Business

Appendix to the Service Description

Live Intelligence – LLMs

1 LLMs (February, 6 2026)

The table below lists the AI models (Large Language Models, or LLMs) currently available or soon available through Live Intelligence as of the date of this document.

It also outlines the corresponding calculation method for unit consumption.

- For each user request (comprising both the question and the answer) units or partial units are consumed. The total unit consumption is calculated by adding together the following components, as detailed in the table below: **Input Processing Cost:** Units consumed for processing the user's input.
- **Web Search Cost (if applicable):** Additional units consumed if a web search is performed.
- **Output Processing Cost:** Units consumed for generating the output.

Editor	LLM	Location of LLM processing	Cost of processing 1000 input tokens (questions + contexts)	Additional cost of processing a request including a web search	Cost of processing 1000 output tokens (responses)
OpenAI	**GPT-4.1**	**European Economic Area and Switzerland**	2.74 units	14,00 units	10.97 units
OpenAI	**GPT-4.1-mini**	**European Economic Area and Switzerland**	0.55 units	14,00 units	2.19 units
OpenAI	**GPT-4.1-nano**	**European Economic Area and Switzerland**	0.14 units	*Not available*	0.55 units
OpenAI	**GPT-5-mini**	**European Economic Area and Switzerland**	0.35 units	14,00 unités	2.74 units
OpenAI	**GPT-5**	**European Economic Area and Switzerland**	1.72 units	14,00 unités	13.71 units
OpenAI	**GPT-5.1 (expected availability : end of february 2026)**	**European Economic Area and Switzerland**	1.72 units	14,00 unités	13.71 units
OpenAI	**o3**	**European Economic Area and Switzerland**	13.71 units	*Not available*	54.86 units

OpenAI	**o3-mini**	**European Economic Area and Switzerland**	1.51 units	*Not available*	6.03 units
Google	**Gemini 2.0 Flash**	**Europe-west4 (Netherlands)***	0.21 units	*Not available*	0.82 units
Google	**Gemini 2.5 Flash**	**Europe-west1 (Belgium)***	0.42 units	*Not available*	3.42 units
Anthropic	**Sonnet 3.7**	**Europe-west1 (Belgium)***	4,11 units	*Not available*	20,57 units
Anthropic	**Sonnet 4**	**Europe-west1 (Belgium)***	4,11 units	*Not available*	20,57 units
Anthropic	**Sonnet 4.5**	**Europe-west1 (Belgium)***	4,11 units	*Not available*	20,57 units
Mistral AI	**Small 3.1 (expected availability : end of february 2026)**	**Europe-west4 (Netherlands)***	0,14 units	*Not available*	0,41 units
Mistral AI	**Medium 3 (expected availability : end of february 2026)**	**Europe-west4 (Netherlands)***	0,55 units	*Not available*	2,74 units
Mistral AI	**Codestral 2 (expected availability : end of february 2026)**	**Europe-west4 (Netherlands)***	0,41 units	*Not available*	1,23 units

Calculations of Units consumption examples:

- A query using GPT-4.1-mini with 1300 input tokens and 250 output tokens, without web search, consumes:
  (1300/1000) x 0.55 + (250/1000) x 2.19 = 1.2625 units

- A query using GPT-4.1 with 4000 input tokens, with web search enable, and 300 output tokens, consumes:
  (4000/1000) x 2.74 + 14 + (300/1000) x 10.97 = 28.251 units

Update history:

- **28/01/2026:** Add Mistral AI models (Small 3.1, Medium 3, Codestral 2), Add OpenAI AI model (GPT-5.1), remove OpenAI AI models (GPT-4o, GPT-4o-mini).
- **05/01/2026:** Addition of details regarding the calculation of Unit consumption (usage quota). Adding information regarding the Sovereign Space (LLM and usage quota consumption calculation).
- **12/05/2025:** addition of Anthropic AI models. Addition of location information for treatments.
- **08/10/2025:** First version of the document (extract from the Service Description in the form of an appendix).