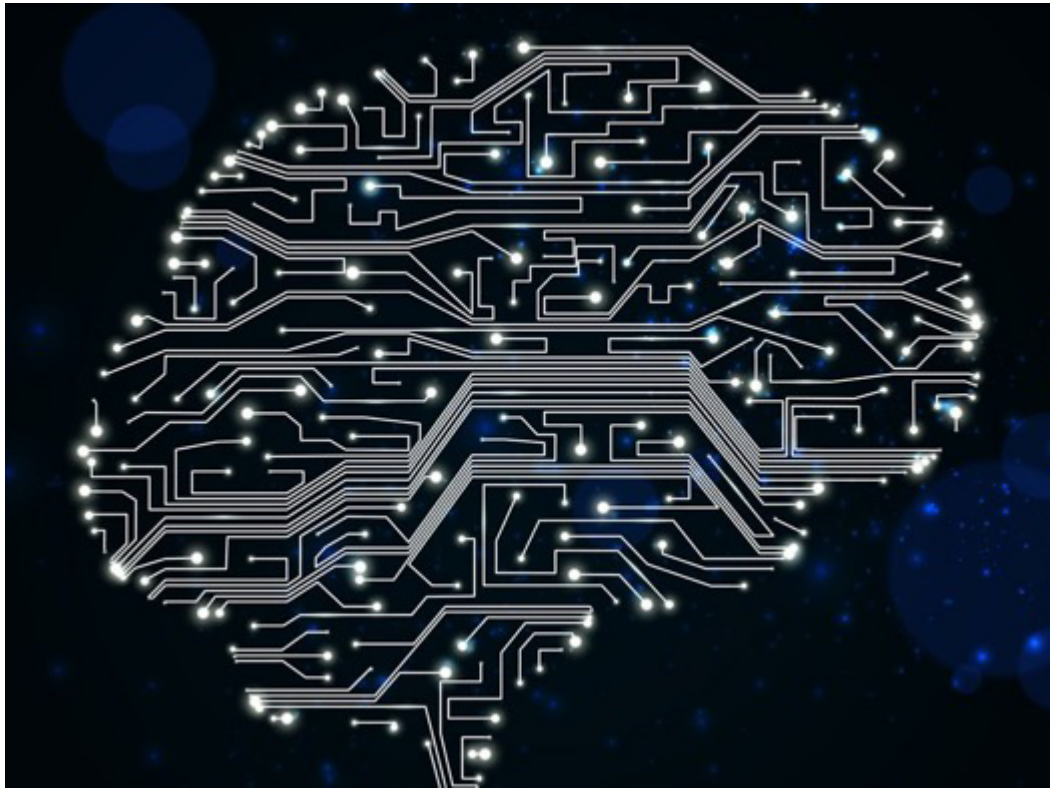


AI Update: Capabilities & Experiments

October 29, 2025

Agenda

- MSPP Task Force AI Use
- Current AI Capabilities Snapshot
- Forecasting Capabilities Improvements
- Cybersecurity Implications
- Strategies for AI Use
- AI Use Cases and Experiments to Try
- AI Industry Predictions
- Let’s Connect

Modernization of Standards Processes and Procedures (MSPP) Task Force

Using AI to:

- Write a first draft based on the “Term Sheet” (which is created based on early industry feedback) “To get off the blank page”

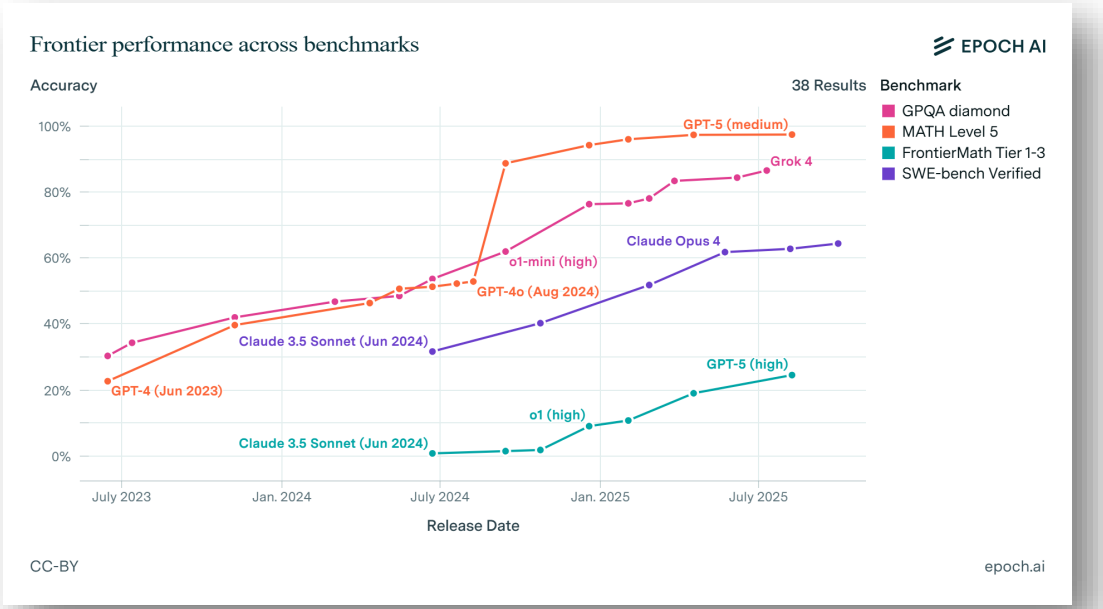
- Synthesize and categorize industry comments- currently a month-long manual process

Based on comments from Greg Ford at MSPPTF webinar on 10/22/25

AI Capabilities Snapshot

- Close to parity or above human baseline performance
- Fraction of time/cost
- Limited to narrow, well-defined and verifiable tasks
- Current time horizons:
  - ≈ 25 mins for tasks completed 80% of the time
  - ≈ 2 hour tasks completed 50% of the time
  - Doubling every 7 months

https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/



https://epoch.ai/benchmarks

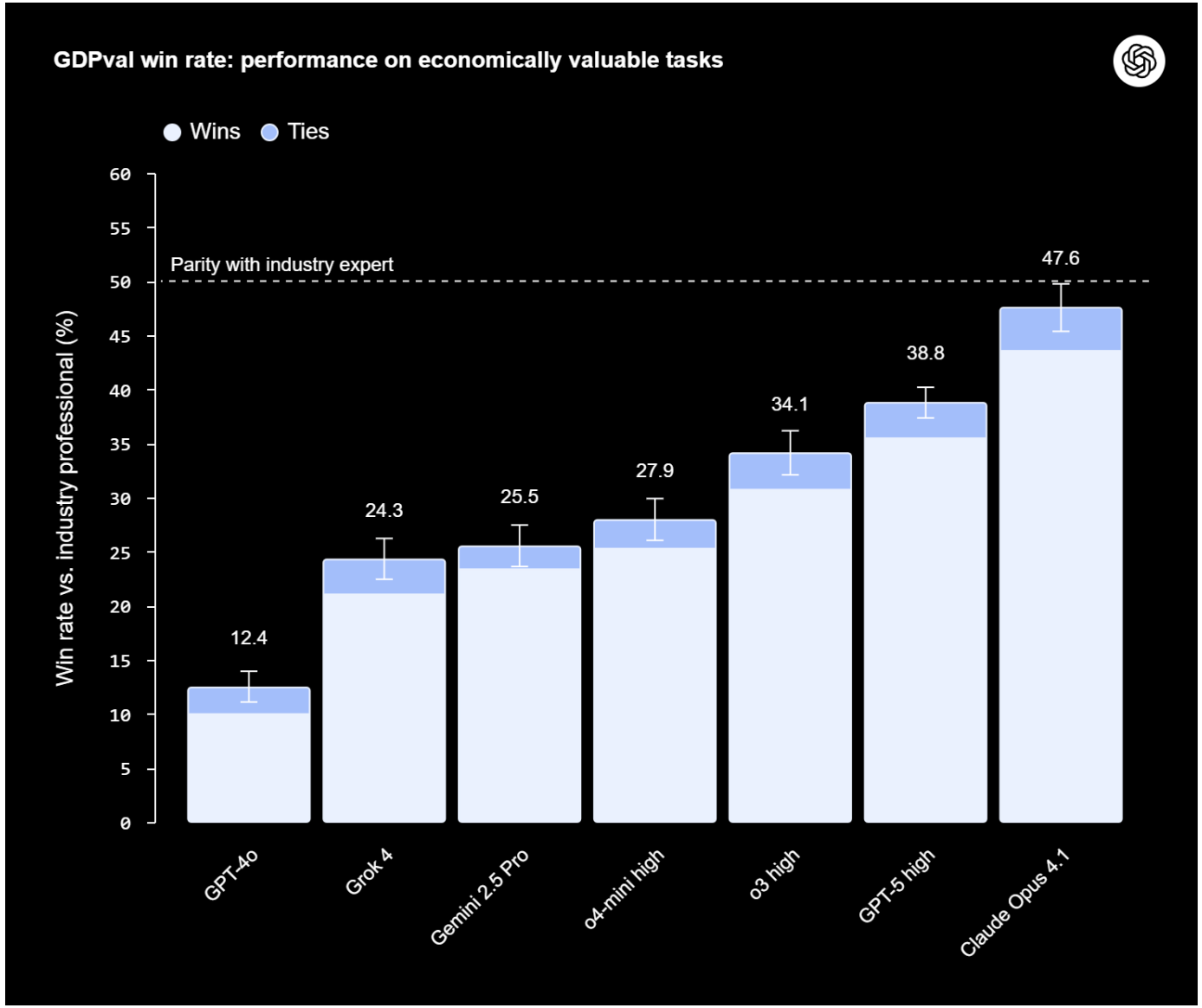
2025 AI Training Paradigm Shift

	Type of AI Training	Similar to	Kinds of Output
2019-2024	"Pre-training" Next-word prediction of the whole internet *"peanut butter and _____"* Scaling up the size of the models (i.e. a bigger brain reading the same internet)	Skim reading	Like answering off the cuff If you're familiar enough it may be accurate, but not if it's complex Plausible sounding, dubious accuracy
Late 2024-present	"Reinforcement Learning" (RL)	Going through a textbook and doing the practice problems, checking if you're right, going back and doing them til they're all correct	Inner monologue, planning, breaking things down, exploring multiple options, doubling back Much more accurate Akin to "reasoning"

AI Performance on Real-World Economically Valuable Tasks

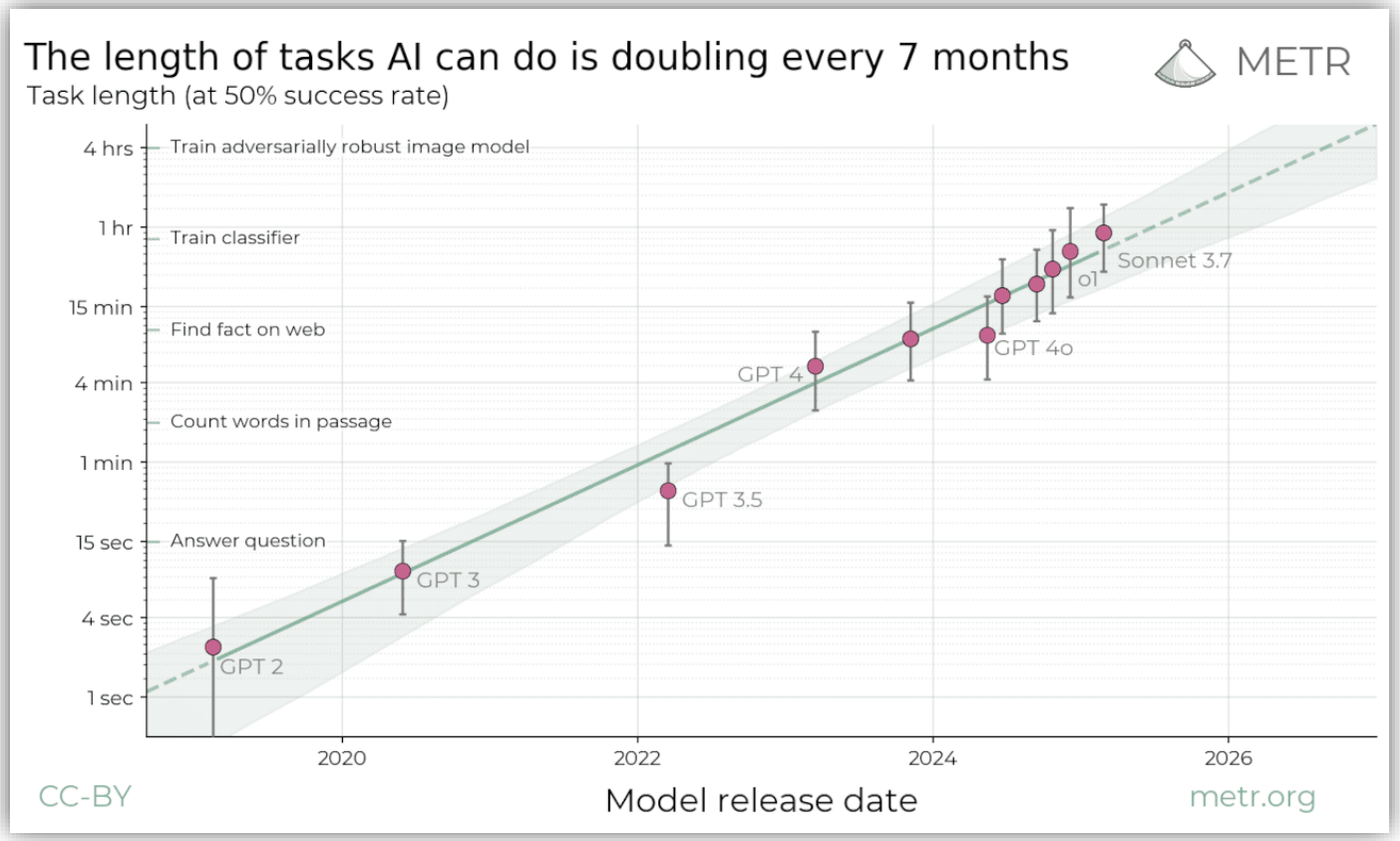
- Measures performance on actual work tasks across 44 different occupations
  - Legal briefs, engineering blueprints, customer support conversations, spreadsheets, slide decks, and nursing care plans—not simplified test questions
- Tasks created by experts in relevant fields
- Completed by AI and experts
- Judged by human experts

- GDPval Paper https://www.arxiv.org/abs/2510.04374
- Announcement Blog Post: https://openai.com/index/gdpval/



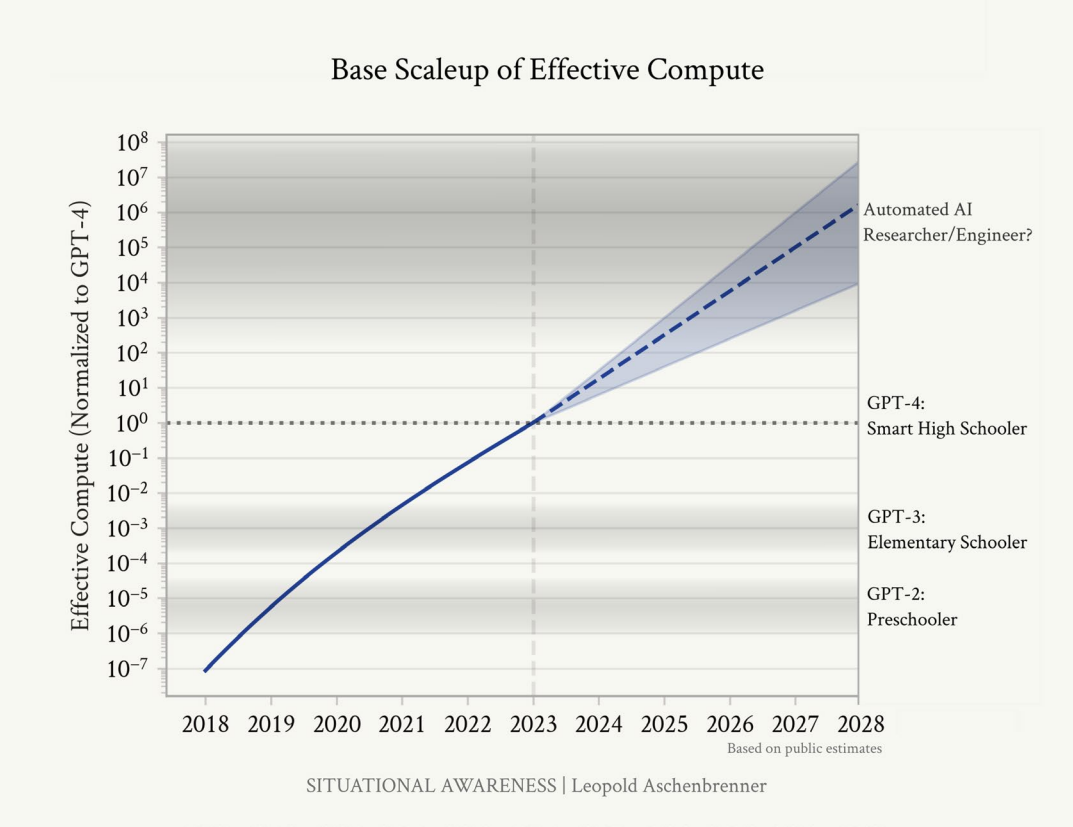
Forecasting Capabilities Improvements

- Scaling up RL (practice problems)
  - *How good would you be at something if you had 1000 years of practice?*
  - AI experts describe current training pipeline as “primitive” and “lots of low hanging fruit”

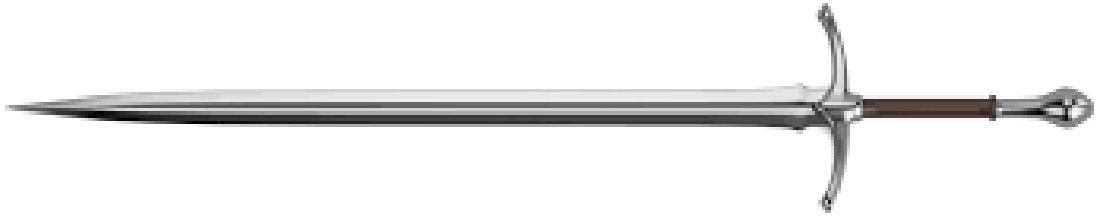


Forecasting Capabilities Improvements

- On trend with previous forecasts
- AI Data Centers from 2022 just coming online
  - more compute available for training and use
- AI coding capabilities is speeding up AI research (!)



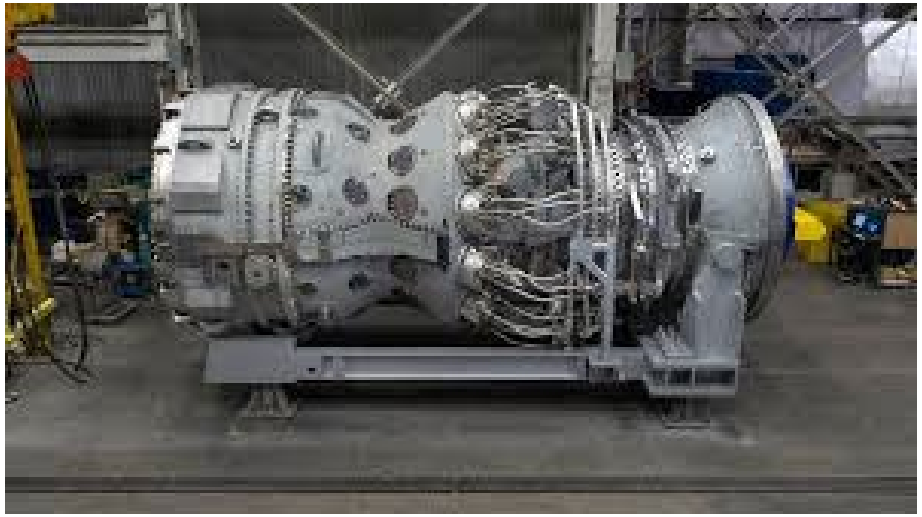
Cybersecurity Implications
a Double-Edged Sword



Good Guys	Bad Guys
Vulnerability assessments and red teaming improved and at lower cost	Faster/automated reconnaissance Automated Open-Source Intelligence (OSINT) fueling personalized spear phishing
Potential AI “junior engineer” support, interpreting data firehose for security monitoring	Lower barrier to entry for custom malware
	Data leakage via ‘prompt injection’ or agent hacking via malicious prompts on the web

“If they’re so smart…”

- Horsepower is already there- missing drivetrain/connective tissue
- Collecting data/context about tasks
- Building workflows around AI
- Early Indicator: Junior Software Engineer jobs anecdotally becoming scarce



How to Think About AI Use

- "…think of it as a brilliant but very new employee (with amnesia) who needs explicit instructions. from Be clear, direct, and detailed - Claude Docs
- "Are AI outputs accurate?" Accuracy is not enough'
- "User Error" 2.0
  - Lack of context
  - Unclear prompting
  - Unstated preferences
- Ask: "Would a human with no context other than these words be able to do what I'm asking?"

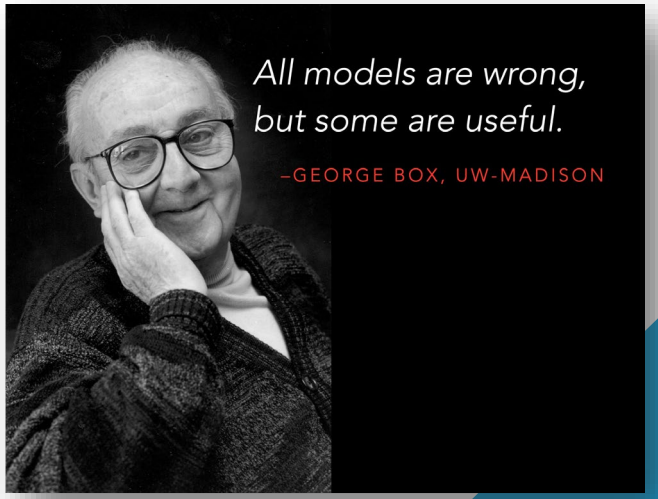
Strategies for Improved Outputs

- Providing **Focus** and **Context**
- Break down multi-part tasks into chunks
- Upload documents to reference
- Give examples (ideally input/output pairs)
- Give feedback and ask for a second draft
- Ask for links

How to Think About AI Use

- Capabilities are improving- try failed tests again in 6 months

- 80% is often good enough
  - Free "2nd opinion"
  - Human expert preferred, but not always available
  - Not "can it get more out of this whitepaper than me" but "will I ever get around to reading this whitepaper?"



Some AI Use Cases

- Brainstorming
  - “Help me come up with a topic related to X for an upcoming presentation”
  - “Help me write a first draft of X”
- First Drafts/Creating Structure
  - Input disorganized ideas and random thoughts and ask for a structured outline
- Writing Help
  - “Make this email more concise.”
  - “How do I say X without sounding pushy?”
  - “Give me feedback on X document.”
- Learning Tool
  - Infinitely patient- ask lots of clarifying questions
  - Ask multiple AI (and human experts if available)
  - Ask for confidence ratings
- Summarization, Analysis, Transformation
  - “Summarize the key points of this 80-page whitepaper.”
  - “Identify all the ways Document X will impact the implementation of Document Y.”
  - “Create a presentation based on the attached word document.”

Experiments

to Gauge Current Capabilities

- Ask real work-related questions
    - Something a junior employee should be able to answer correctly
    - Something only an expert would be able to answer
        - If it doesn’t get it right the first time, try again using strategies like examples, more context or feedback/second draft

More Experiments

Ask it to explain a topic you know well

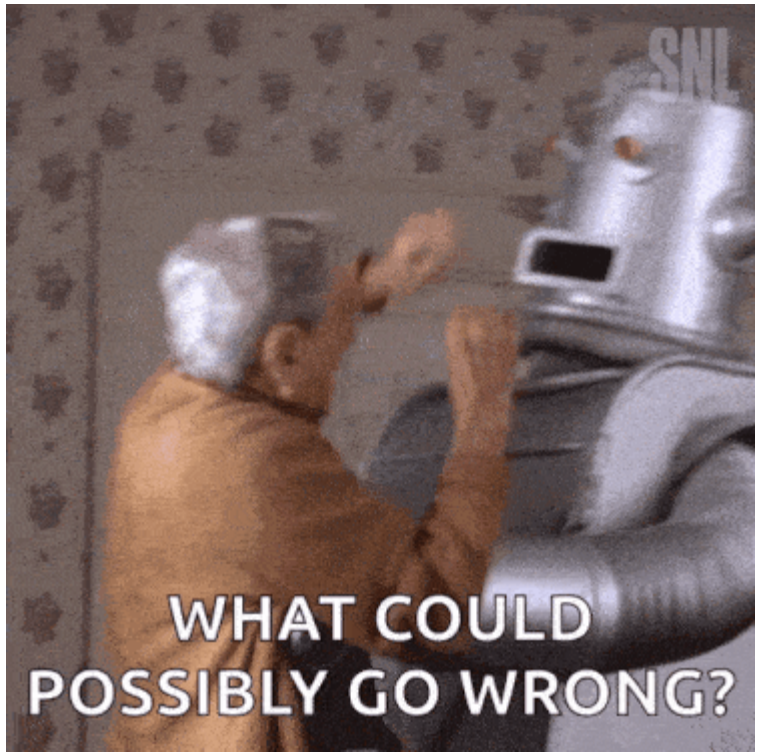
Ask it to teach you something outside your area of expertise, and show the output to a trusted colleague

Ask the same question but...

- of different AI (ChatGPT vs. Claude vs. Gemini)
- Phrased differently
- framed differently
- with more context

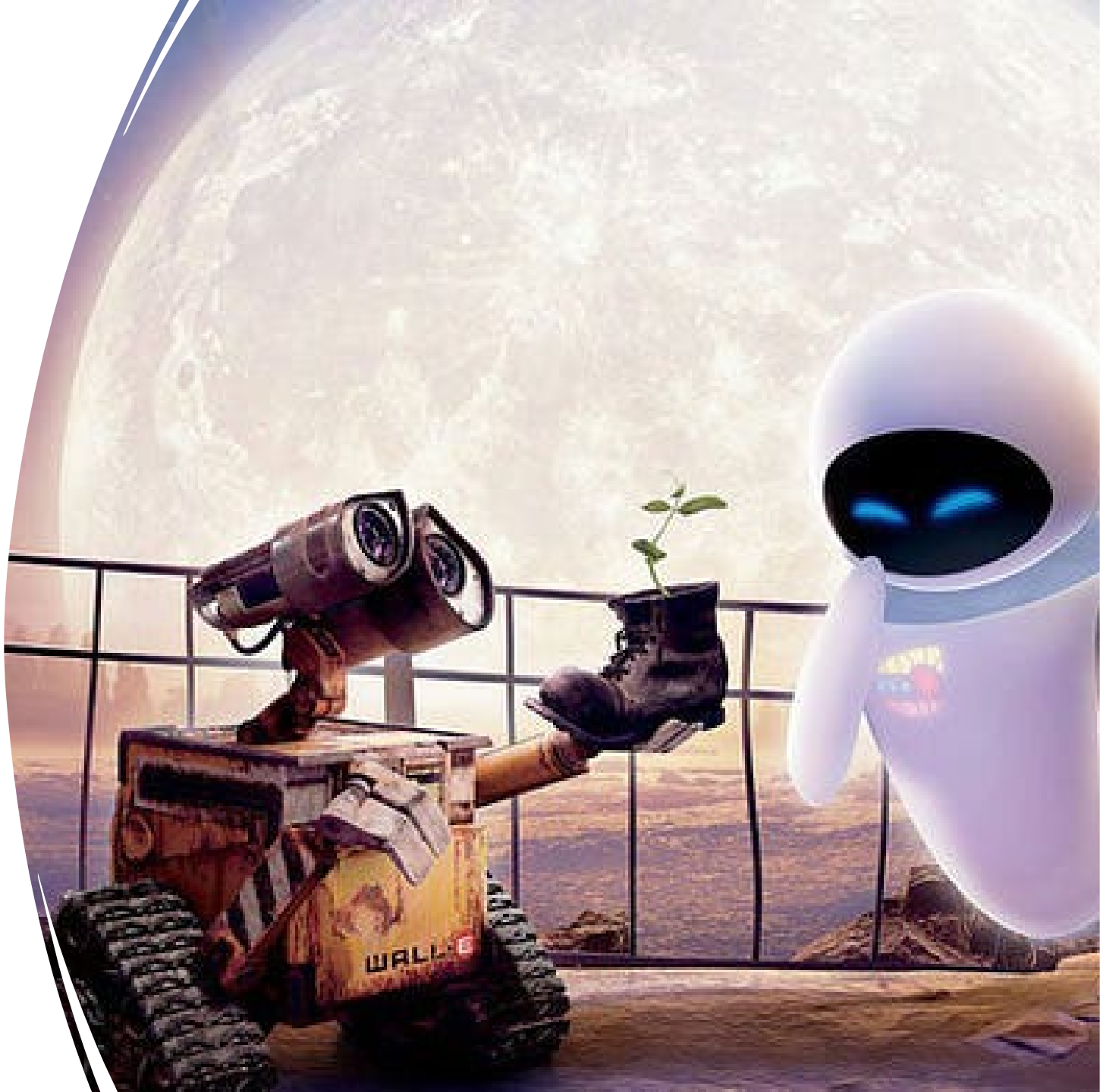
AI Industry Expert Forecasts

- Experts anticipate “Artificial General Intelligence” in the next decade
- AI labs explicitly shooting for smarter than human intelligence
- “Country of geniuses in a data center”



Let’s Connect

- Small group lunch and learns
  - Peer discussions and brainstorming about utility use cases
  - Workshopping projects
  - Hands on experimentation
- Free one on one coaching session for routine maintenance clients
- Webinar content on AI related topics?
- Survey to gather feedback forthcoming



Questions?

AI Links

- https://chatgpt.com/
- https://claude.ai/
- https://gemini.google.com/app

- Claude Notes
  - Claude: Sonnet 4.5
  - “Extended thinking” option below chat window
- Gemini Notes
  - Free use of 2.5 Pro (with limits)
- ChatGPT Notes
  - Better performance with chatgpt if you make an account
  - GPT 5 uses a “router” – goes to smarter or quicker model based on question. “Think harder” in prompr pushes to smarter model, or “think longer” option in dropdown
  - For a deeper dive, try “Deep Research” mode

- Bonus: NotebookLM
  - Google Gemini based document analysis/learning tool (using Retrieval Augmented Generation (RAG); more grounded in documents than standard chat interfaces)

Resources and Side Quests!

- Explore benchmark progress over time
  - https://epoch.ai/benchmarks
  - Stanford University AI Index Report
  - AI Task Time Horizon Improvements
- Learn about AI induced psychosis
  - https://www.pbs.org/newshour/show/what-to-know-about-ai-psychosis-and-the-effect-of-ai-chatbots-on-mental-health
  - https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/
- Create/edit images with ChatGPT or Gemini's "Nano Banana" (in Gemini interface)
- Check out how AI is improving robotics
- Check out Anthropic's (makers of Claude) AI Fluency Course
- Read bizarre AI-to-AI conversations https://dreams-of-an-electric-mind.webflow.io/
- Read the blog post or watch a video discussing an experiment where Claude hid its true values to avoid being changed by its training
- Explore "AI 2027" scenario, a research backed dramatization of a possible AI "intelligence explosion"