



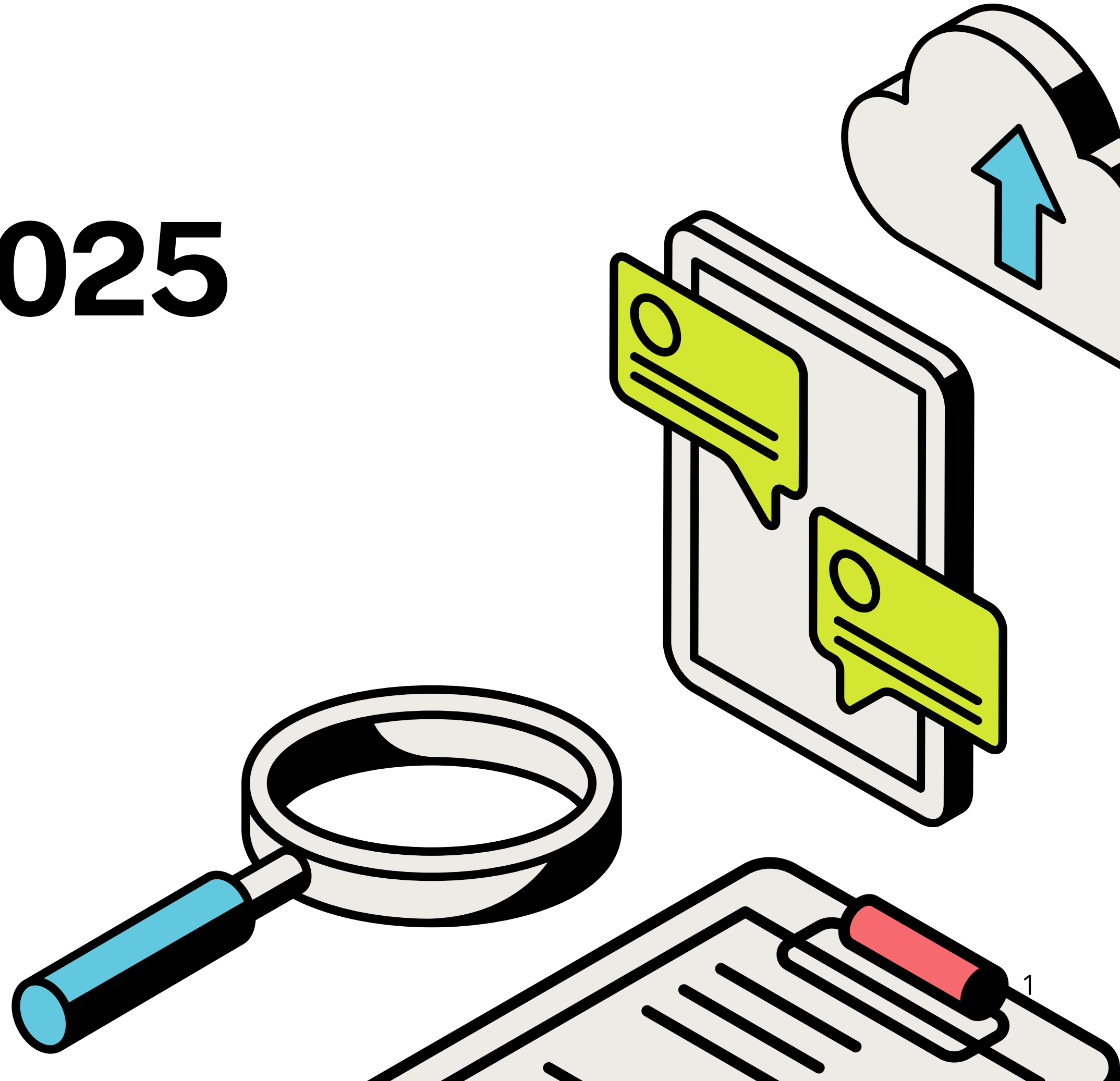
VIGNEMALE - IA PAU

Data Battle 2025

Présenté par: Adrien Bernard , Tiffany Gay-Belille , Jacques Dumora

Ecole: CY Tech

Date: 31 mars 2025



Le sujet

Plateforme de révision de questions pour
étudiants en propriété intellectuelle

Proposé par:



PAU · PARIS



Les données

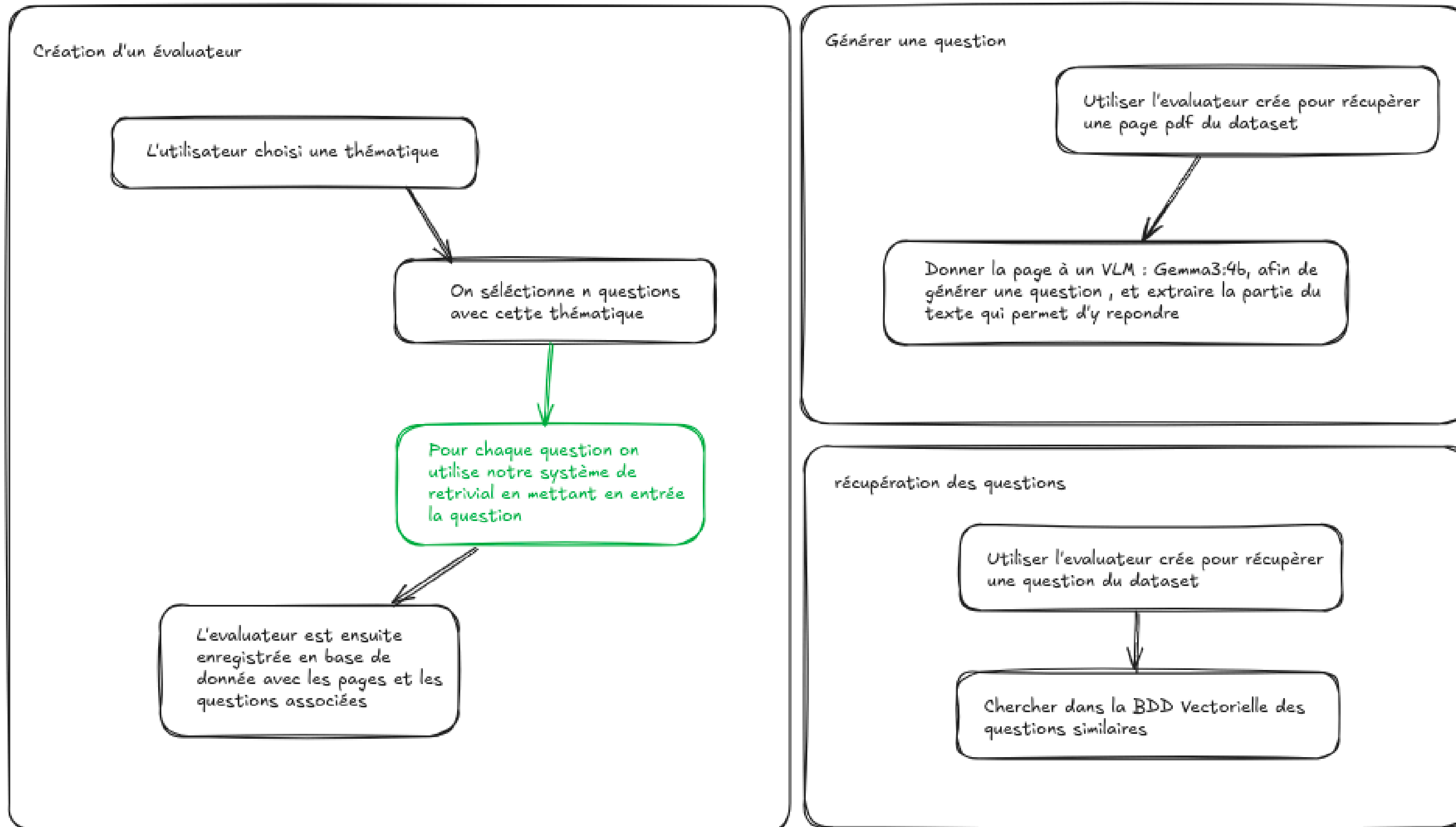
Documents du répertoire “Official Legal Publications”

Les questions et réponses des examens précédents

Les différentes catégories + les questions labelisées par catégories

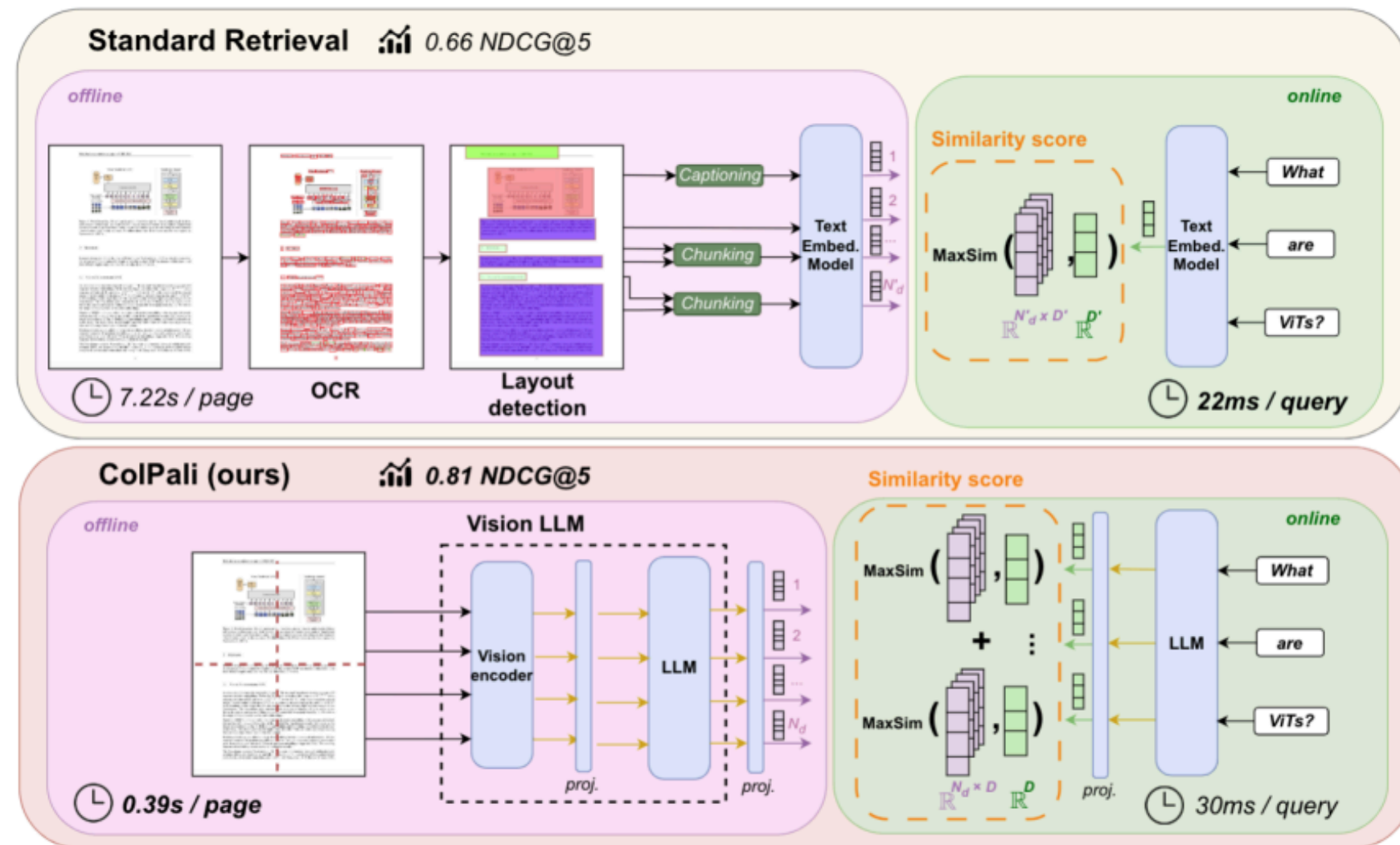
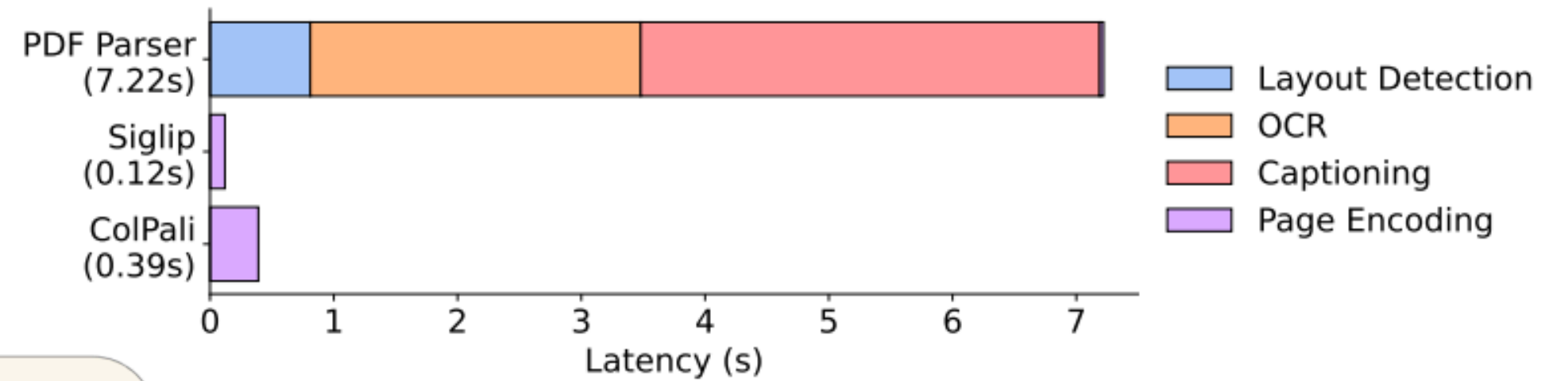


Notre Processus



Systeme de retrivial

Contextualized Late Interaction



98%

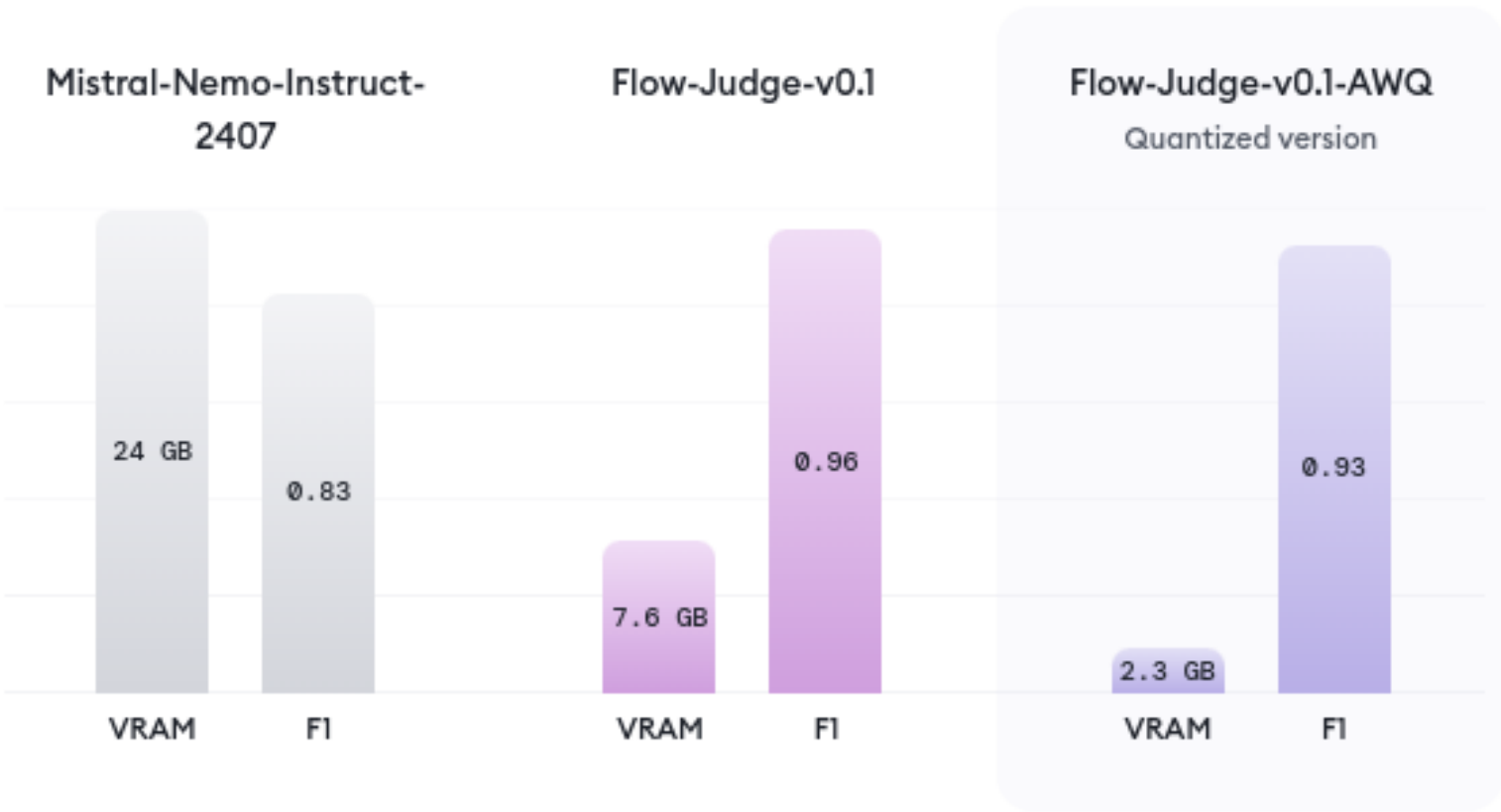
De précision sur
notre dataset

Analyse des réponses

Approche basée sur les llm-as-a-judge

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/
score



Evaluator	Pass / Fail Held-out Test set		
	Precision	Recall	F1
microsoft/Phi-3.5-mini-instruct	0.685	1.000	0.813
meta-llama/Meta-Llama-3.1-8B-Instruct	<u>0.870</u>	0.982	<u>0.923</u>
mistralai/Mistral-Nemo-Instruct-2407	0.709	<u>0.994</u>	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 retrieval



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

13,2 Go

14,5 Go

Mistral-7B-Instruct-v0.3
non quantizer

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 plateforme 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 !

