

ANCORIA

by Ancoris Solutions

OUTILS & FORMATIONS IA · GUIDE N° 04

Mistral Studio

la plateforme API de Mistral

Ce que Studio sait faire, ce que ça coûte vraiment, et la seule question qui compte avant d'écrire une ligne de code : faut-il vraiment une API ?

0,15 \$

PAR MILLION DE TOKENS, MODÈLE
D'ENTRÉE

-50 %

EN TRAITEMENT PAR LOT

4 \$

POUR 1 000 PAGES PASSÉES À L'OCR

OUVERTURE

À qui s'adresse ce guide

Aux directions des systèmes d'information de collectivités, chefs de projet numérique territorial, chargés de mission données, et à tous ceux qui doivent arbitrer une demande qui commence par « on pourrait brancher l'IA sur... » — y compris quand ils ne coderont pas eux-mêmes.

Mistral Studio est la plateforme développeur de Mistral : l'endroit d'où l'on appelle les modèles par programme, au lieu de discuter avec eux dans une fenêtre. C'est une bascule de nature, pas de degré. Un abonnement Vibe, c'est un outil qu'on donne à un agent. Une clé d'API, c'est un **projet logiciel** — avec son développement, son hébergement, sa maintenance, son marché et son registre des traitements.

Ce guide ne vous apprendra pas à programmer, et ne remplace pas la documentation officielle, qui est bonne. Il répond à trois questions qu'aucune documentation d'éditeur ne traite : **faut-il une API ? combien ça coûte réellement ?** et **qu'est-ce que ça engage** pour une personne publique ?

LA PHRASE À GARDER EN TÊTE

Sur presque tous les usages territoriaux, le coût des jetons est dérisoire — vous verrez au chapitre 3 des traitements de masse à moins de vingt dollars. Ce qui coûte cher, c'est tout ce qu'il y a autour : le développement, l'hébergement, la reprise des données, la maintenance sur cinq ans, et l'agent qui saura encore expliquer le système dans deux ans. Un devis d'API qui ne parle que de tokens ne décrit pas le projet.

Studio, Vibe, Code : trois portes, un compte

Depuis la refonte du printemps 2026, Mistral expose trois surfaces qui partagent la même organisation et la même facturation. Les confondre est la première source de malentendu dans un dossier.

Surface	Ce que c'est	Qui s'en sert
Vibe	L'assistant, en modes Work, Code et Chat. Une interface, un abonnement par personne.	Les agents du service, tous les jours
Studio	La plateforme développeur : API, console, clés, agents, traitement par lot, workflows.	Une équipe technique, ou un prestataire
Vibe Code	L'agent de développement, en terminal ou dans l'éditeur.	Les développeurs, sur leur propre code

Le guide n° 03 de cette collection traite de Vibe pour la réponse aux appels à projets. Celui-ci ne le répète pas : il commence là où l'interface ne suffit plus.

CHAPITRE 1

Vibe ou Studio : la question d'avant

Une API se justifie quand l'une de ces trois conditions est vraie : le **volume** rend le geste humain impossible, la **répétition** est parfaitement stable, ou le traitement doit s'**insérer dans un logiciel** existant. Si aucune ne l'est, un abonnement et une bonne méthode coûtent moins cher et vont plus vite.

La situation	Le bon outil	Pourquoi
Rédiger et contrôler un dossier de subvention	Vibe	Volume faible, jugement humain à chaque étape, aucune insertion logicielle
Faire une veille économique hebdomadaire	Vibe	Une tâche planifiée suffit ; l'API n'ajouterait que de la maintenance
Passer 60 000 pages de délibérations à l'OCR	Studio	Volume impossible à la main, traitement identique sur chaque page
Classer et router 200 000 demandes citoyennes par an	Studio	Répétition stable, gain massif, sortie structurée exploitable
Ajouter une recherche intelligente au site ou à l'intranet	Studio	Le traitement doit vivre dans un logiciel, pas dans une fenêtre
Aider dix agents à mieux rédiger	Vibe	Un besoin de formation déguisé en besoin technique
Répondre à un élu qui a vu une démo	Ni l'un ni l'autre	D'abord écrire le problème. La moitié des projets meurent là, et c'est bien

INSIGHT · LE TEST DES TROIS COLONNES

Avant tout chiffrage, remplir trois colonnes sur une feuille : **combien de fois par an** ce traitement se produit, **combien de minutes** il prend aujourd'hui à un agent, et **ce qui se passe si le résultat est faux**. Un traitement rare, court, et à conséquence grave ne doit pas être automatisé. Un traitement massif, long, et à conséquence réversible est un candidat évident. Tout le reste demande une décision, pas un calcul.

CHAPITRE 2

Ce que Studio contient

Studio se présente comme « la plateforme développeur de Mistral pour les applications d’IA », accessible à la fois par API et par une console interactive. Voici la carte, avec pour chaque brique ce qu’elle vaut dans un contexte territorial.

Brique	Ce qu’elle fait	Intérêt territorial
API de conversation	Génération de texte, complétion, sorties structurées, appel de fonctions	Élevé — c’est le socle de tout le reste
Agents & Conversations	Agents autonomes qui planifient, appellent des outils, et se passent la main entre eux	Moyen — utile quand un traitement enchaîne plusieurs étapes
Document AI	OCR, extraction structurée annotée, questions-réponses sur document	Très élevé — le fonds documentaire d’une collectivité est du papier scanné
Audio	Transcription, synthèse vocale	Moyen — comptes rendus de séance, accessibilité
RAG et embeddings	Libraries gérées par Mistral, ou index maison avec un modèle d’embedding	Élevé — recherche interne, assistance aux agents d’accueil
Connecteurs	Serveurs MCP enregistrés comme outils utilisables par les agents	Moyen — brancher un logiciel métier existant
Sécurité et modération	Classification des contenus à risque	Élevé dès qu’un texte produit est exposé au public
Traitement par lot	Inférence asynchrone de masse, à moitié prix	Très élevé — c’est là que les gros volumes deviennent abordables
Workflows	Orchestration durable, reprise après panne, validation humaine, planification	Élevé sur un processus long, mais en public preview
Administration	Organisations, workspaces, clés, limites, mesure d’usage	Indispensable — voir le chapitre 7

À SAVOIR · LE FINE-TUNING EST DÉPRÉCIÉ

La documentation de Mistral classe le fine-tuning parmi ses fonctions dépréciées, avec un avertissement sans ambiguïté — « *This feature is deprecated and is no longer actively supported* » — et sans nommer de remplacement.

Conséquence directe, et elle est importante : **ne pas écrire « modèle spécialisé par apprentissage sur nos données » dans un cahier des charges**. C’était la formule attendue il y a deux ans ; elle vous engagerait aujourd’hui sur une voie que l’éditeur n’entretient plus. L’équivalent moderne s’obtient par le contexte : documents fournis à la requête, recherche documentaire, et consignes explicites.

CHAPITRE 3

Les modèles, et ce qu’ils coûtent

La facturation se fait au million de jetons — grossièrement, un jeton vaut trois quarts de mot en français. Les prix ci-dessous sont ceux affichés par Mistral en dollars ; une facturation en euros est proposée.

Modèle	Entrée / M jetons	Sortie / M jetons	Quand le choisir
Ministral 3 (3B)	0,10 \$	0,10 \$	Tri, étiquetage, extraction simple, à très gros volume
Mistral Small 4	0,15 \$	0,60 \$	Le rapport qualité-prix par défaut pour un traitement de masse
Ministral 3 (8B / 14B)	0,15 / 0,20 \$	0,15 / 0,20 \$	Sortie longue à coût plat
Mistral Large 3	0,50 \$	1,50 \$	Raisonnement solide, contexte de 256 000 jetons, prix contenu
Mistral Medium 3.5	1,50 \$	7,50 \$	Le haut de gamme multimodal et agentique — à réserver
Codestral	0,30 \$	0,90 \$	Code
Mistral Embed	0,10 \$	—	Indexation documentaire
Mistral Moderation 2	gratuit	—	Filtrage des contenus à risque

Hors jetons	Prix	Le piège
OCR 4.1	4 \$ / 1000 pages	—
Document AI (OCR + extraction structurée)	5 \$ / 1000 pages	1 \$ de plus la page mille : n’hésitez pas
Recherche web (outil)	30 \$ / 1000 appels	0,03 \$ l’appel : mille fois le prix d’une requête texte
Exécution de code (outil)	30 \$ / 1000 appels	Idem — un outil n’est pas un jeton
Génération d’image	100 \$ / 1000 images	0,10 \$ l’image
Transcription audio	0,003 \$ / minute	Une séance de 3 h coûte moins d’un centime
Synthèse vocale	0,016 \$ / 1000 caractères	—

Deux remises changent les ordres de grandeur : le **traitement par lot divise le prix par deux**, et la **mise en cache des jetons d’entrée réduit leur coût jusqu’à 90 %** — ce qui compte dès qu’une longue consigne se répète à chaque requête, c’est-à-dire presque toujours.

Trois calculs de coin de table

Les hypothèses sont explicites pour que vous puissiez les refaire avec les vôtres. Elles sont volontairement grossières : l'objectif est l'ordre de grandeur, pas le devis.

Traitement	Hypothèses	Coût des appels
Passer le fonds délibératif à l'OCR	5 000 délibérations, 12 pages en moyenne, soit 60 000 pages, avec extraction structurée	≈ 300 \$, une fois
Classer et router les demandes citoyennes	200 000 demandes par an, 400 jetons en entrée et 60 en sortie, sur Small 4, en lot	≈ 10 \$ par an
Assistant de recherche interne	3 000 documents indexés, 100 requêtes par jour, 6 000 jetons de contexte et 400 de réponse, sur Large 3	≈ 30 \$ par mois, plus 1 \$ d'indexation

CE QUE CES CHIFFRES DISENT VRAIMENT

Router deux cent mille demandes citoyennes pour le prix de deux cafés n'est pas une bonne nouvelle budgétaire : c'est une information sur **où se trouve la difficulté**. Elle n'est pas dans le modèle. Elle est dans la récupération des demandes, la définition des catégories, la reprise dans le logiciel de GRC, le taux d'erreur acceptable, la procédure de recours quand le classement se trompe, et la formation des agents d'accueil. Un projet dont le poste principal est l'API est un projet mal chiffré. Attendez-vous à ce que les appels pèsent moins de 5 % du coût complet sur cinq ans — et construisez le dossier là-dessus.

CHAPITRE 4

Les quatre briques qui portent l'essentiel

1. Document AI : le papier scanné

Mistral décrit son offre documentaire comme du « traitement documentaire de niveau entreprise, combinant OCR et extraction structurée ». Trois services derrière un même point d'entrée `/v1/ocr` : l'**OCR** proprement dit, les **annotations** — l'extraction en sortie structurée —, et le **questions-réponses sur document**. Le multilingue est pris en charge.

C'est la brique la plus rentable dans une collectivité, pour une raison simple : le patrimoine informationnel d'une commune ou d'un EPCI est massivement du PDF scanné non cherchable. Délibérations, arrêtés, comptes rendus de séance, permis, dossiers d'urbanisme, conventions. Rendre ce fonds interrogeable change la vie d'un service juridique bien plus sûrement qu'un assistant conversationnel.

2. Sortie structurée : transformer du texte en données

L'usage territorial le plus sous-estimé. Un modèle peut recevoir un texte libre et rendre un objet strictement conforme à un schéma que vous imposez — donc directement insérable en base, sans relecture de format.

```
# Ce qu'on demande, en substance : une demande citoyenne
# transformee en ligne de base, jamais en prose.

schema = {
  "type": "object",
  "properties": {
    "categorie": {"enum": ["voirie", "dechets", "urbanisme",
                          "etat-civil", "scolaire", "autre"]},
    "urgence": {"enum": ["faible", "moyenne", "haute"]},
    "commune": {"type": "string"},
    "adresse": {"type": ["string", "null"]},
    "resume": {"type": "string", "maxLength": 200},
    "service": {"type": "string"},
    "confiance": {"type": "number"},
    "a_relire": {"type": "boolean"}
  },
  "required": ["categorie", "urgence", "resume", "confiance",
               "a_relire"]
}

# Deux champs font tout le travail de securite :
#  confiance -> en dessous d'un seuil, on ne route pas
#  a_relire -> le modele signale lui-meme ce qu'il n'a pas su
#              trancher, et un agent le reprend
```

INSIGHT · LES DEUX CHAMPS QUI RENDENT UN CLASSEMENT DÉFENDABLE

Un classement automatique sans indice de confiance est ingérable : soit on fait confiance à tout, soit on relit tout, et dans les deux cas l’automatisme n’a rien apporté. Un seuil de confiance et un drapeau « à relire » transforment un système opaque en une **file de tri** : 85 % passent, 15 % arrivent chez un agent avec la raison du doute. C’est ce découpage qui rend le dispositif explicable à un élu, et acceptable par les agents concernés.

3. RAG : interroger ses propres documents

Deux voies, et le choix entre elles structure tout le projet.

	Libraries gérées	Index maison
Principe	Mistral « ingère, vectorise et cherche » les documents pour vous ; un outil intégré permet aux agents d’y puiser	Vous découpez, vous appelez <code>mistral-embed</code> , vous stockez dans votre propre base vectorielle
Mise en œuvre	Rapide	Un vrai chantier
Maîtrise	Le découpage et le classement vous échappent	Total : taille des fragments, filtres par métadonnée, recherche lexicale en complément
Réversibilité	Les documents vivent chez l’éditeur	Vos vecteurs restent chez vous ; changer de modèle demande de réindexer
Bon choix quand	Le corpus est modeste et le besoin urgent	Le corpus est le patrimoine du service, et doit lui survivre

Un détail de méthode que la documentation souligne et que les projets oublient : la **taille des fragments** se règle par l’expérience. Des fragments plus petits sont souvent meilleurs à la recherche, parce qu’un long passage noie le sens utile dans du remplissage. Commencer petit, mesurer, agrandir si nécessaire.

4. Agents, conversations et connecteurs

Au-dessus de la simple requête, Studio propose des **agents** — des systèmes autonomes qui, sur une consigne de haut niveau, planifient, utilisent des outils et enchaînent des étapes — et une **API de conversations** qui gère les échanges avec un modèle seul, un agent, ou plusieurs agents qui se passent la main.

Les outils intégrés sont la recherche web, l’exécution de code, la génération d’images, la bibliothèque documentaire, l’appel de fonctions maison, et les **connecteurs** : n’importe quel serveur compatible **MCP** — le standard ouvert de connexion des modèles aux outils — peut être enregistré et devenir un outil disponible. Un connecteur peut rester privé ou être partagé au workspace, et une clé d’API peut être limitée aux seuls connecteurs partagés. Cette partie est en *public preview* : l’interface peut changer.

CHAPITRE 5

Le traitement par lot, ou la moitié du prix

Le **traitement par lot** exécute un très grand nombre de requêtes en asynchrone, avec « une remise de 50 % ». C'est le mécanisme qui rend abordables exactement les usages où une collectivité a quelque chose à gagner : reprise d'un fonds, requalification d'un stock, classement d'un historique.

Point	Ce que la documentation précise
Remise	50 % sur le coût de calcul
Volume maximal	1 million de requêtes par lot
Sans fichier	En dessous de 10 000 requêtes , le lot peut être passé en ligne, sans téléverser de fichier
Format	JSONL, une requête par ligne : <code>{"custom_id": "...", "body": {...}}</code>
Points d'entrée acceptés	conversations, complétions de chat, complétions à trous, embeddings, modération, OCR, classifications, transcription audio
États d'un lot	QUEUED, RUNNING, SUCCESS, FAILED, TIMEOUT_EXCEEDED, CANCELLATION_REQUESTED, CANCELLED
Récupération	Téléchargement du fichier de sortie par son identifiant, une fois le lot terminé

```
# Un lot, c'est un fichier texte. Une ligne = une requete.
{"custom_id":"delib-2019-0412","body":{"model":"mistral-small-latest",
  "messages":[{"role":"user","content":"Resume en 3 phrases ..."}]}}
{"custom_id":"delib-2019-0413","body":{"model":"mistral-small-latest",
  "messages":[{"role":"user","content":"Resume en 3 phrases ..."}]}}

# Le custom_id est votre identifiant metier : c'est lui qui vous
# permettra de recoller chaque reponse a la bonne deliberation.
# Le choisir stable et signifiant, jamais un numero de ligne.
```

LA BONNE SÉQUENCE POUR UNE REPRISE DE FONDS

1. Prendre **50 documents** représentatifs, y compris les moches : le scan de travers, le tableau sur deux pages, le manuscrit.
2. Les traiter en direct, en itérant sur la consigne jusqu'à ce que le résultat tienne. C'est ici que se passe le travail.
3. Relire les 50 sorties **à la main, en entier**. Compter les erreurs, les nommer, décider lesquelles sont acceptables.
4. Alors seulement lancer le lot complet, et échantillonner 1 % du résultat pour vérifier que le taux d'erreur tient à l'échelle.
5. Conserver le fichier d'entrée et le fichier de sortie. C'est votre preuve de ce qui a été fait, et votre point de reprise si un modèle change.

CHAPITRE 6

Workflows : quand il faut de la durabilité

Mistral Workflows est une plateforme d'orchestration pour « processus en plusieurs étapes combinant appels au modèle, usage d'outils, API externes et intervention humaine ». Elle est en **public preview**, et elle mérite d'être connue parce qu'elle répond à un besoin très administratif : un traitement qui s'étale, qui attend une validation, et qui ne doit rien perdre en route.

Propriété	Ce qu'elle apporte	Traduction pour une collectivité
Durabilité	« Les pannes ne perdent pas le travail » : chaque étape est enregistrée, un autre processus reprend à la dernière étape terminée	Un traitement de nuit interrompu ne se relance pas de zéro
Reprises	Les réessais sont natifs, avec un délai configurable par activité	Une API tierce indisponible ne casse pas la chaîne
Attente longue	Un workflow se met en pause sur un signal humain ou un événement externe ; il « s'exécute de quelques secondes à plusieurs mois »	Le circuit de validation d'un agent ou d'un élu entre dans le processus
Planification	Tâches récurrentes ou ponctuelles, façon cron	Le traitement du lundi matin
Exécution hybride	Mistral héberge l'orchestrateur, votre code s'exécute chez vous — poste de développement, Kubernetes, machines virtuelles	Argument utile en comité de sécurité : la logique métier ne quitte pas votre infrastructure
Observabilité	Traces OpenTelemetry sans câblage supplémentaire, visibles dans Studio ou exportables vers votre propre outil	Auditabilité, et pas de dépendance à la console de l'éditeur

PUBLIC PREVIEW : CE QUE ÇA VEUT DIRE DANS UN MARCHÉ

« Public preview » signifie que l'interface peut changer sans préavis contractuel. Pour un prototype ou un service interne, c'est acceptable. Pour une brique dont dépend un service au public, cela demande une clause : que se passe-t-il si l'API change, qui paie la reprise, et sous quel délai. Poser la question au moment du marché coûte une phrase ; la poser après coûte un avenant.

CHAPITRE 7

Administrer : clés, workspaces, limites

C'est le chapitre que tout le monde saute et que tout le monde regrette. Trois faits structurent l'organisation d'un compte Studio.

Fait	Conséquence pratique
Une clé est rattachée au workspace où elle est créée	Créer un workspace par environnement — développement, recette, production — chacun avec ses clés
Les limites d'appel sont au niveau du workspace et partagées par toutes ses clés	Deux applications dans le même workspace se volent mutuellement leur débit. C'est la panne difficile à diagnostiquer
La clé complète n'est montrée qu'une fois	Elle va directement dans un coffre à secrets. Pas dans un courriel, pas dans un ticket, pas dans un dépôt de code
Les clés peuvent porter une date d'expiration	Une politique d'organisation fixe une durée maximale ; un administrateur de workspace peut être plus strict, jamais plus laxiste
Il n'y a pas de rotation en place	Faire tourner une clé, c'est en créer une nouvelle puis supprimer l'ancienne. À prévoir dans la procédure, pas à improviser
La portée connecteur se choisit à la création	Par défaut, une clé n'atteint que les connecteurs partagés — le bon réglage pour un traitement automatique

Sur les **limites de débit**, la documentation reste volontairement générale : elles s'expriment en requêtes par seconde, en jetons par minute et en jetons par mois, elles varient selon l'offre et le modèle, et les valeurs propres à votre organisation s'affichent dans le panneau d'administration, à la page des limites. Aucun barème public ne permet de les anticiper — c'est une vérification à faire **avant** de dimensionner un traitement de masse, pas pendant.

TROIS WORKSPACES, PAS UN

La faute la plus fréquente et la plus coûteuse : une seule clé, créée le premier jour par la première personne, utilisée par le prototype, la recette et la production. Quand elle fuit, tout s'arrête en même temps ; quand un test déraile, il consomme le débit du service en ligne ; et quand son créateur quitte la collectivité, personne ne sait ce qui en dépend. Trois workspaces, trois clés, une date d'expiration : dix minutes au début, des semaines économisées ensuite.

CAS D'USAGE

Un EPCI et ses demandes citoyennes

Communauté d’agglomération fictive · 95 000 habitants · 28 communes · un guichet unique qui reçoit environ 180 000 sollicitations par an, par formulaire web, courriel et téléphone retranscrit. Deux agents à plein temps sur le tri et le routage, et un délai de première réponse qui dérive à 9 jours en été.

Le besoin n’était pas « faire de l’IA ». Il était : **ramener le délai de premier routage sous 24 heures sans recruter**. C’est cette formulation qui a permis de chiffrer, puis d’arbitrer.

Ce qui a été mis en place

Brique	Choix retenu	Pourquoi
Modèle	Mistral Small 4	La tâche est un classement, pas une rédaction. Large 3 testé : pas de gain mesurable sur cette tâche
Format	Sortie structurée à schéma imposé, avec confiance et drapeau de relecture	Insérable directement dans la GRC existante
Volume courant	Appels en direct, au fil de l’eau	Le délai compte : un lot nocturne aurait tué l’objectif des 24 heures
Reprise de l’historique	Traitement par lot sur 3 ans d’archives	Moitié prix, et aucune urgence
Seuil	Confiance sous 0,80 ou drapeau levé → file humaine	Décidé après relecture manuelle de 300 cas
Workspaces	Trois : bac à sable, recette, production	Le débit de la production n’est pas partagé
Hors périmètre	Aucune réponse rédigée automatiquement au citoyen	Décision politique, prise avant le premier appel d’API

Résultats à six mois

Indicateur	Avant	Après
Délai médian de premier routage	3 jours, jusqu'à 9 en été	moins de 2 heures
Part des demandes routées sans intervention	0 %	83 %
Erreurs de routage détectées	≈ 6 % (mesuré sur un échantillon)	≈ 4 %, et tracées
Temps agent consacré au tri	2 ETP	0,6 ETP, redéployé sur le traitement au fond
Coût annuel des appels d'API	—	≈ 15 \$, plus ≈ 200 \$ de reprise d'historique, une fois
Coût complet du projet, première année	—	≈ 42 000 € : conception, développement, intégration GRC, accompagnement des agents, hébergement

LES DEUX LIGNES QUI COMPTENT DANS CE TABLEAU

15 \$ contre 42 000 €. Le rapport est de un à deux mille. Toute discussion, tout comité, toute négociation qui porte sur le premier chiffre se trompe de sujet.

Le taux d'erreur n'est pas tombé à zéro, il est passé de 6 % à 4 %. Et c'est le second gain réel, moins visible que le délai : avant, les erreurs existaient sans être comptées. Elles sont désormais tracées, échantillonnées et discutées en revue mensuelle. Un dispositif qui se mesure vaut mieux qu'un dispositif qu'on croit juste.

CE QUI A ÉTÉ DIFFICILE, ET QUE PERSONNE N'AVAIT CHIFFRÉ

La définition des catégories. Le service en avait 14 dans le logiciel, dont quatre que personne n'utilisait de la même façon. Trois semaines d'atelier ont été nécessaires avant la première ligne de code — et c'est ce travail, pas le modèle, qui explique les 83 %.

L'accueil par les agents. Présenté comme un tri automatique, le dispositif inquiétait ; présenté comme une file de lecture qui leur envoie les cas douteux avec la raison du doute, il a été adopté. Le système n'avait pas changé, seulement la manière de le décrire — et la décision, prise dès le départ, de ne jamais laisser une réponse partir sans agent.

DONNÉES

RGPD, rétention et marché public

L’API a un avantage net sur l’abonnement, et c’est le seul endroit de l’écosystème Mistral où la **rétention zéro** est atteignable. Encore faut-il en connaître les conditions exactes.

Sujet	Ce qui est garanti	La condition
Rétention par défaut	Les entrées et sorties sont conservées 30 jours glissants pour la détection d’abus, puis supprimées	S’applique à l’usage courant de l’API
Rétention zéro	Possible : rien n’est conservé au-delà du temps de la réponse, et la fenêtre de 30 jours ne s’applique plus	Uniquement en paiement à l’usage, sur les appels sans état — complétions, embeddings, modération, OCR, audio. Sur demande motivée, accordée à la discrétion de l’éditeur
Ce qui en est exclu	—	Le traitement par lot, les conversations, les Libraries, les agents, le stockage de fichiers. Et l’offre gratuite
Entraînement	Un réglage de retrait empêche l’usage des données éligibles pour améliorer les modèles	Distinct de la rétention : ne pas confondre les deux dans une note au DPO
Hébergement	Infrastructure européenne, éditeur français, soumis au RGPD par construction	La qualification SecNumCloud concerne certaines offres d’hébergement, à vérifier au cas par cas — elle ne se déduit pas de la nationalité de l’éditeur
Certifications	ISO 27001 et 27701, SOC 2 Type II, rapports disponibles sur demande	Aucune certification ISO 42001 revendiquée publiquement à ce jour

LE POINT QUI SURPREND · LE LOT N’EST PAS COUVERT

La rétention zéro exclut le traitement par lot. Or c’est précisément le mode qu’on emploie pour reprendre un fonds documentaire — donc le moment où le volume de données traité est le plus grand.
La conséquence n’est pas qu’il faut renoncer, c’est qu’il faut le savoir et l’écrire : sur une reprise d’archives comportant des données personnelles, la fenêtre de 30 jours s’applique, elle doit figurer dans l’analyse d’impact, et elle se discute avec le DPO **avant** le lancement du lot, pas après.

Cinq phrases à faire figurer dans un marché

Non pas du droit — votre service juridique fera mieux — mais les cinq points que les cahiers des charges d'IA oublient le plus souvent, et qui coûtent cher ensuite.

- **Le modèle n'est pas nommé.** Spécifier une famille et un niveau de qualité mesuré sur vos données, pas « Mistral Medium 3.5 » : les versions changent plus vite que les marchés.
- **Les prompts et les jeux d'évaluation appartiennent à la collectivité.** C'est le vrai actif du projet, et le seul qui vous suit chez un autre prestataire ou un autre éditeur.
- **La réversibilité est décrite concrètement** : format d'export des données et des index, délai, et démonstration de restitution en fin de marché.
- **Le taux d'erreur acceptable est chiffré et mesuré** par un protocole écrit, sur un échantillon défini, à une fréquence définie. Sans cela, « ça marche » n'est pas vérifiable.
- **Aucun envoi automatique à un administré** sans validation humaine, sauf décision explicite et tracée de la collectivité, prise à un niveau qui l'engage.

GARDE-FOUS

Bonnes pratiques pour un usage durable

Risque	Mitigation
Le POC qui ne devient jamais un service	Décider dès le départ du critère de passage en production, et de qui le prononce. Un POC sans critère de sortie est une dépense, pas une expérimentation
Un modèle nommé dans un document contractuel	Spécifier un niveau de qualité mesuré, pas une référence commerciale
Le fine-tuning inscrit au cahier des charges	La fonction est dépréciée chez Mistral. Passer par le contexte : documents fournis, recherche documentaire, consignes explicites
Le coût des outils confondu avec celui des jetons	Un appel de recherche web coûte 0,03 \$, mille fois une requête texte. Compter les appels d'outils séparément
Une clé unique pour tous les environnements	Trois workspaces, trois clés, une date d'expiration, un coffre à secrets
Des limites de débit découvertes en production	Les relever dans le panneau d'administration avant de dimensionner, et prévoir le comportement en cas de dépassement
Une brique en <i>public preview</i> sous un service au public	Clause explicite sur le changement d'interface : qui reprend, à quel délai, à quel prix
Un classement sans indice de confiance	Seuil et drapeau de relecture, file humaine pour le reste, taux d'erreur échantillonné et discuté
La dépendance à un seul prestataire	Prompts, schémas et jeux d'évaluation versionnés côté collectivité, dans un dépôt qu'elle possède

Règle d'or

Une API ne résout pas un problème : elle industrialise une solution déjà trouvée. Si le traitement n'a jamais été fait correctement à la main sur cinquante cas, il n'y a rien à automatiser — il y a un travail de définition à faire d'abord. C'est ce travail-là, et non le choix du modèle, qui décide du sort du projet.

Un mois pour savoir si le projet tient

Semaine	Ce qu'on fait	Ce qu'on obtient
1	Écrire le problème en une page : volume annuel, minutes par cas, conséquence d'une erreur. Sans API, sans code	La moitié des idées s'arrêtent ici, et c'est le meilleur retour sur investissement du mois
2	Créer un workspace « bac à sable », une clé expirante, traiter 50 cas réels en direct, itérer sur la consigne	Une mesure, pas une impression
3	Relire les 50 sorties à la main. Compter et nommer les erreurs. Définir le seuil et le drapeau de relecture	Le taux d'erreur réel, et le partage entre automatique et humain
4	Chiffrer le projet complet — intégration, accompagnement, maintenance sur cinq ans — et remettre les appels d'API à leur place, en bas du tableau	Un dossier défendable, ou un abandon argumenté

Un abandon en semaine 4 n'est pas un échec : c'est un mois passé au lieu d'une année. Les projets qui coûtent cher aux collectivités ne sont pas ceux qu'on arrête, ce sont ceux qu'on n'a jamais su arrêter.

SOURCES

Ce qui a été vérifié, et où

Tous les prix, limites et noms de fonctions de ce guide ont été relevés le **31 août 2026** sur les pages officielles ci-dessous. Les tarifs et les modèles de Mistral bougent vite : avant de citer un chiffre de ce guide dans un document engageant, le revérifier à sa source.

Ce qui en vient	Page officielle
Périmètre de Studio et carte des briques	docs.mistral.ai/studio
Agents, conversations, outils intégrés, passage de main	docs.mistral.ai/studio/agents/introduction
Prix par modèle, OCR, outils, remises lot et cache	mistral.ai/pricing/api
Traitement par lot : remise, plafonds, format, états	docs.mistral.ai/studio/batch-processing
Document AI : OCR, annotations, questions-réponses	docs.mistral.ai/studio/document-processing/overview
RAG : Libraries gérées, embeddings, découpage	docs.mistral.ai/studio/knowledge-rag/rag_quickstart
Connecteurs et MCP, portées privée et partagée	docs.mistral.ai/studio/connectors
Workflows : durabilité, reprises, mode hybride, traces	docs.mistral.ai/studio-api/workflows/getting-started/overview
Clés d'API, workspaces, expiration, rotation, portées	docs.mistral.ai/admin/identity-access/api-keys
Limites de débit et leur emplacement	docs.mistral.ai/admin/user-management-finops/tier
Fine-tuning déprécié	docs.mistral.ai/resources/deprecated/finetuning
Rétention zéro : périmètre et exclusions	help.mistral.ai — « Can I activate Zero Data Retention (ZDR)? »
Certifications ISO 27001/27701 et SOC 2 Type II	help.mistral.ai ; trust.mistral.ai

POUR ALLER PLUS LOIN

Continuer avec Ancoria

- Guide n° 03 — Mistral Vibe pour la réponse aux appels à projets — ancoria.tech/ressources
- Le Brief IA des Territoires, la lettre — ancoria.tech
- Échanger avec l'équipe — ancoria.tech/a-propos

Ce guide a été conçu par **Ancoria — by Ancoris Solutions**, plateforme française dédiée à l'IA appliquée aux territoires et à l'immobilier d'entreprise. Veille stratégique, bibliothèque de prompts, analyses comparatives.

Édition 1 — août 2026. Le cas d'usage est une reconstitution : les proportions et les ordres de grandeur sont réalistes, la collectivité ne l'est pas. Prochaine relecture de fond prévue en février 2027.