Google has published a rare analysis of the influence of AI on energy consumption based on real data from its data centers. The company monitors the energy used by the processors, AI-accelerators, memory, as well as cooling and water. Carbon dioxide emissions are also taken into account in the production of equipment. Only user energy, network and models are not taken into account.

According to Google, one text request through their AI uses only 0.24 watt hours, selects 0.03 g $CO_2$ and consumes 0.26 ml of water-about the same as 9 seconds of TV viewing in the United States.

The main achievement is optimization. The use of the MIXTURE-OF-Experts approach allows you to activate only part of the model, reducing the load by 10-100 times. The company also creates its own AIS-accelerators and optimizes their work.

These measures, as the company said, allowed to reduce energy consumption for a request by 33 times a year. Even taking into account a huge number of requests, the total load on the ecology is now much less.