CONDITIONS PARTICULIÈRES APPLICABLES AUX SERVICES IA

au 07/04/2026

Les présentes conditions particulières de services (ci-après les « Conditions Particulières ») sont conclues entre d'une part, toute personne physique ou morale souscrivant aux Services IA proposés par Scaleway (ci-après désignée par le « Client ») et d'autre part l'entité Scaleway, telle que désignée dans les Conditions Générales de Services liant les parties et identifiée comme entité contractante (ci-après désignée par « Scaleway »).

Elles ont pour objet de définir les conditions applicables dans le cadre de la souscription et de l'utilisation des Services IA par le Client.

Les Conditions Particulières complètent les stipulations des Conditions Générales de Services. En cas de contradiction entre les Conditions Générales de Services et les Conditions Particulières, les stipulations des présentes prévalent.

ARTICLE 1. DÉFINITIONS

Dans le cadre des présentes Conditions Particulières, les mots ou expressions ci-après ont la signification suivante :

- **End Of Life (EOL)** : désigne le statut d'un Modèle qui n'est plus accessible via l'API.
- **Generative API** : désigne le Service IA fourni par Scaleway.
- **Instance GPU** : désigne un processeur graphique ayant pour fonction le calcul d'images.
- **Input** : désigne les données (personnelles ou non) transmises par le Client (texte, instruction, audio, image, vidéos) via le Service IA, afin d'obtenir un Output.
- **Managed Inference** : désigne le Service IA fourni par Scaleway pour lequel une infrastructure est dédiée au Client et dont les caractéristiques sont détaillées au sein des présentes Conditions Particulières.
- **Modèle** : désigne un modèle d'intelligence artificielle pré-entrainé mis à disposition dans le cadre du Service IA via une API, capable d'effectuer une génération de texte et de chat, d'analyse d'image ou de transcription audio.
- **Output** : désigne les résultats générés par le Service IA à partir des Input du Client (texte, fichier d'image, etc.).
- **Tokens** : désignent les unités de base du texte ou d'image qu'un modèle génératif traite. Selon la méthode de calcul (tokenisation) propre au Modèle, un Token peut représenter un mot, une partie de mot, un signe de ponctuation ou un caractère. Le nombre de Tokens traités détermine (i) la capacité d'analyse du Modèle (fenêtre de contexte) et (ii) l'assiette de facturation du Service conformément à la grille tarifaire en vigueur.
- **Quota** : désigne les seuils maximaux de fréquence d'utilisation des Services, définis par unité de temps.

- **Services IA** : désignent, dans le cadre des présentes Conditions Particulières, les services dédiés à l'intelligence artificielle : Generative API et Managed Inference.

Sauf si le contexte implique clairement le contraire, les mots indiqués au singulier incluent le pluriel et réciproquement, la référence à un genre inclut les autres genres, la référence à une personne physique inclut les personnes morales, associations etc. et réciproquement, et les mots parents ont les significations correspondantes.

Les titres des articles des Conditions Particulières figurent à titre indicatif uniquement et ne doivent affecter en aucune mesure l'interprétation des stipulations des Conditions Particulières.

Si une définition n'est pas mentionnée au sein de cet article, les mots ou expressions commençant par une majuscule auront la signification leur étant donnée dans les Conditions Générales de Services.

ARTICLE 2. DESCRIPTION DES SERVICES

2.1 Généralités

Le Client peut souscrire, via sa Console de Gestion de Compte, par API, par l'interface en ligne de commande (CLI), par le kit de développement logiciel (SDK) ainsi que par le fournisseur Terraform, à un ou plusieurs types de Services IA, parmi :

- Generative API ;
- Managed Inference.

2.1.1 Generative API

Le Service Generative API est un service de mise à disposition de Modèles d'intelligence artificielle accessibles via une API. Ce Service permet au Client de soumettre des Input (prompts, texte, fichiers audio, etc.) afin de générer, par un traitement automatisé, un Output (texte, prompts, fichier d'image, classifications, etc.).

Scaleway assure la gestion de l'infrastructure et le déploiement des Modèles au sein d'un environnement mutualisé. Afin de garantir la stabilité globale du Service, Scaleway met en œuvre des dispositifs de supervision et de régulation de charge. Le Client reconnaît que cette mutualisation implique une absence d'engagement de performance fixe ou de débit garanti. En cas de pic d'activité, Scaleway se réserve la faculté d'activer des mécanismes de limitation de débit (*Rate Limiting*) afin de préserver l'intégrité du Service.

Il est recommandé de réorienter les requêtes vers un « traitement par lots » (ci-après « Batch Processing ») pour des volumes importants de requêtes ou recourir à des infrastructures dédiées telles que le Service Managed Inference. En effet, dans le cadre du Batch Processing, Scaleway assure la planification et l'ordonnancement des tâches afin d'optimiser à la fois l'allocation des ressources et les délais de traitement. Leur exécution sera achevée dans un délai inférieur à vingt-quatre (24) heures.

Les API proposées dans le cadre du Service Generative API sont conçues pour être compatibles avec les bibliothèques et les SDK d'OpenAI®, y compris la bibliothèque client Python d'OpenAI® et les SDK LangChain. Cette compatibilité est fournie à titre indicatif et peut évoluer, notamment en raison des modifications apportées par les éditeurs tiers concernés, sans que Scaleway ne puisse garantir une correspondance totale, permanente ou exhaustive avec ces bibliothèques et SDK.

2.1.2 Managed Inference

Le Service Managed Inference désigne le service de déploiement et d'exécution de Modèles au sein d'un environnement de calcul dédié et isolé pour le compte du Client. Le Service permet au Client d'interroger, via une API, soit des Modèles mis à disposition par Scaleway, soit des Modèles tiers importés par le Client. Le traitement des Inputs et la génération des Outputs s'effectuent sur des ressources privatives, garantissant au Client une performance constante et une isolation logique complète des ressources par rapport aux autres clients de Scaleway.

Dans ce contexte, le Client peut configurer des options (telles que le type d'Instance GPU, dimension du Modèle, mémoire) selon ses cas d'usage, besoins et compatibilités.

2.2 Solutions tierces

Dans le cadre du Service IA, Scaleway met à disposition des Modèles tiers et s'engage à respecter les licences et toute politique d'utilisation sur son périmètre de responsabilité, ainsi qu'à mettre à disposition les informations fournies par les fournisseurs de Modèles lorsque celles-ci sont disponibles. Le Client s'engage à prendre connaissance des termes de la licence encadrant l'utilisation du Modèle sélectionné. Le Client reconnaît que les conditions d'utilisation de ces Modèles peuvent contenir des restrictions d'utilisation.

Conformément aux normes de l'industrie (telles que la compatibilité avec l'API OpenAI®), Scaleway fait ses meilleurs efforts afin d'assurer une compatibilité avec la plupart des écosystèmes et outils d'intelligence artificielle par défaut.

2.3 Maintenances

Scaleway peut être amenée à effectuer des opérations de maintenance des différentes infrastructures utilisées dans le cadre de la fourniture des Services afin d'en assurer le maintien en conditions opérationnelles.

Nonobstant ce qui précède, le Client reste responsable de signaler à l'Assistance Technique tout Incident Technique qui impacterait l'utilisation des Services IA. Il revient en outre au Client de prendre toutes mesures visant à se prémunir contre les risques inhérents aux opérations de maintenance, quelle qu'en soit leur origine, en procédant notamment (mais non limitativement) à des opérations de sauvegarde régulières de ses Contenus.

ARTICLE 3. CONDITIONS DE SOUSCRIPTION

Préalablement à toute souscription de Service, le Client s'engage à prendre connaissance de l'ensemble des informations mis à sa disposition par Scaleway, en ce compris la Documentation, afin de s'assurer de l'adéquation du Service IA sélectionné, à ses besoins et ses cas d'usages. Selon le type de Service IA sélectionné, il lui appartient d'être en conformité technique conformément à la Documentation et/ou de sélectionner les caractéristiques techniques au moment de sa souscription.

L'activation et le maintien des droits d'usage du Service sont conditionnés à (i) la création d'un Compte Client valide, (ii) la vérification d'identité (« KYC ») et (iii) la validation des moyens de paiement. Scaleway se réserve le droit de suspendre l'accès au Service en cas de non-conformité à l'une de ces étapes de vérification.

ARTICLE 4. CONDITIONS D'UTILISATION

4.1 Généralités

Dans le cadre du Service Generative API, le Client peut utiliser les points d'accès serverless préconfigurés de modèles d'IA sans qu'il soit nécessaire de configurer l'infrastructure sous-jacente.

Par opposition, dans le cadre du Service Managed Inference, le Client est responsable de la configuration des paramètres réseau applicables à chaque déploiement. Il dispose des options suivantes :

- Réseau privé : le Client peut activer le réseau privé afin de garantir une communication sécurisée et une disponibilité restreinte aux seuls environnements internes des réseaux privés.
- Réseau public : le Client peut activer l'accès via réseau public afin de permettre l'accès aux ressources depuis l'Internet public. Dans ce cas, l'accès aux ressources exposées est protégé par une clé API du Service IAM par défaut.

Scaleway se réserve le droit d'appliquer toute mesure de sécurité ou contrainte technique nécessaire au maintien de l'intégrité, de la confidentialité et de la disponibilité de l'infrastructure réseau associée à ces Services.

Le Service Managed Inference permet au Client de déployer les Modèles mis à sa disposition ou de déployer ses propres Modèles. A cette fin, Scaleway met à disposition du Client, au sein de sa Documentation, un catalogue répertoriant l'ensemble des Modèles pouvant être déployés par le Client dans le cadre du Service.

Dans le cadre du Service Generative API, le Client ne peut en aucun cas augmenter le nombre maximal de Tokens Output au-delà des limites fixées pour chaque Modèle. Ces limites ont notamment pour objet de protéger le Client contre :

- des temps de génération excessifs susceptibles d'entraîner l'interruption de la requête en raison d'un dépassement de délai de réponse HTTP. Les limites sont conçues pour

garantir qu'un modèle enverra sa réponse HTTP dans un délai inférieur à cinq (5) minutes ;

- un risque de facturation non maîtrisée, certains Modèles étant susceptibles d'entrer dans des boucles de génération infinies, dans lesquelles le contenu généré se répète sans interruption en fonction de prompts spécifiques. Dans l'hypothèse où le Client a besoin d'un plus grand nombre de Tokens Output, il lui est recommandé de souscrire au Service Managed Inference.

Il est expressément précisé que, dans les cadre des Services IA, Scaleway ne garantit pas que les Outputs sont exempts d'erreur, totalement fiables, pertinents ou conformes aux attentes du Client.

Scaleway met à disposition du Client, dans sa Documentation applicable aux Services, un modèles de partage de responsabilité (« Shared Responsibility Model » ou « SRM ») précisant les rôles, périmètres d'intervention et responsabilités respectifs de Scaleway et du Client. Le Client reconnaît avoir connaissance de l'existence de ces modèles et s'engage à les consulter et à en tenir compte dans le cadre de son utilisation des Services.

4.2 Quotas

Dans le cadre du Service Generative API, le Client peut être soumis à certains Quotas. Le dépassement de ces seuils est susceptible d'entraîner une suspension temporaire du Service ou une limitation des performances. Le niveau de Quota est déterminé selon : (i) le Modèle sélectionné, (ii) la validation préalable d'un moyen de paiement et (iii) de la vérification préalable de l'identité du Client (KYC).

Chaque requête est soumise à une limite de fenêtre contextuelle, c'est-à-dire, limitée en nombre de données (par exemple : longueur de l'Input) que le Modèle va pouvoir traiter en une seule fois.

Afin de faire une demande d'augmentation de ces Quotas, le Client peut solliciter l'Assistance Technique.

Par opposition, le Service Managed Inference n'est pas soumis à Quota dans le nombre de Tokens générés, dans la mesure où le Service fournit une capacité dédiée, les Ouputs étant limités par la taille de la ressource sélectionnée et déployée par le Client.

4.3 Cycles de vie des Modèles

Au titre du Service Generative API, Scaleway met régulièrement à jour les Modèles avec de nouvelles versions officielles via la Console de Gestion de Compte. Les Modèles peuvent donc être soumis à différents statuts, tels que : Preview, Active, Deprecated ou EOL.

En fonction du statut du Modèle concerné, le Service Generative API peut bénéficier, ou non, d'un engagement de niveau de service (SLA) ainsi que des Services de Support associés :

- **Preview** : Les Modèles en statut Preview sont disponibles pour effectuer des tests. Ils bénéficient d'un support d'un (1) mois et d'un préavis d'un (1) mois avant la suppression d'un Modèle en mode Preview.

- **Active** : Les Modèles en statut Active bénéficient d'un support pendant au moins huit (8) mois à compter de leur lancement. Un préavis de trois (3) mois est communiqué avant que le Modèle ne passe au statut Deprecated ou EOL.

- **Deprecated** : Les Modèles en statut Deprecated dispose d'une version du Modèle plus récente. Un Modèle Deprecated signifie qu'un statut EOL est planifié. Bien que les Modèles en statut Deprecated restent utilisables, Scaleway recommande de migrer vers une version Active.

- **EOL** : Le Modèle n'est plus accessible. Préalablement au statut EOL, le Client reçoit un préavis de la part de Scaleway pour l'avertir que le Modèle ne sera plus disponible via le Service Generative API. Certains Modèles pourront cependant continuer à être utilisés, pour une durée plus longue, via le Service Managed Inference. Si un Modèle équivalent est disponible, Scaleway se réserve le droit de rediriger le trafic vers ce Modèle alternatif afin d'éviter toute interruption de Service, bien qu'il n'existe aucune garantie sur la similarité des Outputs générés à ceux du Modèle initial.

4.4 Mesures de sécurité

Dans le cadre des Services IA, Scaleway met en œuvre des mesures de sécurité pour garantir la confidentialité et l'intégrité des Contenus du Client. Le trafic entre le Client et le Service est chiffré via le procédé TLS, garantissant la protection des données en transit.

Scaleway ne conserve aucune requête ni aucun Contenu généré après le traitement. Les données du Client, y compris les prompts, complétions et données d'entraînement :
- ne sont pas utilisées pour entraîner, réentraîner ou améliorer les Modèles de base ;
- ne sont pas accessibles par les fournisseurs de Modèles LLM ;
- ne sont pas utilisées pour améliorer le Service ;
- ne sont pas accessibles par les services tiers.

Les Inputs et les Ouput traités lors de l'inférence sont considérés comme des données appartenant au Client. Scaleway ne journalise pas ces données. Toutefois, en cas de détection d'une activité malveillante ou qui a un impact sur la qualité du Service, Scaleway peut conserver temporairement le contenu de la requête HTTP afin d'identifier la cause racine (*rootcause*) ou une éventuelle vulnérabilité de sécurité.

Dans le cadre du Service Generative API, lorsque le mode Batch Processing est utilisé, les Input sont stockés uniquement pendant la durée du traitement, c'est-à-dire vingt-quatre (24) heures. Au-delà de cette durée, les données sont supprimées.

Le Client reconnaît que le Service se limite à la génération d'Outputs. Il appartient au Client de transférer et de conserver ces données vers un service de stockage externe tel que le Service Object Storage. A ce titre, il demeure exclusivement responsable de la gestion du

cycle de vie de ses données (et notamment de leur sauvegarde, conservation, suppression ou archivage).

4.5 Utilisation prohibée des Services

Dans le cadre de l'utilisation des Services, Scaleway assure la fourniture et la gestion de l'infrastructure nécessaire à l'exécution des Modèles tandis que le Client reste responsable de l'utilisation qu'il fait des Services et de la gestion de ses données. A ce titre, le Client s'engage notamment à :

- respecter la réglementation applicable en matière d'intelligence artificielle et notamment le règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle en tant que « Fournisseur » ou « Déployeur » au sens du règlement précité ;
- ne pas contourner les Quotas ou limitations techniques des Services ;
- ne pas utiliser les Services à des fins illicites, détournées ou prohibées par les présentes Conditions Particulières et plus largement par les Conditions Générales de Services Scaleway ;
- ne pas créer, générer ou faciliter la création de codes malveillants, de logiciels malveillants, de virus informatiques ou entreprendre toute autre action qui pourrait surcharger, interférer ou nuire au bon fonctionnement et à l'intégrité des Services, ou d'un système informatique ;
- ne pas utiliser les Services en y intégrant des contenus contrefaisants et/ou illicites ;
- ne pas promouvoir, contribuer, encourager, inciter ou utiliser les Services à des fins violentes, diffamatoires, offensantes, illégales, discriminatoires, racistes ou inappropriées, et toutes autres activités pénalement répréhensibles sous quelque forme que ce soit.

Scaleway se réserve le droit de suspendre les Services en cas de comportement abusif, fraude, surconsommation, ou violation des présentes Conditions Particulières.

ARTICLE 5. DURÉE ET RÉSILIATION

5.1 Durée

Les présentes Conditions Particulières entrent en vigueur à compter de la souscription du Client aux Services IA et restent applicables toute leur durée d'utilisation par le Client.

5.2 Résiliation

Sous réserve de toute durée minimum d'utilisation ou durée d'engagement applicable, la résiliation des Services IA peut être opérée à tout moment à l'initiative du Client depuis la Console de Gestion de Compte ou par les biais décrits dans les Conditions Générales de Services. Au terme de ladite résiliation, la facturation cesse de prendre effet.

La suppression d'un déploiement est une action irréversible qui efface toutes les données et ressources associées.

ARTICLE 6. CONDITIONS FINANCIERES

6.1 Service Generative API

Le Service est généralement facturé sur la base d'un modèle « à l'usage » (pay-as-you-go). Chaque tranche d'un million de Tokens « Input » ou « Output » donne lieu à facturation au tarif en vigueur selon le Modèle concerné. Toutefois, en fonction du type de Modèle, il peut exister d'autres méthodes de facturation (et notamment la facturation à la seconde) indiquées au sein de la Console de Gestion de Compte et/ou de la Documentation que le Client s'engage à consulter.

A la fin de chaque mois, une facture est émise à l'adresse indiquée au sein de la Console de Gestion de Compte et réglée par le Client selon le moyen de paiement indiqué par ce dernier.

Le Client peut consulter la consommation de Tokens par l'intermédiaire du Service Cockpit de Scaleway et y accéder depuis la Console de Gestion de Compte.

6.2 Managed Inference

La facturation débute uniquement à compter du moment où le déploiement est opérationnel et peut faire l'objet de requêtes. Le Service Managed Inference est facturé par heure et/ou par mois, en fonction du type de ressources provisionnées pendant la durée du déploiement indépendamment de l'utilisation ou non de la ressource sous-jacente. Le Client reconnaît qu'il lui revient de consulter les conditions financières propres à chaque Modèle.

A la fin de chaque mois, une facture est émise à l'adresse indiquée au sein de la Console de Gestion de Compte et réglée par le Client selon le moyen de paiement indiqué par ce dernier.