SPECIFIC CONDITIONS APPLICABLE TO AI SERVICES

Version of 07/04/2026

These specific conditions of service (hereinafter the "Specific Conditions") are entered into between, on the one hand any natural or legal person subscribing to the AI Services offered by Scaleway (hereinafter referred to as the "Client") and, on the other hand, the Scaleway contracting entity, as designated in the General Terms and Conditions of Service (hereinafter referred to as 'Scaleway

The purpose of these Specific Conditions is to define the terms and conditions applicable to the subscription and use of AI Services by the Client.

The Specific Conditions supplement the provisions of the General Terms of Service. In the event of any contradiction between the General Terms of Service and the Specific Conditions, the provisions herein shall prevail.

ARTICLE 1. DEFINITIONS

In the context of these Specific Conditions, the following words or expressions shall have the meaning set forth below:

- **AI Services**: means, within the framework of these Specific Conditions, to the following services dedicated to artificial intelligence: Generative API and Managed Inference.
- **Generative API** : means the AI Service provided by Scaleway.
- **GPU Instance**: means a graphical processing unit used for image processing and calculation.
- **Input**: means the data (personal or otherwise) transmitted by the Client (text, instruction, audio, image, video) via the AI Service in order to obtain an Output.
- **Managed Inference**: means the AI Service provided by Scaleway for which an infrastructure is dedicated to the Client, the characteristics of which are detailed in these Specific Conditions.
- **Model**: means a pre-trained artificial intelligence model made available as part of the AI Service via an API, capable of performing text and chat generation, image analysis, or audio transcription
- **Output**: means the results generated by the AI Service based on the Client's Input (text, image file, etc.).
- **Tokens**: means the basic units of text or image that a generative model processes. Depending on the tokenization method specific to the Model, a Token may represent a word, part of a word, a punctuation mark, or a character. The number of processed Tokens determines (i) the Model's analysis capacity (context window) and (ii) the billing basis for the Service in accordance with the price list in force.
- **Quota**: means the maximum usage frequency thresholds for the Services, defined per unit of time.

Unless the context clearly implies otherwise, words in the singular include the plural and vice versa, reference to one gender includes other genders, reference to a natural person includes legal entities, associations, etc., and vice versa, and related words have corresponding meanings.

Article headings in these Specific Conditions are for convenience only and shall not affect the interpretation of the provisions herein.

If a definition is not mentioned in this article, words or expressions beginning with a capital letter shall have the meaning given to them in the General Terms of Service.

ARTICLE 2. DESCRIPTION OF SERVICES

2.1 General information

The Client may subscribe, via its Account Management Console, by API, by Command Line Interface (CLI), by the Software Development Kit (SDK), as well as via the Terraform provider, to one or more types of AI Services, including:

- Generative API ;
- Managed Inference.

2.1.1 Generative API

The Generative API Service provides access to AI Models via an API. This Service allows the Client to submit Input (prompts, text, audio files, etc.) to generate, through automated processing, an Output (text, prompts, image files, classifications, etc.).

Scaleway manages the infrastructure and the deployment of Models within a multi-tenant (shared) environment. To guarantee the overall stability of the Service, Scaleway implements monitoring and load regulation mechanisms. The Client acknowledges that this multi-tenancy implies no fixed performance commitment or guaranteed throughput. In the event of peak activity, Scaleway reserves the right to activate rate-limiting mechanisms to preserve the integrity of the Service.

For large volumes of requests, it is recommended to redirect queries to "Batch Processing" or to use dedicated infrastructures such as the Managed Inference Service. Indeed, under Batch Processing, Scaleway handles task scheduling and orchestration to optimize both resource allocation and processing times. Execution will be completed within a period of less than twenty-four (24) hours.

APIs offered under the Generative API Service are designed to be compatible with OpenAI® libraries and SDKs, including the OpenAI® Python client library and LangChain SDKs. This compatibility is provided for information purposes only and may evolve, particularly due to changes made by the relevant third-party editors, without Scaleway being able to guarantee full, permanent, or exhaustive correspondence with these libraries and SDKs.

2.1.2 Managed Inference

The Managed Inference Service refers to the deployment and execution of Models within a dedicated and isolated computing environment for the Client. The Service allows the Client to query, via an API, either Models provided by Scaleway or third-party Models imported by the Client. The processing of Inputs and the generation of Outputs take place on private resources, ensuring the Client constant performance and complete logical resource isolation from other Scaleway clients.

In this context, the Client can configure options (such as GPU Instance type, Model size, memory) according to its use cases, needs, and compatibility requirements.

2.2 Third-party Solutions

As part of the AI Services, Scaleway provides third-party Models and undertakes to comply with licenses and any usage policies within its scope of responsibility, and to make available information provided by Model providers when available. The Client agrees to review the terms of the license governing the use of the selected Model. The Client acknowledges that the terms of use for these Models may contain usage restrictions. In accordance with industry standards (such as OpenAI® API compatibility), Scaleway uses its best efforts to ensure compatibility with most AI ecosystems and tools by default.

2.3 Maintenances

Scaleway may perform maintenance operations on the various infrastructures used to provide the Services to ensure they remain in operational condition.

Notwithstanding the foregoing, the Client remains responsible for reporting any Technical Incident impacting the use of AI Services to Technical Support. Furthermore, the Client is responsible for taking all measures to protect against risks inherent in maintenance operations, regardless of their origin, notably (but not limited to) by performing regular backups of their content.

ARTICLE 3. SUBSCRIPTION CONDITIONS

Prior to any Service subscription, the Client agrees to review all information made available by Scaleway, including the Documentation, to ensure that the selected AI Services is suitable for its needs and use cases. Depending on the type of AI Services selected, it is the Client's responsibility to be technically compliant with the Documentation and/or to select the technical characteristics at the time of subscription.

Activation and maintenance of Services usage rights are subject to (i) the creation of a valid Client Account, (ii) identity verification ("KYC"), and (iii) validation of payment methods. Scaleway reserves the right to suspend access to the Service in the event of non-compliance with any of these verification steps.

ARTICLE 4. CONDITIONS OF USE

4.1 General information

In the context of the Generative API Service, the Client may use preconfigured serverless access points for AI models without the need to configure the underlying infrastructure.

In the context of the Managed Inference Service, the Client is responsible for configuring the network parameters applicable to each deployment. The Client has the following options:

- Private Network: The Client may activate the private network to ensure secure communication and availability restricted solely to the internal environments of private networks.
- Public Network: The Client may activate access via a public network to allow access to resources from the public Internet. In this case, access to exposed resources is protected by an IAM Service API key by default.

Scaleway reserves the right to apply any security measures or technical constraints necessary to maintain the integrity, confidentiality, and availability of the network infrastructure associated with these Services.

The Managed Inference Service allows the Client to deploy Models provided to them or to deploy its own Models. For this purpose, Scaleway provides the Client, within its Documentation, with a catalog listing all Models that can be deployed by the Client as part of the Service.

In the context of the Generative API Service, the Client may under no circumstances increase the maximum number of Output Tokens beyond the limits set for each Model. These limits are intended to protect the Client against:

- Excessive generation times likely to lead to request interruption due to HTTP response timeout. Limits are designed to ensure that a model sends its HTTP response within a period of less than five (5) minutes;
- Risk of uncontrolled billing, as some Models may enter infinite generation loops, where generated content repeats uninterruptedly based on specific prompts. If the Client requires a larger number of Output Tokens, it is recommended to subscribe to the Managed Inference Service.

Scaleway does not guarantee that these results are error-free, fully reliable, relevant, or compliant with the Client's expectations.

Scaleway provides the Client, in its Documentation applicable to the Services, with a Shared Responsibility Model ("SRM") specifying the respective roles, scopes of intervention, and responsibilities of Scaleway and the Client. The Client acknowledges awareness of these models and agrees to consult and consider them when using the Services.

4.2 Quotas

In the context of the Generative API Service, the Client may be subject to certain Quotas. Exceeding these thresholds may result in temporary suspension of the Service or performance

limitations. The Quota level is determined based on: (i) the selected Model, (ii) prior validation of a payment method, and (iii) prior identity verification of the Client (KYC).

Each request is subject to a context window limit, i.e., limited by the amount of data (e.g., Input length) that the Model can process at one time.

To request an increase in these Quotas, the Client may contact Technical Support.

Conversely, the Managed Inference Service is not subject to a Quota on the number of generated Tokens, as the Service provides dedicated capacity, with Outputs being limited by the size of the resource selected and deployed by the Client.

4.3 Model Lifecyles

For the Generative API Service, Scaleway regularly updates Models with new official versions via the Account Management Console. Models may therefore be subject to different statuses, such as: Preview, Active, Deprecated, or EOL.

Depending on the status of the Model concerned, the Generative API Service may or may not benefit from a Service Level Agreement (SLA) and associated Support Services:

- **Preview**: Models in Preview status are available for testing. They benefit from one (1) month of support and a one (1) month notice period before the removal of a Model in Preview mode.

- **Active**: Models in Active status benefit from support for at least eight (8) months from their launch. A three (3) month notice period is provided before the Model moves to Deprecated or EOL status.

- **Deprecated**: Models in Deprecated status have a newer version of the Model available. A Deprecated status means an EOL status is planned. Although Models in Deprecated status remain usable, Scaleway recommends migrating to an Active version.

- **EOL**: Model is no longer accessible. Prior to the EOL status, the Client receives notice from Scaleway warning that the Model will no longer be available via the Generative API Service. However, some Models may continue to be used for a longer period via the Managed Inference Service. If an equivalent Model is available, Scaleway reserves the right to redirect traffic to this alternative Model to avoid any service interruption, although there is no guarantee that generated Outputs will be similar to those of the initial Model.

4.4 Security Measures

As part of the AI Services, Scaleway implements security measures to ensure the confidentiality and integrity of Client Content. Traffic between the Client and the Service is encrypted via TLS, ensuring data protection in transit.

Scaleway does not retain any requests or generated Content after processing. Client data, including prompts, completions, and training data:

- is not used to train, retrain, or improve the base Models;
- is not accessible by LLM Model providers;
- is not used to improve the Service;
- is not accessible by third-party services.

Inputs and Outputs processed during inference are considered data belonging to the Client. Scaleway does not log this data. However, in the event of detected malicious activity or activity impacting Service quality, Scaleway may temporarily retain the content of the HTTP request to identify the root cause or a potential security vulnerability.

In the context of the Generative API Service, when Batch Processing mode is used, Inputs are stored only for the duration of processing, i.e., twenty-four (24) hours. Beyond this duration, data is deleted.

The Client acknowledges that the Service is limited to generating Outputs. It is the Client's responsibility to transfer and store this data to an external storage service such as the Object Storage Service. As such, the Client remains exclusively responsible for managing the lifecycle of its data (including backup, retention, deletion, or archiving).

4.5 Prohibited Use of the Services

In the context of using the Services, Scaleway ensures the provision and management of the infrastructure necessary to run the Models, while the Client remains responsible for its use of the Services and the management of its data. As such, the Client specifically undertakes to:

- comply with applicable regulations regarding artificial intelligence, notably Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024, laying down harmonized rules on artificial intelligence, as a "Provider" or "Deployer" within the meaning of said regulation;
- not circumvent the Quotas or technical limitations of the Services;
- not use the Services for purposes that are unlawful, diverted, or prohibited by these Specific Conditions and, more broadly, by the Scaleway General Terms of Service;
- not create, generate, or facilitate the creation of malicious code, malware, or computer viruses, nor undertake any other action that could overload, interfere with, or harm the proper functioning and integrity of the Services or any computer system;
- not use the Services by integrating infringing and/or illegal content;
- not promote, contribute to, encourage, incite, or use the Services for purposes that are violent, defamatory, offensive, illegal, discriminatory, racist, or inappropriate, or for any other criminally punishable activities in any form whatsoever;

Scaleway reserves the right to suspend the Services in the event of abusive behavior, fraud, overconsumption, or violation of these Specific conditions.

ARTICLE 5. TERM AND TERMINATION

5.1 Term

These Specific Conditions shall come into force upon the Client's subscription to the AI Services and shall remain applicable for the entire duration of its use by the Client.

5.2 Termination

Subject to any applicable minimum period of use or commitment period, the termination of the AI Services may be carried out at any time at the Client's initiative via the Management Console or through the means described in the General Terms of Service. Upon the effective date of said termination, billing shall cease.

The deletion of a deployment is an irreversible action that erases all associated data and resources

ARTICLE 6. FINANCIAL CONDITIONS

6.1 Generative API Service

The Service is generally billed based on a "pay-as-you-go" model. Each increment of one million "Input" or "Output" Tokens is subject to billing at the current rate for the relevant Model. However, depending on the type of Model, other billing methods may exist (notably per-second billing) indicated within the Management Console and/or the Documentation, which the Client undertakes to consult.

At the end of each month, an invoice is issued to the address indicated within the Management Console and settled by the Client according to its specified payment method.

The Client can monitor Token consumption through Scaleway’s Cockpit Service, accessible from the Management Console.

6.2 Managed Inference

Billing begins only once the deployment is operational and ready to receive requests. The Managed Inference Service is billed per hour and/or per month, depending on the type of resources provisioned for the duration of the deployment, regardless of whether or not the underlying resource is actually used. The Client acknowledges that it is its responsibility to consult the specific financial conditions for each Model.

At the end of each month, an invoice is issued to the address indicated within the Management Console and settled by the Client according to its specified payment method.