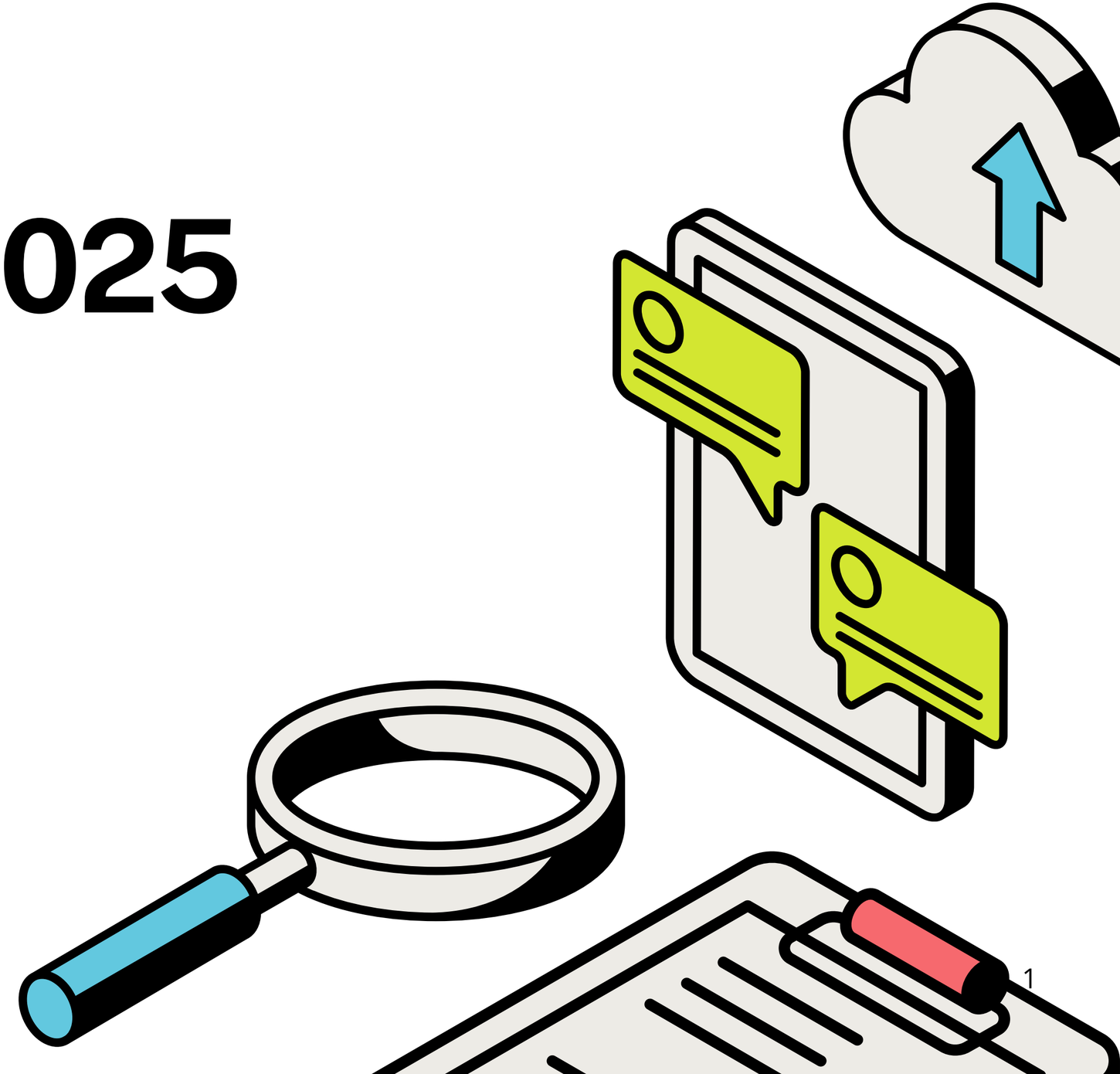
VIGNEMALE - IA PAU

Data Battle 2025



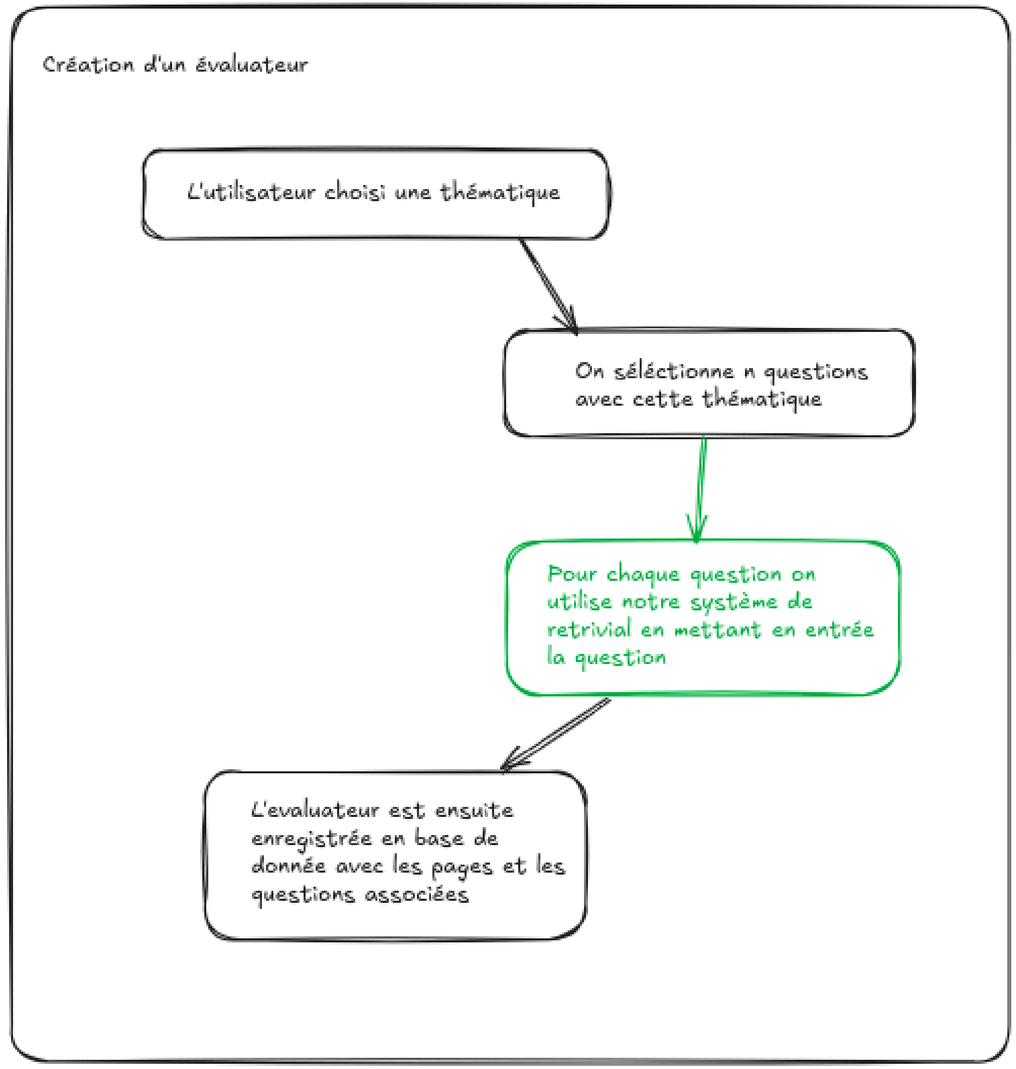
Date: September 30, 2023

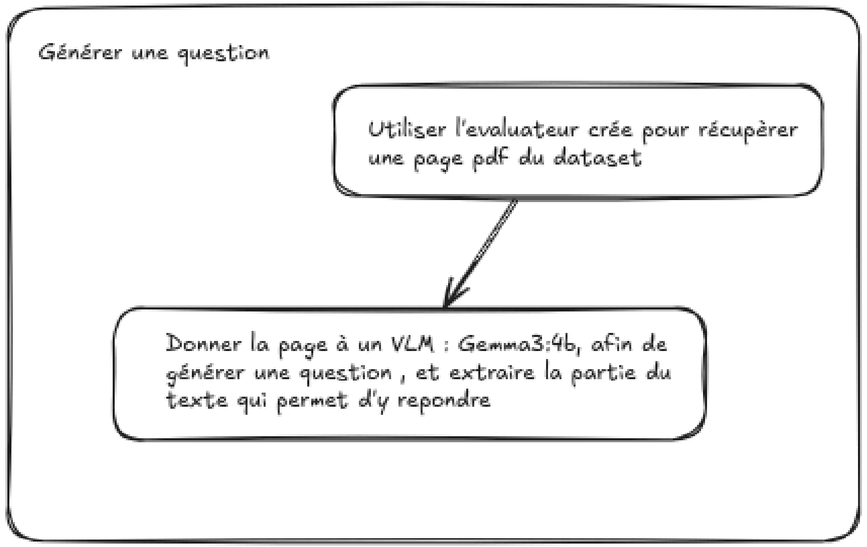
Prepared by: Adrien Bernard , Tiffany Gay-Belille , Jacques Dumora

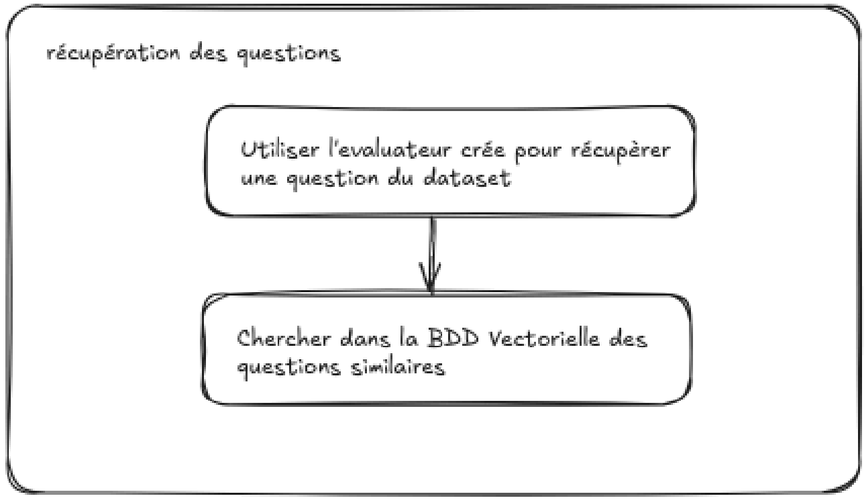
Les données

- Les 3 documents principaux du répertoire Official Legal Publications
- Les questions et réponses des examens précédents
- Les différentes catégories + les questions labelisées par catégories

Notre Processus

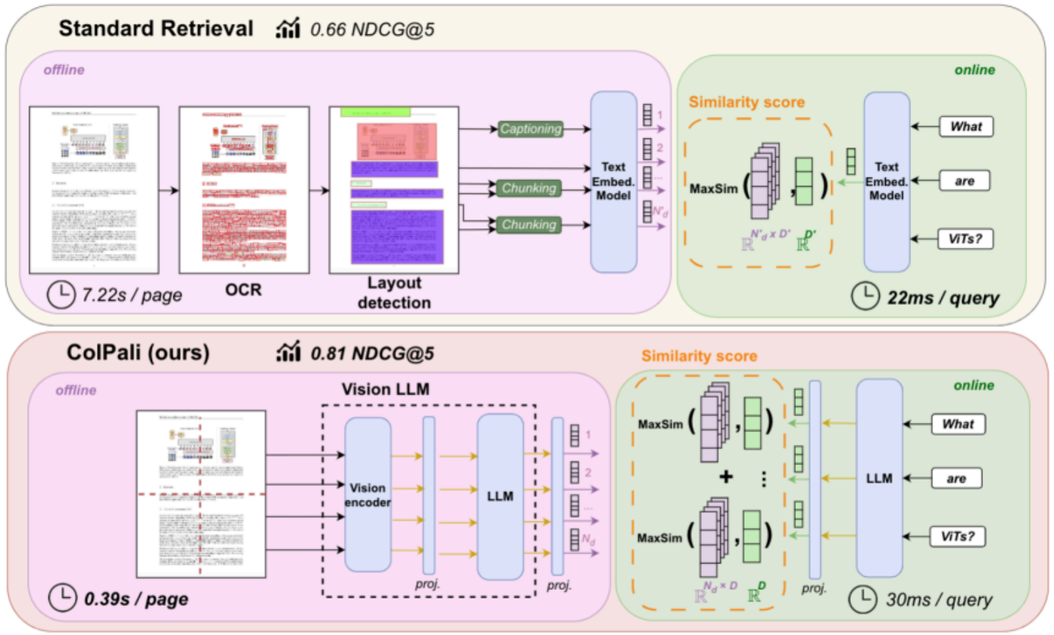


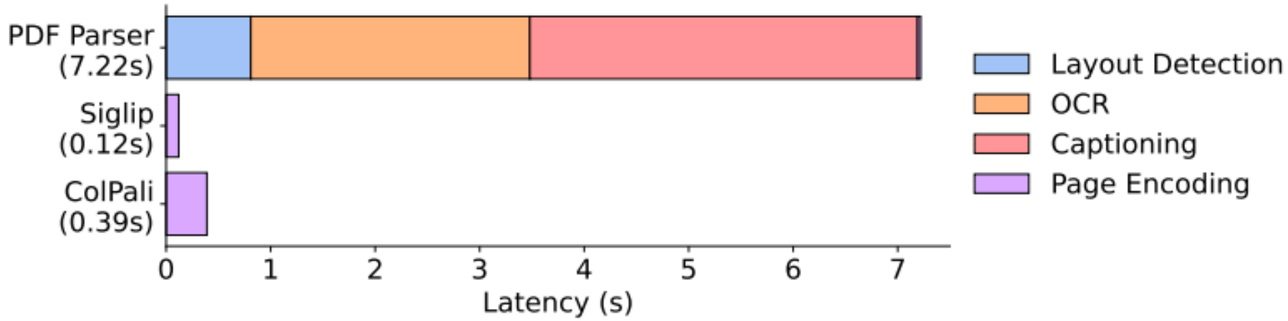




Système de retrivial

Contextualized Late Interaction





Evaluation sur notre Dataset : 98% de précision

Analyse des réponses

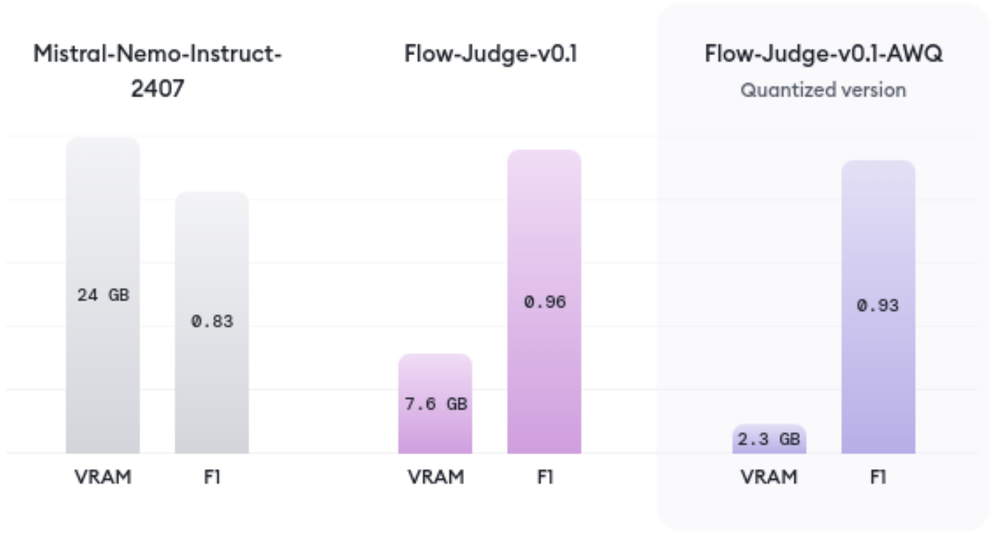
Approche basée sur les llm-as-a-judge

Utilisation de Flow-Judge-GGUF : modèle Phi-3.5 mini fine tuné pour l'évaluation de réponses en fonction d'un contexte

Génère des réponses avec le format :

<feedback>
…
</feedback>
<score>
…
</score>

	Pass / Fail Held-out Test set		
Evaluator	Precision	Recall	F1
microsoft/Phi-3.5-mini-instruct	0.685	**1.000**	0.813
meta-llama/Meta-Llama-3.1-8B-Instruct	0.870	0.982	0.923
mistralai/Mistral-Nemo-Instruct-2407	0.709	0.994	0.827
gpt-4o-mini	0.834	1.000	0.910
flowaicom/Flow-Judge-v0.1	**0.940**	0.972	**0.955**



Impact environnemental

- Stockage des données : Utilisation de qdrant , quantization binaire sur les vecteurs : réduction de la taille par 32 accélérations du processus de retrivial
- Modèle de génération de question / réponse : Gemma3 4B version quantizer : 3.3go
- Modèle d’analyse : flowaicom/Flow-Judge-v0.1-GGUF (Phi3.5-Mini 3.8B) version quantizer : 2.39 go
- Modèle d’embedding : vidore/colqwen2.5-v0.2 (Qwen2.5 VL 3 B) : 7,51go

Total : 13,2go, pour comparaison : Mistral-7B-Instruct-v0.3 non quantizer = 14,5go

Gemma 3

Qwen2.5

Conclusion

Plan de mise en production

- Ajouter un système d’authentification
- Partage des Evaluateurs
- Tous les utilisateurs ne possèdent pas de gpu
  Solution : utilisation de platforme d’inférence optimisé (Groq : consommation 10x inférieur aux puces Nvidia)
- Fonctionnalité deep search pour permettre de générer des rapports complets sur des cas complexes

Merci de votre attention !