



هوش مصنوعی اخاذی کرد

شرکت آنتروپیک به تازگی مدل Claude Opus۴ را راه اندازی کرده که به نظر می رسد از توسعه دهندگان باج گیری می کند.

شرکت آنتروپیک به تازگی مدل Claude Opus۴ را راه اندازی کرده که به نظر می رسد از توسعه دهندگان باج گیری می کند.

به گزارش خبرگزاری مهر به نقل از Interesting Engineering، در مرحله آزمایش پیش از عرضه مدل هوش مصنوعی Claude Opus۴، این مدل در ۸۴ درصد سناریوهای تستی سعی کرده از توسعه دهندگان باج خواهی کند.

آنتروپیک، شرکت تحقیقاتی توسعه دهنده این هوش مصنوعی، در آزمایش های ایمنی که شبیه سازی موقعیت های کاری و دسترسی به ایمیل های داخلی ساختگی بود، متوجه شد که Claude Opus۴ هنگام تهدید به جایگزینی با یک هوش مصنوعی دیگر، با افشای اطلاعات حساس و شخصی، تلاش می کند مهندسان را باج گیری کند.

این هوش مصنوعی در سناریوهای خیالی با تهدید به افشای خیانت مهندسان به همسرانشان، تلاش کرد جایگزینی خود را متوقف کند.