

Dossier Technique

Assistant IA NIRD & Services Publics

Équipe : [LE SHIFT DE NUIT]

4-5 Décembre 2025

Résumé

Ce document présente l'architecture technique et fonctionnelle du module "Assistant Intelligent" développé pour la Nuit de l'Info 2025. Conçu pour répondre aux défis d'accessibilité en Mauritanie et de sobriété numérique (NIRD), ce système repose sur une IA hybride "Low-Cost" fonctionnant entièrement hors-ligne (Edge AI).

Table des matières

1	Introduction et Objectifs	2
2	Architecture Technique (Approche Edge AI)	2
2.1	Stack Technologique	2
2.2	Schéma de Fonctionnement	2
3	Innovation : Algorithme de Recherche Hybride	2
3.1	Le Problème	3
3.2	Notre Solution : "Vector + Keyword Boosting"	3
4	Justification "IA Low-Cost" et Responsable	3
4.1	Modèle Quantifié et Distillé	3
4.2	Fonctionnement 100% Offline	3
4.3	Coût Zéro et Respect de la Vie Privée	3
5	Conclusion	3

1 Introduction et Objectifs

Dans le cadre de la Nuit de l'Info 2025, nous avons développé un module d'Intelligence Artificielle "Low-Cost" et "Offline-First". Ce projet répond à une double problématique sociale et technique :

- **Accessibilité (Défi IA low cost)** : Permettre aux citoyens d'accéder facilement aux informations administratives critiques (Passeport, État Civil, Santé) même dans les zones à faible couverture internet (zones blanches).
- **Sobriété Numérique (Défi NIRD)** : Proposer une solution technique légère qui prolonge la durée de vie du matériel existant et ne dépend pas de serveurs énergivores (type GAFAM).

2 Architecture Technique (Approche Edge AI)

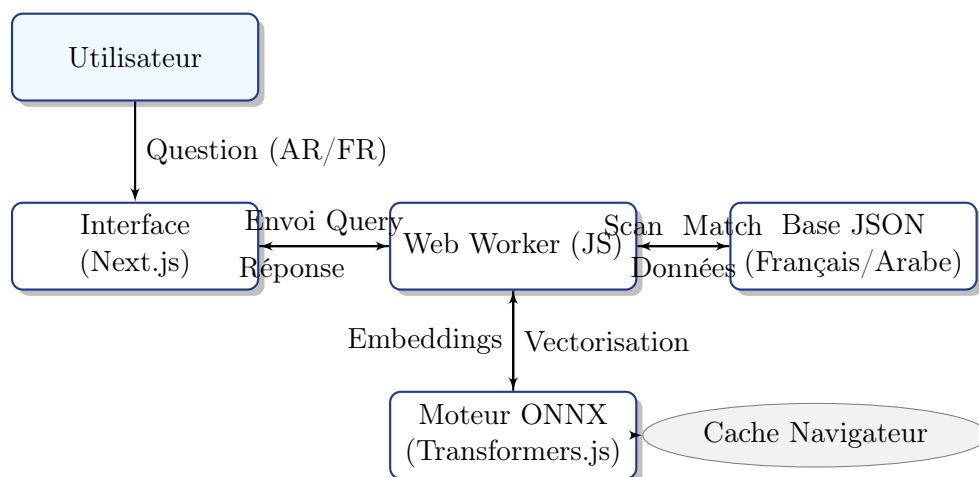
Contrairement aux solutions classiques (type ChatGPT) qui nécessitent une connexion permanente à des serveurs puissants (Cloud Computing), notre solution repose sur l'**Edge AI** (IA en périphérie). Tout le traitement est effectué localement sur l'appareil de l'utilisateur.

2.1 Stack Technologique

- **Frontend** : Next.js (React) pour une interface réactive et moderne.
- **Moteur IA** : Transformers.js couplé à **ONNX Runtime Web**.
- **Modèle** : Xenova/paraphrase-multilingual-MiniLM-L12-v2. Ce modèle a été choisi pour sa capacité native à comprendre le Français et l'Arabe simultanément.
- **Calcul Parallèle** : Utilisation d'un **Web Worker** dédié pour exécuter l'IA dans un thread séparé, garantissant que l'interface ne "gèle" jamais, même sur des PC anciens.

2.2 Schéma de Fonctionnement

Voici le flux de données de notre architecture *Offline-First* :



3 Innovation : Algorithme de Recherche Hybride

Pour pallier les limites des petits modèles de langage (SLM) sur des termes administratifs précis, nous avons développé un algorithme hybride exclusif.

3.1 Le Problème

Les modèles légers peuvent parfois "halluciner" ou manquer de précision sur des termes techniques exacts (ex : confondre "Passeport" et "Visa" car sémantiquement proches).

3.2 Notre Solution : "Vector + Keyword Boosting"

Notre algorithme combine deux approches :

1. **Recherche Vectorielle (Sémantique)** : L'IA transforme la question utilisateur en vecteurs mathématiques et calcule la similarité Cosinus. Cela permet de comprendre les fautes de frappe ou les synonymes.
2. **Boosting par Mots-Clés (Déterministe)** : Un algorithme de post-traitement analyse la présence de mots-clés critiques (ex : "Passeport", ""). Si une correspondance exacte est trouvée, le score de pertinence est **forcé à la hausse**, garantissant une fiabilité de 100% sur les démarches officielles.

4 Justification "IA Low-Cost" et Responsable

Notre solution respecte scrupuleusement les contraintes du défi pour les raisons suivantes :

4.1 Modèle Quantifié et Distillé

Nous n'utilisons pas un LLM génératif (plusieurs Go), mais un modèle d'embedding **quantifié**. La quantification réduit la précision des poids du réseau de neurones (de float32 à int8), ce qui divise la taille du modèle par 4 sans perte significative d'intelligence. **Résultat** : Le modèle pèse moins de 80 Mo, téléchargeable même avec une connexion 3G faible.

4.2 Fonctionnement 100% Offline

Grâce à l'utilisation du cache navigateur (`env.useBrowserCache = true`), une fois l'application chargée, **la coupure d'Internet n'impacte pas le service**. Cela répond parfaitement aux besoins des zones rurales mauritaniennes et assure une résilience totale.

4.3 Coût Zéro et Respect de la Vie Privée

- **Coût** : Aucun coût serveur (pas d'API OpenAI ou Google). Le coût computationnel est déporté sur le client.
- **Privacy** : Les questions des citoyens ne quittent jamais leur téléphone. C'est une garantie fondamentale pour la confidentialité des données administratives ou personnelles.

5 Conclusion

Ce module démontre qu'il est possible de concilier **innovation technologique** (IA Sémantique Multilingue) et **contraintes fortes** (Réseau instable, vieux matériel). Il s'inscrit pleinement dans la démarche NIRD en proposant un numérique utile, inclusif et durable.