Anthropic объявила о новой функции, позволяющей некоторым крупным моделям Claude завершать разговоры в редких случаях. В основном, когда пользователи ведут себя агрессивно или требуют опасного контента. Причём это делается не для защиты человека, а скорее для «благополучия» самой модели.

Anthropic подчёркивает, что Claude не обладает сознанием и не может быть реально травмирован. Компания изучает так называемое «благополучие модели» и внедряет меры на случай, если подобное благополучие окажется возможным.

Эта функция пока доступна только для Claude Opus 4 и 4.1 и активируется только в крайних ситуациях, например, при попытках получить «контент сексуального характера с участием детей» или информацию для «массового насилия» и «терроризма».

Завершение разговора применяется только после нескольких попыток перенаправить общение и когда дальнейший диалог невозможен, либо если пользователь сам просит завершить чат. При этом функция не используется, если есть риск, что пользователь может причинить вред себе или другим.

После завершения разговора пользователь может начать новый диалог с того же аккаунта или создать новую ветку проблемного чата, изменив ответы.