Maîtrise des tokens.

Version 1.0 · en vigueur le 15 novembre 2026

1. Objet et portée

1.1. Ce que couvre cette page. Le coût d'usage des outils d'IA que nous livrons, exploitons ou utilisons pour vous : ce qu'est un token, ce qui fait monter la consommation, comment le budget est fixé, plafonné, coupé, payé, et comment nous vous en rendons compte et la réduisons.

1.2. À qui elle s'applique. À tout outil, agent ou automatisation de BONBAY et de ses quatre marques qui appelle un modèle facturé à l'usage, par API. Les abonnements par utilisateur (une licence mensuelle par personne) ne se comptent pas en tokens : leur prix figure au devis, et seules les règles de clés et de relevé de cette page s'y appliquent.

1.3. Articulation avec les autres documents. L'article 4 des conditions générales de vente pose le cadre : les consommations tierces sont distinguées de nos honoraires, le budget autorisé est écrit, aucune marge non annoncée n'est ajoutée, et un quota n'est un plafond garanti que si le devis décrit son mécanisme. L'annexe A des CGV prévoit, dans sa matrice IA, qui paie, le budget, l'alerte et le plafond réellement disponible. Cette page dit comment nous l'appliquons. Ce qui part vers le modèle relève de IA et vos données ; la garde des clés d'API, d'Accès à vos systèmes.

2. Un token, en termes simples

2.1. Définition. Un modèle d'IA ne lit pas des mots, mais des morceaux de texte appelés tokens (ou jetons). Un mot court et courant fait souvent un token ; un mot long ou rare en fait deux ou trois ; la ponctuation compte aussi. En français, 1 000 tokens représentent de l'ordre de 600 à 750 mots, selon le modèle. Les CGV en donnent la définition : le découpage dépend du modèle et du fournisseur, et un token n'équivaut pas à un mot.

2.2. Entrée et sortie. Le fournisseur compte deux flux. Les tokens d'entrée sont tout ce qui est envoyé au modèle : les instructions de l'outil, la question, les documents joints, l'historique de la conversation. Les tokens de sortie sont la réponse. Chez la plupart des fournisseurs, un token de sortie coûte plusieurs fois plus cher qu'un token d'entrée.

2.3. Ce qui se compte aussi. Certains modèles raisonnent par étapes avant de répondre : ces étapes sont facturées comme de la sortie, même si vous ne les voyez pas. Une image ou un PDF envoyé est converti en tokens. Un agent qui travaille en plusieurs étapes renvoie à chaque étape ce qu'il a déjà lu (voir IA générative et LLM).

2.4. Le prix. Chaque fournisseur publie un prix par million de tokens, propre à chaque modèle, pour l'entrée et pour la sortie, parfois en dollars. Ces prix changent souvent. C'est pourquoi le devis indique le modèle, la date du tarif de référence et l'unité de facturation, et pourquoi les montants de cette page sont des illustrations.

---

3. Ce qui fait monter la facture

Les instructions renvoyées à chaque question. Les consignes de l'outil (son rôle, ses règles, ses exemples) partent avec chaque requête.
Les documents entiers. Joindre un contrat de 40 pages pour une question sur une seule clause multiplie l'entrée par dix ou plus.
L'historique. Dans une conversation, chaque nouvelle question renvoie toutes les précédentes.
Les agents en boucle. Un agent qui relit à chaque étape un contexte de 20 000 tokens consomme 300 000 tokens d'entrée en 15 étapes, pour une seule tâche.
Le plus gros modèle par défaut. Au sein d'une même gamme, le prix par token varie fortement d'un modèle à l'autre, souvent d'un facteur 5 ou plus.
Les réponses longues. Une réponse de 1 500 mots quand 150 suffisent paie dix fois la sortie.
Les relances automatiques. Une erreur qui fait relancer la même requête en boucle consomme sans rien produire.

Pour estimer le budget complet d'un projet, au-delà de la consommation, voir ROI et budget d'un projet IA.

---

4. Un exemple chiffré

4.1. Le cas. Illustration : un assistant interne répond aux questions d'une équipe sur son manuel de procédures, à raison de 200 questions par jour ouvré, 22 jours par mois, soit 4 400 questions par mois. Les volumes sont des ordres de grandeur. Les prix sont hypothétiques, choisis pour rendre le calcul lisible : 3 € par million de tokens d'entrée, 0,30 € par million de tokens d'entrée relus depuis le cache, 15 € par million de tokens de sortie. Ce ne sont pas les tarifs d'un fournisseur.

POSTE, PAR QUESTION	SANS MAÎTRISE	AVEC NOS RÈGLES
Instructions de l'outil	2 000 tokens, renvoyés à chaque fois	2 000 tokens, relus depuis le cache
Documents joints	8 000 tokens (sections entières)	2 500 tokens (passages utiles)
Historique de la conversation	2 000 tokens (tout l'historique)	500 tokens (résumé)
Question	100 tokens	100 tokens
Réponse	600 tokens	400 tokens (longueur plafonnée)

POSTE, PAR QUESTION	SANS MAÎTRISE	AVEC NOS RÈGLES
Total par question	12 700 tokens	5 500 tokens, dont 2 000 en cache
Total par mois	55,9 millions de tokens	24,2 millions, dont 8,8 en cache
Coût mensuel d'illustration	environ 199 €	environ 70 €

4.2. Lecture. Les mêmes réponses coûtent près de trois fois moins. Si les questions simples, environ 70 % ici, passent à un modèle cinq fois moins cher, le total d'illustration descend vers 31 €. À l'inverse, le même outil branché sans ces règles sur un modèle cinq fois plus cher approcherait 1 000 € par mois. Ce que nous maîtrisons, ce sont les volumes et le choix du modèle ; le prix unitaire appartient au fournisseur.

4.3. Limite de l'exemple. Le cache n'est relu que si les questions se suivent à quelques minutes d'intervalle, et certains fournisseurs facturent plus cher sa première écriture. Les volumes réels se mesurent après la mise en service (article 5.3).

5. Le budget fixé au devis

5.1. L'estimation. Avant le devis, nous estimons la consommation à partir de vos chiffres : nombre d'utilisateurs, requêtes par jour, tokens d'entrée et de sortie par requête, modèle prévu. Les hypothèses sont écrites au devis, pour que chacun puisse refaire le calcul.

5.2. Un budget par outil. Chaque outil a son propre budget mensuel, en euros, avec le mécanisme d'alerte et d'arrêt décrit à l'article 6 de cette page. L'environnement de test a un budget séparé et réduit : les essais ne consomment pas le budget de production.

5.3. Le recalage. Après 30 jours d'usage réel, nous comparons la consommation mesurée à l'estimation et vous proposons par écrit de garder ou d'ajuster le budget. Rien ne change sans votre accord.

5.4. Une hausse du fournisseur. Elle ne modifie pas un forfait ferme en cours. Pour une dépense variable, nous vous informons du nouveau prix avant d'engager la dépense (article 4 des CGV).

6. Alertes, plafonds et arrêt

SEUIL DU BUDGET MENSUEL	CE QUI SE PASSE	QUI EST PRÉVENU
50 %	Alerte d'information, avec la projection de fin de mois	Votre interlocuteur désigné et BONBAY
80 %	Alerte ; si la projection dépasse 100 %, BONBAY en analyse la cause et vous l'écrit sous 2 jours ouvrés	Votre interlocuteur désigné et BONBAY

SEUIL DU BUDGET MENSUEL	CE QUI SE PASSE	QUI EST PRÉVENU
100 %	Arrêt des appels au modèle ; l'outil affiche un message prévu à l'avance	Votre interlocuteur désigné et BONBAY
Au-delà	Seulement sur votre accord écrit (un courriel suffit), pour un montant ou une durée fixés	Votre interlocuteur désigné

6.1. Plafonds par requête. Chaque requête a un maximum de tokens d'entrée et de sortie. Une requête qui le dépasse est refusée ou découpée.

6.2. Plafonds par tâche d'agent. Chaque tâche d'agent a un nombre maximal d'étapes, une durée maximale et un maximum de tokens, fixés au devis (par exemple 20 étapes et 10 minutes). Au-delà, l'agent s'arrête et rend compte de ce qu'il a fait.

6.3. Notre propre compteur. Les limites réglées dans la console d'un fournisseur s'appliquent parfois avec retard. Nous plaçons donc un compteur dans l'outil lui-même, qui coupe les appels au seuil convenu sans attendre le fournisseur.

6.4. Le bouton d'arrêt. Vous et BONBAY pouvez couper l'outil à tout moment. La procédure d'arrêt est écrite avant l'activation (article 8.2 des CGV) ; la remise en marche est faite par une personne. Les traitements automatiques en attente sont suspendus, pas effacés.

---

7. Qui paie le fournisseur, et le relevé mensuel

MODE	COMMENT	CE QUE VOUS GARDEZ
Compte et clé à votre nom (par défaut)	Vous ouvrez le compte chez le fournisseur, avec notre aide, et le payez directement ; nous recevons une clé aux droits limités, rangée dans notre coffre d'accès	La main : vous voyez la consommation dans sa console et pouvez couper la clé quand vous voulez
Compte de BONBAY, refacturé	Quand le fournisseur ne permet pas un compte à votre nom, ou pour un essai court ; une clé dédiée à votre projet isole votre consommation	Refacturation au coût réel, sans marge, facture du fournisseur consultable

7.1. Une clé par outil et par environnement. Chaque outil a sa clé, distincte en test et en production, pour que chaque token soit attribué. Une clé ouverte pour vous ne sert ni à un autre client ni à nos propres besoins. Elle est révoquée en fin de mission.

7.2. Le relevé mensuel. Avant le 10 de chaque mois, vous recevez pour le mois écoulé, par outil et par modèle : le nombre de requêtes, les tokens d'entrée, de cache et de sortie, le coût, la comparaison au budget, les cinq tâches les plus coûteuses, les alertes déclenchées et la cause des écarts. Il est remis en PDF et en CSV.

7.3. Devise. Quand le fournisseur facture en dollars, la conversion suit le taux réellement appliqué au paiement, sans marge de change.

7.4. Si le compte est à votre nom. Vous en restez titulaire : le paiement, le renouvellement et la surveillance des alertes de la console vous reviennent (article 8.5 des CGV). Notre compteur et notre relevé s'ajoutent à ces alertes, ils ne les remplacent pas.

---

8. Comment nous réduisons la consommation

Le plus petit modèle qui fait le travail. Nous comparons plusieurs modèles sur vos cas réels, puis réservons le plus grand aux questions qui le demandent.
Le cache. Les instructions et les documents stables sont mis en cache quand le fournisseur le propose.
Les résumés. Un long historique est remplacé par son résumé.
Les passages utiles. Une recherche dans vos documents n'envoie au modèle que les extraits utiles, pas le document entier.
Des réponses à la bonne longueur. La longueur de sortie est plafonnée selon l'usage.
Les traitements groupés. Les tâches non urgentes passent par lots, un mode souvent facturé moins cher par les fournisseurs.
Pas d'IA pour rien. Quand une règle simple ou un script suffit, nous n'appelons pas de modèle.
Une revue chaque trimestre. Nous vérifions si un modèle plus récent ou moins cher fait le même travail, et vous le proposons.

Ces règles servent aussi la sobriété numérique de notre charte RSE (section 5, une IA sobre).

---

9. Limites

9.1. Les prix et le comptage changent. Un fournisseur peut modifier ses tarifs ou sa façon de compter. Un changement de modèle peut changer le nombre de tokens pour le même texte.

9.2. L'arrêt arrête le service. À 100 % du budget, l'outil ne répond plus. Si l'outil est critique pour votre activité, nous prévoyons avec vous un mode dégradé ou un budget plus large ; nous ne décidons pas seuls de continuer.

9.3. Les requêtes déjà parties. Notre compteur coupe les nouveaux appels ; une requête déjà envoyée au moment du seuil se termine. Le dépassement possible se limite à ces requêtes.

9.4. Ce qui passe hors de l'outil. Notre compteur ne voit que ce qui passe par l'outil. Une consommation faite par vos équipes sur le même compte, en dehors de l'outil, n'y figure pas : nous recommandons un projet et une clé séparés.

9.5. L'estimation n'est pas une promesse. La consommation réelle dépend de l'usage de vos équipes. Ce qui vous engage et nous engage, c'est le budget et son plafond, pas l'estimation (article 3.1 des CGV).

---

10. Contact et mise à jour

10.1. Responsable. Pierre Marsanne, président de BONBAY, porte la sécurité, la confidentialité et la protection des données. Pour toute question sur cette page ou sur un relevé : contact@bonbay.fr, ou BONBAY, SAS au capital de 4 500 €, 38 cours de l'Intendance, 33000 Bordeaux. Réponse écrite sous 2 jours ouvrés.

10.2. Version. Version 1.0, en vigueur le 15 novembre 2026. Cette page est revue chaque année (prochaine revue : novembre 2027) et à chaque changement important. Vous êtes prévenu de tout changement qui vous concerne ; une nouvelle version ne modifie pas rétroactivement un contrat en cours (article 2 des CGV).

10.3. Les autres engagements. L'ensemble des pages est réuni dans le Centre de confiance, dont Conformité et audits pour vérifier ces engagements.
