IA LOCALE · CERVEAU MÉTIER

Assistant IA ADN

Le « cerveau métier » d'Artisan du Numérique : une IA qui tourne sur le PC, connaît les fiches de l'entreprise (tarifs, méthodes, conformité) et prépare plans d'action, devis et prospection, sans qu'aucune donnée ne parte dans le cloud.



CRÉÉ LE
7 juillet 2026

REFONTE
12 octobre 2026 : version allégée, 100 % locale

MOTEUR
LM Studio + un modèle au choix (Gemma 4 12B par défaut)

CONNAISSANCES
18 fiches métier ADN, cherchées à chaque question

INTERFACE
page à la charte ADN, fenêtre dédiée, un clic

CODE
64 fichiers, 5 600 lignes environ

DANS CE DOSSIER

01 En bref
02 L'idée de départ
03 À quoi il sert
04 Les tenants et les aboutissants
05 Les dates
06 Ce qu'il contient
07 La façon de procéder
08 Le rendu final
09 Où il en est

01

En bref

L'Assistant IA ADN est l'IA de travail de Thibault : il répond **avec les vraies données d'Artisan du Numérique** (grille tarifaire, méthodes, conformité, savoir-faire technique) et cite les fiches dont il tire chaque réponse.

Il sert à préparer un plan d'action client, une trame de devis, un mini-audit RGPD, un mail de démarchage conforme, ou à enquêter sur des entreprises du secteur à partir de sources publiques (registre des entreprises, annonces de créations, carte OpenStreetMap, diagnostic de souveraineté de leur site).

Tout tourne sur la carte graphique du PC. C'est aussi le **prototype d'une offre** : l'« IA locale pour PME » proposée sur le site, un assistant métier qu'une petite entreprise peut garder chez elle.

22 / 22
faits justes au banc d'essai avec le modèle retenu

2,2 s
avant le premier mot (médiane, modèle par défaut)

18
fiches de connaissances ADN

10 min
sans question : le modèle libère seul la carte graphique

02

L'idée de départ

En juillet 2026, Artisan du Numérique vend notamment de l'IA locale. Thibault veut s'en servir lui-même, pour son propre cerveau métier, plutôt que de confier ses tarifs, ses clients et ses méthodes à un service en ligne. C'est la même conviction que celle du site, appliquée à l'intelligence artificielle.

Le 7 juillet, le choix technique est fait : **une recherche dans des fiches (RAG) plutôt qu'un réentraînement du modèle**. Les connaissances restent des documents lisibles, qu'on corrige en éditant une fiche ; le modèle, lui, peut changer.

> CE QUE L'ASSISTANT NE DOIT JAMAIS FAIRE
>
> Inventer un tarif, un chiffre ou un fait daté. Le prompt système lui impose de citer ses fiches ou le web, et de dire qu'il ne sait pas plutôt que d'improviser. Au banc d'essai, une question « hors grille » vérifie précisément qu'il refuse d'inventer un prix.

03

À quoi il sert

- **Répondre juste sur ADN** : tarifs exacts de la grille, méthodologie de plan d'action, conformité RGPD et IA Act, charte visuelle, impression 3D, matériel PC, Meshtastic, droit à l'image… chaque réponse cite ses fiches.
- **Préparer des livrables** : plan d'action client, trame de devis, mini-audit RGPD, « faut-il y aller ? » pour décider d'un projet.
- **Prospecter localement** : prospects dans un rayon autour du Trait, nouvelles entreprises du mois, enquête sur une entreprise (registre officiel, ancienneté, activité), contacts complétés par OpenStreetMap et le web, tableau exportable en CSV.
- **Transformer un diagnostic en accroche** : la note A–E du site d'un prospect et ses points faibles deviennent un argument chiffré, dans un mail qui respecte le RGPD (base légale, identification, lien de désinscription).
- **Lire des documents** : texte, Word, PDF, et PDF scannés grâce à une reconnaissance de caractères en français.
- **Garder la main sur la machine** : jauges GPU, mémoire vidéo, processeur et mémoire vive ; commandes `/eco`, `/max`, `/off` ; bouton STOP qui coupe tout et rend la carte graphique.

04

Les tenants et les aboutissants

>_ Les choix

- **Local d'abord.** Le modèle tourne dans LM Studio sur le PC ; le serveur de l'assistant n'écoute que la machine. Seules sortent les requêtes web et les appels aux API publiques françaises (registre des entreprises, géographie, BODACC, météo), et seulement quand la question l'exige.
- **Des sources officielles et gratuites.** Annuaire des entreprises de l'État, API Géo, BODACC, OpenStreetMap, Open-Meteo : aucune clé, aucun abonnement, et un cache local pour ne pas interroger deux fois la même chose.
- **Le même moteur de diagnostic que le site.** L'assistant appelle le diagnostic de souveraineté du site tel quel, sans en copier le barème : une seule source de vérité.
- **Les prix viennent du catalogue des offres.** La fiche tarifaire de l'assistant se régénère d'une commande depuis le catalogue du back-office : un prix changé une fois change partout.

>_ Ce que la première version a appris

La version de juillet était puissante mais lourde : un modèle de 24 milliards de paramètres occupait 13,6 Go sur les 16 de la carte graphique, avec un logiciel de RAG séparé et deux serveurs. Résultat constaté le 12 octobre : l'assistant n'avait plus été rouvert depuis l'été. D'où la refonte du jour même.

MODÈLE (BANC DU 12/10/2026)	FAITS JUSTES	1ER MOT	VITESSE	MÉMOIRE VIDÉO
Gemma 4 12B (retenu)	22 / 22	2,2 s	46 jetons/s	7,7 Go
Ministral 3 14B	22 / 22	1,8 s	40 jetons/s	9,7 Go
Mistral Small 3.2 24B	6 / 6 (partiel)	3,9 s	9 jetons/s	13,8 Go

05

Les dates

07/07/2026 — Le kit assistant
LM Studio, Mistral Small 3.2 24B, RAG sur 6 fiches ; premier plan d'action avec les tarifs exacts.

08/07 — L'interface ADN
Page maison à la charte, pièces jointes (Word, PDF, OCR), lancement en un clic, historique des conversations.

09/07 — Le web et les données publiques
17 fiches ; recherche web planifiée par le modèle ; registre des entreprises, BODACC, API Géo, météo ; réponses en direct ; jauges de la machine.

10/07 — La prospection outillée
Chaînage avec le diagnostic de souveraineté, mails conformes au RGPD, cache et journal des prospects, bouton STOP, choix du modèle, fermeture = tout libéré.

12/10 — Allègement et banc d'essai
Plus de logiciel de RAG séparé : recherche dans les fiches intégrée ; un seul serveur ; déchargement automatique ; banc comparant trois modèles ; code découpé en petits fichiers.

06

Ce qu'il contient

PARTIE	RÔLE
base-connaissances/	les 18 fiches (.md) : identité, grille tarifaire, catalogue, méthode, RGPD, IA Act, charte, audit, IA locale, infrastructure libre, Meshtastic, impression 3D, robotique, matériel PC, photo-vidéo, micro-entreprise
system-prompt-adn.txt	la personnalité et les règles de l'assistant, modifiables sans code
interface/api/	un point d'entrée par action : question, santé, modèles, contexte, déchargement, fiches, jauges
interface/lib/rag/	découpage des fiches, index, recherche hybride (sens + mots-clés)
interface/lib/prospection/	registre, BODACC, géographie, OpenStreetMap, journal des prospects
interface/lib/llm/	le seul fichier qui parle à LM Studio (remplaçable par un autre moteur)
interface/assets/	l'interface : modules JavaScript, feuille de style, orbe de réflexion
tools/banc-essai/	questions réelles, banc et rapport de comparaison des modèles

07

La façon de procéder

- **Valider avec de vraies questions** : plan d'action pour un plombier, prospects autour du Trait, enquête sur une entreprise, diagnostic d'un site réel ; chaque fonction testée de bout en bout avant d'être gardée.
- **Surveiller la machine avant d'accuser le code** : quand l'assistant « rampait », la cause était une carte graphique saturée (moins de 500 Mo libres) ; d'où les jauges, le mode éco et le déchargement automatique.
- **Comparer les modèles sur pièces** : le banc du 12 octobre pose les mêmes 11 questions à chaque modèle et mesure faits justes, délai et vitesse ; le choix du modèle par défaut en découle.
- **Conçu par Thibault, développé avec un assistant de code IA (Claude Code).**

08

Le rendu final

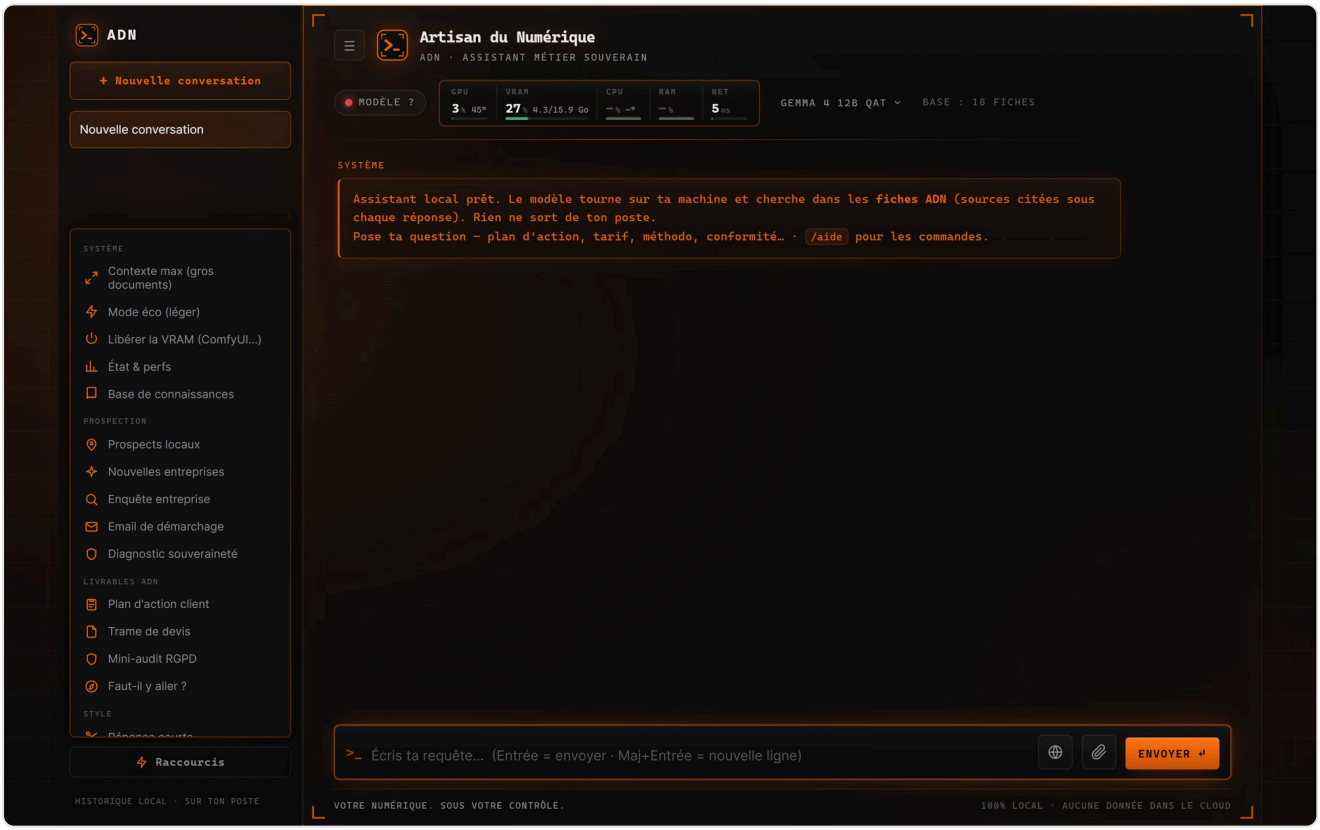


fig. 1 — L'interface, lancée pour ce dossier sans modèle chargé : raccourcis (système, prospection, livrables), jauges de la machine, choix du modèle, base de 18 fiches.

En usage, une réponse s'écrit mot à mot sous une « orbe » de réflexion, avec le journal des étapes (planification, registre, web, rédaction) et les fiches citées en dessous. Les tableaux s'exportent en CSV.

09

Où il en est

La version allégée est prête depuis le 12 octobre. Prochaines étapes identifiées :

— décliner la version légère en offre « assistant IA métier » pour une petite entreprise sans grosse carte graphique.

Sources : fichiers du projet, rapport du banc d'essai des modèles (12/10/2026), capture faite le 25/10/2026 sans modèle chargé