



HACKATHON ASSEMBLÉE NATIONALE · 2026

Le Rapporteur

Zéro article inventé. Chaque citation vérifiée.

L'assistant juridique citoyen qui refuse d'halluciner

Le problème

- Les LLM généralistes **inventent des articles de loi** avec un aplomb parfait
- Des avocats ont déjà été sanctionnés pour des **jurisprudences fictives**
- Le citoyen n'a **aucun moyen de vérifier**
- En droit, une réponse fausse est **pire que pas de réponse**

✗ UNE IA GÉNÉRALISTE, AUJOURD'HUI

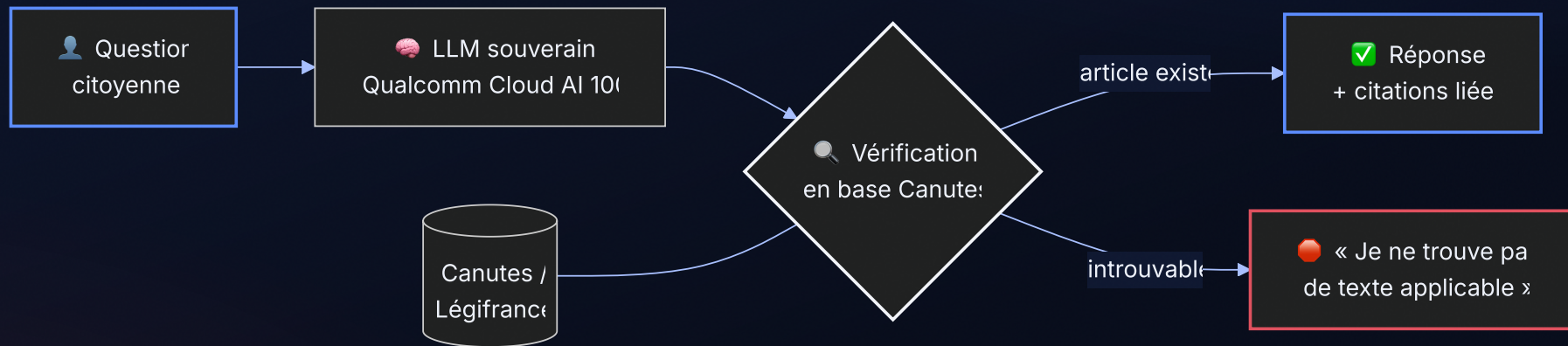
« Selon l'article **L. 4321-7 du Code du travail**, votre employeur doit... »

Cet article **n'existe pas**.

Réponse plausible ≠ réponse vraie.

La solution

Un assistant qui **prouve** chaque citation — ou refuse de répondre



Comment ça marche

1

Génération souveraine

Un LLM open source sur silicium souverain : **Qualcomm Cloud AI 100** en direct, AMD via **Modular MAX** en cible. Hors NVIDIA.

2

Vérification en base

Chaque référence confrontée à **Canutes / Légifrance** en direct : existe, dans ce code, en vigueur. On expose aussi un **serveur MCP**.

3

Réponse sourcée

Citations **cliquables vers le texte consolidé**, ou refus honnête si rien ne s'applique.

Données **Tricoteuses** (assemblée · sénat · légifrance) · Backend **OpenAI-compatible**, swappable

Prouvé, pas promis

La garantie anti-hallucination est **encodée en Gherkin**,
rejouée en CI sur un benchmark de questions citoyennes.

100 %
citations vérifiées

0
article fictif toléré

« Je ne
sais pas »
est une réponse
valide

```
Scénario: Pas de citation inventée
Étant donné une question citoyenne
sur l'article X du code Y
Quand le système répond
Alors chaque article cité doit exister
dans Canutes / Légifrance
(vérification en base)
Et si aucun texte applicable n'est trouvé,
la réponse doit être un refus explicite
Et aucune jurisprudence absente des
sources ne doit être mentionnée
```


Fiable · Frugal · Portable

- **Fiable** — vérification systématique, refus explicite, benchmark reproductible
- **Frugal** — modèle compact quantisé, une requête MCP par citation
- **Portable** — `docker pull` et c'est parti : AN, préfecture, mairie

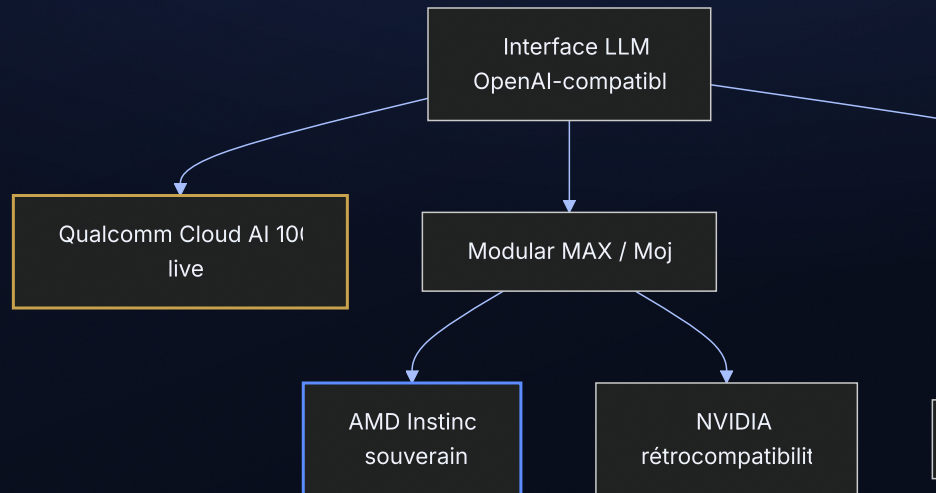
Vision

- **Aujourd'hui** — questions citoyennes sur les codes en vigueur
- **Demain** — un produit pour les **services de l'AN** : vérifier les références des amendements et questions écrites
- **Après-demain** — brancher **Catala** : des réponses *calculées*, pas seulement citées

Sans verrou NVIDIA

Une interface OpenAI-compatible. On change de backend, le pipeline ne bouge pas. Faire tourner un LLM open source rime aujourd'hui avec NVIDIA et CUDA. Nous le prouvons autrement.

- **Qualcomm Cloud AI 100** : backend souverain, prouvé en direct
- **Modular MAX** porte l'inférence sur **AMD**, et sur NVIDIA pour la rétrocompatibilité
- Demain, les puces **européennes** : VSora (France), Axelera (Europe)



Chaque watt compte

Un datacenter d'IA, c'est d'abord une facture d'électricité.
Une puce d'inférence dédiée fait le même travail pour bien moins d'énergie qu'un GPU généraliste.

dédié

accélérateur
d'inférence, pas un
GPU détourné

perf /
watt

souveraineté et
écologie

3 var.

changer de
backend, une
variable d'env

MODULAR MAX & MOJO

La couche de **portabilité** : écrire l'inférence une fois, l'exécuter sur AMD, NVIDIA, Apple. C'est ce qui rend les cibles souveraines réalistes.

Un hot path en **Mojo**, benchmarké contre Python, alimente le panneau de frugalité.

Mesuré en direct sur Qualcomm Cloud AI 100 : latence et débit affichés à la démo. Benchmark AMD et NVIDIA via Modular MAX : la prochaine étape.

Un outil, pas seulement une démo

A

Serveur MCP

Le Rapporteur **expose**
`repondre_question` et
`verifier_article` : tout agent peut
brancher notre **vérification juridique**

B

Vérification en base

Article confronté à **Canutes** /
Légifrance en direct : existe, dans ce
code, **en vigueur**, lien cliquable

C

Carte du schéma

Cartographie de Canutes **lisible par**
les IA (JSON, DBML, SchemaSpy),
reversible à Tricoteuses

On ne se contente pas de répondre : on devient une **brique de confiance** que l'écosystème peut brancher.

Ce qui tourne, ce qui vient

DÉJÀ EN PLACE

- Vérification en base réelle, refus explicite, preuve Gherkin
- Backend souverain en direct (Qualcomm Cloud AI 100)
- Serveur MCP, API de fact-check, carte du schéma Canutes

À VENIR

- **RAG** : ancrer le contenu, pas seulement la citation (recherche plein-texte Canutes ou MCP Moulineuse)
- **Mistral en direct** via Modular MAX sur AMD
- **Benchmark perf / watt** AMD contre NVIDIA, hot path **Mojo**
- Unifier le **serveur MCP** sur le moteur de vérification live



Le Rapporteur

*La confiance dans le droit, ça ne s'improvise pas.
Ça se vérifie.*

mcp.hackathon2026.leximpact.dev

db.code4code.eu/canutes

parlement.tricoteuses.fr