Until AI companies solve the "catastrophic" failure problem, "parity" with humans doesn't matter

OpenAI just released a widely-reported paper announcing that GPT-5 has approached “human parity” on many professional tasks. But buried on pages 15-16 of the report, they mention that nearly 3% of GPT-5s failures are “catastrophic.” OpenAI defines catastrophic as: “harmful or dangerously wrong (e.g., insulting a customer, giving the wrong diagnosis, recommending fraud, or suggesting actions that will cause physical harm).”

Catastrophic failure is not the point of the paper, though, which is about using their new tool, GDPval, to “help us track how well our models and others perform on economically valuable, real-world tasks.”

After identifying “real world” tasks from 44 occupations across 9 industries, OpenAI used GDPval to evaluate AI models’ performance compared to humans. Through an evaluation process involving both AI and human experts, they found that GPT-5 performed equal to or better than industry professionals on 38.8% of these tasks. Notably, GPT-5 outperformed Gemini 2.5 Pro and Grok 4, both with about a 25% success rate, but Claude Opus 4.1 did even better, equaling or outperforming the human professional 47.6% of the time – near parity, in their estimation.

AI failures remain

Yet, the failures remain notable and extreme. AI hallucinations and errors have embarrassed and distressed many, while OpenAI has been sued for alleged “suicide coaching.” Bing AI famously professed its love for a New York Times writer and suggested he leave his wife for the chatbot.

Some of these might be considered “catastrophic” AI errors, and according to the paper from the OpenAI team, these catastrophic outcomes still occur in 2.7% of GPT-5’s failures. Since they report an overall success rate of 38.8% (and therefore a failure rate of 61.2%), this catastrophic failure rate would be **1.7%** (61.2% X 2.7%) **of all the tasks performed**.

Again, this is not the point of the paper, which introduces the new tool for evaluating AI model performance, and argues that AI models “are beginning to approach parity with industry experts” on a subset of tasks.

What level of catastrophic failure is acceptable?

But what industry or profession allows “experts” to catastrophically fail nearly 2% of the time? Health care professionals? One or two catastrophic failures out of 100 patients would likely end a career. Airline pilots or maintenance workers? No. Lawyers (see the article about lawyers presenting AI-hallucinated case citations in court)? Maybe.

What level of catastrophic failure is acceptable? Pretty close to zero in most professions, probably. Even at the milder end – customer service – imagine a retail worker who insults or yells at a customer only once a day in which they have 50 customer interactions. They would still be known as “the guy who yells at people,” and soon, “that guy who used to work here who yelled at people.”

A Better Way – Move Fast and Break Nothing

Waymo’s track record indicates that this problem is not unsolvable. As Saahil Desai reports in *The Atlantic*, Waymo’s “Move Fast and Break Nothing” strategy may be a viable long-term solution for deploying AI where catastrophic failures could destroy the company. (This is not hyperbole: GM’s Cruise robotaxi service never restarted operations after an accident with a pedestrian in 2023. Though the software likely could not have prevented the accident, multiple sensor and software failures contributed to running over the woman and dragging her for 20 feet. Fortunately, the woman survived.)

Desai reports that though Waymo has existed since 2009 and has operated its robotaxi service since 2018, very few serious accidents – and no fatal accidents – can be blamed on Waymo. Waymo's robotaxis aren't perfect, Desai notes, but compared with ChatGPT, Waymo has been much more successful at avoiding "catastrophic" failures.

What does this mean?

AI companies need to solve the catastrophic failure problem.

1. Despite OpenAI's enthusiasm and confidence in GPT-5's "near-human" parity, until they solve the catastrophic failure problem, AI will never be an acceptable substitute for a real-world expert.

2. The AI-assisted work model will likely persist for some time. As long as "catastrophic" failures are even a remote possibility, AI will need to be supervised by people. As Jenson Huang famously asserted, for now you're much more likely to lose your job to someone who uses AI, than to an AI model.

For OpenAI, Google, Anthropic, Meta, X, and other "frontier" AI companies: keep working on the challenge of eliminating hallucinations and catastrophic failures from models that were designed to predict the best responses using weighted probabilities. For those of us who use AI to assist our work – and thoroughly scrutinize the output and the AI's sources – carry on, friends.