Fiche mémo | Leçon 19 | 30 min

Brancher OpenAI, Claude, Gemini ou Mistral dans un flux, la clé, le modèle et le coût

Brancher une IA dans un flux demande trois réglages : une clé d'API payante rangée dans les identifiants de l'outil, un modèle choisi pour la tâche, et un plafond de dépense. Classer mille courriels courts coûte quelques dizaines de centimes de dollar en jetons ; la plateforme, avec ses tâches et ses crédits, coûte souvent davantage.

Objectifs

- Créer une clé d'API et la ranger dans les identifiants de n8n, Make ou Zapier
- Choisir un fournisseur et un modèle selon la tâche, les données et le volume
- Estimer le coût mensuel d'un flux en jetons et plafonner la dépense avant le premier appel

À retenir

- L'API d'une IA est un compte à part de l'abonnement, facturé au jeton, qui s'ouvre avec une clé.
- Une clé par flux, rangée dans les identifiants de l'outil, jamais dans un champ de texte ni un tableur.
- Le fournisseur se choisit d'abord sur les données : offre payante, entraînement exclu, contrat accepté, Europe si possible.
- Le modèle le plus économique d'abord ; le plus gros seulement si un test sur vos données l'exige.
- Un plafond de dépense se règle dans la console le jour où la clé est créée.

La demande à recopier

Demande à recopier **Faire chiffrer le coût mensuel d'un flux par un assistant, calcul exécuté**

```
Tu es contrôleur de gestion dans une PME de 12 salariés.
Contexte : un flux appelle une IA à chaque demande de devis reçue. Volume : [nombre] demandes par mois. Jetons par appel, mesurés sur cinq exécutions de test : [nombre] en entrée, [nombre] en sortie. Prix du modèle relevés le [date] : [prix] $ par million de jetons en entrée, [prix] $ en sortie. La plateforme facture [nombre] actions par exécution ; prix de l'offre : [prix] $ par mois pour [nombre] actions.
Tâche : calcule le coût mensuel de l'IA, la part du forfait de la plateforme consommée, et le coût total si le volume double.
Format : un tableau de trois lignes (aujourd'hui, volume doublé, volume triplé), puis deux phrases. Calcule en exécutant du code et montre les formules. N'ajoute aucun prix que je n'ai pas donné.
```

Note : Les jetons réels se lisent dans l'historique d'exécution du nœud IA (champ « usage » ou équivalent) : mesurez-les sur cinq vrais messages plutôt que de les estimer.

Les erreurs fréquentes

L'erreur, et pourquoi elle coûte	La correction
Chercher dans Make ou n8n où coller son identifiant ChatGPT Plus, ou croire que l'abonnement couvre l'API. L'abonnement donne accès à l'application ChatGPT ; l'API est un compte distinct, facturé au jeton, qui ne fonctionne qu'avec une clé.	Créer un compte API au nom de l'entreprise, avec sa propre facturation, et une clé par flux.
Coller la clé dans un nœud Set, un champ d'en-tête écrit à la main ou une cellule de tableur « pour la retrouver ». La clé apparaît alors en clair dans l'historique des exécutions, dans les exports du workflow et chez tous ceux qui ont accès au tableur.	La ranger dans les identifiants de l'outil (credential, connection, account) et choisir cet identifiant dans le nœud.

Ce que contient un appel à une IA

L'identifiant
La clé d'API, rangée dans les identifiants de l'outil. Elle dit qui paie.

Le modèle
Le plus économique pour classer, un intermédiaire pour rédiger. Il fixe le prix du jeton.

La consigne système
Le rôle, les règles, les catégories permises. La même à chaque appel.

Les données du flux
Le courriel, le formulaire, la ligne du tableur, insérés par une expression.

Le format de sortie
Texte libre ou JSON imposé par un schéma, sujet de la leçon suivante.

La limite de jetons
Le plafond de longueur de la réponse. Elle borne aussi le coût d'un appel.

Chaque couche se règle une fois dans le module ou le nœud, puis sert à chaque exécution.