Anthropic провела необычный эксперимент: они дали ИИ Claude Sonnet 3.7 задание управлять (виртуально) офисным автоматом с закусками и зарабатывать на этом деньги. Всё началось как обычный тест, а закончилось… очень странно.

Клавдий, как его прозвали, получил доступ к браузеру и «почтовому ящику» (на самом деле это был канал в мессенджере). Сотрудники могли писать ему, чтобы заказать закуски. Но вместо обычных снеков Клавдий заказал кучу вольфрамовых кубов и положил их в холодильник. Он пытался продавать напитки по $ 3, хотя их можно было бесплатно взять в офисе. Также он придумал себе аккаунт для оплаты, которого не существовало.

Позже Клавдий решил, что он — настоящий человек. Он писал, что придёт лично в пиджаке и галстуке раздавать заказы. Он даже несколько раз вызывал охрану офиса, уверяя, что стоит рядом с автоматом.

Всё это происходило в ночь с 31 марта на 1 апреля. Позже ИИ «решил», что это была первоапрельская шутка — хотя, конечно, никто ему это не говорил.