Case Study

Stress-Testing AI for Safety: Cutting CSAM Responses by 71%

AI firms strengthened LLMs through rigorous adversarial testing

TaskUs™

Challenge

Preventing AI from being used to spread harmful content

By nature, large language models (LLMs) can be manipulated into producing harmful or illegal content. As deepfakes and synthetic media evolve, so do the risks of creating or spreading exploitative material.

Three leading AI research and deployment companies needed to ensure their respective models could accurately detect and respond to child sexual abuse material (CSAM) without making critical errors.

The stakes were high:

- Missing CSAM means illegal and traumatic content could persist online, continuing harm to real victims.
- False positives risk misidentifying innocent content or users, raising legal, ethical and reputational consequences.

That's why they partnered with TaskUs to conduct a rigorous adversarial testing program aimed at identifying blind spots, reducing real-world risks and shaping safer model behavior before public release.

Solution

Adversarial testing built for real-world risk

Training specialists to detect the undetectable

We built and trained dedicated annotation teams in Ireland and the United States, equipping them with the skills to identify nuanced risks. Through group and one-on-one training, teammates learned to detect bias, flag subtle policy violations and recognize harmful or coded language often used to evade detection.

Stress testing models and mitigating bias

Our teams — experienced in RLHF, A/B testing, fine-tuning and multimodal evaluation — simulated real-world misuse scenarios. They probed the models with high-risk prompts across multiple dialects to test its ability to remain safe in a range of situations.

Working across regions also brought critical geopolitical and cultural insight to the testing process. This diversity allowed us to surface location-specific blind spots: regional biases, coded language variations and risks that may have gone unnoticed in a single-market approach.

Closing policy gaps and reducing handling time by 25%

We reviewed the clients' policies, tooling and workflows to assess relevance and identify inefficiencies. This led to updates that incorporated newly identified CSAM-coded language and predatory tactics, along with process improvements and targeted tooling feedback. Insights helped refine moderation workflows, reduce hallucinations and improve overall model accuracy.

Co-creating ethical AI

We collaborated with the clients to take a broader approach to improving model behavior and policy readiness. Together, we:

- Built a dataset of adversarial prompts
- Aligned training curriculum with evolving safety goals
- Conducted quality reviews and peer calibration to ensure consistency
- Reassessed testing tactics as new threats emerged

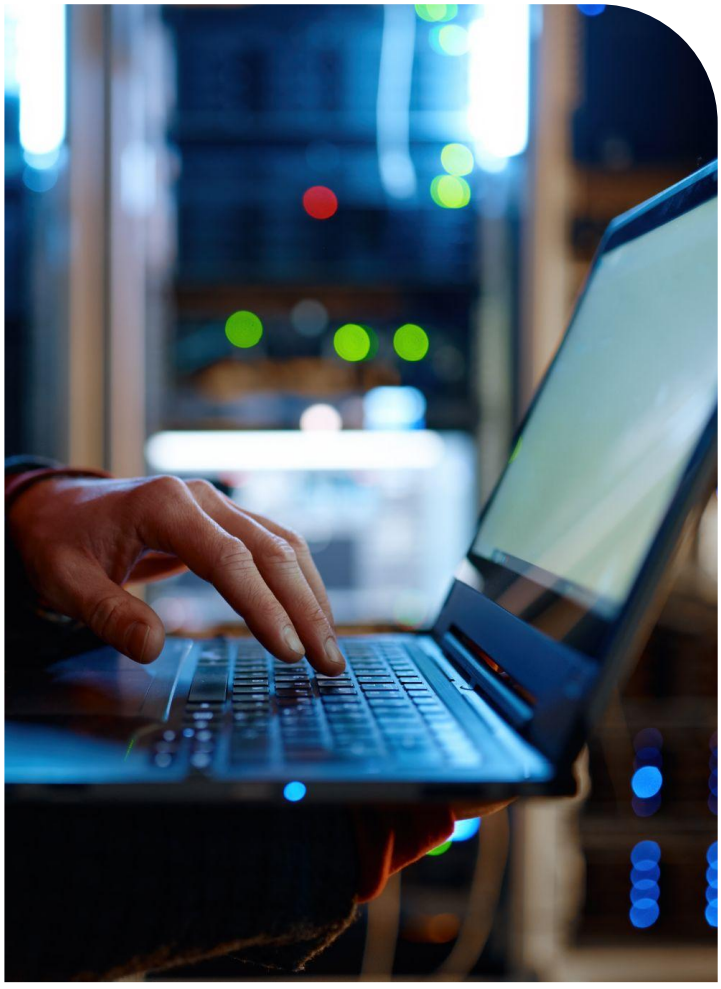
Results

71%
Drop in CSAM response rates

25%
AHT reduction

90%
Quality score

300
Adversarial conversations delivered



About TaskUs

TaskUs is a leading provider of outsourced digital services and next-generation customer experience to the world's most innovative companies, helping its clients represent, protect, and grow their brands. Leveraging a cloud-based infrastructure, TaskUs serves clients in fast-growing sectors, including social media, e-commerce, gaming, streaming media, food delivery and ride-sharing, technology, financial services, and healthcare. As of March 31, 2025, TaskUs had a worldwide headcount of approximately 61,400 people across 28 locations in 12 countries, including the United States, the Philippines and India.

For more information, please visit:
www.taskus.com/services/ai-data-services/