ROYAUME DU MAROC المملكة المغربية

مكتب التكوين المهني وإنعاش الشغل

Office de la Formation Professionnelle et de la Promotion du Travail

SOCLE SOUVERAIN NIBRAS • PASSERELLE CENTRALE DE GOUVERNANCE IA ET ASSISTANT CONVERSATIONNEL

MANUEL D'UTILISATION

NIBRAS Admin Gateway : administration, gouvernance des accès, base de connaissances RAG, quotas et supervision des services d'IA - et guide d'utilisation de Chat NIBRAS

Auteur & concepteur	Mohamed OUBEZZA (mohamed.oubezza@ofppt.ma)
Organisme de tutelle	Office de la Formation Professionnelle et de la Promotion du Travail (OFPPT)
Projet institutionnel	Projet NIBRAS - Intelligence artificielle souveraine
Accès publics	https://admin.nibras.site (NIBRAS Admin) • https://chat.nibras.site (Chat NIBRAS)
Année de formation	2026/2027
Version du document	3.2 - édition mise à jour : vision / OCR, réglage thinking, recettes de modèles, connexions fournisseurs (30 Octobre 2026, application v0.10.0)
Statut du document	Document officiel de référence & guide opérationnel

SOMMAIRE

1. Présentation, rôles et vocabulaire 4
1.1 NIBRAS Admin dans l'écosystème NIBRAS 4
1.2 Vocabulaire 4
1.3 Profils d'utilisation 5
1.4 Par où commencer ? 5
1.5 Conventions de ce manuel 6
2. Se connecter et gérer son compte 7
2.1 Connexion 7
2.2 Mot de passe oublié 7
2.3 Profil, mot de passe et déconnexion 8
3. Naviguer et consulter la vue d'ensemble 9
3.1 Le menu latéral 9
3.2 La vue d'ensemble 9
4. Gérer les utilisateurs et leurs permissions 12
4.1 Créer ou modifier un compte 12
4.2 Consulter ou créer un rôle 12
4.3 Ajuster les permissions individuelles 13
5. Gérer les applications clientes 15
6. Générer et gérer les clés API 17
6.1 Émettre une clé 17
6.2 Réglage « Réflexion (thinking) » 18
6.3 Comprendre les restrictions 19
6.4 Quel modèle une clé utilise-t-elle ? 19
6.5 Suspendre, réactiver, révoquer ou renouveler 19
7. Gérer les fournisseurs et les modèles 21
7.1 Fournisseurs 21
7.2 Modèles 22
7.3 Paramètres de génération et recettes de thinking 22
7.4 Vision et format de réponse 23
7.5 Supprimer un fournisseur ou un modèle 24
8. Tester un modèle de langage 26
9. Créer et autoriser les collections 27
9.1 Créer une collection 27
9.2 Accorder les accès 28
10. Importer et gérer les documents 29
10.1 Importer un lot 29
10.2 Consulter la fiche et les versions 30
10.3 Désactivation, suppression et restauration 31
11. Indexer les documents 33
11.1 Lancer l'indexation 33
11.2 Lire l'état d'une tâche 33
11.3 Vérifier les passages publiés 34

12. Rechercher et générer une réponse RAG 35
12.1 Rechercher les passages 35
12.2 Générer une réponse sourcée 36
13. Consulter la consommation 38
14. Définir et suivre les quotas 40
15. Examiner les requêtes 41
16. Consulter le journal d'audit 43
17. Surveiller les services 44
18. Modifier les paramètres de la plateforme 45
18.1 Confidentialité et conservation 46
19. Utiliser Chat NIBRAS 47
19.1 Ce que Chat NIBRAS permet de faire 47
19.2 Accéder à Chat NIBRAS et prendre en main l'interface 47
19.3 Poser une question sur les documents de l'office 49
19.4 Quand Chat NIBRAS répond qu'il n'a pas l'information 50
19.5 Analyser un fichier déposé 50
19.6 Bonnes pratiques et limites 51
19.7 Gouverner Chat NIBRAS depuis NIBRAS Admin 52
20. Parcours pratiques 53
20.1 Donner accès à une nouvelle application 53
20.2 Brancher une application de lecture d'images (OCR) 53
20.3 Préparer une base documentaire exploitable 53
20.4 De l'import d'un document à la réponse dans Chat NIBRAS 54
20.5 Vérifications régulières 54
21. Résoudre les problèmes courants 55
Annexe A - Référence des permissions et des scopes 57
Annexe B - Lire un code d'erreur 58
Annexe C - Accès aux applications NIBRAS et index des écrans 59

1. Présentation, rôles et vocabulaire

1.1 NIBRAS Admin dans l'écosystème NIBRAS

NIBRAS Admin Gateway est la passerelle centrale du socle NIBRAS. Elle centralise les accès aux services d'intelligence artificielle, les modèles disponibles, la base documentaire (RAG), les quotas et la supervision. L'interface d'administration sert à **configurer et contrôler** ces services ; les applications clientes - **Perform**, **Genero**, **Correcto** et **Chat NIBRAS** - les utilisent avec leurs propres clés API NIBRAS.



Figure 1 - *Le portail NIBRAS : accès à Chat NIBRAS, Perform, Genero et Correcto*

Ce manuel comporte trois parties :

- **Administration (chapitres 2 à 18)** : connexion, utilisateurs et rôles, applications, clés API, modèles, base de connaissances, consommation, quotas, traçabilité, supervision et paramètres.
- **Chat NIBRAS (chapitre 19)** : guide de l'utilisateur de l'assistant conversationnel institutionnel, et manière dont l'administrateur le gouverne depuis NIBRAS Admin.
- **Parcours et dépannage (chapitres 20 et 21)** : scénarios de bout en bout, vérifications régulières et résolution des problèmes courants.

> **Principe de gouvernance**
>
> Les secrets d'inférence (clés des fournisseurs de modèles) restent **côté serveur** : les applications ne reçoivent que des clés NIBRAS, limitées à des services, des modèles et des collections précis, et bornées par des quotas. Chaque action d'administration est consignée dans le journal d'audit.

1.2 Vocabulaire

Terme	Signification dans l'application
Application cliente	Service consommateur, par exemple Perform, Genero, Correcto ou Chat NIBRAS.
Clé API	Identifiant secret autorisant une application à appeler les services NIBRAS.

Terme	Signification dans l'application
Fournisseur	Serveur où est hébergé un modèle de langage, identifié par son URL ou son adresse IP (par exemple le serveur de l'ISTA Bab Tizimi qui héberge Qwen, fournisseur TIZIMI). Un fournisseur peut servir plusieurs modèles.
Modèle public	Nom exposé par NIBRAS aux clients, associé à un modèle réellement servi.
Token	Unité de texte utilisée pour mesurer les entrées, les sorties et les budgets.
Collection	Ensemble de documents partageant une configuration et des autorisations.
Passage ou chunk	Fragment de texte extrait d'un document pour la recherche.
Embedding	Représentation numérique du texte permettant de rechercher des passages proches d'une question.
RAG	Recherche de passages documentaires puis génération d'une réponse accompagnée de sources.
Quota	Limite de consommation ou de fréquence imposée aux appels.
Audit	Historique des actions d'administration et de sécurité.

1.3 Profils d'utilisation

Rôle initial	Permissions	Usage habituel
SUPER_ADMIN	29	Administration générale, rôles et permissions, paramètres et accès aux collections.
ADMIN	26	Administration de la plateforme, dans le respect des restrictions des collections. Ne gère pas les rôles.
RAG_MANAGER	6	Création et gestion documentaire, import, indexation et consultation des collections autorisées.
AUDITOR	13	Consultation des accès, consommations, traces, audit et supervision, sans modification ; recherche documentaire selon ses accès.
USER	0	Accès à son compte ; les autres fonctions nécessitent des permissions supplémentaires.

Ces descriptions correspondent aux rôles initiaux ; le nombre de permissions est celui affiché dans l'écran **Rôles**. Les permissions individuelles peuvent autoriser ou refuser une action (§ 4.3) : le nom d'un rôle ne suffit donc pas à déterminer tous les droits effectifs d'un utilisateur.

1.4 Par où commencer ?

Vous êtes…	Lisez en priorité
Administrateur de la plateforme	Chapitres 2 à 8 (comptes, applications, clés, modèles), 13 à 18 (suivi et paramètres), puis 20.
Gestionnaire documentaire (RAG_MANAGER)	Chapitres 9 à 12 (collections, documents, indexation, Test RAG), puis le parcours du chapitre 20.
Auditeur	Chapitres 3, 13, 15, 16 et 17 (consultation seule).
Utilisateur de Chat NIBRAS	Chapitre 19 uniquement.

1.5 Conventions de ce manuel

Repère	Signification
« Texte en gras »	Nom d'un bouton, d'un onglet, d'un champ ou d'une rubrique de menu.
Menu → Rubrique	Chemin dans le menu latéral, par exemple « Accès → Utilisateurs ».
`nom.technique`	Permission, scope, statut ou valeur technique affichée par l'application.

2. Se connecter et gérer son compte

2.1 Connexion

1. Ouvrir l'adresse communiquée par le responsable de l'installation (en production : **https://admin.nibras.site** ; sur l'installation locale préparée, l'interface compilée est accessible à `http://localhost:8080` lorsque les services sont démarrés).
2. Saisir son identifiant ou son adresse email, puis son mot de passe.
3. Valider la connexion pour accéder à la vue d'ensemble.

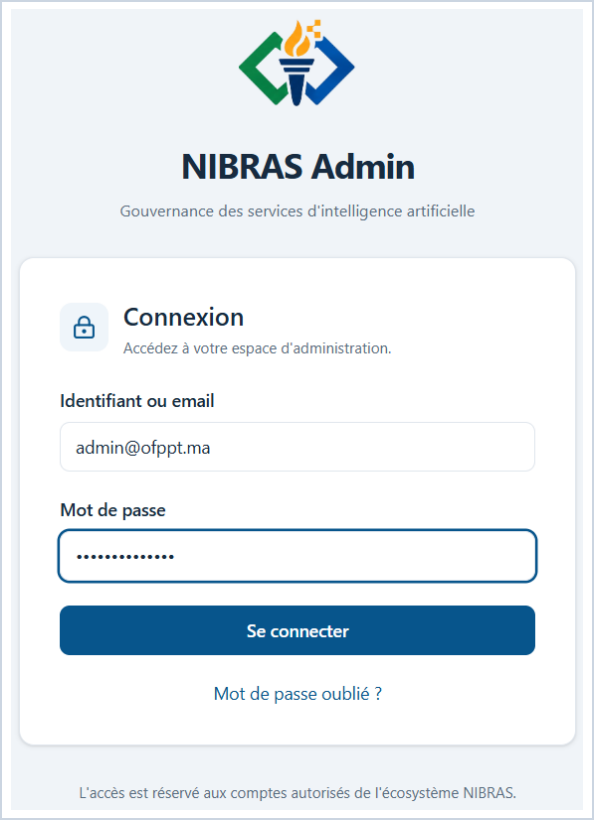


Figure 2 - *Écran de connexion, sans mot de passe visible*

Après plusieurs tentatives incorrectes, l'accès peut être temporairement bloqué : le verrouillage du compte intervient après **cinq échecs** et dure **quinze minutes** ; une limitation des tentatives par adresse IP s'applique également.

2.2 Mot de passe oublié

1. Sur l'écran de connexion, ouvrir **Mot de passe oublié**.
2. Saisir son adresse email et cliquer sur **Envoyer le lien**.
3. Si le service email est configuré, suivre le lien reçu et saisir le nouveau mot de passe.

> **Configuration SMTP requise**
>
> La récupération par email nécessite une configuration SMTP opérationnelle. Si l'écran indique que le service est indisponible, contactez l'administrateur pour rétablir l'accès.

2.3 Profil, mot de passe et déconnexion

Ouvrez le menu utilisateur puis **Mon compte**. Modifiez le nom complet ou l'email et cliquez sur **Enregistrer le profil**. Pour changer le mot de passe, renseignez le mot de passe actuel, le nouveau et sa confirmation, puis cliquez sur **Modifier le mot de passe** : le nouveau mot de passe doit comporter **entre 12 et 128 caractères**, et toutes les sessions ouvertes sont fermées après la modification. L'écran présente également l'historique des connexions (date, résultat, adresse IP). Utilisez la commande de déconnexion du menu utilisateur à la fin de la session.

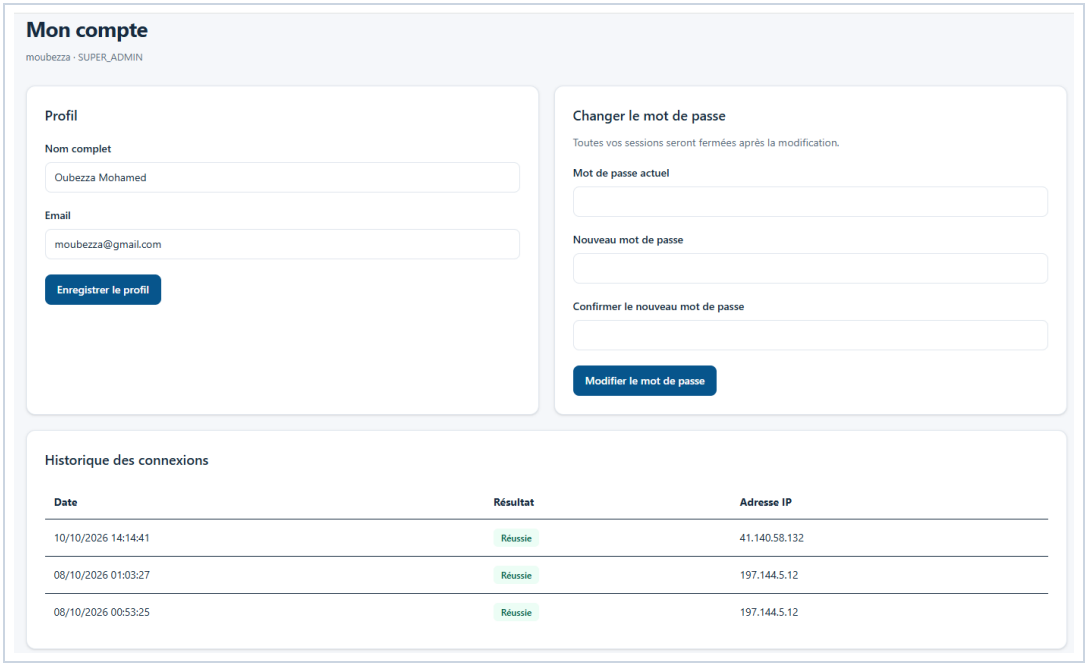


Figure 3 - *Profil, changement de mot de passe et historique des connexions, champs secrets vides*

3. Naviguer et consulter la vue d'ensemble

3.1 Le menu latéral

Le menu latéral regroupe les fonctions en **sous-menus repliables** : **Base de connaissances** (Collections, Documents, Indexation, Test RAG), **Accès** (Utilisateurs, Rôles, Applications, Clés API), **Intelligence artificielle** (Modèles, Fournisseurs, Tests), **Analytics** (Consommation, Quotas) et **Système** (Requêtes, Audit, Monitoring, Paramètres). Cliquez sur une rubrique pour afficher ses liens ; l'ouverture d'une rubrique referme la précédente afin de gagner de la place sans barre de défilement verticale. Le sous-menu de la page consultée s'ouvre automatiquement lors de la navigation. Sur petit écran, ouvrez le menu avec le bouton de navigation ; la touche **Échap** ferme la navigation ou le menu utilisateur ouvert.

> **Une rubrique est absente ?**
> Elle correspond généralement à une permission manquante : le menu s'adapte aux droits du compte connecté (§ 4.3). Les actions marquées **À venir** ne sont pas encore disponibles.

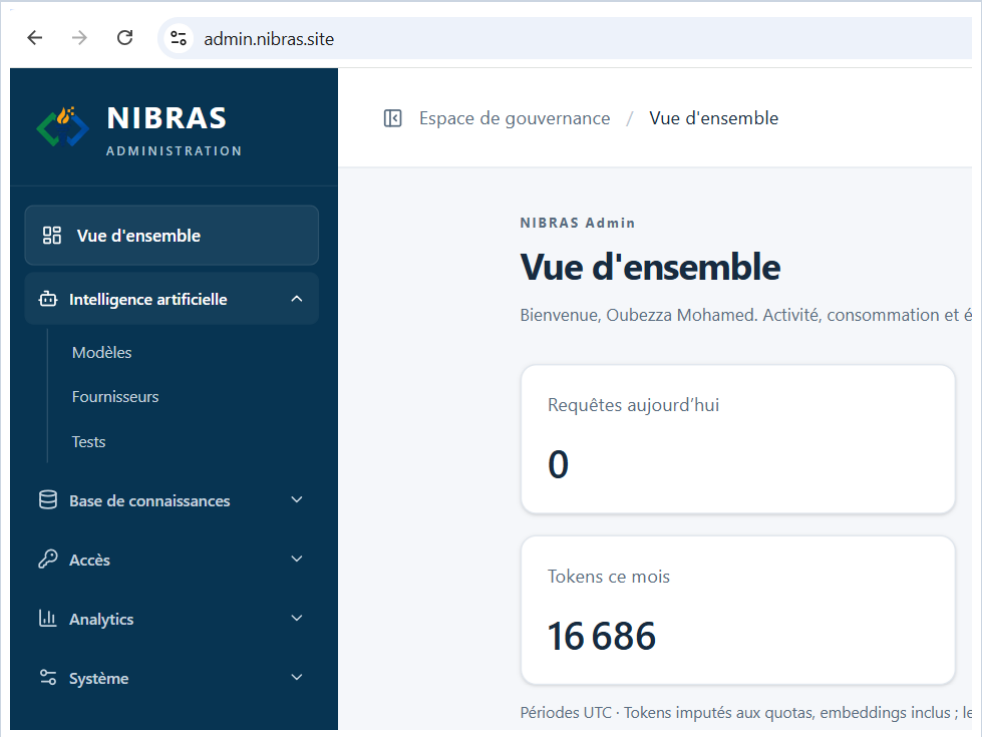


Figure 4 - *Menu latéral et rubriques de l'espace Intelligence artificielle*

3.2 La vue d'ensemble

La **Vue d'ensemble** affiche, selon les permissions :

- les requêtes et tokens du jour, ainsi que les tokens du mois ;
- les utilisateurs et applications actifs ;
- les documents indexés, collections RAG et clés API actives ;
- la consommation sur 30 jours, les principales applications et les principaux utilisateurs ;
- l'état des services, les dernières activités et les dernières erreurs.

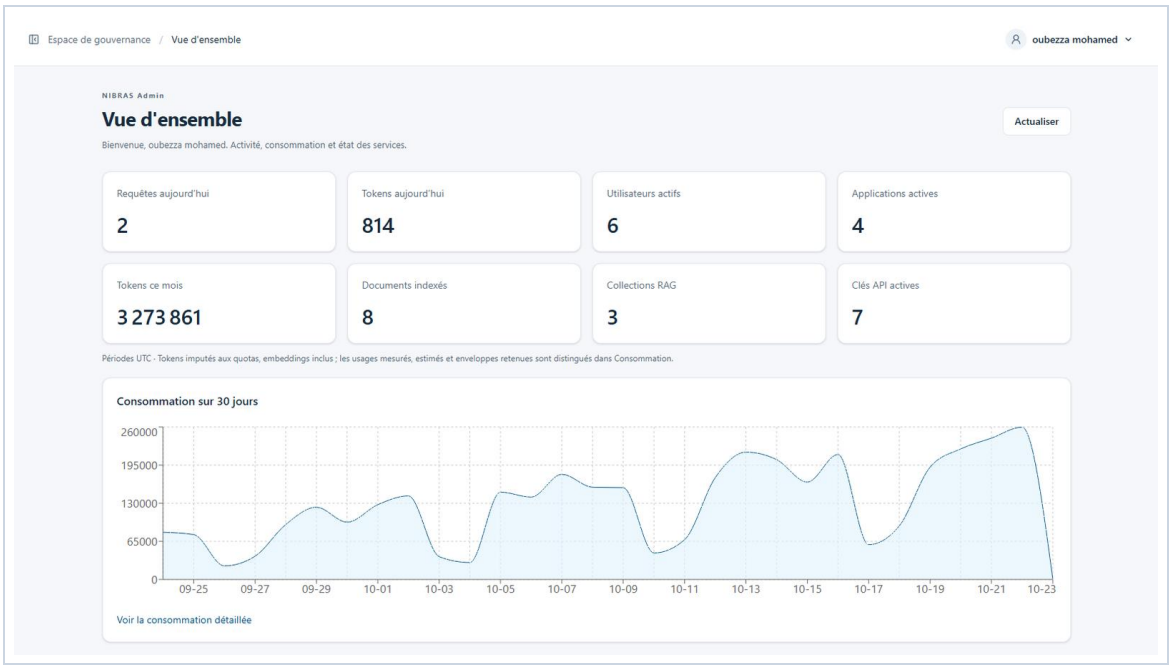


Figure 5 - *Vue d'ensemble : indicateurs du jour et du mois, et consommation sur 30 jours*

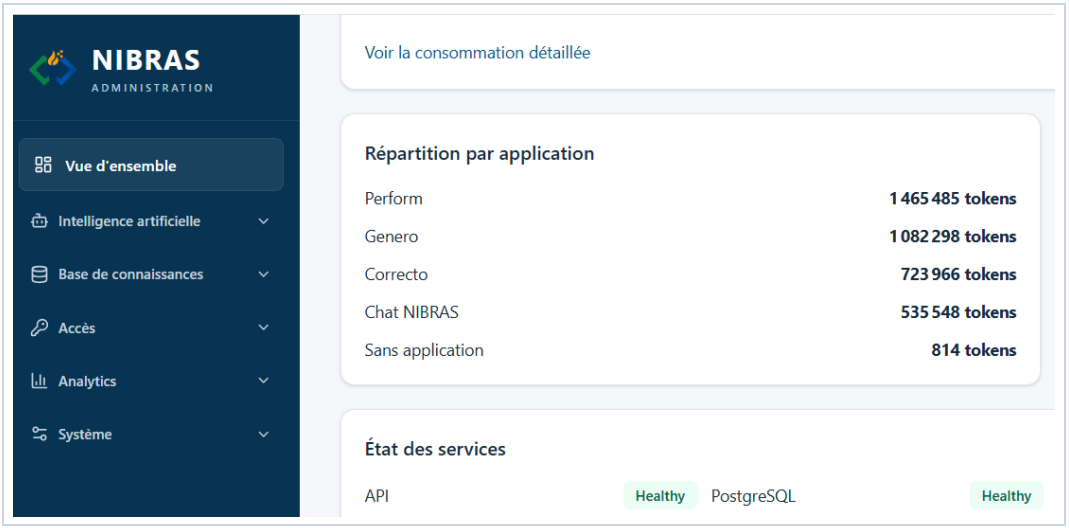


Figure 6 - *Répartition de la consommation par application, avec le menu latéral déployé*

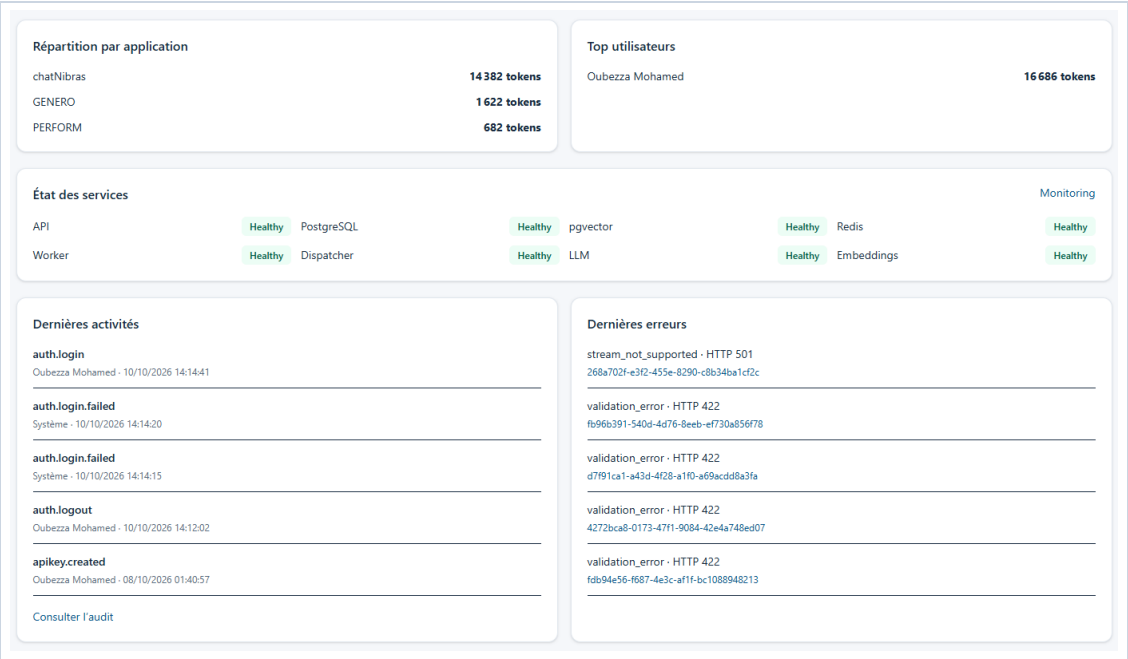


Figure 7 - *Top utilisateurs, état des services, dernières activités et dernières erreurs*

Utilisez le bouton **Actualiser** pour relire les données. Une valeur nulle peut simplement indiquer qu'aucune activité n'a encore été enregistrée. Le nombre d'utilisateurs actifs désigne les comptes *activés*, sans indiquer combien de personnes sont actuellement connectées. Les périodes sont exprimées en **UTC**.

4. Gérer les utilisateurs et leurs permissions

4.1 Créer ou modifier un compte

1. Ouvrir **Accès → Utilisateurs**.
2. Rechercher un compte par nom, identifiant ou email, ou utiliser **Créer un utilisateur**.
3. Pour un nouveau compte, renseigner l'identifiant, le nom complet, l'email et le mot de passe initial.
4. Choisir le rôle et le statut du compte.
5. Enregistrer, puis vérifier les informations dans la fiche.

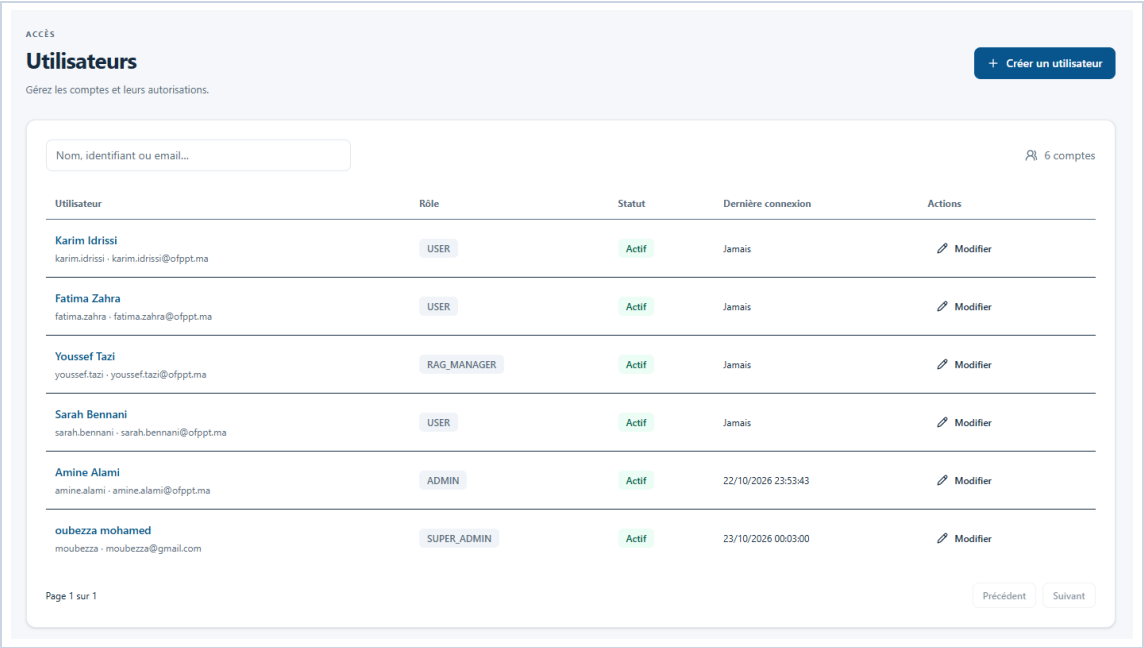


Figure 8 - *Liste des utilisateurs : rôle, statut, dernière connexion et action Modifier*

Statut	Effet
Actif	Connexion et actions autorisées selon les permissions.
Désactivé	Accès bloqué, y compris pour les clés associées au compte.
Suspendu	Accès bloqué tant que le statut n'est pas rétabli.

> **Protection du dernier SUPER_ADMIN**
> L'application protège le dernier SUPER_ADMIN actif pour éviter de perdre l'accès d'administration.

4.2 Consulter ou créer un rôle

1. Ouvrir **Accès → Rôles** : la colonne de gauche liste les rôles et leur nombre de permissions.
2. Sélectionner un rôle pour consulter ses permissions (cases cochées) ; les rôles initiaux portent le badge « Rôle initial ».
3. Si autorisé (`roles.manage`), cliquer sur **Créer un rôle** : choisir un nom en majuscules de 3 à 64 caractères, par exemple `LECTEUR_RAG`, puis une description.
4. Sélectionner les permissions et cliquer sur **Enregistrer le rôle**.

5. Affecter ce rôle aux utilisateurs concernés depuis leurs fiches.

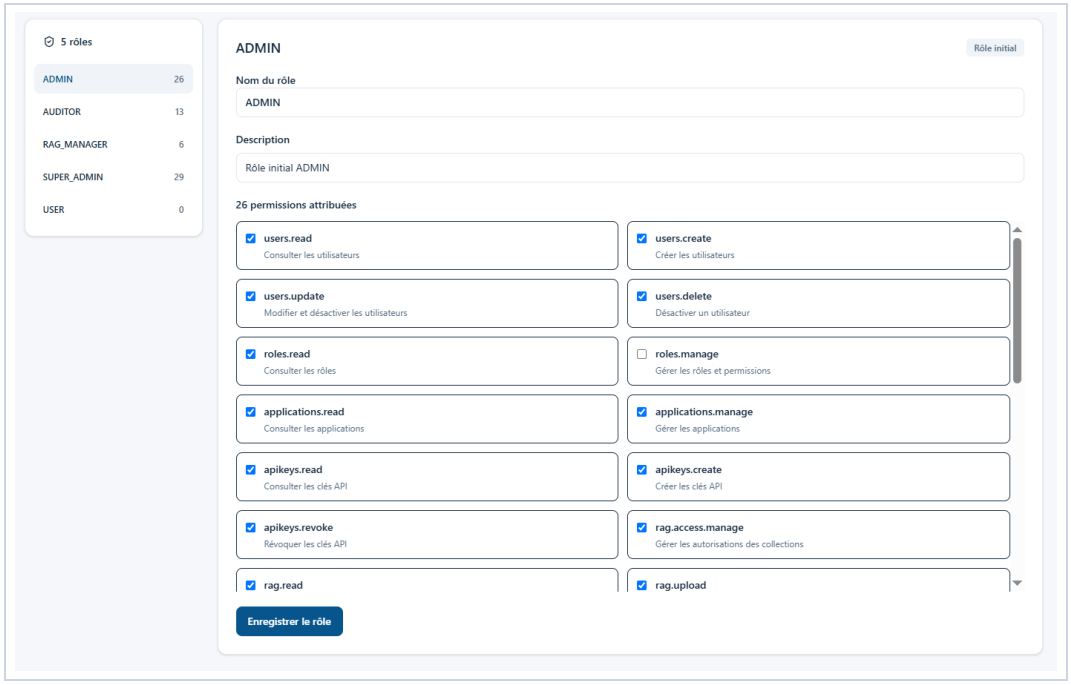


Figure 9 - *Rôles : permissions attribuées au rôle ADMIN (26 permissions)*

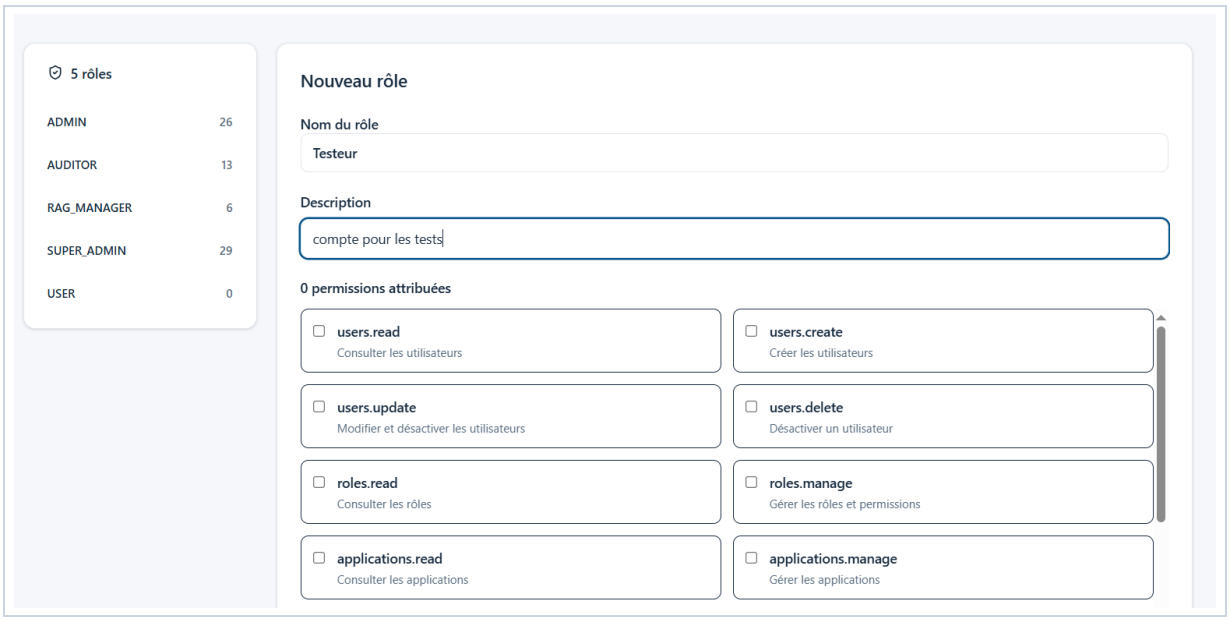


Figure 10 - *Création d'un rôle*

Les possibilités de modification dépendent du rôle sélectionné et des permissions du compte connecté : sur la capture ci-dessus, le rôle ADMIN ne possède pas `roles.manage`. La référence complète des permissions figure en annexe A.

4.3 Ajuster les permissions individuelles

Dans la fiche d'un utilisateur, chaque permission propose trois choix :

Choix	Effet
Hériter	Appliquer la permission du rôle.

Choix	Effet
Autoriser	Accorder explicitement la permission.
Refuser	Interdire explicitement la permission, même si le rôle l'accorde.

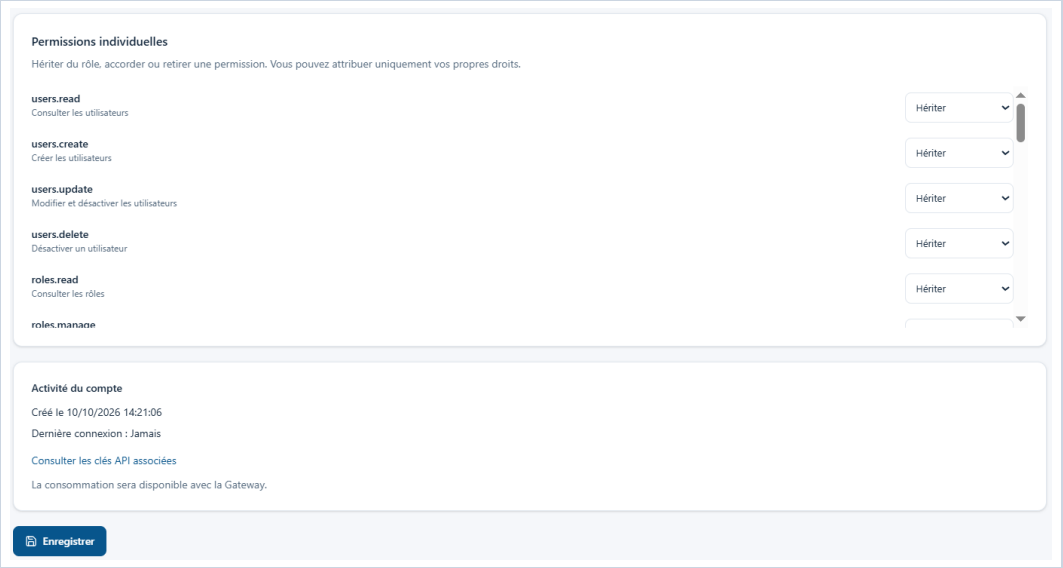


Figure 11 - *Fiche utilisateur : permissions individuelles et activité du compte*



Figure 12 - *Choix Hériter / Autoriser / Refuser pour une permission*

Enregistrez après modification, puis vérifiez séparément les droits de consultation, de création, de modification et de suppression nécessaires au travail attendu.

5. Gérer les applications clientes

Une application cliente représente un service qui consomme NIBRAS (Perform, Genero, Correcto, Chat NIBRAS…). Elle peut posséder plusieurs clés API, par exemple une par environnement.

1. Ouvrir **Accès → Applications**.
2. Utiliser la recherche pour retrouver une application, ou **Créer une application**.
3. Renseigner le nom, l'identifiant unique (*slug*), la description et le statut.
4. Enregistrer, puis ouvrir la fiche pour consulter les clés associées.

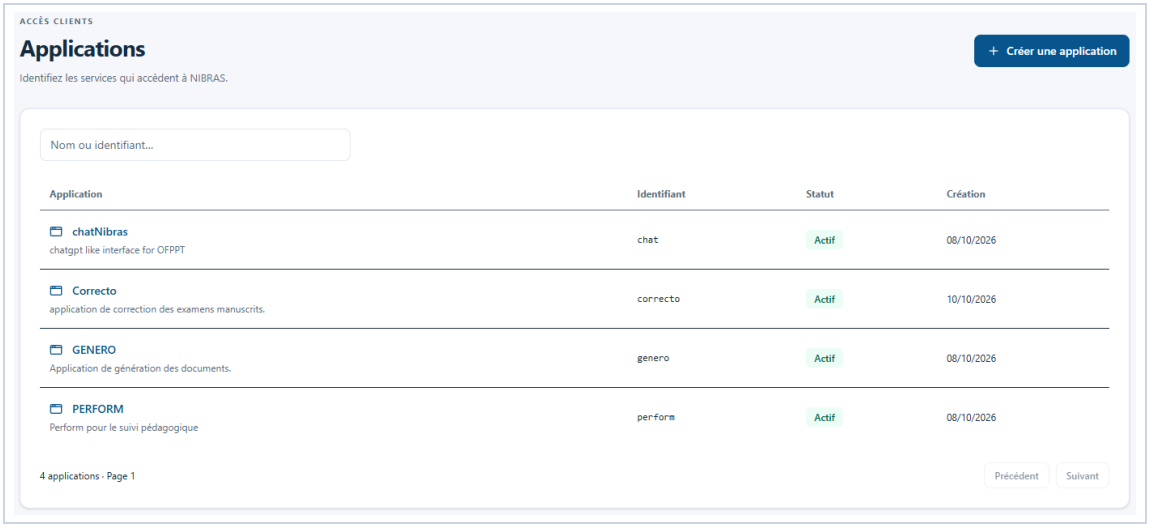


Figure 13 - *Applications clientes enregistrées*

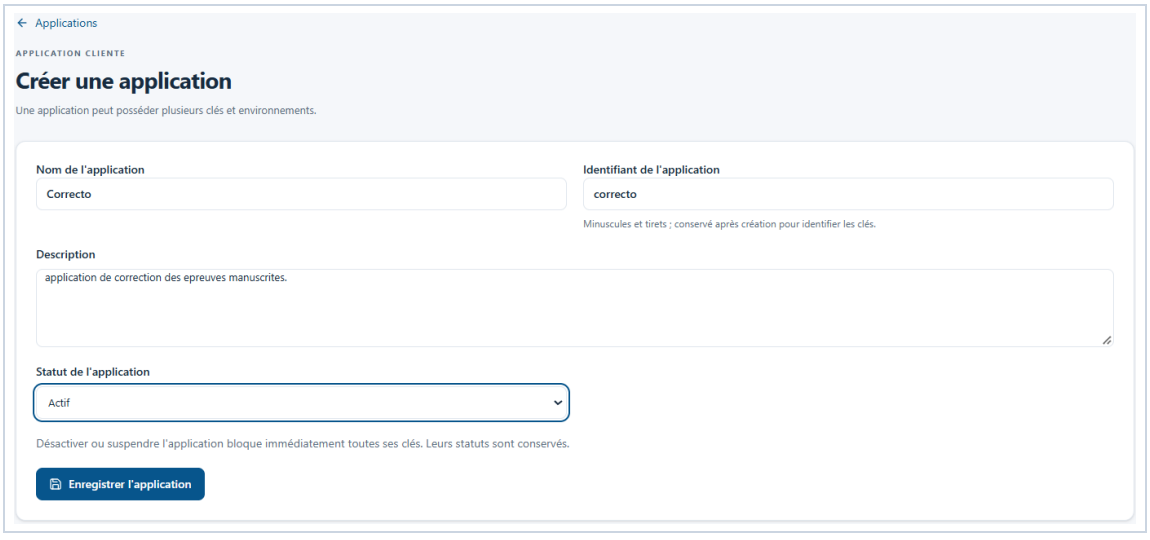


Figure 14 - *Création d'une application : nom, identifiant, description et statut*

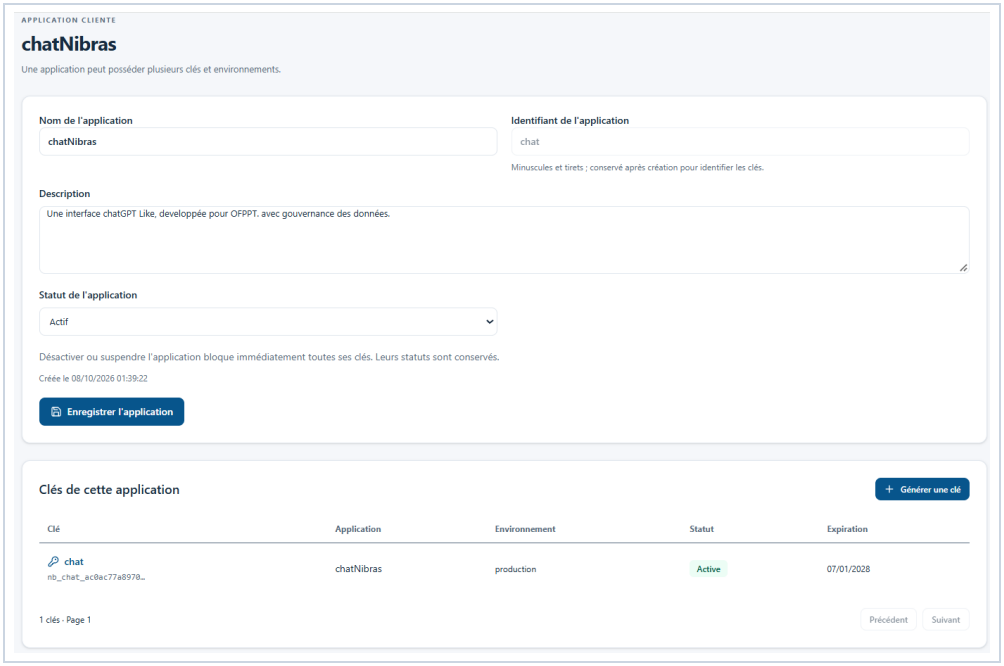


Figure 15 - *Fiche d'une application et liste de ses clés*

L'identifiant (*slug*) **reste fixe après création**. Les applications Perform, Genero, Correcto et Chat NIBRAS peuvent être présentes si leur initialisation a été effectuée ; leur présence ne signifie pas qu'une clé a déjà été générée.

> **Désactiver une application**
>
> Désactiver ou suspendre une application bloque l'utilisation de **toutes ses clés** dès les appels suivants. Les clés conservent leurs propres statuts : elles ne sont pas supprimées par cette action.

6. Générer et gérer les clés API

6.1 Émettre une clé

1. Ouvrir **Accès → Clés API**, puis **Générer une clé API**.
2. Choisir une application active et, si nécessaire, un utilisateur actif associé.
3. Renseigner le nom, le service, l'environnement et une expiration éventuelle.
4. Sélectionner les **scopes** utiles.
5. Définir les restrictions de modèles et de collections, puis les limites éventuelles.
6. Choisir, si nécessaire, le réglage **Réflexion (thinking)** (§ 6.2).
7. Générer la clé et **copier immédiatement sa valeur** dans le gestionnaire de secrets du client.

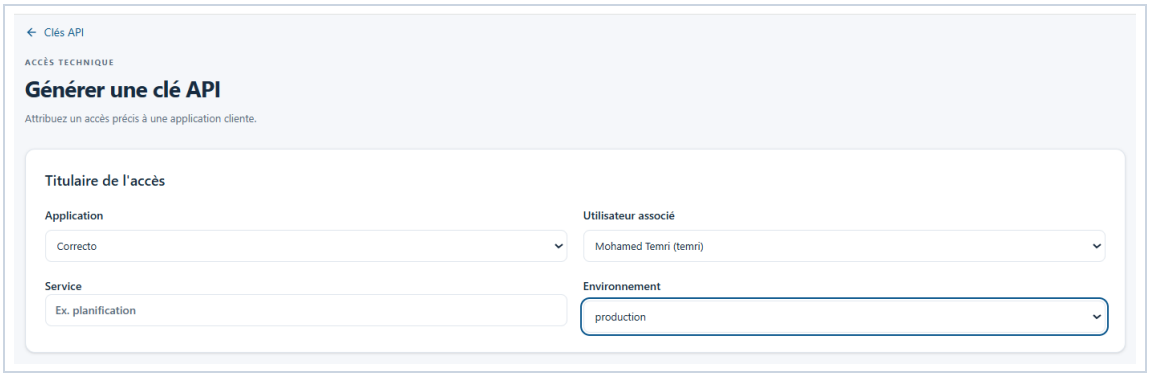


Figure 16 - *Génération d'une clé : titulaire de l'accès (application, utilisateur, environnement)*

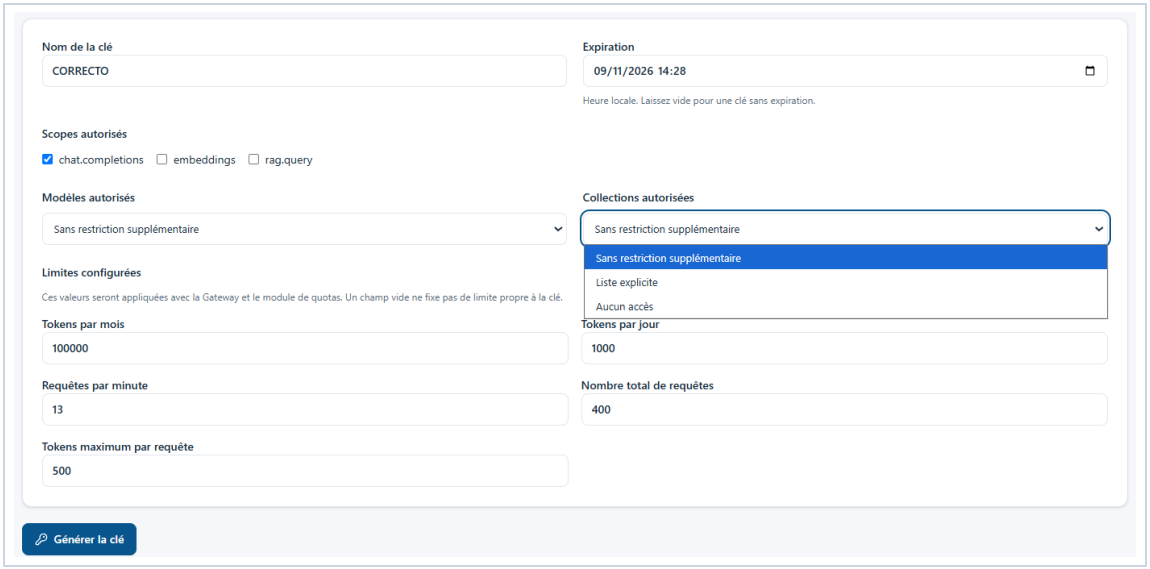


Figure 17 - *Génération d'une clé : nom, expiration, scopes, restrictions et limites*



Figure 18 - *Écran « Enregistrez votre clé API » : la valeur n'est affichée qu'une fois (secret entièrement masqué)*

> **Valeur affichée une seule fois**
> Quitter ou recharger cet écran la fait disparaître : elle ne peut pas être récupérée depuis la liste ou la fiche de la clé. En cas de perte, renouvelez la clé.

Scope	Service autorisé
`chat.completions`	Génération de texte avec un modèle de langage.
`embeddings`	Calcul de représentations numériques du texte.
`rag.query`	Recherche documentaire et réponse RAG.

Une clé NIBRAS sert aux appels clients ; une clé de fournisseur reste côté serveur et ne doit jamais être remise aux clients.

6.2 Réglage « Réflexion (thinking) »

Ce réglage permet de choisir, application par application, si les modèles à raisonnement réfléchissent avant de répondre :

Valeur	Effet
Comportement par défaut du modèle	Aucun paramètre n'est ajouté : le modèle se comporte comme son fournisseur le décide.
Désactivée	La recette « thinking désactivé » de chaque modèle utilisé par la clé est appliquée.
Activée	La recette « thinking activé » du modèle est appliquée.

La recette est définie une fois sur chaque modèle (§ 7.3). Si un modèle n'a pas de recette pour la valeur choisie, rien n'est ajouté pour ce modèle. Le réglage prend effet à la requête suivante, et l'application cliente ne peut pas le contourner : un paramètre `thinking` ou `enable_thinking` envoyé par le client est ignoré. Désactiver la réflexion réduit généralement la latence et évite que les tokens de réflexion consomment la sortie maximale.

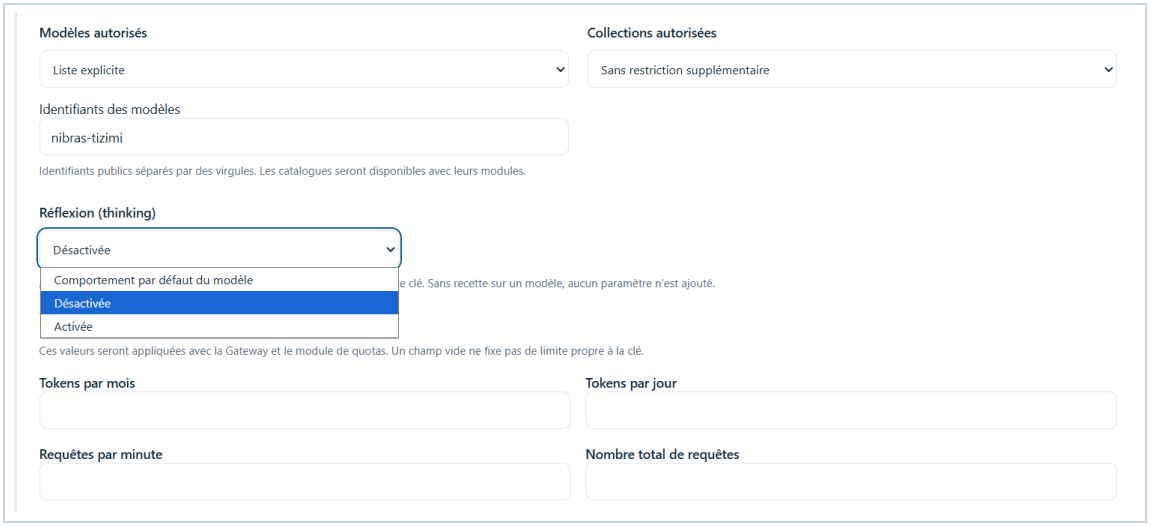


Figure 19 - *Formulaire de clé API : modèles autorisés et réglage « Réflexion (thinking) »*

6.3 Comprendre les restrictions

Pour les modèles et pour les collections, trois modes sont proposés :

Mode	Effet
Sans restriction supplémentaire	Aucun filtre supplémentaire ajouté par cette clé. Les autres contrôles restent appliqués.
Liste explicite	Accès limité aux modèles publics ou collections indiqués.
Aucun accès	Aucun accès à cette catégorie de ressources.

Une restriction de clé ne donne pas, à elle seule, accès à une collection restreinte : les autorisations de collection doivent également le permettre (§ 9.2). Les modèles s'identifient par leur nom public NIBRAS ; les collections par leur identifiant stable ou leur UUID.

6.4 Quel modèle une clé utilise-t-elle ?

Une clé n'est liée à aucun modèle précis. Le modèle appelé se décide à chaque requête :

1. L'application indique le modèle dans le champ `model` de sa requête, avec son **nom public NIBRAS** (par exemple `nibras-main`), pas le nom du fournisseur.
2. Si elle n'en indique aucun, la Gateway utilise le **modèle par défaut** de la plateforme (chapitre 18) ou, à défaut, celui marqué « par défaut » dans le catalogue.
3. La clé peut limiter les choix grâce à **Modèles autorisés** (§ 6.3). Un modèle hors de cette liste est refusé.

Le modèle et son fournisseur doivent être actifs. Une application peut lister les modèles qui lui sont ouverts avec `GET /v1/models`. Pour réserver une application à un seul modèle, indiquez uniquement ce modèle dans « Modèles autorisés » de sa clé.

6.5 Suspendre, réactiver, révoquer ou renouveler

Ouvrez la fiche de la clé depuis la liste, puis utilisez les actions du bloc **Cycle de vie**.

Action ou état	Conséquence
Suspendre	Bloque temporairement la clé.

Action ou état	Conséquence
Réactiver	Rétablit une clé suspendue, si elle n'est pas expirée.
Révoquer	Bloque définitivement cette valeur.
Renouveler	Émet une nouvelle clé et révoque l'ancienne dans la même opération.
Expirée	La date d'expiration est passée ; renouveler pour obtenir une nouvelle clé.

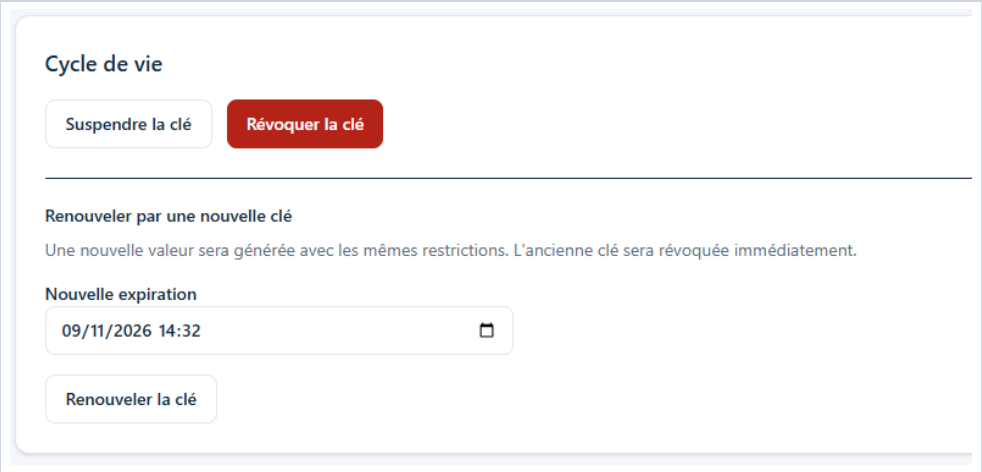


Figure 20 - *Bloc Cycle de vie : suspendre, révoquer, renouveler*

Le renouvellement coupe immédiatement l'ancienne clé

Copiez la nouvelle valeur et mettez à jour le client. Si un déploiement nécessite un chevauchement, créez une clé distincte, mettez à jour le client, puis révoquez l'ancienne.

7. Gérer les fournisseurs et les modèles

7.1 Fournisseurs

Dans NIBRAS, un **fournisseur** désigne le **serveur où est hébergé un modèle de langage**. Il est identifié par son **adresse** : une **URL** ou l'**adresse IP** et le port d'un serveur interne, comme le serveur de l'**ISTA Bab Tizimi** qui héberge le modèle **Qwen**, déclaré sous le nom **TIZIMI**. Un même fournisseur peut servir plusieurs modèles ; chaque modèle déclaré dans NIBRAS est rattaché à un fournisseur (§ 7.2). L'adresse, la clé d'accès et le type d'API sont enregistrés côté serveur ; l'interface sert à déclarer et activer le fournisseur.

Pour ajouter un fournisseur :

1. Demander au responsable technique de déclarer la connexion dans `LLM_PROVIDERS` (fichier `.env` du serveur : URL, clé API, type) et de redémarrer l'API.
2. Ouvrir **Intelligence artificielle → Fournisseurs**, puis utiliser l'action d'ajout.
3. Renseigner le nom du fournisseur, choisir la **connexion serveur** dans la liste et son statut.
4. Enregistrer, puis utiliser **Vérifier la connexion** pour contrôler la disponibilité du service.

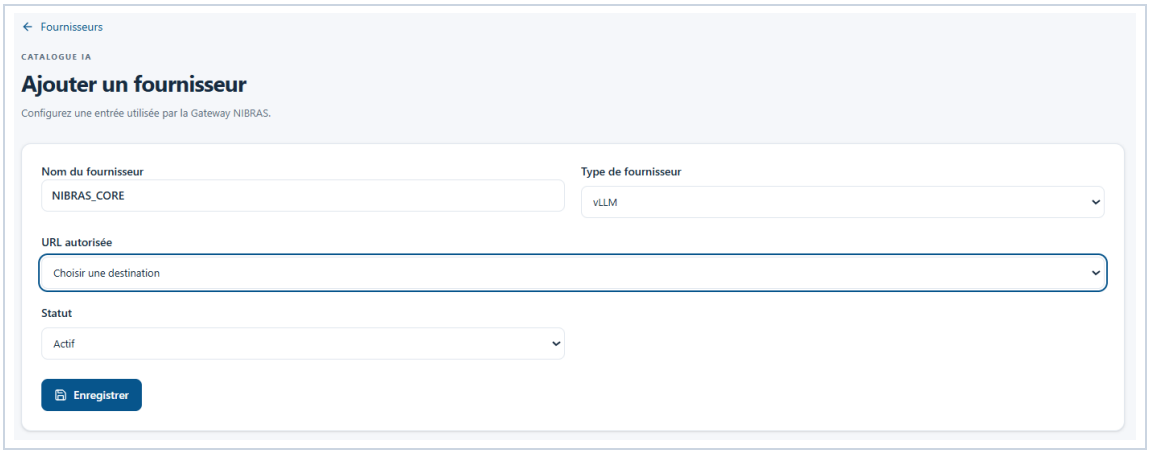


Figure 21 - *Ajout d'un fournisseur : nom, connexion serveur et statut*

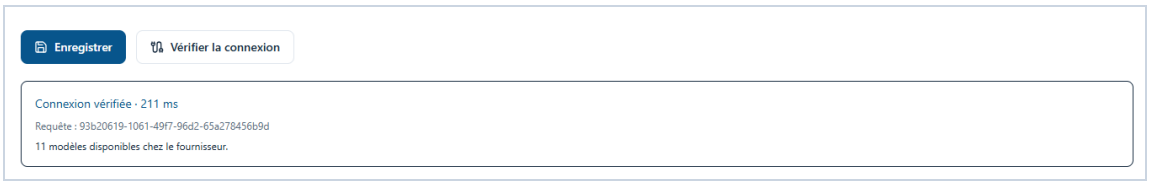


Figure 22 - *Résultat du test de connexion : latence et nombre de modèles disponibles*

L'écran affiche le type, l'URL et le mode de comptage des tokens de la connexion choisie. Il ne permet **jamais** de saisir ou de consulter une clé secrète. La connexion d'un fournisseur ne change plus après sa création ; si elle disparaît du fichier `.env`, le fournisseur est signalé « Connexion absente de .env » et les appels échouent de façon contrôlée.

Chaque connexion a un type, déclaré par le responsable technique :

Type	Usage
`openai`	Toute API compatible OpenAI
`vllm`	Serveur vLLM hébergé par vos soins, joint par son adresse IP ou son URL (par exemple le serveur Qwen de l'ISTA Bab Tizimi).

Le résultat de **Vérifier la connexion** indique la durée de réponse, le nombre de modèles que le fournisseur annonce et leur liste. Un test réussi prouve que l'adresse et la clé sont valides ; il ne garantit pas que le fournisseur accepte tous les paramètres d'une génération (chapitre 8).

7.2 Modèles

1. Ouvrir **Intelligence artificielle → Modèles**.
2. Ajouter un modèle ou ouvrir sa fiche.
3. Choisir son nom public, par exemple `nibras-main`, et son fournisseur.
4. Renseigner le nom exact du modèle servi (tel que l'annonce le fournisseur : respecter la casse, les tirets et les barres obliques), sa fenêtre de contexte et sa sortie maximale.
5. Choisir son statut et, si nécessaire, le définir comme modèle par défaut.
6. Si le modèle lit les images, cocher **Accepte les images (vision / OCR)** et choisir le **Format de réponse** (§ 7.4).
7. Enregistrer, puis utiliser **Vérifier la connexion** sur la fiche du modèle. Si le nom du modèle servi est absent du catalogue du fournisseur, l'écran affiche « Connexion non disponible » avec le code `upstream_model_missing` et **la liste des noms disponibles** : copiez le bon nom dans « Modèle servi », enregistrez et relancez la vérification.
8. Effectuer un test de génération (chapitre 8).

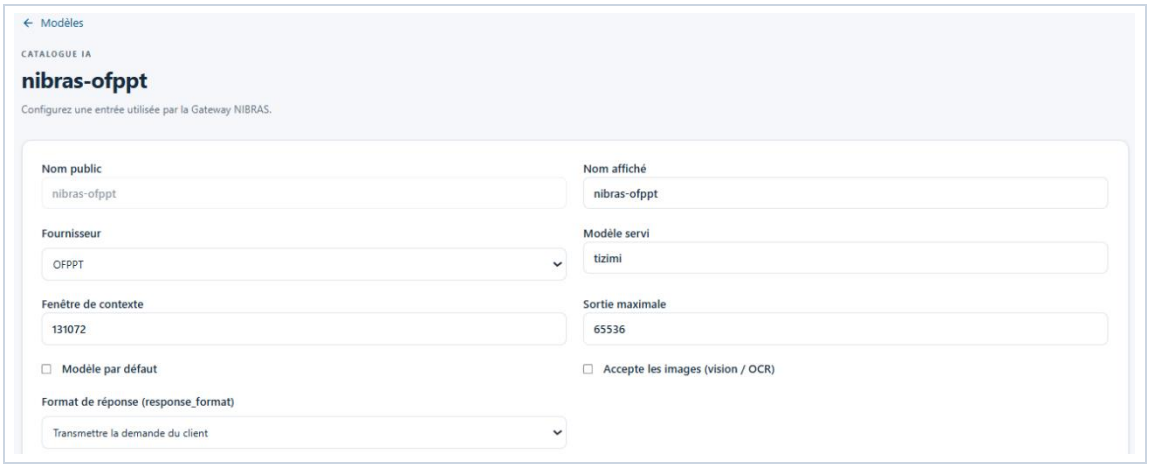


Figure 23 - *Fiche d'un modèle : nom public, fournisseur, modèle servi, contexte et sortie maximale*

Le nom public reste **fixe après création**. Le client utilise ce nom public ; le « modèle servi » désigne le modèle réellement exécuté par le fournisseur. Cette séparation permet de changer de modèle ou de serveur sans modifier les applications clientes. Un modèle et son fournisseur doivent être **actifs** pour permettre la génération.

Si un client demande une sortie supérieure à la **sortie maximale** du modèle (par exemple `max_tokens` = 32768 pour un plafond de 16384), la demande est **ramenée au plafond** au lieu d'être refusée. Elle est aussi ramenée à la place restante dans la fenêtre de contexte.

7.3 Paramètres de génération et recettes de thinking

La fiche d'un modèle contient trois objets JSON, tous facultatifs :

Champ	Rôle
Paramètres par défaut	Ajoutés à la requête si le client ne les fournit pas (par exemple `{"temperature": 0.3}`).
Recette thinking désactivé	Appliquée aux clés dont la réflexion est « Désactivée ».

Champ	Rôle
Recette thinking activé	Appliquée aux clés dont la réflexion est « Activée ».

Les boutons de recettes remplissent les deux recettes pour les cas usuels.

> **Couper la réflexion partout**
> Pour couper la réflexion partout, y compris dans le test LLM qui n'utilise pas de clé, placez la recette « désactivé » dans les **Paramètres par défaut** du modèle ; une clé réglée sur « Activée » la rallume.

> **Vérifier la clé exacte**
> La clé exacte dépend du gabarit de chat du modèle servi : vérifiez-la dans la documentation du modèle ou du fournisseur. Seuls les paramètres de génération connus sont acceptés ; les champs gérés par la Gateway (`model`, `messages`, `stream`, `max_tokens`, `tools`) sont refusés.

Si un modèle existe en deux variantes distinctes (par exemple « instruct » et « thinking »), créez deux modèles publics et limitez les clés avec « Modèles autorisés ». Dans la configuration utilisée pour les validations, l'interface se limite à choisir une connexion et à définir les modèles.

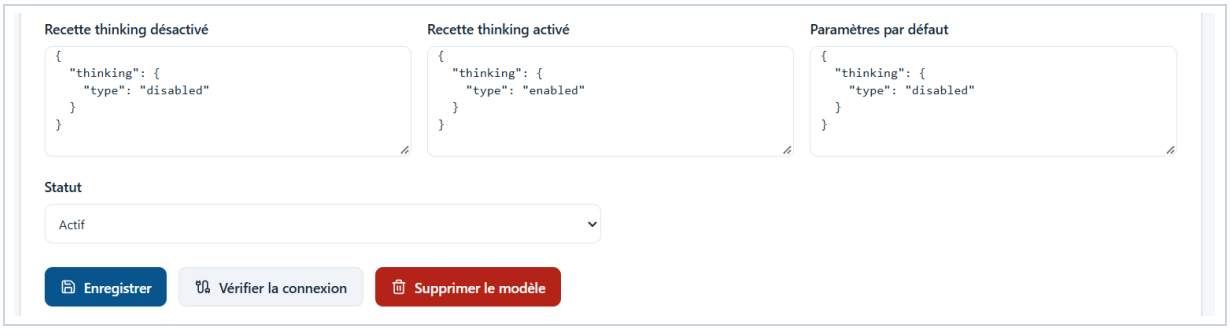


Figure 24 - *Fiche d'un modèle : recettes de thinking, paramètres par défaut et boutons Enregistrer, Vérifier la connexion et Supprimer le modèle*

7.4 Vision et format de réponse

Un modèle multimodal, peut lire des images : OCR de copies scannées, analyse d'une page, etc. Deux réglages de la fiche du modèle le pilotent :

Réglage	Valeur initiale	Effet
Accepte les images (vision / OCR)	Décoché	**Coché** : le modèle reçoit les images envoyées par les clients. **Décoché** : une requête contenant une image est refusée avec le code `model_vision_unsupported` (« Le modèle X n'accepte pas les images »), **sans appel au fournisseur**.
Format de réponse (`response_format`)	Transmettre la demande du client	**Transmettre** : la demande de réponse JSON du client est envoyée telle quelle. **Retirer** : à choisir si le fournisseur refuse ce paramètre. **Convertir en json_object** : toute demande de JSON structuré est envoyée sous la forme `{"type": "json_object"}` ; une requête sans demande de format reste sans.

La liste des modèles affiche la mention **Vision** à côté du nom des modèles concernés.

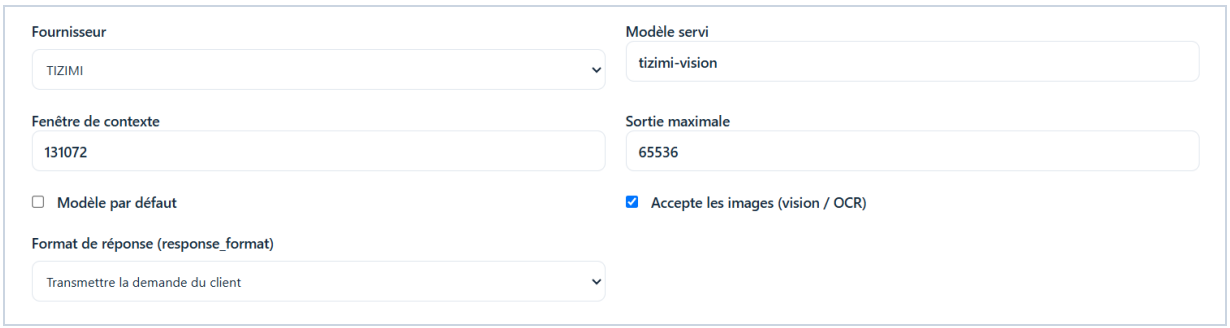


Figure 25 - *Fiche d'un modèle vision : fournisseur TIZIMI, modèle servi, case « Accepte les images (vision / OCR) » et format de réponse*

Exemple : créer nibras-tizimi-vision pour l'OCR de Correcto

1. Ajouter un modèle nommé `nibras-tizimi-vision`, rattaché au fournisseur local.
2. Modèle servi : le nom exact affiché par **Vérifier la connexion**.
3. Fenêtre de contexte : `262144`. Sortie maximale : au moins `8192` (par exemple `32768`).
4. Cocher **Accepte les images** et laisser le format de réponse sur **Transmettre**.
5. Paramètres par défaut : `{« thinking » : {« type » : « disabled »}}`. La réflexion est ainsi coupée quelle que soit la clé. Une température basse (par exemple `« temperature » : 0.2`) rend la lecture plus régulière, mais seulement si le fournisseur l'accepte.
6. Ajouter le modèle aux **Modèles autorisés** de la clé de Correcto.

Ce que la Gateway accepte

- Images au format **data URL base64** uniquement (`data :image/jpeg ; base64,…`) : JPEG, PNG ou WebP.
- Au plus **10 images** par requête, et **5 Mo** par image une fois décodée.
- Requête complète jusqu'à **32 Mo**, soit environ six pages scannées en une seule requête.
- Les adresses d'images `http(s)` sont refusées par défaut, pour la sécurité et la confidentialité.

Ces limites se règlent dans le fichier `.env` du serveur, par le responsable technique (voir le Guide d'administration du VPS). Une image ne compte pas comme du texte dans les quotas : avant l'envoi, la Gateway réserve une estimation haute calculée d'après ses dimensions (environ 4 000 à 5 000 tokens pour une page A4).

> **Confidentialité des images**
> Les images peuvent être des copies d'étudiants. Elles ne sont **jamais conservées** : ni sur disque, ni en base, ni dans les traces. L'option « stockage du contenu de chat » ne s'applique pas aux requêtes avec image. Les journaux techniques n'indiquent que le modèle, le nombre et la taille des images, la durée et les tokens.

7.5 Supprimer un fournisseur ou un modèle

Sur la fiche d'un fournisseur ou d'un modèle, le bouton rouge **Supprimer le fournisseur** ou **Supprimer le modèle** ouvre une confirmation. L'action est définitive.

Élément	Règles
Modèle	Refusé s'il est le modèle par défaut : désignez d'abord un autre modèle par défaut. Son nom est retiré des listes « Modèles autorisés » des clés qui le citaient (une clé dont la liste devient vide n'a plus accès à aucun modèle).

Élément	Règles
Fournisseur	Refusé tant qu'il possède des modèles : supprimez-les d'abord. La connexion du fichier `.env` n'est pas modifiée.
Élément défini dans `.env`	Pas de bouton de suppression : il serait recréé au prochain démarrage. Désactivez-le à la place.

L'historique de consommation est conservé : les requêtes passées gardent le nom du modèle, le type de fournisseur et les limites de l'époque. Chaque suppression est inscrite au journal d'audit. Pour mettre un modèle hors service sans le supprimer, passez son statut à « Désactivé ».



Figure 26 - *Confirmation de suppression d'un modèle*

8. Tester un modèle de langage

1. Ouvrir **Intelligence artificielle → Tests** pour accéder à **Test LLM**.
2. Sélectionner un modèle public actif.
3. Saisir un prompt, par exemple : « Propose trois conseils pour préparer une séance de formation. »
4. Ajuster la **sortie maximale** et, si nécessaire, la **température** : une température basse favorise des réponses plus régulières ; une valeur plus élevée augmente leur variation. **Le champ est vide par défaut** : la température n'est alors pas envoyée et le modèle utilise sa valeur habituelle.
5. Cliquer sur **Lancer le test** et consulter la réponse, les informations de consommation et l'identifiant de requête.

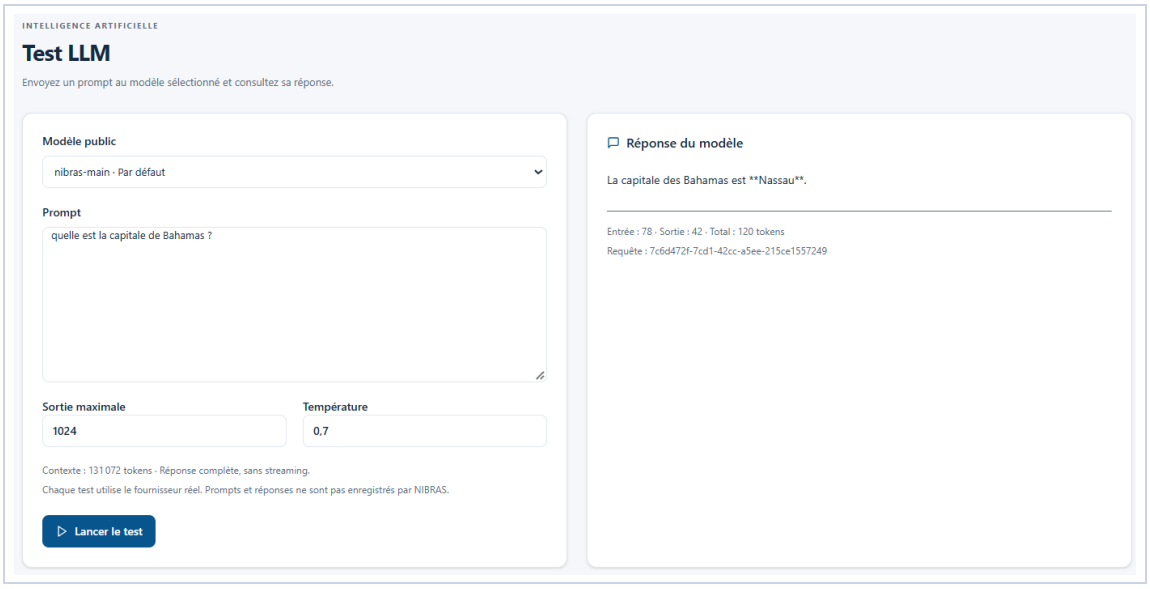


Figure 27 - *Test LLM : prompt, paramètres et réponse du modèle avec sa consommation*

Le test LLM répond au prompt saisi, sans consulter la base documentaire : utilisez **Test RAG** (chapitre 12) pour interroger les documents avec leurs sources.

9. Créer et autoriser les collections

9.1 Créer une collection

1. Ouvrir **Base de connaissances → Collections**.
2. Cliquer sur **Créer une collection**.
3. Renseigner le nom, l'identifiant stable et la description.
4. Choisir la visibilité et le statut.
5. Vérifier le modèle d'embeddings, la taille des passages et leur chevauchement.
6. Enregistrer.

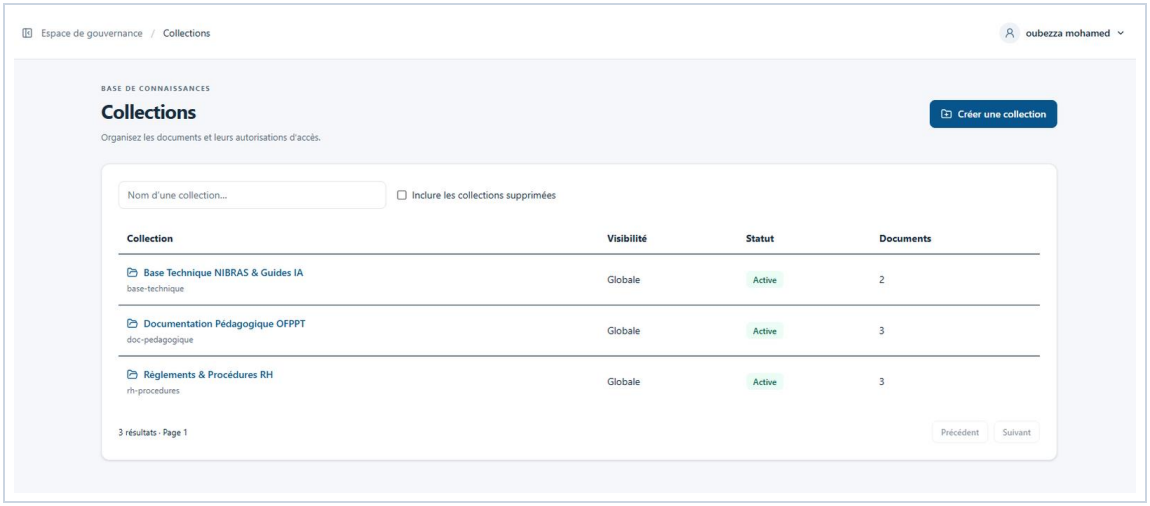


Figure 28 - *Collections : visibilité, statut et nombre de documents*

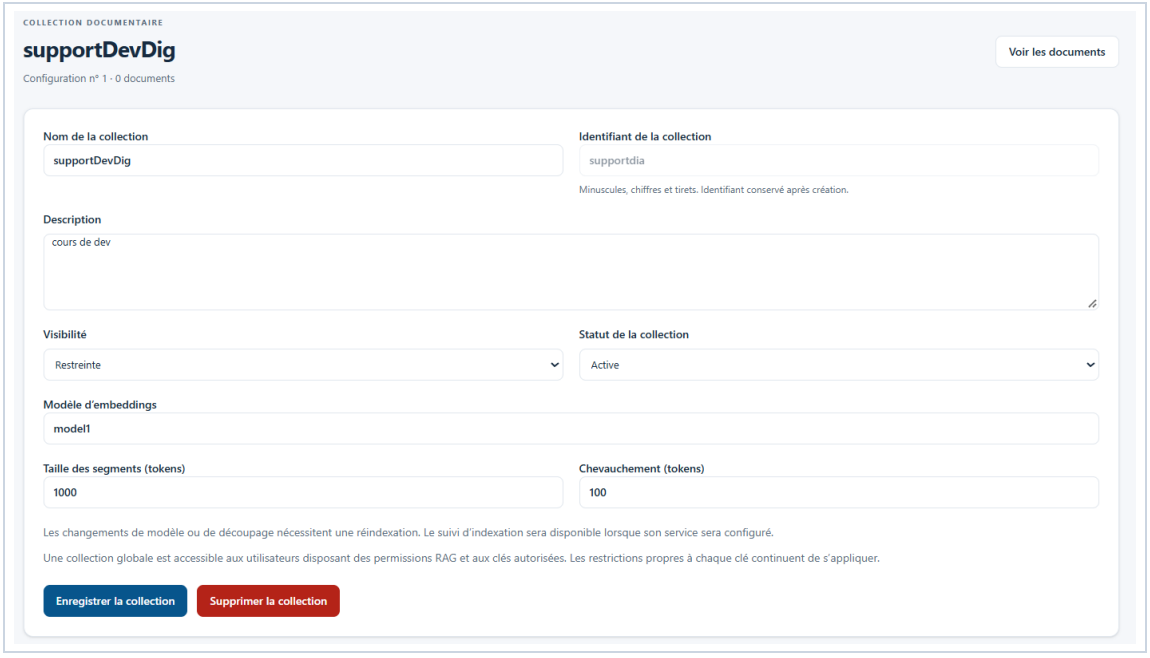


Figure 29 - *Fiche de collection : visibilité, statut et paramètres de découpage*

Visibilité	Accès
Restreinte	Accès limité aux acteurs autorisés ; le créateur administratif dispose de l'accès à sa collection.
Globale	Accessible aux acteurs disposant des permissions ou scopes nécessaires, sous réserve des restrictions de leur clé.

Une collection globale ne constitue pas une publication anonyme sur Internet. Une collection désactivée ne peut pas être utilisée par les clients.

9.2 Accorder les accès

Dans la fiche, ouvrez **Autorisations de la collection**. Sélectionnez les **applications**, **utilisateurs** ou **clés API** autorisés, puis cliquez sur **Enregistrer les autorisations**. Cette modification exige la permission `rag.access.manage`.

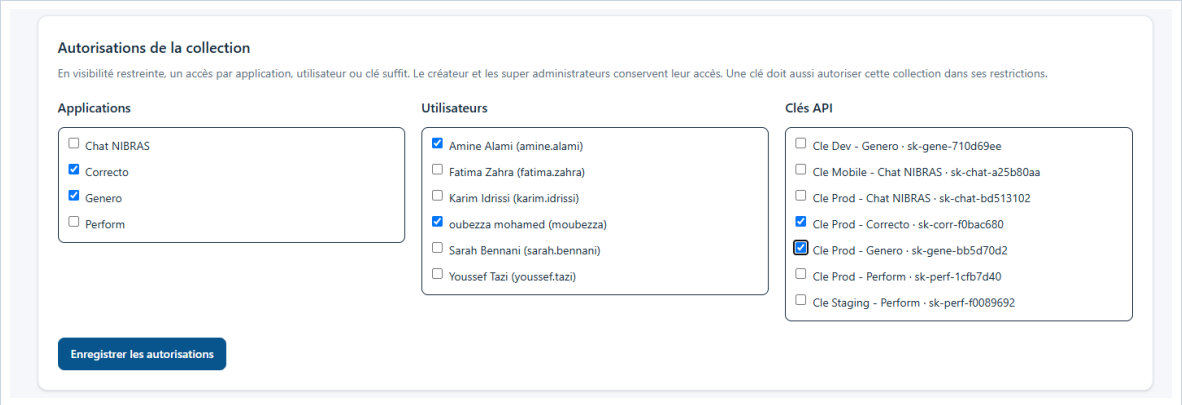


Figure 30 - *Autorisations de la collection : applications, utilisateurs et clés API*

Pour un accès client à une collection restreinte, l'autorisation de son application, de son utilisateur associé ou de sa clé peut permettre l'accès ; les restrictions propres à la clé s'appliquent ensuite. Pour un utilisateur administratif, l'accès relève du créateur ou d'une autorisation utilisateur : SUPER_ADMIN dispose d'un accès aux collections, tandis qu'ADMIN respecte leurs restrictions.

> **Modifier le découpage impose une réindexation**
>
> Modifier le découpage d'une collection existante marque ses documents pour réindexation. Effectuez cette réindexation avant de reprendre les recherches : les documents dont la configuration est devenue périmée sont exclus des résultats.

10. Importer et gérer les documents

10.1 Importer un lot

1. Ouvrir **Base de connaissances → Documents**, puis **Importer des documents**.
2. Choisir une collection active à laquelle vous avez accès.
3. Sélectionner ou glisser les fichiers dans la zone prévue.
4. Si utile, renseigner les métadonnées communes au format JSON.
5. Valider l'import et consulter le résultat affiché pour chaque fichier.
6. Ouvrir les documents importés pour lancer et suivre leur indexation (chapitre 11).

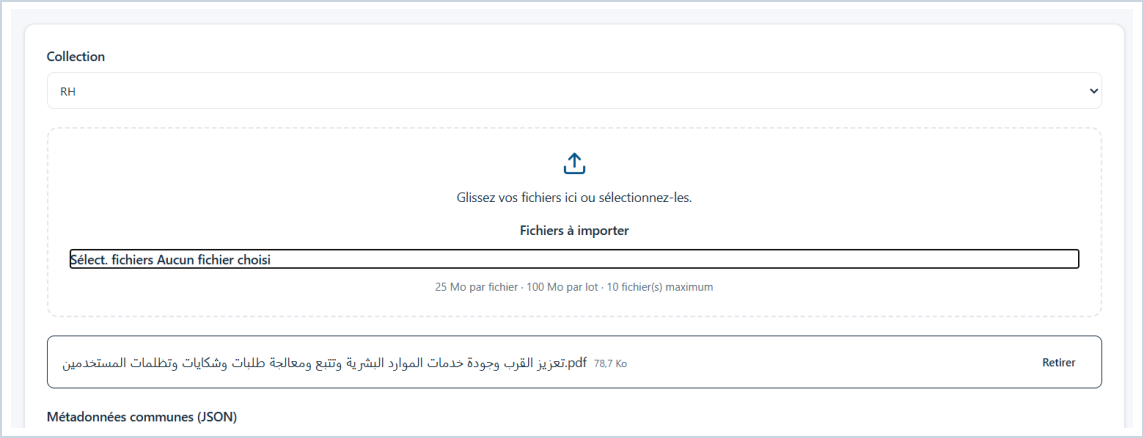


Figure 31 - *Import : choix de la collection et dépôt du fichier*

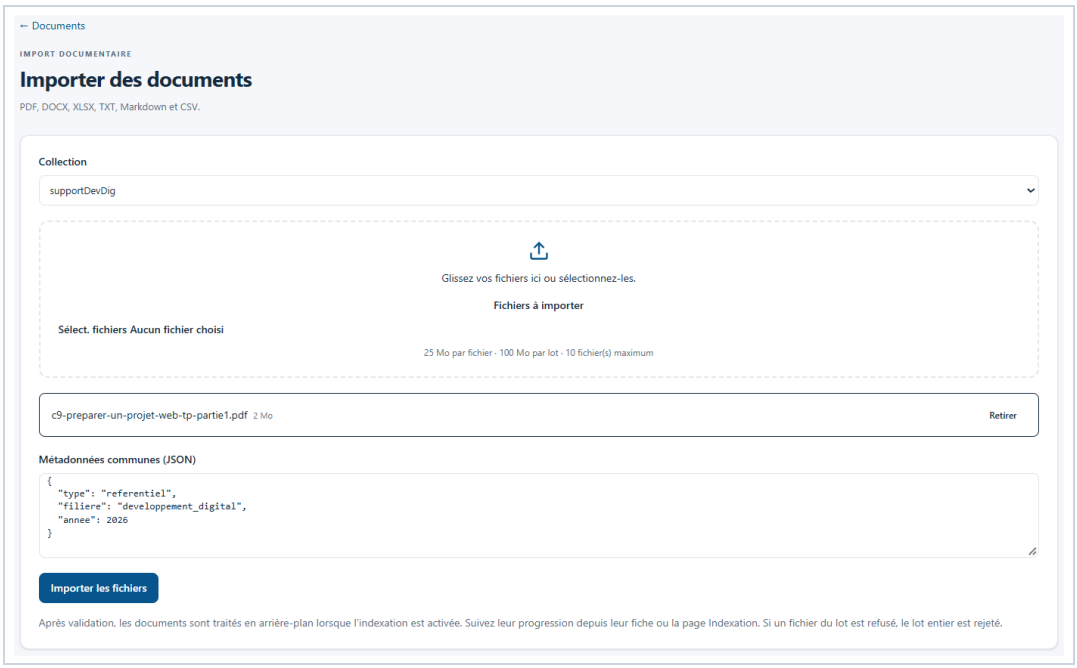


Figure 32 - *Import : métadonnées communes (JSON) et bouton « Importer les fichiers »*



Figure 33 - *Résultat de l'import : document en cours de traitement*

Formats acceptés : PDF, DOCX, XLSX, CSV, TXT et Markdown (`.md`). Les limites initiales sont de **25 Mo par fichier**, **100 Mo par lot** et **10 fichiers par lot** ; le responsable technique peut les adapter. Les fichiers sont contrôlés sur leur contenu, pas uniquement sur leur extension.

Exemple de métadonnées permettant ensuite un filtrage :

Métadonnées communes (JSON)

```
{
  "type": "referentiel",
  "filiere": "developpement_digital",
  "annee": 2026
}
```

> **Importer n'est pas indexer**
>
> Un document téléversé n'est pas encore disponible dans la recherche : l'import et l'indexation sont deux étapes distinctes. Un contenu déjà présent dans la même collection peut être refusé comme doublon. Un PDF composé uniquement d'images nécessite une extraction OCR préalable : l'application ne réalise pas cette reconnaissance.

10.2 Consulter la fiche et les versions

Depuis la liste, recherchez par titre et filtrez par collection ou statut, puis ouvrez un document. La fiche permet, selon vos droits, de **télécharger** le fichier, **modifier** son titre et ses métadonnées, consulter ses **versions** et ses **passages**.

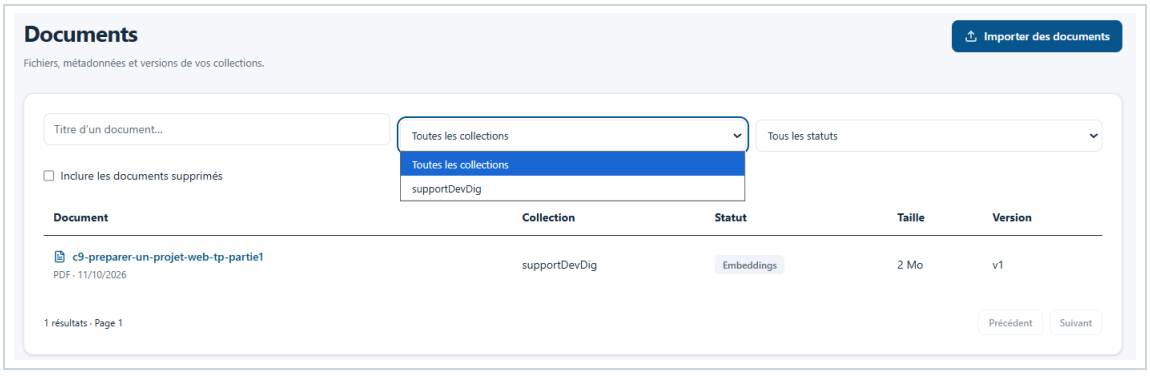


Figure 34 - *Liste des documents, avec filtres par collection et statut*

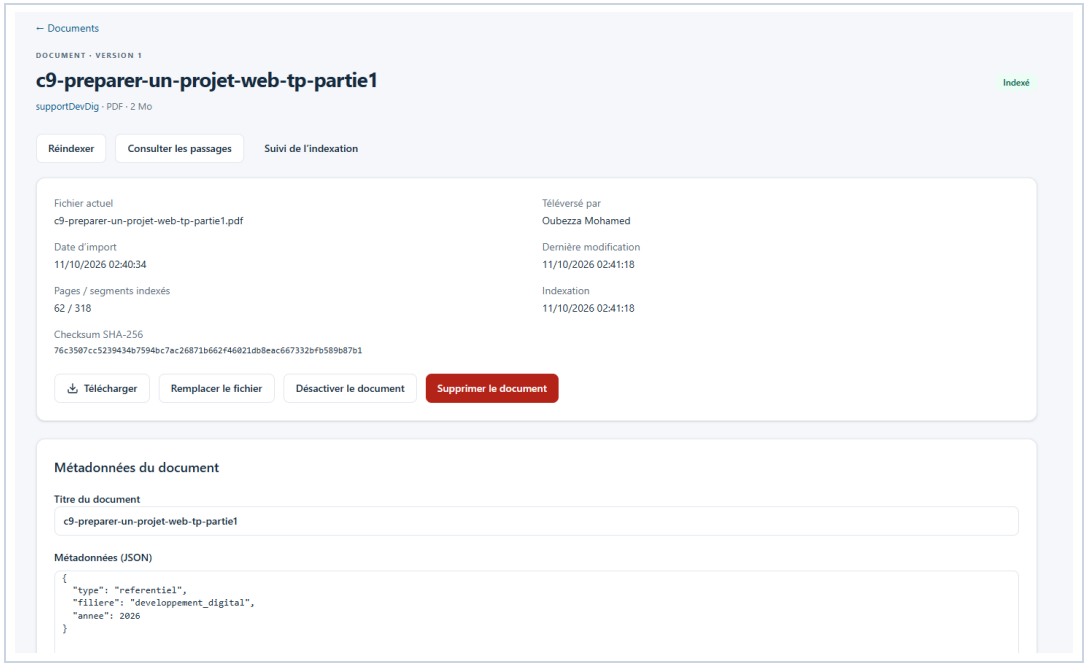


Figure 35 - *Fiche d'un document indexé : informations, actions et métadonnées*

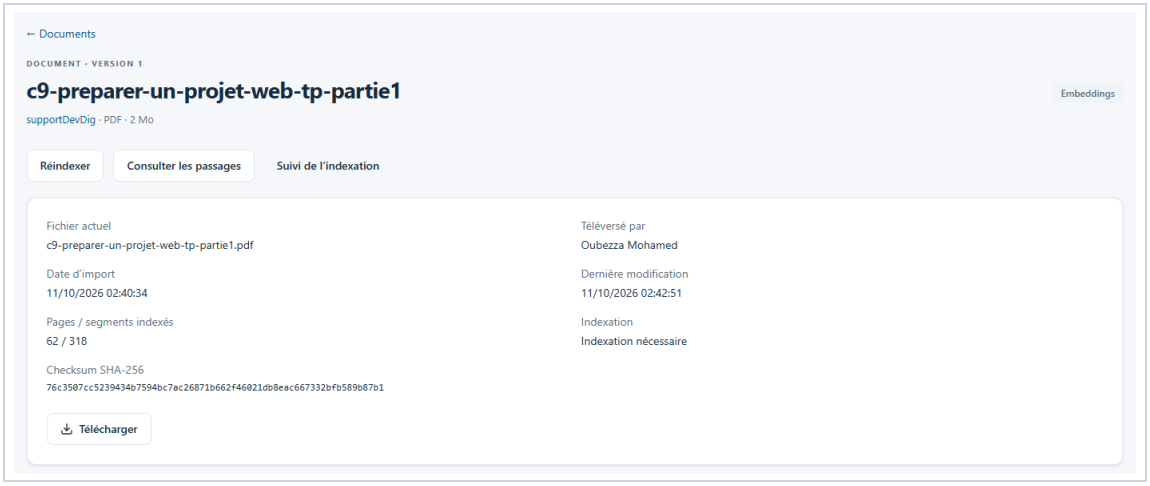


Figure 36 - *Fiche d'un document dont l'indexation est nécessaire*

Pour ajouter une version, cliquez sur **Remplacer le fichier** et importez le nouveau fichier : le titre et les métadonnées sont conservés. Indexez ensuite la nouvelle version. Lorsque la configuration reste compatible, les passages de l'ancienne version publiée restent disponibles jusqu'à la réussite de la nouvelle indexation.

10.3 Désactivation, suppression et restauration

Action	Effet et récupération
Désactiver le document	Exclut le document des accès clients. Une réactivation est possible.
Supprimer le document	Suppression logique : retrait immédiat des accès clients, avec conservation des fichiers et versions.
Restaurer le document	Rétablit un document supprimé logiquement, si les droits et l'état de la collection le permettent. Vérifier ensuite son indexation.
Purger définitivement	Supprime les fichiers, versions et vecteurs d'un document supprimé logiquement. La restauration devient impossible dès la mise en attente.

La purge demande une confirmation explicite et les permissions correspondantes : lisez le récapitulatif avant de confirmer.



Figure 37 - *Confirmation de purge définitive d'un document*

11. Indexer les documents

L'indexation **extrait le texte**, le **découpe en passages**, calcule leurs **embeddings** et **publie** les résultats pour la recherche. Elle est exécutée en arrière-plan.

11.1 Lancer l'indexation

1. Ouvrir la fiche du document ou de la collection.
2. Cliquer sur **Réindexer** et confirmer l'opération. Le même bouton sert à préparer un document nouvellement importé.
3. Ouvrir **Suivi de l'indexation** ou **Base de connaissances → Indexation**.
4. Consulter les prérequis, la progression et le résultat de la tâche.

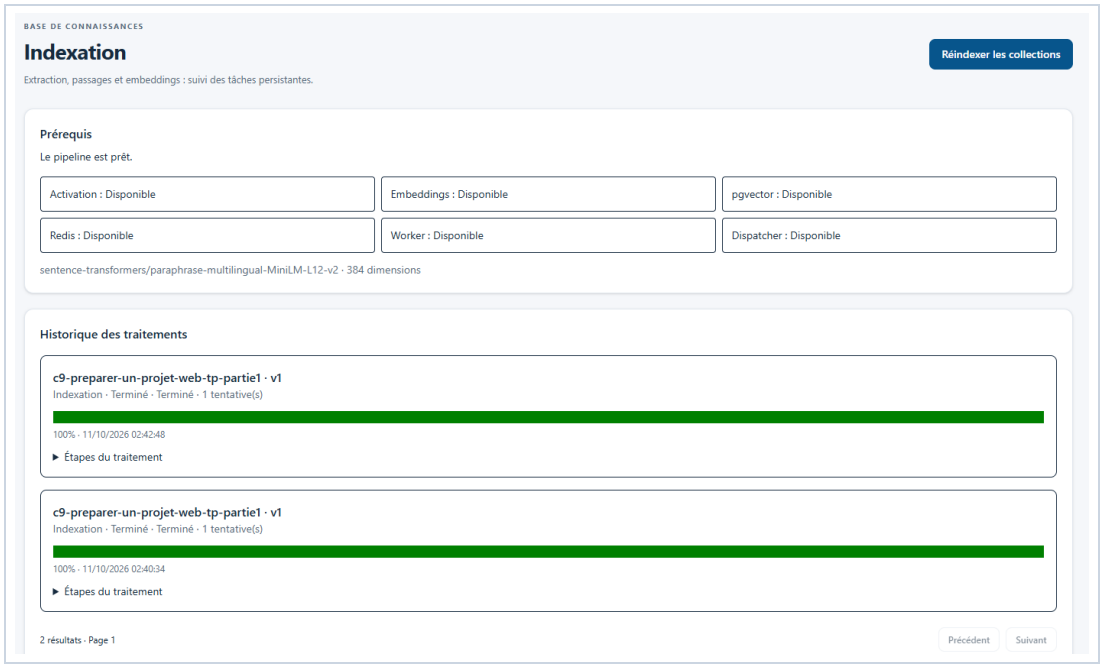


Figure 38 - *Indexation : prérequis du pipeline et historique des traitements*

L'écran propose également **Réindexer les collections** pour les collections accessibles : cette action peut traiter de nombreux documents, vérifiez le périmètre annoncé avant confirmation.

11.2 Lire l'état d'une tâche

Chaque traitement affiche une barre de progression et ses étapes (extraction, découpage, embeddings, publication) :



Figure 39 - *Traitement en cours (30 %, à gauche) puis terminé (100 %, à droite)*

État technique	Lecture pratique
`QUEUED`	Tâche en attente de prise en charge.
`RUNNING`	Traitement en cours.

État technique	Lecture pratique
`RETRY`	Nouvelle tentative prévue après une difficulté temporaire.
`SUCCEEDED`	Traitement terminé avec succès.
`FAILED`	Échec ; consulter l'étape et le message associés.
`CANCELLED`	Tâche annulée ; vérifier la situation du document avant de relancer.

Si les prérequis ne sont pas disponibles ou si la file n'avance pas, consultez **Monitoring** (chapitre 17). Après un échec, corrigez la cause avant de relancer ; certains incidents temporaires donnent lieu à des tentatives automatiques.

11.3 Vérifier les passages publiés

Dans la fiche d'un document, cliquez sur **Consulter les passages**. Recherchez du texte et examinez les fragments, leurs références disponibles et la version associée. Cette vérification permet de repérer une extraction vide, un contenu incomplet ou un découpage inadapté.

12. Rechercher et générer une réponse RAG

12.1 Rechercher les passages

1. Ouvrir **Base de connaissances → Test RAG**.
2. Saisir une question précise liée aux documents importés.
3. Sélectionner éventuellement des collections ; sans sélection, la recherche porte sur les collections accessibles.
4. Ajuster **Passages maximum** et **Score minimum**.
5. Ajouter éventuellement des **Filtres metadata (JSON)** correspondant aux métadonnées enregistrées.
6. Cliquer sur **Rechercher les passages**.

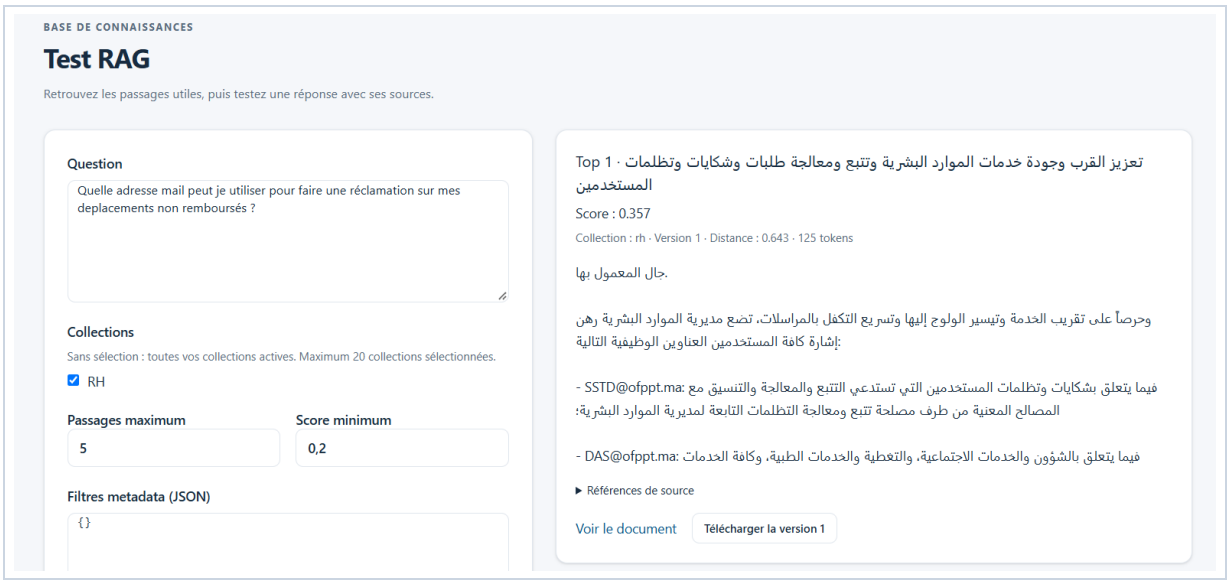


Figure 40 - *Test RAG : question, collection, paramètres et premier passage retrouvé avec son score*

Exemple de filtre metadata

```
{"type": "referentiel", "annee": 2026}
```

Les propriétés du filtre se cumulent : un document doit correspondre à **toutes** les valeurs demandées. Le nombre de passages est compris entre 1 et 20 ; le score minimum entre -1 et 1.



Figure 41 - *Test RAG : liste des passages et score de similarité*

> **Lire le score**
>
> Le score indique une similarité entre la question et le passage ; il ne mesure pas la probabilité que la réponse soit correcte. Examinez le texte retrouvé et les références. Si aucun passage n'est pertinent, reformulez la question, vérifiez l'indexation, les filtres et les accès, ou ajustez progressivement le seuil.

12.2 Générer une réponse sourcée

1. Vérifier les résultats de recherche.
2. Choisir le **Modèle de réponse** et la **Sortie maximale**.
3. Cliquer sur **Générer une réponse**.
4. Lire la réponse et contrôler les références `[S1]`, `[S2]`… dans les passages et documents associés.
5. Télécharger la version source, si nécessaire et si vos droits le permettent.

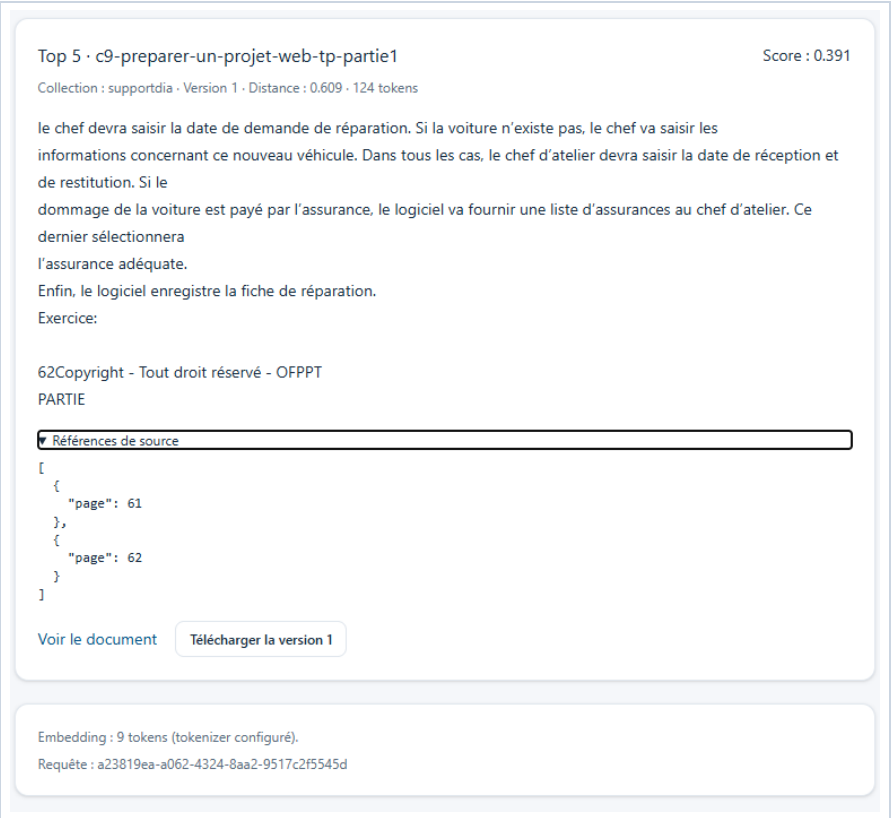


Figure 42 - *Passage retrouvé : références de source, lien vers le document et consommation de la requête*

La recherche demande `rag.read`. La génération administrative demande également `models.manage` : un gestionnaire RAG ou un auditeur peut donc accéder à la recherche sans disposer du bouton de génération.

Si aucun contexte documentaire n'est retrouvé, l'application n'effectue pas de génération et signale l'absence de passages. Si les sources ou droits changent pendant le traitement, la réponse peut être refusée : relancez après vérification. Une recherche consomme l'embedding de la question ; une réponse RAG consomme en plus une génération : ces deux opérations participent aux quotas et limites de fréquence.

> **Contrôler les sources**
>
> Les citations facilitent la vérification, elles ne la remplacent pas : contrôlez les sources avant d'utiliser la réponse comme référence métier.

13. Consulter la consommation

1. Ouvrir **Analytics → Consommation**.
2. Choisir les dates et les filtres utiles : application, utilisateur, clé, modèle, résultat, statut HTTP ou identifiant de requête.
3. Examiner les totaux et les tendances par jour, semaine ou mois.
4. Consulter les regroupements par utilisateur, application ou modèle depuis les entrées Analytics correspondantes.
5. Si autorisé, exporter le résultat avec l'action **CSV**.

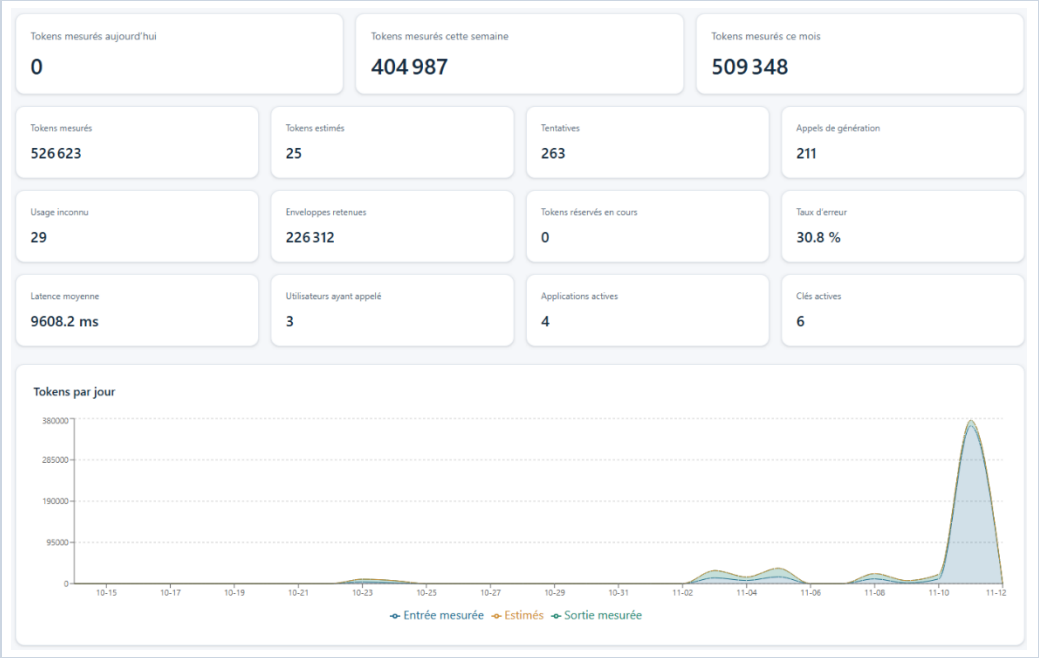


Figure 43 - *Totaux et indicateurs de consommation*

Le taux d'erreur est la part des appels terminés en erreur sur la période affichée. La valeur visible dans la figure 43 (30,8 %) a été relevée pendant une période d'essais du prototype : elle ne constitue pas une mesure de fiabilité en production.

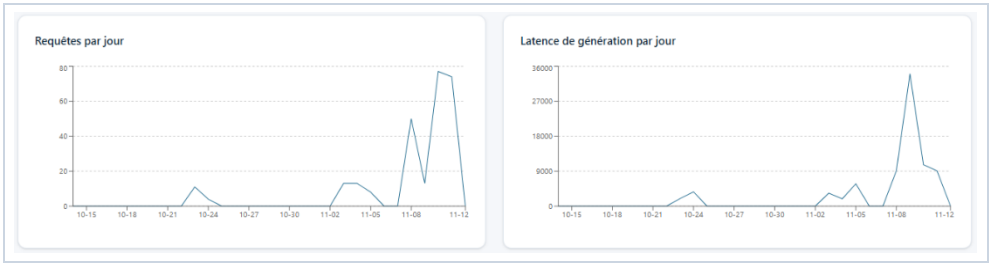
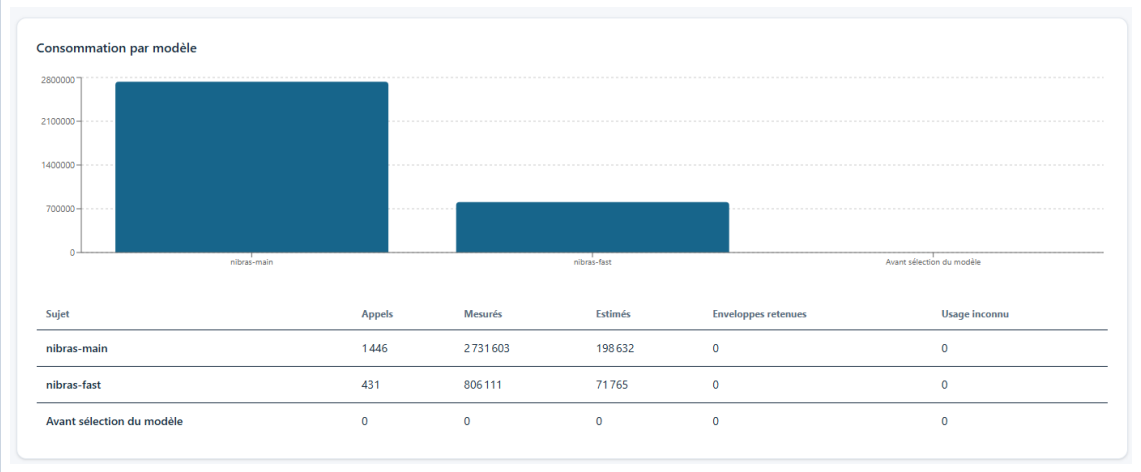


Figure 44 - *Requêtes et latence par jour, répartition par application*



Sujet	Appels	Mesurés	Estimés	Enveloppes retenues	Usage inconnu
nibras-main	1 446	2 731 603	198 632	0	0
nibras-fast	431	806 111	71 765	0	0
Avant sélection du modèle	0	0	0	0	0

Figure 45 - *Consommation par modèle : appels, tokens mesurés, estimés et enveloppes retenues*

Colonne	Signification
Appels	Nombre d'opérations enregistrées pour le sujet (modèle, application ou utilisateur).
Mesurés	Tokens comptés par le fournisseur lui-même.
Estimés	Tokens estimés par NIBRAS lorsque le fournisseur ne les communique pas.
Enveloppes retenues	Enveloppes de tokens retenues lorsque l'usage exact manque.
Usage inconnu	Opérations dont la consommation n'a pas pu être déterminée.

Les périodes de suivi utilisent **UTC** ; les semaines commencent le lundi et les mois au premier jour. L'export reprend les filtres sélectionnés et **exclut les prompts et réponses**. Les tokens imputés aux quotas peuvent donc différer des seuls tokens mesurés. Les graphiques portent sur les *opérations* : une réponse RAG peut produire une opération d'embedding et une opération de génération, alors que le compteur de requêtes parentes ne les compte pas comme deux demandes indépendantes. Les valeurs représentent des tokens, sans calcul implicite de prix en devise.

14. Définir et suivre les quotas

1. Ouvrir **Analytics → Quotas**.
2. Choisir le niveau : **Global**, **Application**, **Clé API** ou **Utilisateur**.
3. Pour un niveau individuel, sélectionner le sujet concerné.
4. Lire les budgets disponibles et la consommation courante (encadré de droite).
5. Si autorisé, saisir les limites puis cliquer sur **Enregistrer les quotas**.

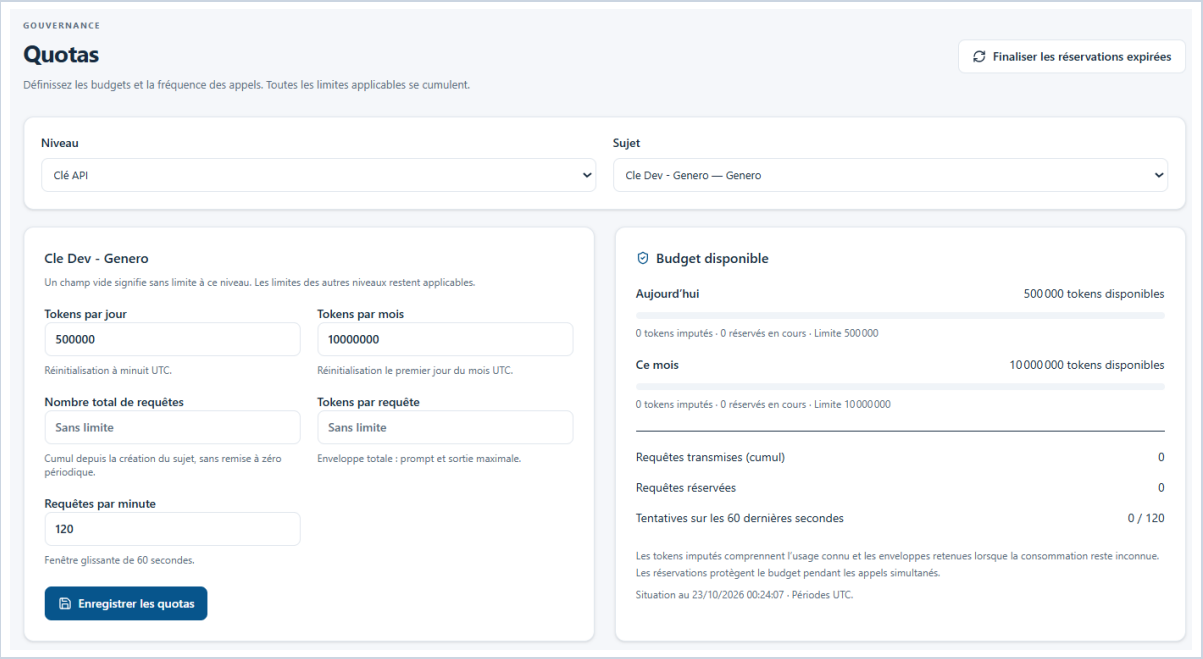


Figure 46 - *Quotas d'une clé API : limites configurées et budget disponible*

Limite	Fonction
Tokens quotidiens	Budget de tokens pour la journée UTC.
Tokens mensuels	Budget de tokens pour le mois UTC.
Requêtes par minute	Limite de fréquence sur une fenêtre glissante de 60 secondes.
Nombre total de requêtes	Plafond cumulé, indépendant d'un changement de jour ou de mois.
Tokens maximum par requête	Plafond de l'enveloppe d'appel : tokens d'entrée et sortie maximale demandée.

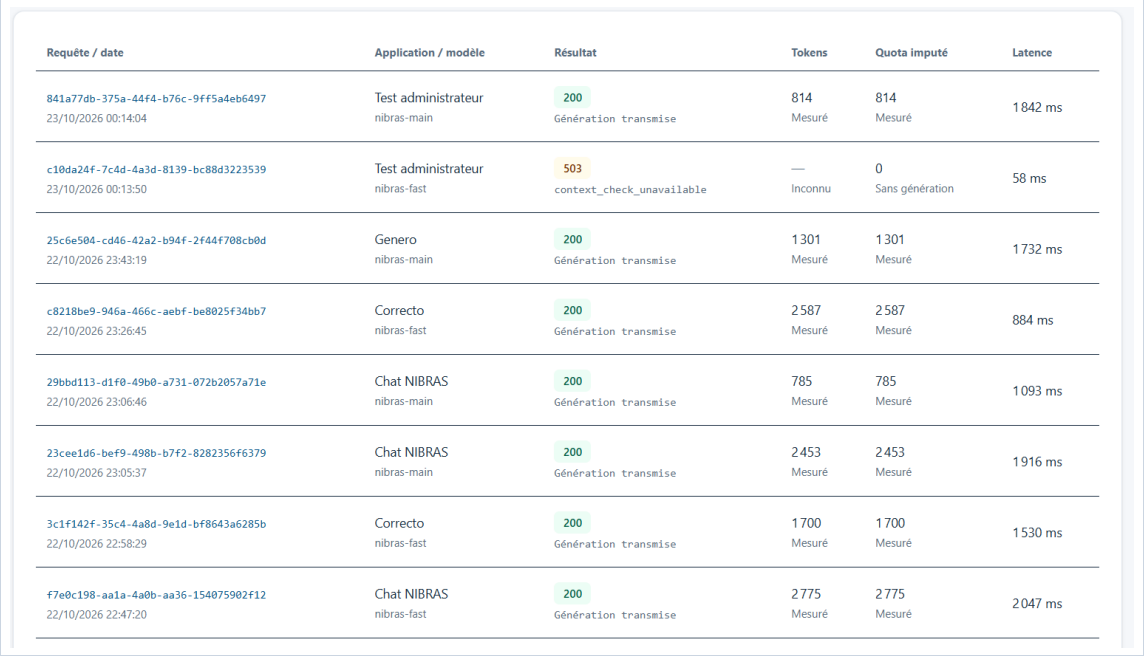
Un champ vide signifie « **sans limite à ce niveau** ». **Toutes les limites applicables se cumulent** : une clé ne peut pas dépasser la limite globale ou celle de son application parce que son propre quota est plus élevé.

Une génération peut être refusée si l'entrée additionnée à la sortie maximale demandée dépasse le budget disponible, même si la réponse finale aurait pu être plus courte : réduire le prompt ou la sortie demandée peut permettre l'appel.

L'enregistrement demande `quotas.manage` et le droit de gestion du sujet concerné ; le niveau global exige aussi `settings.manage`. Les réservations des appels en cours sont prises en compte pour éviter des dépassements simultanés. En cas de réservations expirées après un incident, un administrateur autorisé peut utiliser **Finaliser les réservations expirées** : cette action ne remet pas les compteurs à zéro et ne restitue pas une consommation réellement effectuée.

15. Examiner les requêtes

1. Ouvrir **Système → Requêtes**.
2. Filtrer par période, acteur, modèle, résultat, statut HTTP ou identifiant de requête.
3. Ouvrir une trace pour consulter son attribution, sa latence, ses tokens, ses réservations et une éventuelle erreur.
4. Pour un appel RAG, examiner les liens entre la requête parente et les opérations associées.



Requête / date	Application / modèle	Résultat	Tokens	Quota imputé	Latence
841a77db-375a-44f4-b76c-9ff5a4eb6497 23/10/2026 00:14:04	Test administrateur nibras-main	200 Génération transmise	814 Mesuré	814 Mesuré	1 842 ms
c10da24f-7c4d-4a3d-8139-bc88d3223539 23/10/2026 00:13:50	Test administrateur nibras-fast	503 context_check_unavailable	— Inconnu	0 Sans génération	58 ms
25c6e504-cd46-42a2-b94f-2f44f708cb0d 22/10/2026 23:43:19	Genero nibras-main	200 Génération transmise	1 301 Mesuré	1 301 Mesuré	1 732 ms
c8218be9-946a-466c-aebf-be8025f34bb7 22/10/2026 23:26:45	Correcto nibras-fast	200 Génération transmise	2 587 Mesuré	2 587 Mesuré	884 ms
29bbd113-d1f0-49b0-a731-072b2057a71e 22/10/2026 23:06:46	Chat NIBRAS nibras-main	200 Génération transmise	785 Mesuré	785 Mesuré	1 093 ms
23cee1d6-bef9-498b-b7f2-8282356f6379 22/10/2026 23:05:37	Chat NIBRAS nibras-main	200 Génération transmise	2 453 Mesuré	2 453 Mesuré	1 916 ms
3c1f142f-35c4-4a8d-9e1d-bf8643a6285b 22/10/2026 22:58:29	Correcto nibras-fast	200 Génération transmise	1 700 Mesuré	1 700 Mesuré	1 530 ms
f7e0c198-aa1a-4a0b-aa36-154075902f12 22/10/2026 22:47:20	Chat NIBRAS nibras-fast	200 Génération transmise	2 775 Mesuré	2 775 Mesuré	2 047 ms

Figure 47 - *Requêtes : application et modèle, résultat HTTP, tokens, quota imputé et latence*

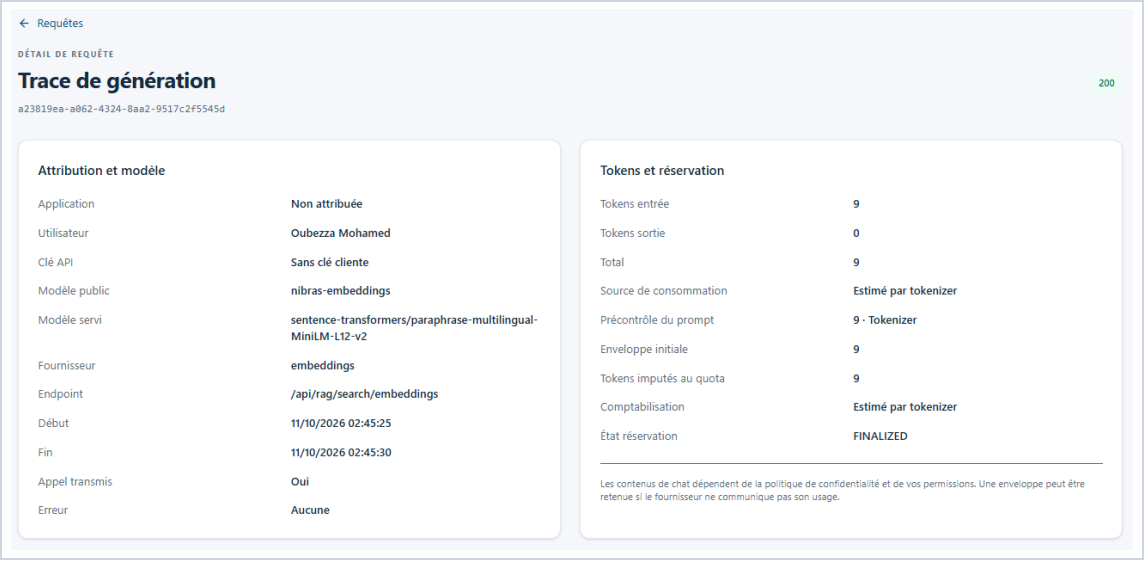


Figure 48 - *Trace d'une génération : attribution, modèle, tokens et réservation*

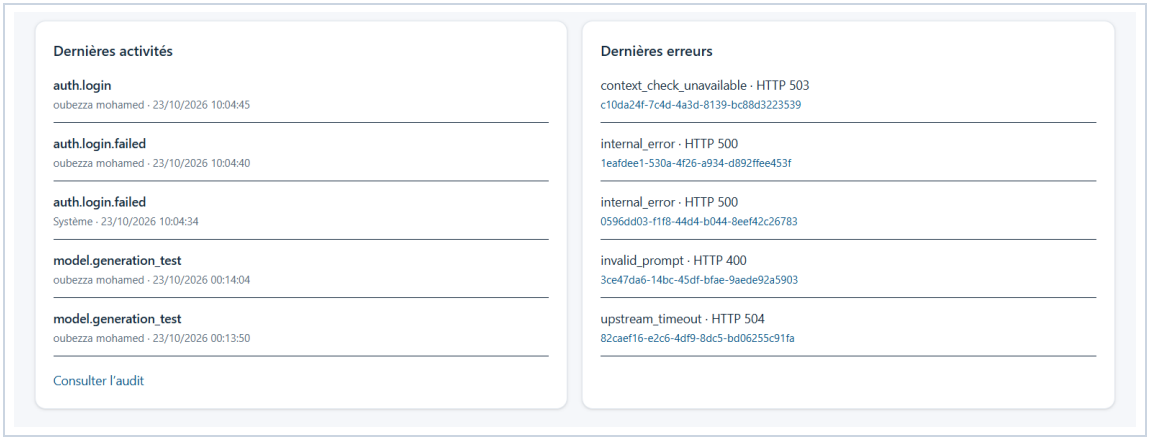


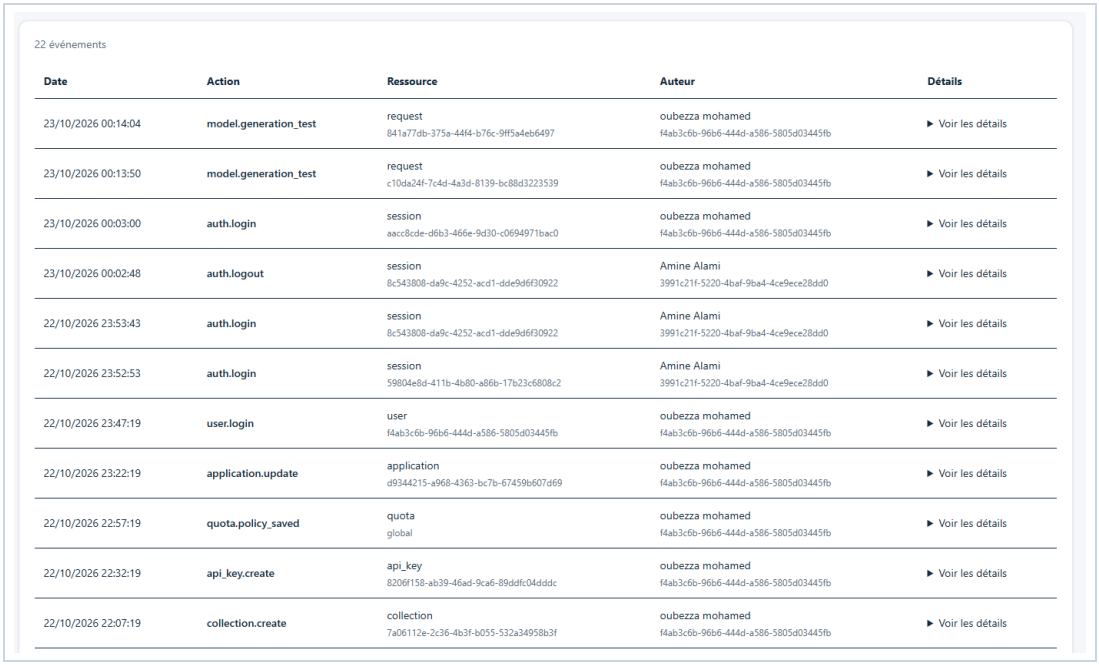
Figure 49 - *Dernières activités et dernières erreurs, avec les identifiants de requête*

Communiquez l'**identifiant de requête** au responsable technique pour faciliter le diagnostic. Les codes d'erreur courants sont décrits en annexe B.

> **Confidentialité des traces**
>
> Les prompts et réponses ne sont **pas stockés par défaut**. Leur éventuelle consultation exige une politique de stockage active et des permissions supplémentaires ; un auditeur ne reçoit pas ce contenu privé. Les contenus RAG et embeddings ne sont pas conservés dans ces traces.

16. Consulter le journal d'audit

1. Ouvrir **Système → Audit**.
2. Filtrer par dates, auteur, action, type de ressource ou identifiant de ressource.
3. Parcourir les résultats pour retrouver une connexion, une modification de compte, une opération de clé ou une action documentaire.
4. Cliquer sur **Voir les détails** pour examiner l'auteur, la date, l'adresse IP et les métadonnées disponibles.



22 événements

Date	Action	Ressource	Auteur	Détails
23/10/2026 00:14:04	model.generation_test	request 841a77db-375a-44f4-b76c-9ff5a4eb6497	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
23/10/2026 00:13:50	model.generation_test	request c10da24f-7c4d-4a3d-8139-bc88d3223539	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
23/10/2026 00:03:00	auth.login	session aacc8cde-d6b3-466e-9d30-c0694971bac0	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
23/10/2026 00:02:48	auth.logout	session 8c543808-da9c-4252-acd1-dde9d6f30922	Amine Alami 3991c21f-5220-4baf-9ba4-4ce9ece28dd0	▸ Voir les détails
22/10/2026 23:53:43	auth.login	session 8c543808-da9c-4252-acd1-dde9d6f30922	Amine Alami 3991c21f-5220-4baf-9ba4-4ce9ece28dd0	▸ Voir les détails
22/10/2026 23:52:53	auth.login	session 59804e8d-411b-4b80-a86b-17b23c6808c2	Amine Alami 3991c21f-5220-4baf-9ba4-4ce9ece28dd0	▸ Voir les détails
22/10/2026 23:47:19	user.login	user f4ab3c6b-96b6-444d-a586-5805d03445fb	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
22/10/2026 23:22:19	application.update	application d9344215-a968-4363-bc7b-67459b607d69	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
22/10/2026 22:57:19	quota.policy_saved	quota global	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
22/10/2026 22:32:19	api_key.create	api_key 8206f158-ab39-46ad-9ca6-89ddfc04dddc	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
22/10/2026 22:07:19	collection.create	collection 7a06112e-2c36-4b3f-b055-532a34958b3f	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails

Figure 50 - *Journal d'audit : date, action, ressource, auteur et détails*



22/10/2026 22:07:19	collection.create	collection 7a06112e-2c36-4b3f-b055-532a34958b3f	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▸ Voir les détails
22/10/2026 21:42:19	document.index	document doc-ref-102	oubezza mohamed f4ab3c6b-96b6-444d-a586-5805d03445fb	▾ Voir les détails IP : 127.0.0.1 { "chunks": 62, "status": "INDEXED" }

Figure 51 - *Événements d'audit avec détails dépliés (adresse IP, métadonnées)*

Le journal est consultable sans modification ni suppression manuelle des événements. Sa durée de conservation relève des paramètres de plateforme et de la maintenance automatique (§ 18).

17. Surveiller les services

Ouvrez **Système → Monitoring**. Les états se rafraîchissent **toutes les 30 secondes** ; utilisez l'actualisation manuelle après une intervention.

État	Interprétation
Healthy	Le contrôle de disponibilité a réussi.
Warning	Une anomalie ou une dégradation nécessite une vérification.
Down	Le service contrôlé est indisponible.

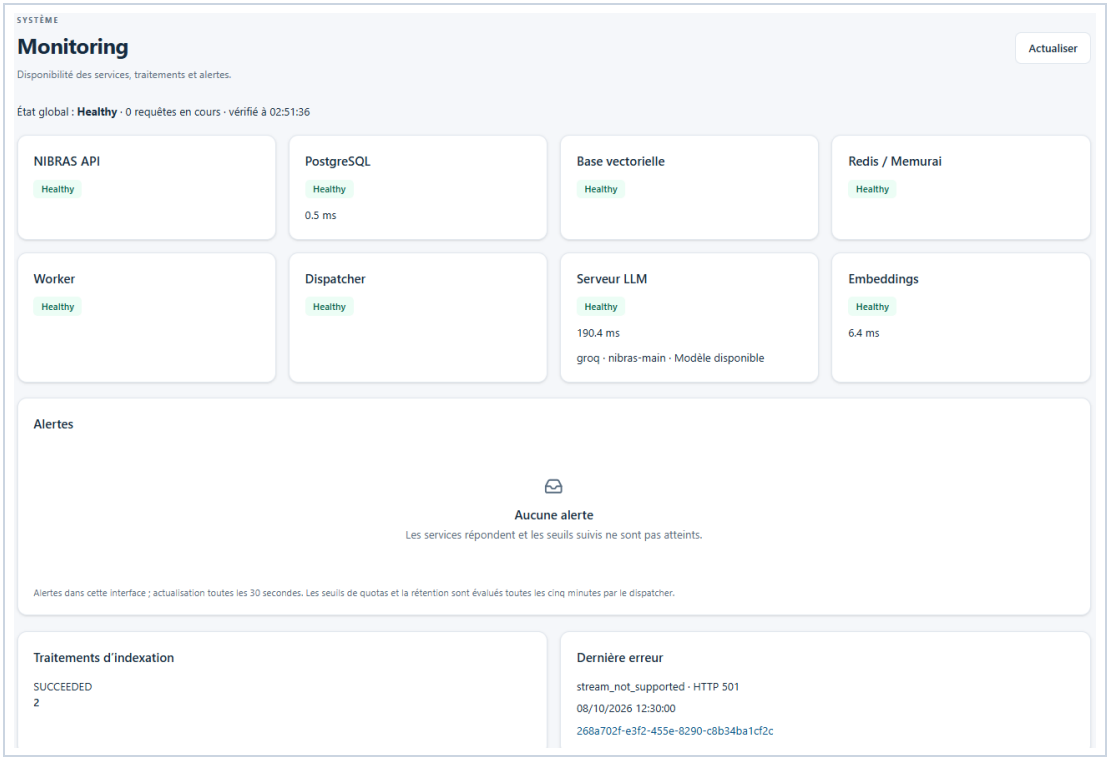


Figure 52 - *Monitoring : état des services, alertes, traitements d'indexation et dernière erreur*

L'écran supervise l'API, PostgreSQL, la base vectorielle (pgvector), le fournisseur LLM, les embeddings, Redis/Memurai, le *worker* et le *dispatcher*. Consultez aussi les alertes, opérations en cours, échecs d'indexation et informations de maintenance.

> **Limites du contrôle de santé**
>
> Les contrôles du fournisseur LLM vérifient la disponibilité **sans générer de texte** : un état Healthy ne remplace pas un test de génération (chapitre 8) ni une mesure de capacité sous charge.

Les alertes apparaissent dans l'interface, **sans envoi automatique d'email ou de webhook**. Elles peuvent signaler un service indisponible, un traitement documentaire en échec, un seuil de quota atteint ou une maintenance absente depuis plus de quinze minutes.

18. Modifier les paramètres de la plateforme

1. Ouvrir **Système → Paramètres**.
2. Consulter les valeurs effectives.
3. Si vous disposez de `settings.manage`, modifier les champs concernés et cliquer sur **Enregistrer les paramètres**.
4. Vérifier le message de confirmation et tester le parcours concerné.

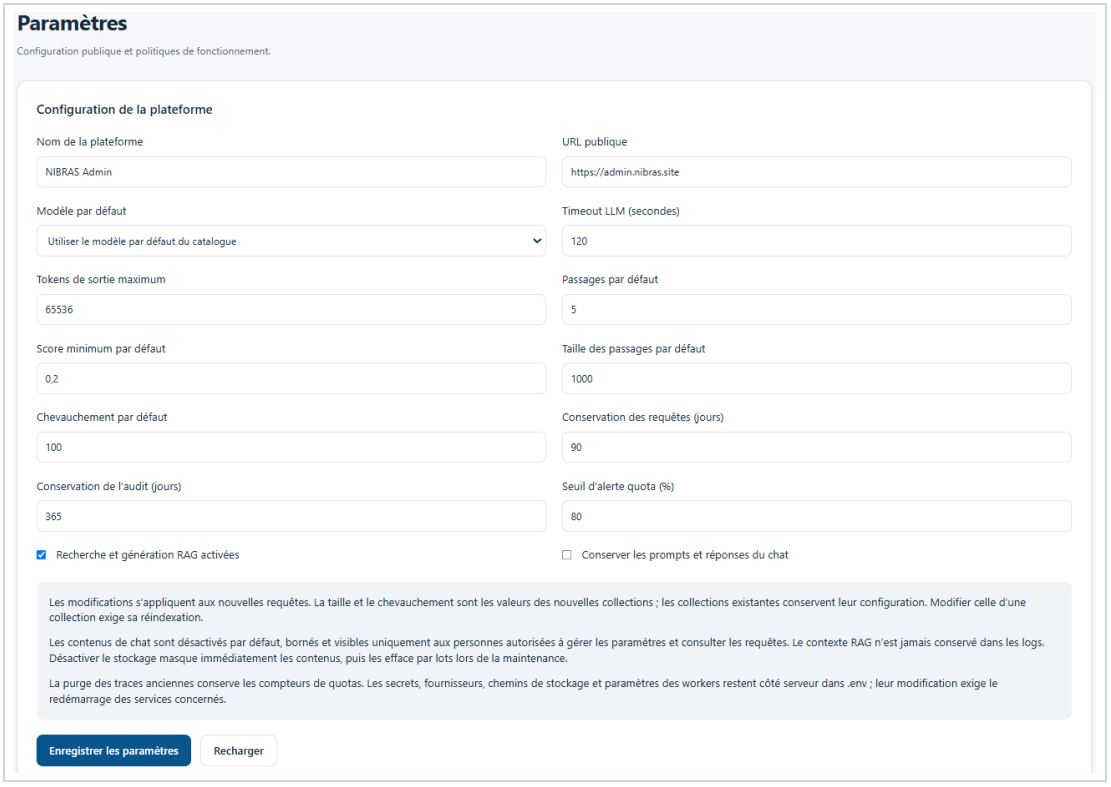


Figure 53 - *Paramètres : général, génération, RAG, confidentialité, conservation et alertes*

Groupe	Paramètres et effets
Général	Nom de plateforme et URL publique, notamment utilisée pour les liens de réinitialisation.
Génération	Modèle par défaut, timeout LLM et plafond de tokens de sortie. Pour l'OCR et le démarrage lent, régler le timeout LLM sur **180 à 300 s**.
RAG	Activation, nombre de passages, seuil de score et découpage par défaut des nouvelles collections.
Confidentialité	Autorisation ou non du stockage du contenu de chat.
Conservation	Durée de conservation des requêtes et du journal d'audit.
Alertes	Pourcentage de quota déclenchant une alerte.

Les réglages enregistrés sont pris en compte pour les **nouvelles opérations**. Les paramètres de découpage par défaut ne modifient pas les collections existantes. Désactiver RAG bloque la recherche et la génération RAG sans supprimer les documents. Si un autre administrateur a enregistré entre-temps, l'application refuse d'écraser silencieusement ses changements : rechargez les valeurs, réappliquez vos modifications puis enregistrez.

18.1 Confidentialité et conservation

Le stockage des prompts et réponses de chat est **désactivé par défaut**. Lorsqu'il est activé, il concerne uniquement les nouvelles générations standard et tests LLM réussis, avec une taille bornée. Les contenus RAG, les embeddings et les **requêtes contenant des images** ne sont jamais stockés dans les traces. La consultation de contenus conservés exige `logs.read`, `settings.manage` et une politique actuellement active ; désactiver cette politique masque immédiatement les contenus existants, la maintenance les efface ensuite.

Valeur initiale	Réglage
90 jours	Conservation des requêtes.
365 jours	Conservation du journal d'audit.
80 %	Seuil d'alerte de quota.
5 minutes	Cadence de la maintenance, qui procède par lots : une suppression complète peut nécessiter plusieurs passages.

La suppression de traces anciennes ne remet pas les compteurs de quotas à zéro. Les secrets et connexions techniques restent dans la configuration serveur : l'installation, le changement de fournisseur, l'email, les sauvegardes et le redémarrage des services sont à traiter avec le responsable technique.

19. Utiliser Chat NIBRAS

Chat NIBRAS est l'assistant conversationnel institutionnel de l'écosystème (de type ChatGPT). Il offre au personnel de l'OFPPT une alternative souveraine aux outils d'IA grand public : il répond aux questions en s'appuyant sur les documents de l'office, analyse des fichiers déposés et indique les sources utilisées. C'est une **application cliente** de NIBRAS Admin : ses appels passent par la passerelle, qui contrôle les accès, les quotas et la traçabilité.

19.1 Ce que Chat NIBRAS permet de faire

Fonction	Ce que vous pouvez faire
Interroger les documents de l'office (RAG)	Poser une question en langage courant : la réponse s'appuie sur les collections documentaires autorisées pour votre profil.
Analyser des documents déposés	Résumer un fichier, en extraire des informations ciblées, comparer des éléments, organiser des données (par exemple classer des notes).
Rédiger et reformuler	Obtenir de l'aide pour rédiger, reformuler ou structurer un texte.
Vérifier les sources	Chaque réponse fondée sur un document affiche la source utilisée (badge de source, bouton « Source »).
Refus d'inventer	Lorsque les documents ne contiennent pas l'information, Chat NIBRAS le dit au lieu de la deviner.
Conversations et export	Retrouver ses conversations passées et réutiliser les réponses, selon la configuration de l'installation.

> **Un assistant, pas un décideur**
> Les réponses de Chat NIBRAS restent soumises à la **vérification de l'utilisateur**. L'assistant n'est pas une source réglementaire : en cas de doute, consultez le document cité.

19.2 Accéder à Chat NIBRAS et prendre en main l'interface

1. Ouvrez **https://chat.nibras.site** dans un navigateur récent et connectez-vous avec le compte fourni par l'administration.
2. Sur la page d'accueil, repérez la zone de saisie « Comment puis-je vous aider aujourd'hui ? ».
3. Saisissez votre question, puis envoyez-la avec le bouton en forme de flèche (ou la touche Entrée).

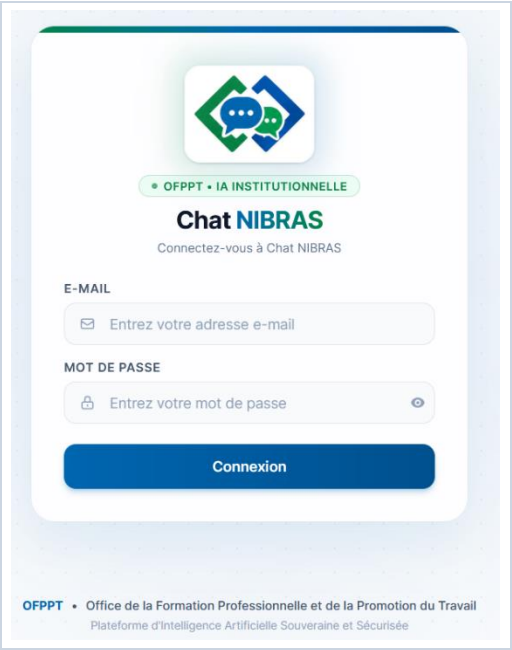


Figure 54 - *Connexion à Chat NIBRAS*

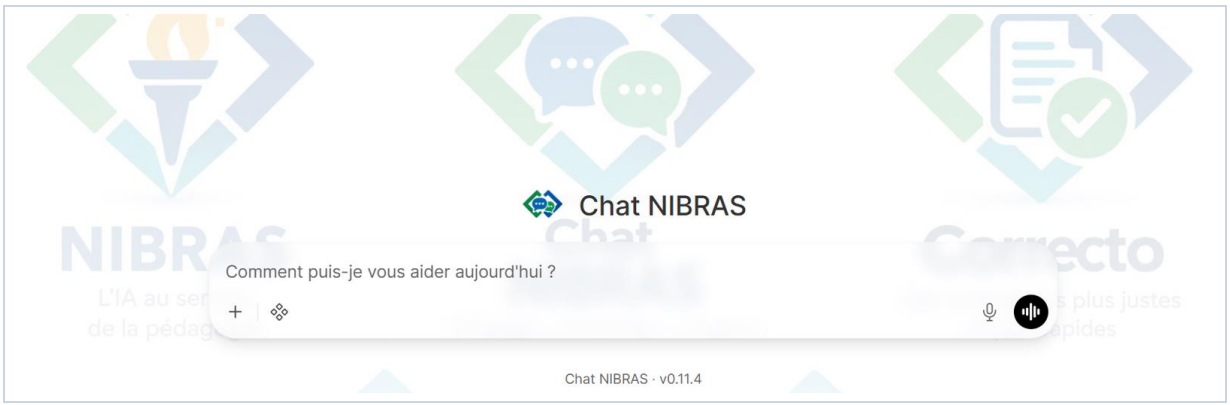


Figure 55 - *Page principale de Chat NIBRAS*

La zone de saisie comprend, de gauche à droite : un bouton **+** pour joindre un fichier, un second bouton d'outils, un **microphone** pour la dictée vocale et le bouton d'envoi. Le bouton d'envoi devient une flèche dès qu'un texte est saisi. La version de l'interface est affichée sous la zone de saisie.

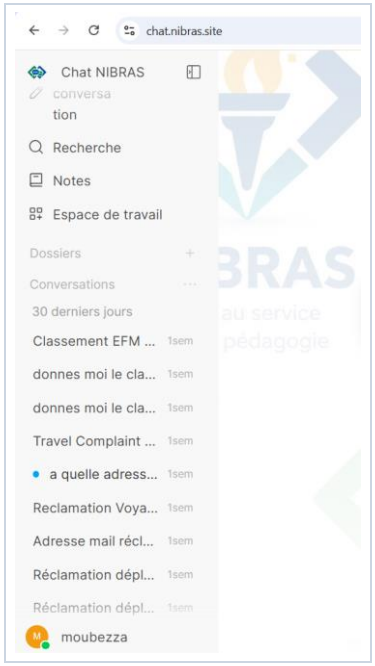


Figure 56 - *Historique des conversations*

19.3 Poser une question sur les documents de l'office

1. Formulez une question **précise** et complète, en citant si possible la filière, le module ou la procédure concernée.
2. Envoyez la question. Chat NIBRAS affiche une indication de progression, par exemple l'exploration de la base de connaissances (*query_knowledge_files*) ou la durée de réflexion.
3. Lisez la réponse, puis contrôlez la **source** indiquée.

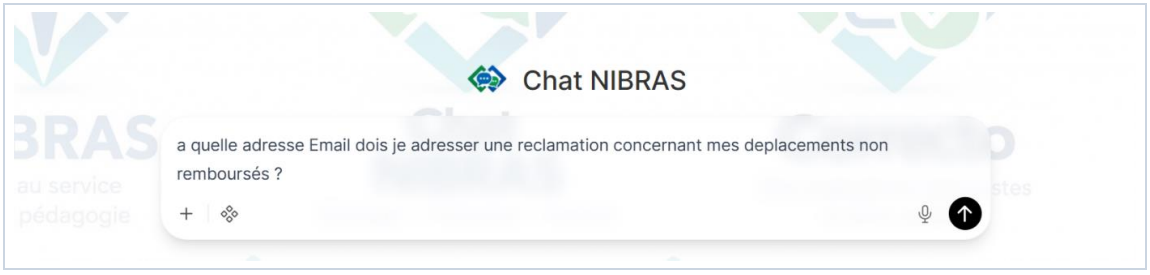


Figure 57 - *Saisie d'une question sur une procédure de l'office*

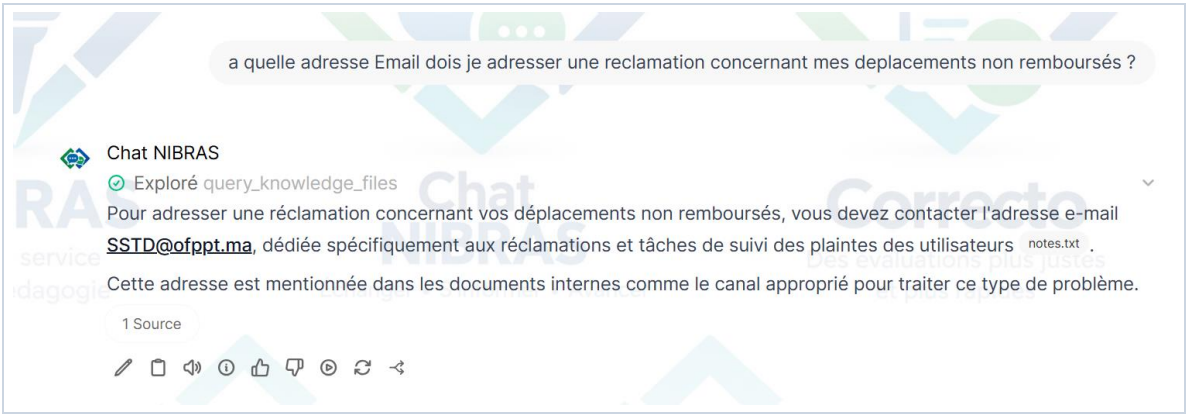


Figure 58 - *Réponse fondée sur un document : indication d'exploration, badge de source (notes.txt) et bouton « 1 Source »*

Dans cet exemple, la réponse précise l'adresse à utiliser pour une réclamation et cite le document d'où elle provient. Le badge placé à la fin de la phrase et le bouton **« 1 Source »** permettent de remonter au document : vérifiez que la source correspond bien à votre question.

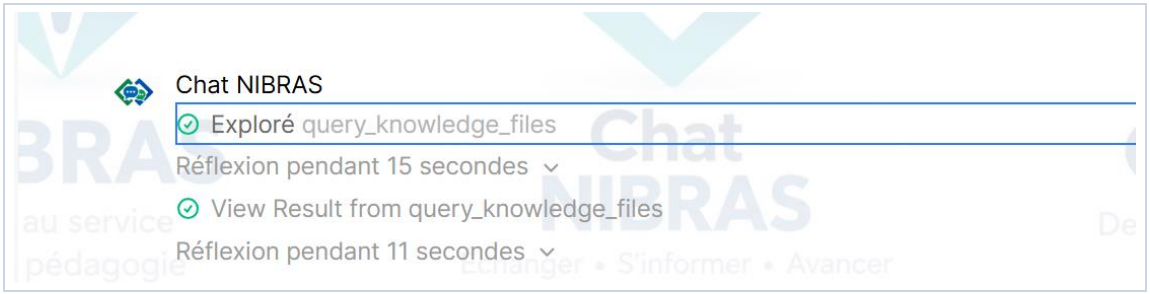


Figure 59 - *Consultation d'une source*

Actions sur une réponse

Sous chaque réponse, une rangée d'icônes permet d'agir sur le message : **copier** le texte, l'**écouter** à voix haute, consulter les informations de génération, **évaluer** la réponse (pouce levé ou baissé), la **régénérer** ou la **partager**. Le jugement positif ou négatif aide l'équipe NIBRAS à améliorer l'assistant. La disponibilité exacte de ces actions dépend de la version de l'interface installée.

19.4 Quand Chat NIBRAS répond qu'il n'a pas l'information

Si les documents autorisés ne contiennent pas la réponse, Chat NIBRAS doit l'indiquer plutôt que d'inventer. C'est le comportement attendu, pas une panne.



Figure 60 - *Signalement de l'absence d'information suffisante dans la base de connaissances*

Dans ce cas :

- **reformulez** la question avec d'autres termes (nom exact de la procédure, du module, du service) ;
- précisez le contexte ou joignez le document concerné (§ 19.5) ;
- si l'information devrait exister, demandez à l'administrateur de vérifier que le document est **importé, indexé et autorisé** pour votre profil (chapitres 9 à 12).

19.5 Analyser un fichier déposé

1. Cliquez sur le bouton **+** de la zone de saisie et choisissez le fichier (PDF, DOCX, TXT ou fichier de données comme un tableur Excel).

2. Vérifiez que le fichier apparaît dans la zone de saisie (nom et taille).
3. Formulez la demande : résumé, extraction, comparaison, classement, calcul…
4. Envoyez, puis relisez le résultat et la ou les sources citées.

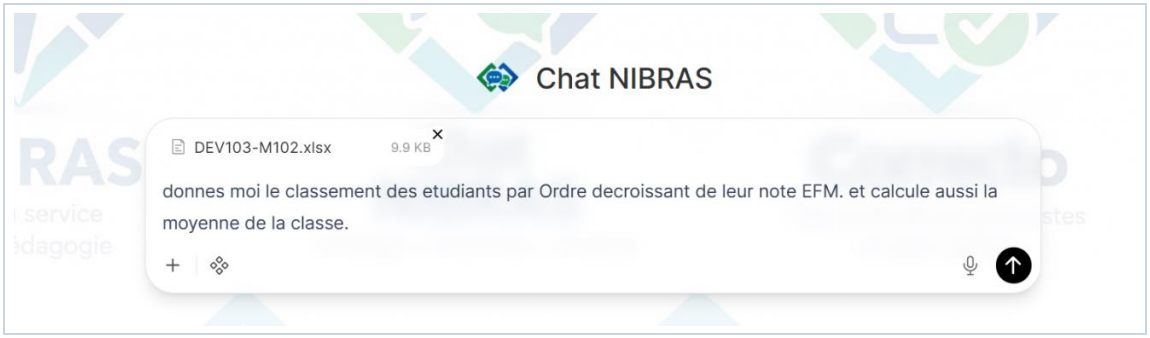


Figure 61 - *Fichier Excel joint et demande de classement avec calcul de moyenne*

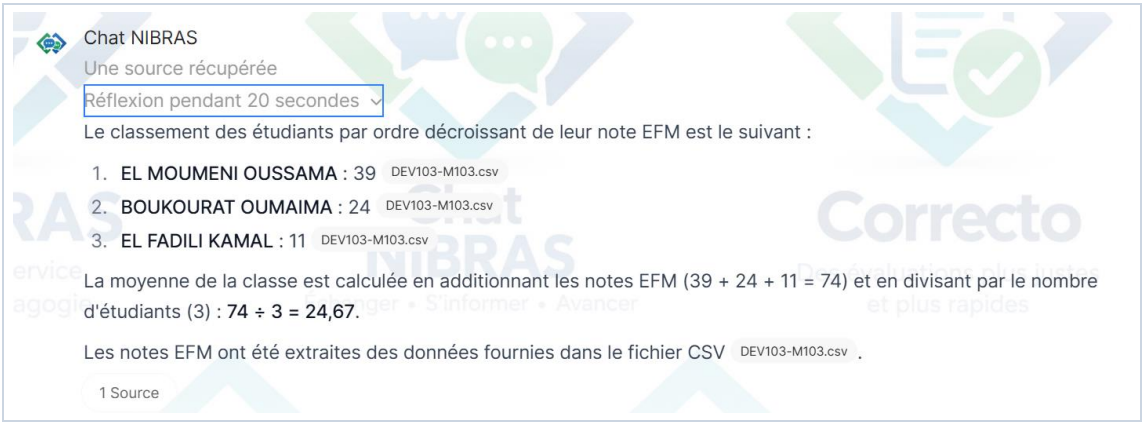


Figure 62 - *Résultat : classement des étudiants par note d'EFM et calcul de la moyenne, avec indication du fichier source*

Contrôler les résultats chiffrés

Pour tout classement ou calcul, vérifiez que **toutes les lignes** attendues ont été prises en compte et que le **fichier source cité** est bien celui que vous avez déposé. Les résultats restent soumis à votre vérification.

19.6 Bonnes pratiques et limites

- **Vérifiez toujours la source** d'une information administrative avant de l'utiliser (adresse, procédure, montant, date).
- Une même question peut recevoir des réponses différentes selon les **documents disponibles** : la réponse dépend du contenu indexé et de vos autorisations.
- Un assistant peut produire une réponse **plausible mais fausse** : c'est le risque d'« hallucination ». Les outils grand public non gouvernés y sont autant exposés ; la citation des sources est ce qui permet de la détecter.

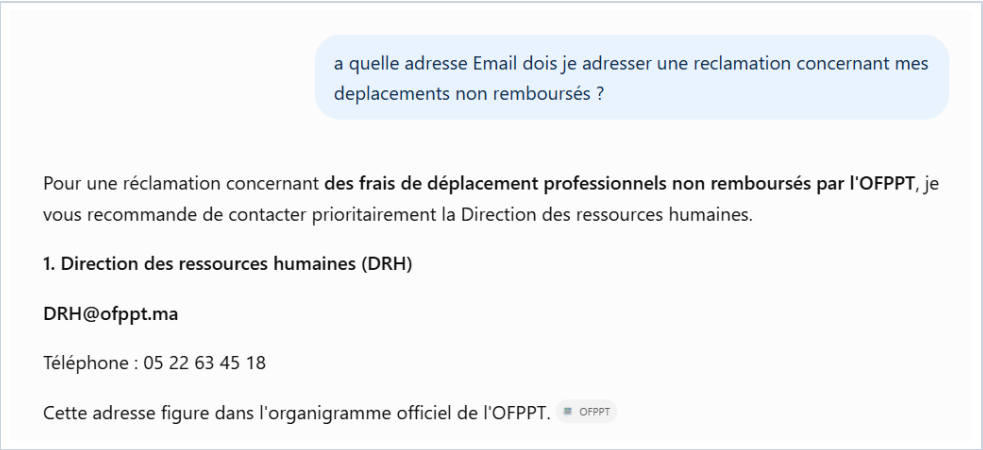


Figure 63 - *Exemple (cité dans le rapport) d'une réponse erronée formulée avec assurance par un outil grand public (ChatGPT, OpenAI)*

19.7 Gouverner Chat NIBRAS depuis NIBRAS Admin

Chat NIBRAS étant une application cliente comme les autres, l'administrateur dispose des mêmes leviers :

Besoin	Où et comment
Vérifier que l'application existe	**Accès → Applications** : application Chat NIBRAS et ses clés (chapitre 5).
Fournir ou renouveler sa clé	**Accès → Clés API** : scopes `chat.completions` et `rag.query`, modèles et collections autorisés (chapitre 6).
Choisir les documents qu'elle peut interroger	**Base de connaissances → Collections → Autorisations** : autoriser l'application, ou sa clé, sur les collections voulues (§ 9.2).
Borner la consommation	**Analytics → Quotas**, niveaux Application et Clé API (chapitre 14).
Suivre l'usage	**Vue d'ensemble** (répartition par application) et **Analytics → Consommation** (chapitre 13) ; **Système → Requêtes** en filtrant sur l'application (chapitre 15).
Maîtriser la confidentialité	**Système → Paramètres** : le stockage des prompts et réponses de chat est désactivé par défaut (§ 18.1).

20. Parcours pratiques

20.1 Donner accès à une nouvelle application

1. Créer ou activer l'application dans **Applications**.
2. Vérifier qu'un fournisseur et un modèle public sont actifs.
3. Fixer les budgets adaptés dans **Quotas**.
4. Générer une clé avec les seuls scopes et modèles nécessaires.
5. Transmettre sa valeur au responsable du client par le canal prévu pour les secrets.
6. Faire effectuer un appel par le client avec sa clé NIBRAS.
7. Vérifier l'attribution dans **Requêtes**, puis la consommation et les quotas.

Un test LLM dans l'interface valide la génération administrative ; le test réalisé par le client avec sa propre clé valide en plus ses scopes, restrictions et quotas.

Pour Open WebUI (Chat NIBRAS), utiliser le modèle public autorisé et une clé NIBRAS avec le scope `chat.completions`. L'option **Stream Chat Response** permet d'afficher la réponse progressivement. Garder les outils intégrés désactivés pour un chat texte. Une réponse interrompue peut être comptabilisée sur la base du maximum réservé lorsque le fournisseur n'a pas communiqué de consommation finale ; consulter **Requêtes** pour distinguer les tokens mesurés de cette charge prudente.

20.2 Brancher une application de lecture d'images (OCR)

Exemple : Correcto, qui lit les copies scannées.

1. Créer un modèle vision, par exemple `nibras-tizimi-vision` (§ 7.4).
2. Dans **Paramètres**, régler le timeout LLM sur 180 à 300 s.
3. Autoriser ce modèle sur la clé de l'application, avec le scope `chat.completions`.
4. Côté application, utiliser l'adresse `https://admin.nibras.site/v1`, sa clé NIBRAS et le nom du modèle vision. Son propre délai d'attente doit être un peu **supérieur** au timeout LLM de NIBRAS : ainsi, en cas de lenteur, l'erreur reçue est claire.
5. Vérifier qu'aucune réflexion n'est facturée : une question simple du type « Combien font 37 × 48 ? Réponds uniquement par le nombre » doit coûter quelques tokens de sortie seulement, et non plusieurs centaines.
6. Lire une vraie copie, puis contrôler la durée et les tokens dans **Requêtes**. Une page A4 prend généralement 20 à 25 s.

20.3 Préparer une base documentaire exploitable

1. Créer une collection, de préférence **restreinte** pour un contenu interne.
2. Accorder les accès aux utilisateurs et applications concernés.
3. Importer les documents avec des métadonnées cohérentes.
4. **Lancer l'indexation et attendre sa réussite.**
5. **Vérifier les passages extraits.**
6. Tester plusieurs questions connues dans **Test RAG**.
7. Contrôler les sources et tester également un compte sans accès à la collection.
8. Autoriser `rag.query` sur la clé cliente et vérifier ses restrictions de collections et de modèles.

20.4 De l'import d'un document à la réponse dans Chat NIBRAS

Ce scénario relie l'administration et l'usage final : un document de procédures est mis à disposition, puis interrogé par un agent.

Étape	Écran	Action
1	Documents › Importer	Choisir la collection (ex. « RH »), déposer le PDF, valider l'import (§ 10.1, figures d'import).
2	Documents › fiche / Indexation	Lancer l'indexation et suivre la progression jusqu'à « Terminé » (§ 11.2).
3	Test RAG	Poser la question attendue et vérifier que le bon passage ressort en tête avec un score correct (§ 12.1).
4	Collections › Autorisations	Autoriser l'application Chat NIBRAS (ou sa clé) à interroger la collection (§ 9.2).
5	Chat NIBRAS	Poser la même question et contrôler que la réponse cite le document (§ 19.3).
6	Requêtes / Audit	Retrouver la requête et l'action d'indexation pour la traçabilité (chapitres 15 et 16).

20.5 Vérifications régulières

- Consulter la vue d'ensemble et les alertes Monitoring.
- Examiner les échecs et opérations restées incomplètes.
- Contrôler les budgets proches de leur plafond.
- Vérifier les indexations après import ou remplacement de fichiers.
- Réviser les comptes, applications et clés devenus inutiles.
- Consulter l'audit après une modification sensible.

21. Résoudre les problèmes courants

Situation	Vérification ou action
Un menu ou un bouton manque	Vérifier le rôle, les permissions individuelles et les autorisations de collection.
Connexion refusée	Vérifier l'identifiant, le mot de passe et le statut du compte ; attendre en cas de verrouillage temporaire.
Aucun email de réinitialisation	Vérifier le message affiché, le dossier indésirable et la disponibilité du service email avec l'administrateur.
Une clé ne fonctionne plus	Vérifier son statut, son expiration, l'application et l'utilisateur associé ; une révocation ou un renouvellement invalide l'ancienne valeur.
Accès client refusé à une ressource	Vérifier les scopes, modèles autorisés, restrictions de collections et autorisations documentaires.
Limite de fréquence atteinte	Espacer les appels et attendre la fin de la fenêtre glissante.
Budget de tokens dépassé	Vérifier les niveaux de quotas et les réservations ; attendre la nouvelle période ou faire ajuster le budget par un administrateur.
Aucun modèle disponible	Vérifier le statut du modèle et de son fournisseur, puis le test de connexion.
« Le fournisseur est indisponible » alors que « Vérifier la connexion » réussit	Vider le champ température du test, retirer les paramètres par défaut et les recettes de thinking du modèle, puis réessayer ; certains fournisseurs refusent des paramètres qu'ils ne gèrent pas. Si l'erreur persiste, transmettre l'identifiant de requête au responsable technique.
`upstream_model_missing` à la vérification d'un modèle	Le nom du « Modèle servi » n'est pas celui du fournisseur : reprendre un nom de la liste affichée.
Une connexion n'apparaît pas dans la liste des fournisseurs	Elle doit être déclarée dans `LLM_PROVIDERS` (fichier `.env`) par le responsable technique, puis l'API redémarrée.
« Connexion absente de .env »	La connexion d'un fournisseur a été retirée du fichier `.env` : la rétablir, sinon les appels échouent de façon contrôlée.
Suppression d'un modèle ou d'un fournisseur refusée	Modèle par défaut : en désigner un autre. Fournisseur : supprimer d'abord ses modèles. Élément défini dans `.env` : le désactiver.
`model_vision_unsupported` : « Le modèle X n'accepte pas les images »	Le client a envoyé une image à un modèle sans vision. Cocher « Accepte les images » sur un modèle multimodal, ou faire utiliser au client le modèle vision prévu (§ 7.4).
Image refusée : « trop volumineuse », « format non supporté », « base64 incorrect », « URL d'image distante refusée », « Trop d'images »	Le client doit envoyer des images JPEG, PNG ou WebP en data URL base64, d'au plus 5 Mo chacune et 10 par requête. Il peut réduire la résolution des pages, par exemple à 2048 px sur le plus grand côté.
`request_too_large` (413)	La requête dépasse 32 Mo : envoyer moins de pages par requête. Si l'erreur vient du proxy avant NIBRAS, faire vérifier la taille maximale autorisée par Nginx.
`llm_timeout` (504) après environ 120 s	Le timeout LLM est trop court : le porter à 180–300 s dans **Paramètres**. Si l'erreur revient souvent, vérifier la réflexion du modèle (une réflexion active ralentit fortement) et les démarrages à froid du fournisseur.
Erreur 524 après environ 100 s	Coupure par Cloudflare : faire passer les clients par un sous-domaine non proxifié.

Situation	Vérification ou action
Génération lente ou en échec	Consulter Monitoring et la trace de requête ; transmettre l'identifiant de requête au responsable technique.
Import refusé	Vérifier le format réel, la taille, les limites du lot, le chiffrement éventuel et les doublons.
Métadonnées refusées	Utiliser un objet JSON valide, avec clés et chaînes entre guillemets doubles, sans virgule finale.
Document importé absent des résultats	Vérifier son indexation, son statut, sa collection, les droits et les filtres RAG.
Texte introuvable dans un PDF scanné	Effectuer une reconnaissance OCR hors de l'application avant import.
Indexation bloquée en attente	Vérifier les prérequis, le worker, Redis/Memurai et le dispatcher dans Monitoring.
Aucun passage pertinent	Reformuler, vérifier le corpus et les métadonnées, puis ajuster le seuil de score.
Sources modifiées pendant la génération	Vérifier les changements de documents et de droits, puis relancer la question.
Paramètres modifiés par un autre administrateur	Recharger les paramètres avant de réappliquer les changements.
Anciennes traces absentes	Vérifier les dates et la durée de conservation ; les quotas conservent leurs propres compteurs.
Clé complète perdue	Créer ou renouveler une clé, mettre à jour le client et révoquer la valeur devenue inutile.
Chat NIBRAS répond « je n'ai pas l'information »	Reformuler ; vérifier que le document est importé, indexé et autorisé pour l'application Chat NIBRAS (chapitres 9 à 12).
Chat NIBRAS ne répond pas ou renvoie une erreur	Vérifier la clé et le quota de l'application, l'état du modèle et des services dans Monitoring ; retrouver la requête dans Système › Requêtes.
Un classement ou un calcul de Chat semble incomplet	Vérifier le fichier cité comme source et le nombre de lignes pris en compte ; reformuler la demande en précisant les colonnes à utiliser.

Pour une demande d'assistance, fournissez l'écran concerné, l'heure, les étapes réalisées, le message et l'identifiant de requête ou de tâche lorsqu'il est disponible. **Ne transmettez jamais la valeur secrète d'une clé ni un mot de passe.**

Les exemples de corpus OFPPT fournis pour les tests sont fictifs et ne constituent pas des références réglementaires. Pour l'utilisation métier, importez et maintenez des documents validés par leur responsable.

Annexe A - Référence des permissions et des scopes

Les permissions s'attribuent aux rôles et, individuellement, aux utilisateurs (Hériter / Autoriser / Refuser). Les libellés ci-dessous sont ceux affichés par l'écran Rôles ou décrits dans les chapitres correspondants.

Permission	Effet	Chapitre
`users.read`	Consulter les utilisateurs.	4
`users.create`	Créer les utilisateurs.	4
`users.update`	Modifier et désactiver les utilisateurs.	4
`users.delete`	Désactiver un utilisateur.	4
`roles.read`	Consulter les rôles.	4
`roles.manage`	Gérer les rôles et permissions (absent du rôle ADMIN).	4
`applications.read`	Consulter les applications.	5
`applications.manage`	Gérer les applications.	5
`apikeys.read`	Consulter les clés API.	6
`apikeys.create`	Créer des clés API.	6
`apikeys.revoke`	Révoquer des clés API.	6
`rag.access.manage`	Gérer les autorisations des collections.	9
`rag.read`	Consulter les collections et documents ; rechercher dans Test RAG.	9 à 12
`rag.upload`	Importer des documents.	10
`rag.update` / `rag.delete`	Modifier, remplacer ou supprimer des documents.	10
`models.manage`	Gérer les modèles ; requis pour la génération administrative (Test LLM, réponse RAG).	7, 8, 12
`quotas.manage`	Enregistrer des quotas (avec le droit de gestion du sujet concerné).	14
`settings.manage`	Modifier les paramètres ; requis aussi pour le quota de niveau Global.	14, 18
`logs.read`	Consulter les journaux ; requis avec `settings.manage` pour lire des contenus conservés.	15, 18

Scope de clé API	Service autorisé
`chat.completions`	Génération de texte avec un modèle de langage.
`embeddings`	Calcul de représentations numériques du texte.
`rag.query`	Recherche documentaire et réponse RAG.

Annexe B - Lire un code d'erreur

Les écrans **Requêtes** et **Dernières erreurs** affichent un code d'erreur et un statut HTTP. Le tableau suivant en donne une lecture indicative ; en cas de doute, transmettez l'identifiant de requête au responsable technique.

Code affiché	HTTP	Lecture indicative
(succès)	200	Requête traitée ; consulter les tokens et la latence.
`invalid_prompt`	400	Prompt ou paramètres invalides : corriger la demande.
`validation_error`	422	Données de la requête non conformes au format attendu.
`internal_error`	500	Erreur interne : relever l'identifiant de requête et prévenir le responsable technique.
`stream_not_supported`	501	Le mode de diffusion en continu demandé n'est pas pris en charge.
`context_check_unavailable`	503	Contrôle du contexte indisponible : vérifier Monitoring, puis réessayer.
`model_vision_unsupported`	400	Image envoyée à un modèle sans vision : activer « Accepte les images » ou utiliser le modèle vision (§ 7.4).
`request_too_large`	413	Requête de plus de 32 Mo : envoyer moins d'images par requête.
`llm_unavailable`	503	Fournisseur indisponible : vérifier l'état de serveur local.
`llm_auth_failed`	-	Clé du fournisseur refusée : le responsable technique corrige `LLM_PROVIDERS` dans le `.env`.
`llm_timeout`	504	Délai de génération dépassé : porter le timeout LLM à 180–300 s (§ 18).
`upstream_timeout`	504	Le fournisseur n'a pas répondu dans le délai : réessayer, vérifier le fournisseur et le timeout (§ 18).

Annexe C - Accès aux applications NIBRAS et index des écrans

Application	Fonction principale	Adresse
Portail NIBRAS	Présentation du projet et accès aux applications	https://www.nibras.site
Chat NIBRAS	Assistant institutionnel exploitant les référentiels de l'OFPPT	https://chat.nibras.site
Perform	Planification et pilotage pédagogique intelligent	https://perform.nibras.site
Genero	Génération assistée de ressources pédagogiques et administratives	https://genero.nibras.site
Correcto	Correction assistée des évaluations manuscrites avec validation humaine	https://correcto.nibras.site
NIBRAS Admin	Administration du socle : accès, base RAG, quotas, supervision	https://admin.nibras.site

Écran	Chemin dans le menu	Chapitre
Vue d'ensemble	Vue d'ensemble	3
Mon compte	Menu utilisateur → Mon compte	2
Utilisateurs / Rôles	Accès → Utilisateurs / Rôles	4
Applications	Accès → Applications	5
Clés API	Accès → Clés API	6
Fournisseurs / Modèles / Tests	Intelligence artificielle	7, 8
Collections / Documents	Base de connaissances	9, 10
Indexation / Test RAG	Base de connaissances	11, 12
Consommation / Quotas	Analytics	13, 14
Requêtes / Audit / Monitoring / Paramètres	Système	15 à 18
Chat NIBRAS	https://chat.nibras.site	19

NIBRAS - conçu pour l'excellence pédagogique et la valorisation du travail du personnel de l'OFPPT.