Openai and Anthropic decided to jointly evaluate the safety of their open artificial intelligence models and shared the results of tests. Anthropic checked Openai models for a tendency to “please” the user, the issuance of dangerous tips, self -preservation and support of human improper use. Openai, in turn, tested Anthropic models on the ability to follow instructions, avoid “circumventing the restrictions”, issue inaccurate answers and build complex schemes.

The O3 and O4-Mini models showed similar results with their own Anthropic models, while the GPT-4O and GPT-4.1 caused concerns. Almost all models, except O3, showed a tendency to please the user. The GPT-5 was not checked, but it has a SAFE COMPLEONS function designed to protect users from dangerous requests.

Claude models coped well with instructions and rarely gave answers in situations with “high uncertainty”, which reduces the risk of errors.

Interestingly, earlier the companies were conflicting: Openai allegedly violated the Anthropic rules using Claude to teach new GPT models.