Einstein Requests

Rate Card

Some of Salesforce's generative AI Services and features consume Einstein Requests. Salesforce Services may have included a quantity of Einstein Requests, with the specific entitlement indicated in the Usage Details table on the Order Form for such Services. In other cases, generative AI features may have been enabled within Services that do not include an entitlement of Einstein Requests, in which case, use of those generative AI features will consume entitlements included with other Services purchased by the customer.

Each API call made through the LLM gateway to a Large Language Model (LLM) consumes Einstein Requests. Usage calculations include:

- a usage type multiplier associated with the LLM used, and
- an API call size factor associated with the size of the API call.

Type of LLM	**Usage Type**	**Multiplier** Sandbox & Production Environments
Bring Your Own Large Language Models	Starter Prompts	4
Salesforce-enabled foundational LLMs	Basic Prompts	4
	Standard Prompts	10
	Advanced Prompts	38

Usage Types and multipliers may be updated from time to time.

Use of generative AI features may also consume Data Cloud credits as well as Flex Credits or Conversations.

Use of any pilot features in conjunction with the Services may consume Einstein Requests, to the extent expressly referenced in the applicable pilot terms agreed to by Customer for such pilot.

The API call size factor is calculated by:

- adding the size of the prompt request plus the size of the LLM-generated response, in tokens
- dividing by 2,000 tokens, and rounding up to the next highest integer.

Examples:

- An API call (prompt request plus LLM response) with 10,000 tokens has an API call size factor of 5.

- An API call with 8,001 tokens has an API call size factor of 5.
- An API call with 10,001 tokens has an API call size factor of 6.

To calculate Einstein Request usage for a given API call, multiply its API call size factor by the applicable usage type multiplier.

Alternative way to represent Einstein Request formula:
Einstein Requests per LLM API Call = Round up (Total input and output tokens metered by the LLM provider in the LLM API call/2000) * (4-10 for Salesforce-managed models and 4 for BYO-LLM)

Consumption examples:
- An API call of the Starter usage type with 1,000 tokens will consume 4 Einstein Requests
- An API call of the Standard usage type with 3,500 tokens will consume 20 Einstein Requests.

To learn more about how Einstein Requests are consumed, and when other types of credits are consumed with generative AI features, please visit: Agentforce and Generative AI Usage and Billing in Salesforce Help.

Note on Tokens, Words, and Files: In language models, the smallest processing units are tokens. A single word can be composed of multiple tokens. For example, the word "fantastic" might be split into multiple tokens like "fan," "tas," and "tic." On average, 2,000 tokens are roughly equivalent to 1,500 words. When processing document files like PDFs, the token count depends on the text content and can vary significantly based on the document's length and complexity. For images, vision-language models can process both text and visual information, converting image content into a format that can be tokenized and analyzed alongside text. The relationship between file sizes and token counts is not linear, as factors such as text density, formatting, and image complexity affect the conversion to tokens. Generally, a 100-page text-heavy PDF document might contain around 50,000 to 75,000 tokens, while a single high-resolution image could be represented by several thousand tokens in a vision-language model.

Updated as of October 24, 2025. Technical restrictions may apply.