IA à l'UdN : ce qu'on peut faire, ce qu'on ne peut pas faire

Document de cadrage pour l'équipe — pour comprendre ce qu'une IA sur notre propre serveur sait faire, ce qui nécessite une IA externe (type Claude), ce que ça coûte, et où se situent les questions qu'on se pose aujourd'hui.

---

1. Panorama : s'y retrouver dans l'IA

Avant d'entrer dans le détail, deux mots de vocabulaire évitent bien des confusions.

Un « modèle » de langage, c'est un programme entraîné sur d'énormes quantités de texte, qui a appris à prédire la suite la plus probable d'une phrase. On mesure sa « taille » en **paramètres** (les réglages qu'il a appris) : plus il y en a, plus il est capable de nuance et de raisonnement — mais aussi plus il est lourd à faire tourner. C'est un repère utile pour la suite : nos petits modèles maison pèsent 7 à 8 **milliards** de paramètres, les grands modèles du cloud en pèsent **des centaines de milliards**.

Assistant ≠ modèle. On confond souvent deux choses. L'**assistant**, c'est le produit qu'on ouvre (ChatGPT, Claude, Le Chat de Mistral…). Le **modèle**, c'est le moteur à l'intérieur (GPT-5.5, Claude Opus, Mistral Large). C'est comme une voiture et son moteur : on peut changer de moteur sans changer de voiture. Quand vous entendez « GPT-5 » ou « Mistral 3 », ce sont des moteurs, pas des marques différentes.

Les principaux assistants grand public (2026) :

- **ChatGPT** (OpenAI) — le couteau suisse, le plus polyvalent. Moteur : GPT-5.5.
- **Claude** (Anthropic) — rédaction longue, code, données sensibles. Moteurs : Opus, Sonnet, Haiku, Fable.
- **Gemini** (Google) — profondément intégré à Gmail, Docs, Drive. Moteur : Gemini 3 Pro.

- **Copilot** (Microsoft) — l'assistant de Word, Excel, Outlook, Teams. S'appuie sur les moteurs d'OpenAI.
- **Mistral / Le Chat** (français) — le « champion souverain », données hébergées en Europe. Moteurs : Mistral Large 3, Medium 3.5.

À retenir : en 2026, environ **trois entreprises sur quatre** utilisent au moins un outil d'IA, et **aucun outil ne domine sur tout**. Le bon choix dépend du contexte — c'est précisément l'objet de ce document pour l'UdN.

Sources : blog-ia.fr · stephanelarue.com

---

2. Les deux mondes de l'IA (à comprendre d'abord)

Il existe deux familles d'IA très différentes, et le choix entre les deux dépend de deux critères : **la sensibilité des données**, et **la complexité de la tâche**.

L'IA sur notre serveur (auto-hébergée — Ollama, Whisper)

- **Avantage : souveraine.** Nos données ne sortent jamais de nos machines. Idéal pour tout ce qui est confidentiel (gouvernance de nos clients, conflits, données personnelles).
- **Limite :** c'est un « petit » modèle (7-8 milliards de paramètres). Il fait bien les tâches simples et cadrées, mais se perd sur le raisonnement complexe. Il est aussi plus lent (quelques dizaines de secondes par réponse).
- **Bon pour :** transcrire, résumer, reformuler, des tâches répétitives et cadrées.

L'IA externe / cloud (type Claude, ChatGPT)

- **Avantage : très puissante** (des centaines de milliards de paramètres). Excellente sur le raisonnement complexe, la rédaction élaborée, les tâches multi-étapes.
- **Limite :** nos données transitent chez le fournisseur (souvent américain). À éviter pour le très sensible, ou alors en anonymisant d'abord.
- **Bon pour :** le travail cognitif complexe, la rédaction fine, l'analyse poussée.

- **Facturation** : Nous avons deux utilisations possibles : Avec le site ia.universite-du-nous.org ou on sera facturé au token (j'ai mis une limite à 20€ par mois pour l'UdN). sinon via les abonnements des plateformes comme Claude ou ChatGpt

La règle simple

- **Donnée sensible + tâche simple → notre serveur.**
- **Tâche complexe + besoin de qualité → IA externe** (en anonymisant si sensible).

Un point important à retenir

Notre IA locale n'« apprend » **pas toute seule** de nos échanges. Elle s'améliore parce que **nous** enrichissons sa base de connaissances (nos méthodes, nos exemples) au fil du temps. Nous avons besoin de lui donner nos fichiers — qualifs, comptes rendus, déroulés, propositions commerciales — pour qu'elle puisse se servir de ces exemples. Le « de mieux en mieux » est **piloté par nous**, pas automatique.

---

3. Ce que l'IA peut faire pour nous — 5 familles d'usage

Plutôt qu'une liste à rallonge, on peut regrouper les usages en cinq familles, avec à chaque fois un exemple qu'on vit déjà (ou qu'on pourrait vivre) à l'UdN.

- **Synthétiser** — produire un compte rendu à partir de photos de paperboard, résumer la transcription d'une réunion (enjeux, objectifs, prochains pas, « ce qui est dit en creux », points à approfondir), analyser un débrief (les questions qu'on se pose, les difficultés rencontrées, les points d'attention).
- **Transformer** — transcrire un audio en texte (Whisper), traduire un document, reformuler ou proposer un « pas de côté » sur une idée.
- **Concevoir** — écrire des consignes d'ateliers, une trame 6 chapeaux de De Bono, des supports pédagogiques (ESC, GPC, objections…).
- **Automatiser** — mettre à jour des indicateurs, remplir ou exploiter des fichiers Excel et Odoo, préparer et trier des mails, fiabiliser un import RoleBase.
- **Analyser** — lire et structurer des documents longs : rapports, conventions, comptes rendus.

Ces familles se recoupent avec le tableau ci-dessous, qui précise **où** faire chaque tâche.

---

4. Nos usages, classés

Légende : □□ notre serveur (souverain) · □□ IA externe recommandée (qualité) · □□ l'humain reste indispensable

Cas d'usage	Où ?	Commentaire honnête
Transcrire une réunion d'1h	□□ Notre serveur	Whisper fait la transcription très bien, en souverain. Une relecture humaine reste utile (noms propres, acronymes mal transcrits).
Synthèse d'une transcription de réunion	□□ Notre serveur (correct) / □□ pour une synthèse fine	Résumé basique : notre IA suffit. Synthèse nuancée qui capte les enjeux : l'IA externe est bien meilleure.
Produire un déroulé de réunion (3h à 1 journée)	□□ IA externe	Faisable en local mais générique. Un bon déroulé demande du raisonnement et de la finesse. L'IA externe aide surtout à la créativité — sachant qu'elle ne peut pas atteindre notre finesse. Notre expertise de facilitation reste le cœur.
Proposer des questions (6 chapeaux de De Bono)	□□ Notre serveur	Tâche cadrée et méthodique, bien adaptée à une IA locale, surtout avec nos exemples en base de connaissance.
À partir de notes, imaginer un séminaire d'une journée	□□ IA externe	Conception créative multi-étapes : un gros modèle est nettement meilleur. L'IA locale donnerait un canevas plat.
Répondre à une offre (cahier des charges + réunion de qualif)	□□ IA externe	Tâche complexe (comprendre un besoin, argumenter, structurer). Données parfois sensibles → anonymiser ou traiter le non-sensible. C'est typiquement ce qu'on fait déjà avec Claude.
Document de proposition commerciale avec notre charte graphique	□□ IA externe (texte) + □□ mise en forme	L'IA rédige le contenu ; la charte graphique se gère avec nos outils de mise en page (l'IA ne « pose » pas une charte graphique complexe seule).
Retrouver des infos dans nos documents (« a-t-on déjà bossé sur les tiers-lieux ? qu'a-t-on fait ? »)	□□ Notre serveur (RAG)	Cas idéal pour le RAG local : indexer nos livrables et comptes rendus, puis interroger en souverain. Très utile et faisable chez nous.
Synthèse de nos accompagnements à partir des livrables	□□ Notre serveur (RAG) / □□ pour une analyse fine	Retrouver et résumer : local. Tirer une analyse transversale profonde : IA externe.

Cas d'usage	Où ?	Commentaire honnête
Retours d'expérience (REX) à partir de nos documents	□□ Notre serveur (RAG)	Souverain (données internes), et c'est un usage par lots où la lenteur du local n'est pas gênante.
Rapport de séminaire à partir de photos (post-its, tableaux)	□□ IA externe	Lire des photos (post-its manuscrits, affiches) demande une IA « multimodale » puissante. Les petits modèles locaux lisent mal les images.
Webinaires (≈50 vidéos) → extraire l'essentiel, faire des shorts / slides réseaux	□□□□ Mixte, gros chantier	Transcription : □□ local. Extraction des moments forts + rédaction des posts : □□ externe. Montage des shorts + visuels : □□ outils dédiés + humain. Projet réaliste, mais en plusieurs briques.
Imaginer ET mettre en œuvre un plan de com	□□ pour imaginer / □□ pour mettre en œuvre	L'IA aide à concevoir un plan. La mise en œuvre (publier, animer, ajuster) reste un travail humain + outils. Ne pas attendre une automatisation complète.
Image de synthèse / facilitation graphique à partir d'un audio	□□ Très limité aujourd'hui	Transformer un audio de réunion en une vraie facilitation graphique de qualité : aucune IA ne le fait bien aujourd'hui. On peut extraire des idées clés (□□/□□) et éventuellement générer des visuels, mais la facilitation graphique reste un savoir-faire humain. (ChatGPT propose des choses si on le guide bien.)
Lire mes mails, préparer des réponses, automatiser des tâches quotidiennes	□□ IA externe uniquement	Hors de portée de l'IA locale : demande un modèle très puissant et des capacités d'agent.
« Lire les dynamiques collectives, ce qui se joue »	□□ Reste humain	L'IA peut repérer des signaux de surface (temps de parole, thèmes récurrents, tensions lexicales) comme support à votre analyse. Mais « ce qui se joue » — le non-dit, l'énergie, le relationnel — reste votre expertise. Aide au repérage, jamais un verdict.

Quelques cas d'usage supplémentaires qui pourraient nous être utiles :

Cas d'usage	Où ?	Intérêt pour nous
Traduire des documents (FR ↔ EN…)	□□ Notre serveur	Souverain, correct pour des traductions de travail.
Préparer un compte rendu selon NOTRE format à partir d'un transcript	□□ Notre serveur (RAG)	En lui donnant nos modèles de CR, elle produit dans notre structure. Gain de temps réel.

Cas d'usage	Où ?	Intérêt pour nous
Générer des variantes (reformuler une raison d'être, plusieurs formulations d'une redevabilité)	Notre serveur	Tâche cadrée, adaptée au local. C'est déjà l'usage IA dans RoleBase.
Anonymiser un document avant de l'envoyer à une IA externe	Notre serveur	Utile comme « sas » : le local retire les infos sensibles, puis on peut utiliser l'IA externe sans risque.
Base de connaissances interne interrogeable (toute notre doc méthodo)	Notre serveur (RAG)	Un « moteur de recherche intelligent » sur nos savoirs UdN. Très structurant à terme.
Brainstorming / idéation cadrée (ice-breakers, métaphores d'inclusion…)	Notre serveur	Léger, souverain, dépanne bien.

---

5. Les enjeux : souveraineté & écologie

L'IA n'est pas neutre. Deux sujets nous concernent directement, nous à l'UdN, et éclairent pourquoi on tient à faire tourner une partie de l'IA chez nous.

Souveraineté

Aujourd'hui, **86 %** des entreprises françaises confient encore leurs données sensibles à des IA américaines soumises au Cloud Act. Le marché reste dominé par des acteurs américains, aussi bien pour les modèles (OpenAI, Google, Anthropic) que pour les processeurs graphiques (**Nvidia ~95 %** du marché) et l'infrastructure (AWS, Azure, Google Cloud). Choisir des outils souverains, ou héberger nous-mêmes, c'est reprendre la main sur nos données et celles de nos clients.

Écologie

L'entraînement et l'usage massif de l'IA ont un coût environnemental réel :

- Les data centers dans le monde ont consommé **environ 415 TWh en 2024**, avec une projection de **800 à 1 000 TWh d'ici 2027** — plus du double en trois ans, l'IA générative en étant le principal moteur.
- En France, la consommation des data centers a augmenté de **38 % depuis 2021**.
- Toujours en France, **160 centres ont prélevé 575 000 m³ d'eau potable en 2024**, l'équivalent de la consommation annuelle d'une commune de **10 000 habitants**.

Ces chiffres invitent à un usage sobre : privilégier le local pour le volume, réserver les gros modèles cloud aux usages qui le justifient vraiment.

Sources : journaldunet.com · macertif.com · Bpifrance · economiematin.fr

---

6. Approfondir : IA locale, Ollama, RAG & skills

Cette section explique, sans jargon, les briques techniques derrière « notre serveur ».

Mistral, l'alternative souveraine

Mistral est l'éditeur français d'IA. Ses atouts : **hébergement européen**, conformité RGPD et AI Act « natives », et des **modèles open weight** (téléchargeables et auditables) — donc utilisables chez nous. À qualité comparable, ses tarifs sont souvent plus bas, et ses modèles sont particulièrement à l'aise en français. La nuance honnête : la souveraineté n'est pas qu'une étiquette. Mistral dépend encore de sous-traitants et de processeurs hors UE, même s'il documente ses transferts. Autrement dit, c'est **la meilleure option souveraine crédible**, pas une garantie absolue.

Une IA sur son PC ? Les besoins matériels

Faire tourner un modèle en local, c'est possible sur une machine ordinaire, à condition d'ajuster ses attentes à son matériel :

- **Minimum** avec Ollama : 8 Go de RAM, un processeur 64 bits, **pas de carte graphique obligatoire**.
- **Confortable** : 16 Go de RAM et une carte graphique de 8-12 Go, qui fait tourner des modèles de 7 à 14 milliards de paramètres à bonne vitesse.
- **Règle simple :** compter environ **1 Go de RAM par milliard de paramètres**. Sans carte graphique, ça fonctionne quand même, mais c'est plus lent.

C'est cette contrainte matérielle qui explique pourquoi « notre serveur » utilise de petits modèles : ils tiennent en mémoire et restent rapides.

Ollama, c'est quoi ?

Ollama est un logiciel qui permet de lancer un modèle open source **en local**, en une seule commande. Il peut aussi utiliser des modèles d'IA d'internet qu'il soit gratuit ou payant (comme Claude Sonnet ou ChatGPT 5.5). L'analogie la plus parlante : Ollama, c'est comme un **lecteur de musique installé sur votre machine (type VLC)**, par opposition à un **service de streaming (comme ChatGPT en ligne)**. Avec le lecteur local, vous possédez les fichiers (ici, les modèles), ça marche sans connexion, et personne ne sait ce que vous « écoutez » — vos conversations restent privées. Concrètement, comme aucune donnée ne quitte la machine, il n'y a **aucun transfert hors UE** : c'est la voie la plus directe vers une IA conforme au RGPD. ça marche seulement pour les IA téléchargé en local.

Une IA sur un VPS pour toute l'équipe — et ses limites

Plutôt que d'installer l'IA sur chaque poste, on peut la faire tourner sur un **serveur mutualisé** (notre VPS dédié IA) accessible à toute l'équipe. Avantage : un seul serveur pour tous, coût partagé. Limites à connaître : la puissance et la mémoire graphique sont **partagées** entre les utilisateurs, le matériel a un coût, et — règle de sécurité — il ne faut **jamais exposer** ce service directement sur internet sans authentification.

RAG : brancher l'IA sur NOS documents

Le **RAG** (Retrieval-Augmented Generation, « génération augmentée par récupération ») est la brique la plus utile pour nous. Le principe, en trois temps :

1. **Chercher** — l'IA retrouve, dans nos propres documents, les passages pertinents pour la question posée.
2. **Augmenter** — elle ajoute ces passages au contexte de la question.
3. **Générer** — elle rédige une réponse en s'appuyant sur ces sources, donc **vérifiable**.

Deux bénéfices concrets : **beaucoup moins d'erreurs inventées** (puisque la réponse s'appuie sur des sources réelles), et une mise en place **5 à 10 fois plus rapide** — et bien moins coûteuse — qu'un ré-entraînement de modèle, avec une base facile à mettre à jour. Pour nous, c'est un assistant branché sur nos comptes rendus, notre wiki (BookStack), nos procédures et RoleBase, notre drive.
C'est un sacré boulot pour faire les RAG. donc ça va prendre du temps mais plus on travail dessus, plus l'IA fera ce qu'on lui demande en fonction de notre métier. Elle apprends de nous, par contre on le l'entraine pas du tout.

Les « skills »

Un **skill**, c'est une « compétence » ou un mode d'emploi qu'on donne à l'IA pour qu'elle exécute une tâche précise **à notre façon** — par exemple produire un compte rendu au format UdN, ou suivre notre méthode de facilitation. C'est le moyen d'inscrire nos savoir-faire dans l'outil, sans reprogrammer quoi que ce soit.

Sources : digitiz.fr · localaimaster.com · tech-insider.org · natural-net.fr (RAG)

---

7. Combien ça coûte ?

Deux modèles économiques très différents.

L'IA sur notre serveur (Ollama/Whisper) — coût FIXE

- On paye le serveur (VPS dédié IA : **~10-20 € HT/mois**).
- Et c'est tout. Que l'on fasse 10 ou 10 000 requêtes, le prix ne change pas. Le modèle est **gratuit** (open source).
- Soit **moins d'1 €/personne/mois** pour 15 personnes. Imbattable — mais uniquement pour les tâches que l'IA locale sait faire.

L'IA externe (Claude, etc.) — coût VARIABLE (à l'usage)

- Soit par abonnement (**~20 €/mois par personne** pour un usage confortable),
- Soit à la consommation (API : on paye selon le volume traité — de quelques euros à bien plus selon l'intensité).
- Le coût est souvent **par utilisateur**, pas mutualisé (il existe des offres équipe, mais individualisées par siège).

Comparaison

	IA locale (notre serveur)	IA externe (cloud)
Coût	~12-15 €/mois fixe, illimité	~20 €/mois/pers (abo) ou variable (API)
Mutualisation	oui (un serveur pour tous)	souvent par personne

	IA locale (notre serveur)	IA externe (cloud)
Données	souveraines	transitent chez le fournisseur
Qualité	correcte (tâches simples)	excellente (tâches complexes)
Vitesse	plus lente (CPU)	rapide

Le modèle réaliste pour l'UdN

Ce n'est pas « l'un OU l'autre », mais **les deux, selon l'usage** :

- **IA locale** pour le gros volume de tâches simples et sensibles (coût fixe mutualisé, très économique).
- **IA externe** pour le travail complexe de ceux qui en ont besoin (coût par personne, à réserver aux usages qui le justifient).

---

8. Nos questionnements actuels

Voici les questions qu'on se pose honnêtement aujourd'hui à l'UdN, avec des éléments pour y voir clair.

Nos données clients - souveraineté ?

C'est la question centrale. Le point souvent mal compris : le **Cloud Act** vise la **nationalité juridique** du fournisseur, pas la localisation des serveurs. Un datacenter en France opéré par un acteur américain **reste soumis** au Cloud Act. Nos leviers de protection :

- choisir un prestataire de **droit européen** (idéalement certifié SecNumCloud par l'ANSSI) ;
- **chiffrer** en gardant nos propres clés ;
- pour le sensible, passer par l'**IA locale (Ollama)** ou **Mistral** ;
- ne **jamais** coller de données personnelles dans une version gratuite grand public ;
- utiliser notre local comme **sas d'anonymisation** avant tout envoi à une IA externe.

Cadre applicable : le RGPD et l'AI Act, avec des sanctions pouvant atteindre 20 M€ ou 4 % du chiffre d'affaires mondial. Autrement dit, ce n'est pas un détail : c'est structurant pour notre responsabilité.

Les IA local ne pose aucun problème puisque tout est sur notre serveur, par contre des qu'on utilise une IA payante ou gratuite sur le web ça transite sur ses serveurs. Je ne crois pas qu'on aura de problème de confidentialité car les données sont trop grande et c'est pas nous avec nos recherches qui entrainons les IA, c'est les boites comme anthropic ou openAI qui entraine les IA. Nous n'avons

pas les moyens. Donc même si nos données transite, ça pose plus des questions éthique que de fuite.

Garde-t-on une vérification humaine ?

Oui, et c'est un principe de travail, pas une option. On traite chaque sortie d'IA comme un **brouillon**, jamais comme une vérité. En pratique, un tri simple : les tâches à faible enjeu et réversibles (résumer un mail, proposer un titre) n'ont pas besoin de relecture systématique ; les tâches à **fort enjeu ou irréversibles** (une réponse à une offre, un document qui nous engage) passent **toujours** par une validation humaine.

Les erreurs de l'IA (« hallucinations »)

Une IA peut produire un contenu **plausible mais faux** : c'est ce qu'on appelle une hallucination. Ce n'est pas un bug rare, c'est un comportement **structurel** des modèles de langage. Le risque, pour nous, n'est pas l'erreur en soi, mais **l'erreur non détectée** dans un document qui nous engage. La parade tient en trois mots : **ancrer** (RAG + recherche web pour appuyer les réponses sur des sources), **vérifier** (relecture humaine sur le sensible), **former** l'équipe à repérer une réponse inventée. Point de vigilance utile : on a tendance à faire davantage confiance à une réponse d'IA bien formulée qu'à une source humaine — d'où l'importance de garder un œil critique, surtout sur les chiffres, les citations et les références.

---

9. Ce qu'il faut retenir

- **Notre IA locale** peut être notre outil pour tout ce qui est simple + sensible : transcription, synthèses cadrées, recherche dans nos documents, reformulations. Elle protège nos données et celles de nos clients.
- **L'IA externe** (Claude / ChatGPT) reste nécessaire pour le travail complexe : conception créative, rédaction fine, réponses à des offres, analyse poussée. On l'utilise en anonymisant quand c'est sensible.
- **L'humain reste au centre.** L'IA prépare, propose, dégrossit. Mais la facilitation, la lecture des dynamiques, la relation, le sens — c'est nous. L'IA est un assistant, pas un remplaçant de notre métier.
- **Attention aux attentes.** Certaines choses qu'on imagine « faciles pour une IA » (lire une facilitation graphique depuis un audio, comprendre finement un collectif) ne sont pas mûres. Mieux vaut viser des usages solides que d'être déçus.

Chiffres et versions de modèles à jour à la date de rédaction (juillet 2026). L'IA évolue vite : à revérifier avant toute décision d'outillage.

Revision #8
Created 2026-07-01 08:57:45 UTC by Yohan Reversat
Updated 2026-07-04 19:05:47 UTC by Yohan Reversat