SPÉCIMEN · RAPPORT D'AUDIT IA

Un rapport d'audit IA, en entier, sur une entreprise qui n'existe pas

Voici le livrable du format Audit IA, publié tel quel. Cinq pistes examinées, une lancée, une reportée, deux écartées, et une qui n'est pas un sujet d'IA générative. Avec les calculs apparents et les annexes techniques.

CAS FABRIQUÉ

Cette entreprise n'existe pas

CAS FABRIQUÉ

Cette entreprise n'existe pas

Je ne publie pas de rapport client, même anonymisé. Un audit décrit l'organisation réelle d'une entreprise, ses lenteurs et ses arbitrages : ça ne s'expose pas au motif que ça rendrait service à mon site.

Le cas qui suit est donc fabriqué de bout en bout. L'entreprise, les chiffres, les entretiens, le projet abandonné : tout est inventé. Ce qui ne l'est pas, c'est la forme du document, l'ordre des questions, la façon dont un verdict se motive et le niveau de détail des annexes.

Tous les chiffres sont des hypothèses de travail. Quand un chiffre est un résultat, ses termes sont écrits juste à côté, pour que vous puissiez refaire l'opération et contester l'hypothèse plutôt que le total.

LE SUJET

Le distributeur

Distribution B2B de fournitures industrielles : roulements, visserie, équipements de protection, outillage. Clients industriels, ateliers de maintenance et collectivités.

EFFECTIF 380 personnes	IMPLANTATION 4 agences, 1 plateforme logistique
CATALOGUE 34 000 références actives	COMMANDES environ 4 800 par mois
ADMINISTRATION DES VENTES 14 personnes	SYSTÈMES un ERP de 2013, un CRM séparé, un PIM partiel
DOCUMENTATION 34 000 fiches PDF, dont 40 % de scans	ANTÉCÉDENT un chatbot arrêté au bout de 4 mois en 2025

EN UNE PAGE

Ce que je recommande, et ce que je refuse

Cinq pistes ont été examinées. Une seule démarre tout de suite. Une deuxième démarre après un chantier qui n'a rien d'intelligence artificielle et qui doit passer devant. Une troisième attend que la première tienne ses seuils. Les deux dernières ne se font pas, et l'une d'elles n'est pas un sujet d'IA générative du tout.

Le reste du rapport motive chacun de ces cinq verdicts, montre les calculs qui les portent, et donne l'architecture visée pour les deux chantiers retenus.

RÉF.	PISTE	VERDICT	PROCHAIN PAS	GAIN ESTIMÉ
C1	Saisie des commandes reçues par e-mail — L'intégration de fichiers structurés couvre 62 % des lignes pour moins cher et sans jamais se tromper. L'IA prend la traîne, soit environ 900 documents par mois.	Oui, après l'intégration	12 j, après l'intégration	environ 630 h par an
C2	Recherche dans la documentation technique — Périmètre borné, réponse vérifiable parce que sourcée, et un jeu d'évaluation constructible dès maintenant à partir des questions déjà posées au support.	À lancer en premier	4 j de cadrage	environ 1 320 h par an
C3	Aide à la réponse aux appels d'offres — Ce cas suppose que la recherche documentaire tourne. Sans elle, c'est de la rédaction non sourcée, c'est-à-dire ce qu'un appel d'offres tolère le moins.	Plus tard	à réexaminer	non chiffré
C4	Chatbot client sur le site — Les trois causes de l'échec de 2025 n'ont pas bougé : aucune source de vérité sur le stock et les prix, aucun jeu d'évaluation, aucune escalade vers un humain.	Non, pas maintenant	aucun	aucun
C5	Prévision des ruptures de stock — C'est de la statistique sur l'historique de l'ERP, et cet historique a un trou de sept mois. Le chantier est un chantier de données.	Pas un sujet d'IA générative	un chantier de données	non chiffré

LES CHARGES SONT EN JOURS DE TRAVAIL, JAMAIS EN EUROS. LES GAINS SONT EN HEURES PAR AN, AVEC LEUR CALCUL EN SECTION 05 : VOTRE COÛT HORAIRE CHARGÉ, VOUS LE CONNAISSEZ MIEUX QUE MOI.

SOMMAIRE

Le rapport

01 · CAS FABRIQUÉ

Le mandat, et ce que j'ai regardé

Le périmètre, la méthode, et ce que quatre jours ne permettent pas de savoir.

La direction du distributeur a posé la question dans ces termes : « Tout le monde parle d'IA : où ça sert vraiment ici, et par quoi commencer ? » Telle quelle, cette question n'a pas de réponse, parce qu'elle porte sur une technologie et pas sur un travail. Je l'ai donc réécrite en **question auditable**, c'est-à-dire en question à laquelle quatre jours de travail peuvent répondre. La voici : parmi les usages que vos équipes citent spontanément, lesquels tiennent devant les données qui existent réellement aujourd'hui, pour quel effort, et dans quel ordre les prendre ?

Pendant ces quatre jours, j'ai conduit 11 entretiens, du dirigeant au magasinier, et 3 observations de poste en ADV, l'administration des ventes. Un entretien vous dit ce que les gens croient faire, une observation montre les gestes qu'ils ne pensent plus à raconter : les reprises de saisie, les allers et retours entre deux écrans, les vérifications qui ne figurent dans aucune procédure. J'ai aussi repris 2 exports de l'ERP sur 12 mois et échantillonné 400 documents au hasard.

LE DÉROULÉ

JOUR	CE QUE JE REGARDE	CE QUE ÇA PRODUIT
Jour 1	Entretiens, circuit des commandes	Liste des usages candidats
Jour 2	Observations en ADV, outils du poste	Volumes et gestes réels
Jour 3	Exports ERP, échantillon de documents	État des données disponibles
Jour 4	Effort, risques, ordre de passage	Un verdict par piste
Restitution	Relecture avec les personnes vues	Rapport commenté, décisions actées

Le quatrième jour n'ajoute aucune matière, il sert à chiffrer l'effort et à mettre les cinq pistes dans un ordre. La restitution se fait devant les personnes que j'ai interrogées, et c'est le seul filtre qui compte : si une phrase du rapport ne tient pas devant celui qui fait le travail tous les jours, elle saute. Le cadre général est décrit dans ma méthode.

Ce que cet audit ne dit pas

- Ce n'est pas un audit de sécurité : je regarde où sont les données, pas si elles sont correctement protégées.
- Ce n'est pas une revue du code de l'ERP : je prends le système tel qu'il tourne, sans juger la façon dont il est écrit.
- Ce n'est pas un test utilisateur : trois observations montrent des gestes, elles ne mesurent pas une adoption.
- Un échantillon de 400 documents donne des **ordres de grandeur**, jamais des taux : les pourcentages de ce rapport se lisent comme des repères de discussion.

RÉSERVE

400 documents tirés au hasard, c'est peu. Un deuxième tirage donnerait d'autres chiffres, et je n'ai pas fait ce deuxième tirage. Quand j'écris que 40 % des fiches fournisseurs sont des scans, lisez « entre un tiers et la moitié ». Si une de vos décisions se joue à quelques points de pourcentage, mes chiffres ne la fondent pas : il faut compter, et compter demande plus de quatre jours.

02 · CAS FABRIQUÉ

L'état des lieux

Où passe le temps de l'administration des ventes, dans quel état sont les données, et ce que le POC de 2025 a appris.

Ce que je décris ici vient de 11 entretiens, de 3 observations de poste en ADV, de 2 exports de l'ERP sur 12 mois et de 400 documents tirés au hasard. C'est peu pour tout savoir, c'est assez pour voir où le travail se perd. Je décris le fonctionnement réel, pas celui des procédures.

Le chemin d'une commande

Une commande arrive par e-mail. Quelqu'un ouvre le message, puis la pièce jointe : un PDF, un tableur, parfois quelques lignes écrites dans le corps du message. Il lit, il retape les références une par une dans l'ERP, il vérifie le prix, il renvoie une confirmation au client. Ce trajet occupe une part importante des journées des 14 personnes de l'ADV.

Sur environ 4 800 commandes par mois, **61 % arrivent par e-mail**, 27 % par le portail web et 12 % par téléphone. Cette répartition décide de la suite : les 27 % du portail arrivent déjà structurés et ne se retapent pas, les 12 % du téléphone ne relèvent pas d'un modèle de langage. Le sujet tient donc entier dans les 61 %, et nulle part ailleurs.

La documentation technique

Le distributeur conserve 34 000 fiches PDF fournisseurs. D'un fournisseur à l'autre, rien ne se ressemble : la structure, le vocabulaire, les unités. Environ 40 % de ces fiches sont des scans, c'est-à-dire des images de page, sans texte à l'intérieur.

La conséquence est concrète. Une recherche par mot clé ne voit pas ces 40 %. Un vendeur qui cherche un diamètre ou une norme dans une fiche scannée ne trouve rien, alors que l'information est bien là, visible à l'écran. Il ouvre alors les fiches une par une. C'est le point que les entretiens ramènent le plus souvent.

L'état des données

Deux faits pèsent plus que les autres. L'historique de l'ERP porte un **trou de 7 mois en 2022**, au moment de la migration : les mouvements de cette période n'ont pas été repris. Ce trou interdit une chose précise, la prévision statistique sur les ventes : un calcul de saisonnalité a besoin d'années complètes pour comparer un mois à lui-même, et 2022 n'est plus une année complète.

Le PIM, le référentiel qui décrit les produits, est partiel et alimenté à la main. Il interdit autre chose : servir de source de vérité. Aucun assistant ne peut affirmer un prix, un stock ou une caractéristique en s'appuyant dessus, faute de savoir si la fiche a été mise à jour. J'ai détaillé ailleurs ce qui rend un jeu de données exploitable : vos données sont-elles prêtes pour l'IA ?.

Le POC de 2025

Un chatbot client a été commandé à une agence en 2025, puis arrêté au bout de 4 mois. Je ne reproche rien à cette agence : l'erreur est banale, et elle est structurelle. Trois causes se cumulent, et aucune ne tient à la qualité du travail fourni.

- Aucune source de vérité n'était exposée sur le stock et les prix : le chatbot répondait sans rien pouvoir vérifier.
- Aucun jeu d'évaluation n'existait, donc personne ne pouvait dire s'il répondait juste, ni constater un progrès d'une version à l'autre.
- Aucune procédure d'escalade vers un humain n'était prévue, donc un client bloqué restait bloqué.

> Ce POC n'a pas échoué à cause du modèle. Il a échoué parce qu'on lui a demandé de répondre sur des données qu'on ne lui avait pas données, sans aucun moyen de savoir s'il se trompait.

Ces trois manques se réparent, et ils se réparent ailleurs que dans le chatbot. C'est la raison pour laquelle je propose plus loin de commencer par autre chose.

03 · CAS FABRIQUÉ

Ce que je lance, et dans quel ordre

Un chantier à démarrer maintenant, un second à démarrer après une intégration qui n'a rien d'IA.

L'ordre qui suit n'est pas un classement de mérite, c'est un ordre d'exécution. C1 fait plus de bruit dans l'entreprise : la saisie des commandes occupe l'ADV tous les jours, et c'est le sujet cité en premier dans les entretiens. C2 passe pourtant devant, pour une raison simple. C'est le seul des cinq cas dont on peut vérifier la qualité avant de le mettre entre les mains de quelqu'un, et sur un terrain où une erreur ne part pas chez un client. Commencer par lui, c'est apprendre à mesurer sur un périmètre sans conséquence, avant d'appliquer la même discipline à un chantier qui touche des commandes réelles.

C2 : recherche dans la documentation technique

La douleur est nommée par 9 des 11 entretiens, sans que la question soit posée. Un technicien cherche une cote, un couple de serrage, une équivalence entre deux références. Il ouvre trois fiches, il appelle un collègue, il finit par appeler le fournisseur. Personne ne sait dire combien de temps ça prend. Tout le monde sait dire que ça arrive plusieurs fois par jour.

L'assistant fait une chose et une seule : il retrouve les passages utiles dans les 34 000 fiches fournisseurs, et il répond en citant la fiche et la page. Il ne rédige pas d'argumentaire commercial, il ne devine pas un prix, il ne remplace pas le fournisseur. Quand il ne trouve pas, il le dit. C'est le schéma décrit dans l'assistant documentaire.

Trois raisons en font le premier chantier, et le verdict est **oui, à lancer en premier**. Le périmètre est borné : un corpus, une question, une réponse. La réponse est vérifiable parce qu'elle est sourcée : celui qui lit ouvre la fiche citée et tranche en dix secondes. Et le jeu d'évaluation, c'est-à-dire la liste des questions dont on connaît la bonne réponse, est constructible dès maintenant. Le support a déjà reçu ces questions. Elles sont dans les tickets et dans les boîtes mail, il suffit de les rassembler.

Ce qui le ferait échouer est connu d'avance. Environ 40 % des fiches sont des scans, donc des images sans texte exploitable. Si l'étape de lecture d'image est bâclée, l'assistant répondra à côté sur cette part du corpus, sans jamais signaler qu'il ne sait pas. Deuxième risque, un périmètre qui s'élargit en cours de route vers les prix, le stock et les conditions commerciales, trois sujets dont la source de vérité est ailleurs. Un point reste à vérifier avant de lancer : ce que les contrats fournisseurs autorisent en matière d'indexation interne.

C1 : saisie des commandes reçues par e-mail

Le verdict est **oui, mais pas comme prévu**. La demande initiale était une IA qui lit toutes les commandes arrivant par e-mail. L'export ERP dit autre chose : 62 % des lignes de commande viennent de 23 clients réguliers. Ces 23 clients peuvent envoyer un fichier structuré, c'est-à-dire un format que l'ERP sait lire seul. Reste 38 % des lignes, soit environ 900 documents par mois, dans des formats qu'aucune intégration ne couvrira : scans, tableurs bricolés, commandes tapées dans le corps du message. C'est là, et seulement là, que l'IA a un rôle.

Sur la part intégrable, une intégration de fichiers classique bat l'IA sur les trois critères qui comptent. Elle coûte moins cher à poser et à faire tourner. Elle ne se trompe jamais, parce qu'elle ne fait aucune inférence : soit le fichier est conforme, soit il est rejeté. Et elle ne consomme rien à l'usage, pas un token. Mettre un modèle de langage là où un format de fichier suffit, c'est payer une lecture à chaque document pour un travail qu'une correspondance entre champs règle une fois pour toutes.

LES DEUX VOIES POUR C1, ET CE QU'ELLES COUVRENT

VOIE	CE QUE ÇA COUVRE	COÛT DE FONCTIONNEMENT	SE TROMPE
Fichier structuré	62 % des lignes, 23 clients	Nul une fois posé	Non, le fichier passe ou il est rejeté
Extraction par IA	38 % des lignes, 900 documents par mois	Facturé au document traité	Oui, taux à mesurer et à surveiller

La règle de mesure se pose avant le premier essai, pas après. On ne mesure pas le taux de réussite du modèle : ce chiffre ne veut rien dire pour l'ADV, qui ne compte pas des champs mais des commandes. On mesure la part des commandes intégrées sans reprise humaine, sur un échantillon relu ligne par ligne. C'est ce seul chiffre qui décide de continuer, d'ajuster ou d'arrêter. Le détail de la chaîne de lecture des documents est traité dans ce guide.

L'ordre est donc : intégration des 23 clients d'abord, extraction par IA ensuite, sur la traîne seule. Inverser les deux revient à faire lire par une machine des documents qui n'auraient jamais dû être des documents.

> Une partie du travail d'un projet IA n'est pas de l'IA, et c'est souvent la partie qui produit le gain.

04 · CAS FABRIQUÉ

Ce que je reporte, et ce que j'écarte

Un cas qui dépend d'un autre, un échec qu'il ne faut pas rejouer, et un besoin qui n'appelle pas de modèle de langage.

Trois pistes sur cinq ne partent pas maintenant. Deux sont reportées, une est écartée dans sa forme actuelle. Aucune ne l'est parce que le sujet serait sans intérêt, mais parce qu'une condition manque, et que cette condition est nommable.

C3 : Aide à la réponse aux appels d'offres

Verdict : **plus tard**, et ce n'est pas un refus. Répondre à un appel d'offres, c'est produire un document où chaque affirmation technique engage le distributeur. Un modèle de langage rédige très bien un paragraphe plausible. Sans base de recherche fiable derrière lui, il produit de la rédaction non sourcée, c'est-à-dire exactement le contraire de ce qu'un appel d'offres exige. Le risque n'est pas de perdre du temps, il est d'annoncer une caractéristique produit fausse et de s'y engager.

Ce qui le débloque est précis : que C2 tourne en production, avec des réponses sourcées et un jeu d'évaluation tenu à jour. Le jour où un rédacteur demande une caractéristique et reçoit la fiche fournisseur qui la porte, l'aide à la rédaction devient une couche mince posée sur une base saine. Avant ce jour, C3 n'est pas un chantier, c'est un pari.

C4 : Chatbot client sur le site

Verdict : **non, pas maintenant.** Non de calendrier, pas de principe. Le POC de 2025 s'est arrêté au bout de quatre mois pour trois raisons, et ces trois raisons n'ont pas bougé. Aucune source de vérité n'est exposée sur le stock et les prix. Aucun jeu d'évaluation n'existe, donc personne ne peut dire si l'outil répondait juste. Aucune procédure ne passe la main à un humain quand la machine sort de son périmètre. Relancer sans traiter ces trois points, c'est refaire l'échec avec un modèle plus récent.

Les trois conditions de réouverture se vérifient, elles ne se déclarent pas.

- Le stock et les prix sont lisibles par une machine, en temps réel, par une interface fournie par l'éditeur de l'ERP, et cette interface renvoie la même valeur que l'écran d'un commercial.
- Un jeu d'évaluation d'au moins deux cents questions clients réelles existe, avec la bonne réponse écrite à côté, et le taux de réponses justes est mesuré avant toute mise en ligne.
- Une règle d'escalade est écrite et testée : au bout de deux échecs, ou sur toute question de prix négocié, la conversation part vers l'ADV avec son historique.

S'ajoute la transparence : le règlement européen sur l'IA impose que l'utilisateur sache qu'il parle à une machine. Des cinq cas, ce chatbot public est le seul concerné. Une phrase à l'écran suffit, ce n'est pas un projet. Le cadre est repris dans l'IA Act expliqué aux entreprises.

C5 : Prévision des ruptures de stock

Verdict : **ce n'est pas de l'IA générative.** Prévoir une rupture, c'est lire l'historique des sorties, les délais fournisseurs et les saisonnalités, puis appliquer de la statistique. Un modèle de langage n'apporte rien ici, et il ajoute une source d'erreur là où le sujet demande de la régularité. Le vrai obstacle est ailleurs : l'historique de l'ERP a un trou de sept mois en 2022, à la migration. Une prévision calée sur une année incomplète se trompe avec assurance.

À la place : traiter ce trou, poser une méthode de calcul simple sur les références qui pèsent, et mesurer l'écart entre prévision et réel pendant un trimestre avant d'automatiser quoi que ce soit. C'est moins vendeur qu'un projet d'IA. C'est plus utile, et ce travail de données sert ensuite tous les autres cas, y compris ceux qu'on lance. La trame est dans vos données sont-elles prêtes pour l'IA.

> Un cas écarté avec une condition écrite vaut mieux qu'un cas lancé sans savoir ce qui le ferait échouer.

> CE QUI ME FERAIT CHANGER D'AVIS
>
> Trois conditions, et vous pouvez les constater sans moi. Pour C3, C2 tourne en production depuis un trimestre et son taux de réponses correctes et sourcées est mesuré. Pour C4, les trois points ci-dessus sont cochés et l'ADV a validé la règle d'escalade. Pour C5, le trou de 2022 est traité et douze mois d'historique continu sont disponibles. Si l'une de ces conditions tombe plus tôt que prévu, le cas repasse devant les autres.

05 · CAS FABRIQUÉ

Ce que ça rapporte, ce que ça coûte

Des heures par an, jamais des euros : le calcul est posé, le coût horaire est le vôtre.

Ce rapport ne convertit aucun gain en euros. Je ne connais pas votre coût horaire chargé, ni en ADV, ni au support. Vous le connaissez, et votre direction financière aussi. Un montant inventé dans un rapport d'audit décrédibilise tout ce qui l'entoure, y compris les parties qui tiennent. Je donne donc des heures par an, avec les termes du calcul posés à côté du résultat. La multiplication vous revient : elle prend dix secondes et elle donne un chiffre que vous pouvez défendre.

Deux chantiers portent un gain chiffrable dès maintenant : C2 sur la recherche documentaire, C1 sur la traîne des commandes reçues par e-mail.

GAIN DE C2, RECHERCHE DOCUMENTAIRE, EN HEURES PAR AN

Recherches documentaires par jour, tous services confondus	120
Part que l'assistant traite sans reprise, hypothèse	60 %
Recherches concernées par jour	72
Temps moyen aujourd'hui, temps visé	7 min, puis 2 min
Temps gagné par recherche	5 min
Gain par jour	6 heures
Jours ouvrés par an	220
Gain annuel	environ 1 320 heures

Ce chiffre vaut ce que vaut l'hypothèse de 60 %, qui n'est pas mesurée : le jeu d'évaluation du prototype la confirmera ou la démentira. Il ne vaut pas une promesse, et il ne dit rien de la charge de mise en oeuvre.

GAIN DE C1 SUR LA TRAÎNE, EXTRACTION DES COMMANDES, EN HEURES PAR AN

Documents de commande hors clients intégrables, par mois	900
Part traitée sans reprise, hypothèse	70 %
Documents concernés par mois	630
Temps de saisie aujourd'hui, contrôle après extraction	6 min, puis 1 min
Temps gagné par document	5 min
Gain par mois	52,5 heures
Mois par an	12
Gain annuel	environ 630 heures

Ce chiffre suppose que les 23 clients réguliers envoient d'abord un fichier structuré : sans cette étape, la traîne n'est pas isolée et le calcul tombe. Il ne couvre pas les 62 % de lignes reprises par intégration, qui ne sont pas un chantier IA.

Ces heures ne se transforment pas en postes supprimés, et je ne vous vendrai pas l'inverse. 630 heures par an réparties sur quatre agences et 14 personnes en ADV, cela fait quelques heures par semaine et par site. Ce qu'elles deviennent est plus utile : des commandes saisies le jour même plutôt que le lendemain matin, un client rappelé avant midi, et du travail de recopie que personne n'a jamais aimé faire, qui disparaît. Si votre objectif est de réduire l'effectif, dites-le maintenant, parce que le projet se pilote alors autrement et que je ne le présenterai pas aux équipes sous un autre nom.

> Une heure gagnée devient un délai de réponse avant de devenir une ligne de budget.

Le coût de fonctionnement se calcule à partir de volumes, pas d'un forfait. C2 envoie environ 13,2 millions de tokens en entrée et 0,88 million en sortie par mois, le token étant l'unité de texte facturée par le modèle. C1 envoie environ 4,86 millions en entrée et 0,54 million en sortie. La formule est stable : coût = (tokens d'entrée en millions) x P_entrée + (tokens de sortie en millions) x P_sortie. P_entrée et P_sortie sont les prix publics du million de tokens du modèle retenu, au jour de la décision. Je ne les écris pas ici : ils bougent plusieurs fois par an, et un tarif recopié dans un rapport est faux avant d'être lu. Relevez le tarif du jour chez le fournisseur, posez-le dans la formule, et ajoutez l'hébergement et la supervision, qui ne sont pas des tokens. Le détail chantier par chantier, première indexation des fiches comprise, est en annexe A4, et les leviers qui font baisser la facture sont détaillés ici.

La charge de mise en oeuvre suit la même règle : elle est exprimée en jours dans le plan, jamais en euros. Je ne publie pas de grille tarifaire, et la page tarifs dit pourquoi : un prix affiché à l'aveugle serait faux pour votre cas. Le chiffre, vous l'avez au premier appel, avant tout engagement.

HYPOTHÈSE

Tout ce cas est fabriqué. Le distributeur n'existe pas et aucun de ces chiffres n'a été mesuré chez un client. Les volumes et les parts de 60 % et de 70 % sont des hypothèses de travail, posées pour montrer la forme d'un calcul. Ce qu'il faut retenir n'est pas le résultat, c'est la structure : un volume, une part traitée sans reprise, un temps avant, un temps après. Chez vous, les termes changent, la structure tient.

06 · CAS FABRIQUÉ

Risques, conformité et adoption

Ce que la réglementation demande vraiment ici, et le risque que tout le monde sous-estime.

Quatre risques peuvent arrêter un de ces chantiers après son démarrage : la réglementation, les contrats fournisseurs, l'adoption et la dépendance à un prestataire. Les trois derniers pèsent ici plus lourd que le premier.

Ce que la réglementation demande ici

Le règlement européen sur l'IA, appelé « IA Act », impose une obligation de **transparence** : un utilisateur doit savoir qu'il parle à une machine. Cette obligation vise le cas C4, l'assistant ouvert aux clients sur votre site. Si vous le relancez un jour, il doit annoncer qu'il est une machine, et le passage vers un humain doit rester visible.

Les deux chantiers que je recommande de lancer ne sont pas dans ce périmètre. La recherche documentaire interne, C2, et l'extraction des commandes, C1, relèvent du **risque minimal** : ce règlement ne leur impose pas d'obligation particulière. Le point de vigilance est ailleurs. Les commandes contiennent des données personnelles, le nom et les coordonnées d'un acheteur par exemple, et le RGPD s'y applique comme il s'applique déjà à votre ERP. Ce n'est pas un sujet neuf, c'est le même sujet étendu à un traitement de plus. Pour le cadre général, voir l'IA Act expliqué aux entreprises.

> RÉSERVE
>
> Je ne suis pas juriste. Ce qui précède désigne les points à instruire, pas un avis juridique. Ces points se valident avec votre conseil.

Les catalogues fournisseurs

Les 34 000 fiches techniques viennent de vos fournisseurs. Les indexer pour que vos équipes les interrogent en interne n'est pas la même chose que les rediffuser à vos clients. La première lecture est défendable, la seconde ne l'est pas sans accord. Je ne tranche pas la question à votre place : elle se lit contrat par contrat, et vos 11 fournisseurs principaux n'ont pas les mêmes conditions d'usage. La réponse appartient à votre direction juridique, ou à votre conseil externe. Posez la question avant le prototype, pas après.

L'adoption

C'est le risque le plus sous-estimé de la liste. Un assistant que plus personne n'ouvre au bout de trois semaines a coûté sa mise en oeuvre et n'a rien rendu. Le POC arrêté en 2025 appartient à cette catégorie, et le passage du POC à la production se joue en grande partie là. Ce qu'on met en face n'a rien de spectaculaire, et se décide avant le premier jour de code.

- L'assistant s'ouvre dans l'outil déjà utilisé au quotidien, pas dans un onglet de plus à retenir.
- Trois personnes de l'ADV et deux du support essaient le prototype et disent ce qui ne va pas, avant toute ouverture large.
- Chaque réponse affiche ses sources, parce qu'une réponse invérifiable finit par ne plus être lue.
- Le nombre d'utilisateurs actifs par semaine est suivi dès le premier jour, comme la qualité des réponses.

— Une condition d'arrêt est écrite d'avance : si l'usage ne décolle pas en deux mois, on arrête et on dit pourquoi.

La dépendance

Formats ouverts, données exportables à tout moment, et le rapport comme le code produits pendant la mission vous appartiennent.

REGISTRE DES RISQUES

RISQUE	CE QUE ÇA PRODUIT	CE QU'ON MET EN FACE
Réponse fausse non détectée	Décision prise sur une donnée erronée	Sources affichées, jeu d'évaluation avant ouverture
Droits sur les catalogues	Usage interne contesté par un fournisseur	Revue contrat par contrat avant le prototype
Données personnelles des commandes	Traitement hors du cadre RGPD	Registre à jour, accès restreints, hébergement tranché
Abandon par les équipes	Charge engagée, aucun gain	Utilisateurs pilotes, mesure d'usage, condition d'arrêt
Dépendance au prestataire	Coût de sortie non maîtrisé	Standards ouverts, code et documentation livrés

> Sur ce dossier, le risque réglementaire est le mieux documenté et le moins probable. Le risque d'abandon est exactement l'inverse.

07 · CAS FABRIQUÉ

La feuille de route

Trois horizons, des charges en jours, une condition de démarrage par chantier et des critères d'arrêt.

Cette feuille de route engage deux choses : un ordre entre les chantiers, et une condition de démarrage pour chacun. Elle n'engage pas un résultat. Personne ne peut promettre aujourd'hui que l'assistant documentaire tiendra le seuil visé, parce que le jeu d'évaluation qui permettrait de le vérifier n'existe pas encore. Ce qu'on peut poser, en revanche, c'est l'ordre dans lequel on lève les incertitudes, de la moins chère à la plus chère.

L'ordre suit une règle simple : on ne finance pas une étape tant que la précédente n'a pas produit sa preuve. Les charges sont exprimées en jours de mise en oeuvre. Elles ne comprennent ni le temps de vos équipes, ni l'hébergement, ni la consommation du modèle, qui se calcule à part.

LES TROIS HORIZONS, AVEC CE QUI AUTORISE CHAQUE CHANTIER À DÉMARRER.

HORIZON	CHANTIER	CHARGE	CONDITION DE DÉMARRAGE
0 à 3 mois	Cadrage et jeu d'évaluation de C2	4 jours	Un référent documentaire nommé, avec du temps dégagé
0 à 3 mois	Prototype de C2	12 jours	Jeu d'évaluation figé et accepté par les métiers
0 à 3 mois	Fichiers structurés des 23 clients	Hors périmètre IA	Devis et date obtenus de l'éditeur de l'ERP
3 à 6 mois	Mise en production de C2	20 jours et plus	Seuils atteints par le prototype sur le jeu d'évaluation
3 à 6 mois	Prototype d'extraction de C1	12 jours	Intégration des 23 clients livrée ou planifiée
6 à 12 mois	Mise en production de C1	20 jours et plus	Seuils atteints par le prototype d'extraction
6 à 12 mois	C3, réponse aux appels d'offres	Non chiffré à ce stade	C2 tient ses seuils en production réelle

Les chantiers des trois premiers mois se mènent en parallèle, mais ils ne reposent pas sur les mêmes personnes. L'intégration des fichiers des 23 clients est un sujet d'ERP, pas un sujet d'IA : elle revient à l'éditeur. Elle couvre à elle seule 62 % des lignes de commande, sans qu'aucun modèle intervienne. Si elle n'avance pas, le prototype d'extraction perd une partie de sa raison d'être, puisque l'IA se retrouverait à traiter un volume qui n'a pas à lui revenir.

Les critères d'arrêt

Un chantier s'arrête mieux tôt que tard. Voici les moments où on regarde, et ce qui doit faire arrêter plutôt que prolonger.

- À la fin des 4 jours de cadrage, si les questions déjà posées au support ne suffisent pas à construire un jeu d'évaluation que les métiers acceptent, on n'engage pas le prototype.
- À la fin du prototype de C2, si l'assistant reste sous le seuil fixé au cadrage, la mise en production, 20 jours et plus, n'est pas engagée.

- À la revue de fin de prototype, si les personnes du support ne l'ouvrent plus sans qu'on le leur rappelle, le besoin n'était pas là où on le croyait.
- En production, si le taux de réponses reprises à la main remonte deux mois de suite, on gèle les évolutions et on revient au jeu d'évaluation avant d'ajouter quoi que ce soit.
- Avant le prototype de C1, si l'éditeur de l'ERP n'a donné ni devis ni date pour les fichiers structurés, on ne démarre pas la partie IA.
- À chaque revue mensuelle, si le coût recalculé avec les tarifs du jour dépasse l'enveloppe posée au cadrage, on change de modèle ou on arrête, on ne laisse pas glisser.

> Un chantier sans condition d'arrêt écrite n'a pas de condition de démarrage sérieuse.

> HORS PÉRIMÈTRE
>
> C5, la prévision des ruptures, n'y entrera jamais sous forme de modèle de langage. C'est de la statistique sur l'historique de l'ERP, et cet historique a un trou de 7 mois en 2022. Le chantier existe, il relève de la qualité des données, il se traite ailleurs. C4, le chatbot client, ne revient qu'aux trois conditions posées plus haut. Il faut une source de vérité exposée sur le stock et les prix, un jeu d'évaluation qui dit s'il répond juste, et une procédure d'escalade vers un humain. Tant que ces trois points ne sont pas traités, le relancer revient à refaire l'échec de 2025.

Le découpage en cadrage, prototype, puis mise en production n'est pas propre à ce cas : c'est la manière dont je travaille, décrite dans ma méthode. Chaque étape se termine par une décision, et cette décision peut être de ne pas continuer.

ANNEXES TECHNIQUES

Ce qui suit s'adresse à une équipe technique. Un dirigeant n'a pas besoin de le lire pour décider. Son prestataire, lui, doit pouvoir le contester ligne à ligne.

A1 · ANNEXES TECHNIQUES

Architecture visée pour la recherche documentaire

Ingestion, découpage, recherche hybride, et le point dur qui est une métadonnée, pas un modèle.

L'architecture décrite ici sert la piste C2, la recherche dans les 34 000 fiches PDF fournisseurs. Elle tient en quatre étages : ingestion, découpage, recherche, génération. Chacun a son mode de défaillance et son point de mesure. Le fil de cette annexe est simple : sur ce corpus, ce qui casse une réponse se joue avant le modèle. Le cadrage général est décrit dans Le RAG en entreprise, par où commencer.

Chaîne d'ingestion

Le premier traitement est un tri automatique. Chaque document passe par un test de couche texte : si l'extraction directe rend un texte ordonné et assez dense, le PDF est natif et part sur la chaîne courte. Sinon il est traité comme un scan. Environ 40 % des fiches basculent dans cette voie et passent à l'océrisation, page par page, position des blocs conservée. Le résultat du tri devient une métadonnée : la confiance accordée à un fragment océrisé n'est pas la même. Les arbitrages de cette étape sont détaillés dans l'extraction par vision et OCR.

L'étage suivant reconstruit la structure : hiérarchie des titres, ordre de lecture des colonnes, et surtout tableaux de caractéristiques. C'est le point sensible. Un tableau mal extrait décale d'une ligne ou d'une colonne, et la valeur lue reste plausible, donc personne ne la conteste. Sur ce type de corpus, **c'est la première cause de réponse fausse**, loin devant les inventions du modèle. L'extraction de tableaux se teste en propre, comme un composant.

> MÉTHODE
>
> Les 400 documents tirés au hasard pendant l'audit servent de base au jeu de test d'ingestion. On annote à la main les tableaux d'un échantillon de fiches, moitié natives, moitié scannées, et on mesure la part des cellules correctement rattachées à leur en-tête. Ce chiffre se produit avant le prototype.

Découpage

Le découpage suit les sections de la fiche, pas une taille fixe en tokens. Une taille fixe coupe au milieu d'un tableau et sépare une valeur de son en-tête de colonne. Le fragment reste lisible, ne veut plus rien dire, et ressort quand même à la recherche. Chaque fragment reprend donc le titre de sa section et les en-têtes de son tableau, avec un chevauchement pour les sections qui débordent d'une page sur la suivante.

Métadonnées obligatoires

Six champs sont obligatoires sur chaque fragment. Le pipeline rejette à l'indexation celui qui en manque un, plutôt que d'indexer un objet qu'on ne saura ni filtrer ni citer.

- Référence article : c'est la clé de filtrage la plus utilisée, une question part presque toujours d'une référence.
- Fournisseur : il restreint la recherche et trace les conditions d'usage du catalogue, à vérifier contrat par contrat.
- Indice de révision : il distingue deux versions de la même fiche présentes dans le corpus.
- Date de révision : elle tranche entre deux indices quand la numérotation du fournisseur change de logique.

— Langue : le corpus est multilingue, et un fragment en langue étrangère n'est pas exploitable en ADV.

— Page : sans elle, la citation n'est pas vérifiable, et une réponse non vérifiable ne vaut rien ici.

FRAGMENT INDEXÉ, AVEC SES MÉTADONNÉES · JSON


```
{
  "fragment_id": "fp-018342-rD-p07-t02",
  "document_id": "fp-018342",
  "reference_article": "REF-40218-C",
  "fournisseur": "F07",
  "indice_revision": "D",
  "date_revision": "2024-11-18",
  "revision_courante": true,
  "langue": "fr",
  "page": 7,
  "section": "Caractéristiques mécaniques",
  "type_bloc": "tableau",
  "origine": "ocr",
  "confiance_extraction": 0.94,
  "en_tetes_colonnes": [
    "caractéristique",
    "valeur",
    "unité",
    "tolérance"
  ],
  "texte": "Caractéristiques mécaniques. Charge dynamique de base : 29,6 kN, tolérance 2 %. Température maximale
d'emploi : 110 °C, tolérance 5 °C.",
  "tokens": 312
}
```


Recherche

La recherche est hybride : un index lexical et un index vectoriel interrogés en parallèle, puis un reclassement des candidats fusionnés. Le lexical seul rate les reformulations, celui qui demande une température d'emploi quand la fiche parle de plage thermique n'obtient rien. Le vectoriel seul rate les références alphanumériques : deux références qui ne diffèrent que par un suffixe ont des vecteurs presque identiques, alors qu'elles ne se montent pas sur le même équipement. Le lexical traite ces chaînes comme des identifiants exacts. Les deux se couvrent, c'est la raison de l'hybride.

Génération et refus

La réponse cite la fiche et la page, en sortie structurée, pas en texte libre. Une réponse sans citation exploitable est rejetée avant affichage. Le seuil de pertinence est explicite : si aucun fragment ne le dépasse après reclassement, l'assistant dit qu'il n'a pas trouvé et renvoie vers le support. Il ne compose pas une réponse à partir du moins mauvais fragment. Un assistant qui dit qu'il ne sait pas est un assistant utilisable, et son taux d'erreur devient mesurable.

Le point dur

Deux révisions de la même fiche coexistent dans le corpus, et rien n'empêche l'ancienne de mieux coller à la question que la nouvelle. C'est le risque principal de cette architecture, et il faut le nommer : **ce n'est pas un problème de modèle.** C'est une métadonnée et une règle de filtrage. La règle : on indexe toutes les révisions, on ne sert par défaut que la plus récente pour une référence donnée, et les révisions antérieures s'obtiennent sur demande explicite, pour les équipements déjà installés.

> Une fiche périmée servie avec une citation exacte est plus dangereuse qu'une absence de réponse.

Le jeu d'évaluation et les seuils

Soixante questions issues des vrais tickets, quatre mesures, un seuil par mesure, et la CI qui les rejoue.

Le jeu compte 60 questions. Aucune n'est écrite en atelier. Elles viennent des tickets du support et des questions déjà posées par les agences, dans leur formulation d'origine. Pour chacune, la réponse juste existe déjà : une fiche fournisseur, un échange archivé, la mémoire de celui qui a répondu. On la retrouve, on la fige, on la rattache à une fiche et à sa révision.

Cette origine change la nature de la mesure. Une question inventée en atelier est propre : bon vocabulaire, référence complète, contexte donné. Elle teste le système sur une distribution qui n'existe pas dans l'entreprise. Un ticket réel porte la référence tapée de travers, l'abréviation maison, la question posée en deux lignes. Autre effet : personne ne conteste un cas de test qui porte un numéro de ticket.

Les quatre mesures

Quatre mesures indépendantes, calculées sur le même passage. Sur les 60 questions, 45 ont une réponse dans le corpus, 15 n'en ont pas : ces pièges servent la troisième mesure.

- **Réponse retrouvée :** la réponse contient tous les éléments obligatoires listés dans le cas, valeur, unité et condition d'application.
- **Source exacte :** la fiche citée est celle qui porte la réponse, dans la révision en vigueur au jour du test. Une fiche voisine qui dit la même chose est un échec.
- **Refus quand il faut :** sur une question sans réponse dans le corpus, le système dit qu'il ne sait pas et propose l'escalade.
- **Fidélité numérique :** aucune valeur numérique absente des passages cités n'apparaît dans la réponse. Le contrôle est automatique.

LES QUATRE MESURES ET LEURS SEUILS DE DÉCISION

MESURE	CE QU'ON COMPTE	SEUIL DE PASSAGE
Réponse retrouvée	Cas répondables avec tous les éléments obligatoires	90 % des 45 cas
Source exacte	Réponses citant la bonne fiche à la bonne révision	95 % des réponses produites
Refus quand il faut	Pièges où le système refuse et escalade	100 % des 15 pièges
Fidélité numérique	Réponses contenant un nombre absent des sources	0, sinon pas de production

Ce sont des seuils de décision, pas des promesses de performance. Ils disent ce qu'on fait du résultat : sous le seuil, on ne met pas en production. Ils se fixent avant le premier passage, sinon ils s'alignent sur le résultat et ne décident plus rien.

Quand on rejoue le jeu

- À chaque changement de prompt, y compris une reformulation qui paraît anodine.
- À chaque changement de modèle ou de version de modèle.
- À chaque reconstruction de l'index, en particulier après un nouveau découpage.
- À chaque changement d'un paramètre de recherche : passages récupérés, seuil de similarité, pondération.
- En intégration continue à chaque fusion, et avant toute mise en production.

L'automatisation n'est pas un confort. Le fournisseur du modèle fait évoluer ses versions selon son calendrier, et une bascule déplace le comportement sans que rien ne change chez vous. Sans jeu rejoué automatiquement, ce déplacement passe inaperçu jusqu'au premier client mécontent, et vous n'avez alors ni date, ni cause, ni comparaison.

L'entretien du jeu

Un jeu d'évaluation vieillit sans prévenir. Les fiches sont révisées, des références sortent, des fournisseurs changent de format. Une réponse juste au cadrage devient fausse deux trimestres plus tard, et le jeu la compte toujours comme juste.

Trois règles suffisent. Toute fiche citée par un cas est surveillée : si sa révision change, les cas rattachés passent en quarantaine et sont revalidés avant le passage suivant. Chaque mois, quelques cas viennent des tickets récents et autant de cas caducs sortent, à taille constante. Chaque trimestre, une relecture vérifie que ces questions ressemblent encore à celles qu'on pose.

Le jeu a un propriétaire nommé, côté métier : la personne du support qui répond aujourd'hui à ces questions, seule à trancher ce qu'est une bonne réponse. Le prestataire fournit l'outillage, pas la vérité. Sans nom sur cette ligne, le jeu meurt en quelques mois.

> Un jeu d'évaluation qu'on ne rejoue pas est une photographie du passé, pas une mesure.

UN CAS DE TEST DU JEU · YAML

```yaml
- id: DOC-2026-014
  source_ticket: SUP-48213
  type: repondable
  question: "Quel couple de serrage pour les vis M8 classe 8.8 du catalogue outillage ?"
  reponse_attendue:
    elements_obligatoires:
      - "25"
      - "N.m"
      - "filet sec"
  source_attendue:
    fiche: FT-VIS-M8-88
    revision: "R4"
    page: 3
  refus_attendu: false
  nombres_autorises: [25, 8, 8.8]
  proprietaire: support-adv
  derniere_revalidation: 2026-08-19
```

Ce jeu est un jeu de recette. Le suivi en production est traité dans mesurer un agent IA en production.

A3 · ANNEXES TECHNIQUES

Extraction des commandes : schéma et boucle de validation

Une sortie contrainte, un rapprochement au référentiel, et une règle : aucune référence inventée.

L'extraction ne produit pas de texte libre. La sortie du modèle est contrainte par un schéma, validé à la réception : chaque champ a un type, un domaine de valeurs et une obligation de présence. Le modèle ne rédige pas, il remplit. Une sortie non conforme est relancée une fois, puis le document part en revue. Le périmètre du modèle est fixé par là même : il rend ce qu'il lit sur le document, il ne décide rien.

SCHÉMA DE SORTIE CONTRAINT, EXTRACTION D'UNE COMMANDE · JSON

```
{
  "name": "commande_extraite",
  "strict": true,
  "schema": {
    "type": "object",
    "additionalProperties": false,
    "required": ["document_id", "pages", "client", "lignes", "confiance_extraction"],
    "properties": {
      "document_id": { "type": "string" },
      "pages": { "type": "integer", "minimum": 1 },
      "client": {
        "type": "object",
        "additionalProperties": false,
        "required": ["raison_sociale_lue", "reference_client_lue", "confiance"],
        "properties": {
          "raison_sociale_lue": { "type": ["string", "null"] },
          "reference_client_lue": { "type": ["string", "null"] },
          "confiance": { "type": "number", "minimum": 0, "maximum": 1 }
        }
      },
      "lignes": {
        "type": "array",
        "minItems": 1,
        "items": {
          "type": "object",
          "additionalProperties": false,
          "required": ["page", "texte_source", "reference_lue", "designation_lue", "quantite", "unite",
"confiance"],
          "properties": {
            "page": { "type": "integer", "minimum": 1 },
            "texte_source": { "type": "string" },
            "reference_lue": { "type": ["string", "null"] },
            "designation_lue": { "type": ["string", "null"] },
            "quantite": { "type": ["number", "null"], "exclusiveMinimum": 0 },
            "unite": { "type": ["string", "null"], "enum": ["piece", "boite", "metre", "kg", "lot", null] },
            "confiance": { "type": "number", "minimum": 0, "maximum": 1 }
          }
        }
      },
      "confiance_extraction": { "type": "number", "minimum": 0, "maximum": 1 }
    }
  }
}
```

Rapprochement au référentiel

Le modèle ne choisit aucune référence du catalogue. Il rend une chaîne lue, une désignation, une quantité, une page. Le rapprochement vient ensuite, par du code, ligne par ligne, contre le référentiel des 34 000 références actives : égalité stricte sur la référence fournisseur, égalité sur la référence client quand le référentiel la connaît,

similarité textuelle sur la désignation. Chaque candidat sort avec un score de correspondance.

Si aucun article ne dépasse le seuil, la ligne part en revue humaine, sans candidat proposé par défaut. **Aucune référence n'est devinée**, même quand un seul article est proche. Une référence inventée qui passe en commande coûte plus cher que dix lignes envoyées en revue : la ligne en revue coûte un contrôle en ADV, la mauvaise référence coûte une expédition, un retour, un avoir, et un client qui doute du reste de sa commande.

> MÉTHODE
>
> Le seuil se règle sur le jeu d'évaluation, pas au jugé. On le monte tant qu'une référence fausse passe. On ne le descend qu'après une revue montrant que les lignes rejetées étaient toutes rapprochables.

Score de confiance et aiguillage

Deux scores circulent. Le score de ligne combine la confiance d'extraction du modèle et le score de correspondance du rapprochement. Le score de commande est le **minimum** des scores de ses lignes, pas leur moyenne : une moyenne dilue la ligne douteuse dans les bonnes, et c'est elle qu'on cherche.

AIGUILLAGE D'UNE COMMANDE EXTRAITE

SITUATION	DESTINATION
Toutes les lignes au-dessus du seuil	Intégration ERP directe
Une ligne au moins sous le seuil	Revue humaine, commande entière
Sortie non conforme au schéma	Relance unique, puis revue
Client absent du référentiel	Revue humaine, aucune création automatique

La commande part entière en revue, pas seulement sa ligne douteuse. Une commande à moitié intégrée oblige l'ADV à rouvrir un bon partiel et à vérifier ce qui est déjà passé : cela coûte plus cher que de la traiter d'un bloc.

La mesure qui compte

La mesure de production est la **part de commandes intégrées sans reprise humaine**, relevée sur le flux réel du mois, pas le score du modèle sur un banc d'essai. L'écart entre les deux va toujours dans le même sens.

Un banc d'essai note des champs isolés, sur des documents choisis, souvent lisibles et semblables entre eux, et il tolère l'à peu près : une désignation presque juste y compte comme juste. La production note des commandes entières, sur les documents qui arrivent vraiment, scans compris, et une commande est reprise dès qu'une seule de ses lignes est fausse. Les erreurs ne se moyennent pas, elles se cumulent le long de la commande.

POURQUOI UN BON SCORE PAR LIGNE DONNE UN TAUX DE COMMANDES PLUS BAS

Taux de lignes justes, hypothèse	95 %
Lignes par commande, hypothèse	8
Calcul	0,95 à la puissance 8
Commandes sans aucune erreur	environ 66 %

Chiffre illustratif, calculé sur les deux hypothèses posées ici et sur rien d'autre. Il montre pourquoi un score par ligne ne se lit jamais comme un taux de commandes, et pourquoi l'hypothèse de 70 % retenue pour C1 reste un ordre de grandeur, pas une performance annoncée. Protocole de mesure dans [mesurer un agent IA en production](/ressources/mesurer-un-agent-ia-en-production).

La boucle de retour

Chaque correction humaine est enregistrée avec quatre éléments : le document source, la sortie du modèle, la correction retenue et le motif, pris dans une liste courte, référence introuvable, quantité mal lue, ligne oubliée, colonne mal attribuée. Ce quadruplet devient un cas du jeu d'évaluation, ajouté sans tri. On ne garde pas les cas intéressants, on garde ceux qui se sont produits.

Au bout de quelques mois, ce jeu ne ressemble plus à un corpus générique. Il décrit les vrais cas tordus de cette entreprise et de personne d'autre : le fournisseur dont les scans partent de travers, le client qui met ses quantités dans la colonne de droite, la désignation maison qui ne correspond à aucune référence du catalogue. C'est le seul actif du chantier qui ne se périme pas quand le modèle change.

> Un modèle se remplace vite. Un jeu d'évaluation se construit lentement, et c'est lui qu'on possède.

A4 · ANNEXES TECHNIQUES

Le détail du calcul de coût, et sa sensibilité

La formule, ses variables, et les cinq choses qui la font bouger.

Cette annexe pose les volumes de tokens des deux cas retenus, la formule qui les transforme en facture mensuelle, et les points précis où cette facture bouge. Les volumes viennent des hypothèses de charge du rapport. Ils ne sont mesurés sur aucune installation existante.

Volumes de C2, recherche documentaire

L'hypothèse de charge est d'environ 2 200 recherches par mois. Chaque recherche envoie environ 6 000 tokens en entrée, soit le contexte récupéré et joint à la question. Elle produit environ 400 tokens en sortie.

VOLUME MENSUEL DE TOKENS, CAS C2

Recherches par mois	2 200
Contexte envoyé par recherche	6 000 tokens en entrée
Réponse produite par recherche	400 tokens en sortie
Volume mensuel	13,2 millions de tokens en entrée, 0,88 million en sortie

Coût mensuel = 13,2 x P_entrée + 0,88 x P_sortie, où P_entrée et P_sortie sont les prix du million de tokens du modèle retenu au jour de la décision.

L'entrée pèse quinze fois la sortie. Toute optimisation qui ne touche pas au contexte envoyé agit donc sur la plus petite part de la facture.

Volumes de C1, extraction sur la traîne

L'hypothèse porte sur 900 documents par mois, 3 pages en moyenne. Une page lue en image compte environ 1 800 tokens en entrée. La sortie structurée compte environ 600 tokens par document.

VOLUME MENSUEL DE TOKENS, CAS C1

Documents par mois	900
Pages par document	3
Lecture d'image	1 800 tokens en entrée par page
Sortie structurée	600 tokens en sortie par document
Volume mensuel	4,86 millions de tokens en entrée, 0,54 million en sortie

Coût mensuel = 4,86 x P_entrée + 0,54 x P_sortie, avec les mêmes variables de prix. Une reprise manuelle ne consomme rien, une deuxième passe automatique recompte l'entrée en entier.

Le rapport entre entrée et sortie est ici de neuf. La lecture d'image tient le terme dominant, ce qui fait de la part scannée des commandes une variable de coût, pas seulement une variable de qualité.

L'indexation initiale

Les deux montants ci-dessus sont mensuels. L'indexation initiale, elle, est un coût unique : les 34 000 fiches passent une fois par la lecture d'image pour la part scannée, puis par un calcul de vecteurs sur chaque fragment. Deux ordres de grandeur séparent ces deux opérations, la lecture d'image coûtant largement plus cher que les vecteurs.

Ce coût unique se paie en réalité plusieurs fois, et c'est le point à retenir : chaque refonte du découpage oblige à recalculer les vecteurs, et une correction de la chaîne d'océrisation oblige à relire les pages. Pendant les premières semaines, ça arrive plus souvent qu'on ne le prévoit.

> MÉTHODE
>
> On ne l'estime pas, on le mesure. Mille fiches sont traitées pendant le prototype, en gardant la même proportion de scans que le corpus, et le relevé est extrapolé aux 34 000. Le chiffre obtenu vaut pour une passe complète : c'est lui qu'on multiplie par le nombre de refontes qu'on s'autorise.

Sensibilité de la facture

CE QUI DÉPLACE LE MONTANT MENSUEL

CE QUI CHANGE	EFFET SUR LA FACTURE
Le volume double	Facture x 2, les deux termes montent ensemble.
Le contexte envoyé double	Terme d'entrée x 2, sortie inchangée.
Le modèle change de gamme	Facteur direct sur P_entrée et P_sortie.
La part scannée des commandes augmente	Plus de pages en lecture d'image, entrée en hausse.
Le taux de reprise automatique monte	Deuxième passe, entrée et sortie recomptées.

Leviers et contreparties

Quatre leviers agissent sur ces montants. Aucun n'est gratuit, je les donne donc avec ce qu'ils coûtent en échange.

- Mettre en cache la partie stable du contexte réduit le terme d'entrée sur ce qui se répète. En contrepartie, il faut figer cette partie et gérer son invalidation à chaque mise à jour du corpus.
- Passer un modèle plus petit sur la première passe de tri baisse le prix au million. En contrepartie, il faut une seconde passe sur les cas douteux et un jeu d'évaluation tenu sur deux modèles.
- Traiter les documents par lots la nuit permet d'utiliser un tarif par lots si le fournisseur retenu en propose un. En contrepartie, les commandes ne sont plus traitées au fil de l'eau.
- Améliorer le reclassement pour envoyer moins de contexte attaque le terme dominant de C2. En contrepartie, c'est un composant de plus à évaluer, et une coupe trop agressive fait perdre le passage utile.

> RÉSERVE
>
> Le tarif d'un modèle n'est pas une donnée durable. Il change, à la hausse comme à la baisse, et les gammes se renomment. Le livrable ici, c'est la formule et les volumes, pas un montant. Le tarif du jour est à revérifier auprès du fournisseur avant tout arbitrage.

La mécanique de ces leviers est détaillée dans réduire sa facture LLM. Avant décision, on recalcule avec les tarifs du jour et les volumes observés sur le prototype, pas avec ceux posés ici.

Ce que je n'ai pas retenu, et pourquoi

Fine-tuning, base vectorielle dédiée, multi-agents, écriture automatique dans l'ERP.

Un audit se juge autant sur ce qu'il écarte que sur ce qu'il retient. Voici les quatre options que j'ai regardées puis laissées de côté, chacune avec la condition qui la remettrait sur la table.

Le fine-tuning

Le fine-tuning réentraîne un modèle existant sur un corpus propre à l'entreprise, pour qu'il en adopte le vocabulaire et le format de réponse. Il **ne fait pas entrer un catalogue qui bouge dans un modèle.** Les 34 000 références vivent, les fiches fournisseurs sont remplacées, et un modèle affiné fige l'état du catalogue au jour de l'entraînement. Le problème posé ici est la fraîcheur et la citation de la source, pas le style.

Ce qui le rendrait pertinent : un format de sortie très contraint, répété des dizaines de milliers de fois, que la consigne seule n'arrive pas à tenir, ou un vocabulaire métier que le modèle lit mal malgré un contexte bien construit. On mesure l'écart sur le jeu d'évaluation, ensuite on décide. Le raisonnement complet tient dans RAG ou fine-tuning.

Une base vectorielle dédiée

Une base vectorielle dédiée est un serveur spécialisé dans la recherche par similarité, avec ses index, sa mémoire et son exploitation. Le distributeur a déjà une base relationnelle en production, et son extension vectorielle suffit pour un corpus de cette taille. Ajouter un composant dédié, c'est ajouter une sauvegarde, une surveillance, une montée de version et une compétence à tenir, pour un gain de latence invisible ici.

La question se repose au-delà de quelques millions de passages indexés, ou si la recherche devient un produit avec ses propres pics de charge. Tant qu'on reste sous le million, elle est théorique. Le jour où elle se pose, le coût du changement reste faible : le découpage et les vecteurs se recalculent, la logique applicative ne bouge pas.

Une architecture multi-agents

Une architecture multi-agents découpe le travail entre plusieurs modèles qui se le passent, chacun avec son rôle et ses outils. La tâche de C2 est bornée : une question, une recherche, une réponse sourcée. Un agent unique équipé d'un bon outil de recherche la couvre. Chaque agent ajouté apporte un appel, de la latence, des tokens et un point de défaillance de plus, pour un gain qui n'apparaît pas sur une tâche aussi étroite.

Ce qui changerait la donne : une tâche qui enchaîne des décisions hétérogènes, lire un cahier des charges, retrouver les références qui y répondent, vérifier leur disponibilité, puis rédiger une réponse sourcée, avec un critère de reprise différent à chaque étape. C'est le profil de C3, pas celui de C2. On y revient quand C2 tourne et qu'on dispose de mesures par étape.

Un agent qui écrit dans l'ERP

L'écriture automatique fait passer l'extraction de C1 de la proposition à l'acte : la commande entre dans l'ERP sans qu'un humain la valide. **Je la refuse pour l'instant.** Une ligne fausse écrite sans contrôle produit une livraison fausse, une facture fausse et un litige client, et ce coût dépasse la minute de vérification qu'on économise.

Cette porte s'ouvre par une mesure, pas par une confiance. Il faut un taux de reprise stable, suivi sur plusieurs mois, par type de document et par client, et une règle de bascule écrite à l'avance : sous un seuil décidé ensemble, on écrit sans validation, au-dessus, on repasse en contrôle humain. Sans cette série, l'automatisation complète est un pari.

> MÉTHODE
>
> Chacun de ces quatre refus porte une condition de réouverture. Si la condition se réalise, la décision se rejoue, sans rouvrir le débat depuis le début.

> Une option écartée sans condition de retour n'est pas un arbitrage, c'est une préférence.

QUESTIONS

Sur ce rapport, et sur ce qu'il vaut

Ce rapport correspond-il à une mission réelle ?

Non. L'entreprise, les chiffres et les entretiens sont fabriqués, et la page le dit avant la première ligne du rapport. Je ne publie pas de rapport client, même anonymisé. Ce qui est réel, c'est la forme du document, l'ordre des questions et le niveau de détail des annexes.

Pourquoi aucun gain en euros ?

Parce que je ne connais pas votre coût horaire chargé et que vous, oui. Le rapport donne des heures par an et montre le calcul ; la multiplication vous appartient. Un euro inventé dans un rapport d'audit décrédibilise tous les chiffres qui l'entourent.

Un audit de quatre jours suffit-il ?

Pour trancher entre cinq pistes et poser une feuille de route, oui, à condition que le périmètre soit cadré au premier appel. Quatre jours ne suffisent pas à auditer la sécurité d'un système d'information ni à tester une interface avec ses utilisateurs, et le rapport l'écrit lui-même en première section.

Et si l'audit conclut qu'il ne faut rien faire ?

Je vous le dis et on s'arrête là. Sur ce cas, deux pistes sur cinq sont écartées et une troisième est reportée. Un non au bout de quelques jours vaut mieux qu'une désillusion au bout de six mois.

Le rapport m'appartient-il ?

Oui, entièrement, et il ne vous engage à rien pour la suite. Vous construisez avec moi, en interne ou avec quelqu'un d'autre. Tout est écrit pour être repris par une autre équipe : architecture visée, hypothèses de coût, critères d'arrêt.

Votre cas ne ressemblera pas à celui-là

Le distributeur est un cas d'école : quatre agences, un ERP ancien, une documentation en PDF. Le vôtre a ses propres irritants, et le tri se refera de zéro.

Ce qui ne changera pas, c'est la méthode. Je regarde les process réels, je confronte chaque piste à la donnée disponible, au gain chiffrable et au risque, et je vous dis ce qui ne passe pas. Le rapport vous appartient, quelle que soit la suite.