

Anthropic announced that its Claude Sonnet 4 model can now process the context with a volume of up to 1 million tokens. These are approximately 750 thousand words or 1800 A4 sheets. The function is already available in beta version through the API Anthropic and on the Amazon Bedrock platform, and will soon appear in the Google Vertex AI.

Previously, Sonnet 4 and the senior model of the OPUS 4 could work with only 200 thousand tokens. For most tasks, this was enough, but competitors – Google and Openai – have already offered the support of a million tokens.

A large volume of context allows you to analyze entire code bases, combine large sets of documents and create Jeenets who “remember” the work course even after hundreds of operations.

However, the new mode has a price: the processing of volumes above the old limit costs twice as much – \$ 6 per million input tokens against the previous \$ 3, and 50% more expensive for the weekend tokens. Anthropic advises using caching and package processing to reduce costs.

About whether the senior model of the OPUS 4 will receive the same opportunity, has not yet been reported.