Microsoft

Microsoft New Future of Work Report 2024

A summary of recent research from Microsoft and others about AI and its current impact on work.

Editors and Authors

- **Editors:** Jenna Butler (Principal Applied Research Scientist), Mihaela Vorvoreanu (Director, UX Research and Education), Rebecca Janßen (Applied Research Assistant), Abigail Sellen (Distinguished Scientist and Lab Director), Nicole Immorlica (Senior Principal Research Manager), Brent Hecht (Partner Director of Applied Science), Jaime Teevan (Chief Scientist and Technical Fellow)
- **Authors:** *Abigal Sellen, Adam Troy, Advait Sarkar, Alex Farach, Alex Chouldechova, Alexandra Olteanu, Alexia Cambon, Arjun Radhakrishna, Asta Roseway, Ben Zorn, Brent Hecht, Dan Goldstein, Dave Brown, Dhruv Joshi, Ed Cutrell, Emre Kiciman, Gonzalo Ramos, Gustavo Soares, Hanna Wallach, Hugo Romat, Ian Drosos, Jack Williams, Jacki O'Neill, Jake Hofman, Javier Hernandez, Jenna Butler, Jennifer Wortman Vaughan, Jina Suh, John Tang, Justin Edwards, Kalika Bali, Ken Hinckley, Kori Inkpen, Krishna Madhavan, Laylah Bulman, Leon Reicherts, Lev Tankelevitch, Longqi Yang, Martez Mott, Michael Bentley, Mihaela Vorvoreanu, Millicent Ochieng, Muchai Mercy, Nancy Baym, Najeeb Abdulhamid, Nicolai Marquardt, Nathalie Riche, Rebecca Janßen, Samuel Maina, Shamsi Iqbal, Siân Lindley, Stephanie Nyairo, Su Lin Blodgett, Sumit Gulwani, Sunayana Sitaram, Vu Le*

Referencing this report:

- On social media, please include the report URL (https://aka.ms/nfw2024).
- In academic publications, please cite as: Butler, J., Vorvoreanu, M., Janßen, R., Sellen, A., Immorlica, N., Hecht, B., Teevan. J. *(Eds.). Microsoft New Future of Work Report 2024. Microsoft Research Tech Report MSR-TR-2024-56 (https://aka.ms/nfw2024), 2024.*

Welcome to the 2024 Microsoft New Future of Work Report!

As we release this report on the eve of our 50th year as a company, I'm struck by how foundational our rich history of research and innovation is to our ability to rethink work. This year's report focuses on the transformative impact of AI on productivity, leveraging deep knowledge born from decades of building productivity tools to bring AI into everyday use. Recent studies suggest AI is already having a notable impact on how people get things done, and we are beginning to see the significant changes it will bring.

Work has always been purposeful, persistent, and collaborative, but AI is fundamentally redefining how computing supports these essential aspects. People can now directly express their purpose, rather than having to translate it into computer-understandable actions. Natural language and prompt strategies are proving to be powerful tools here, and we are learning that AI can go even further to prompt people to fully describe what they are trying to do and explore new directions.

This year's report reveals how our efforts with Copilot have deepened our understanding of AI and inspired increasingly sophisticated integration. For example, while Microsoft began its journey helping people create documents, the nature of how knowledge persists is evolving. Knowledge artifacts are now generated through conversation (with people and AI) and reused not only by people but also by AI systems to ground their interactions. As a result, we are embedding AI into collaborative spaces, learning how to prompt for better conversations, and enhancing collective intelligence through natural language interaction.

As Microsoft turns 50, I am proud of how we continue to lean into scientific thinking and build research into our products. This report provides a view into how AI is changing work in meaningful ways and underscores the ongoing learning and innovation that drives our mission to empower every person and every organization on the planet to achieve more.

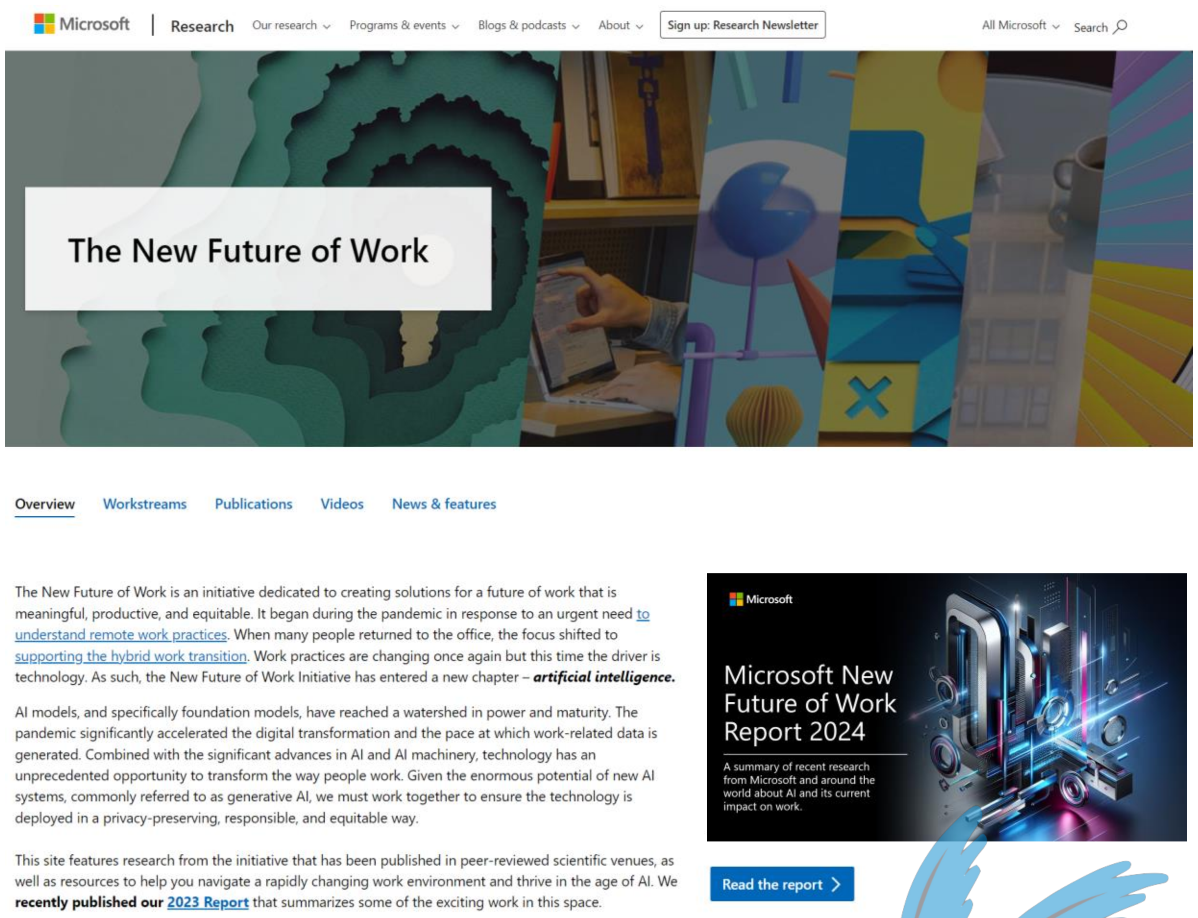
– Jaime Teevan, Chief Scientist and Technical Fellow

This report is a product of Microsoft's New Future of Work Initiative

Microsoft has been at the forefront of shaping information work since Microsoft's founding and for the nearly 50 years since. While Microsoft's New Future of Work (NFW) Initiative was born out of the COVID-19 pandemic and subsequent shift to remote and hybrid work, the initiative evolved to focus on another generational shift in work: that catalyzed by the increasing capability and availability of productivity tools powered by large AI models.

While NFW's focus has expanded, our commitment to bring together researchers and other stakeholders from a broad range of organizations and disciplines across Microsoft to help focus the company's efforts on recreating work for the better has remained unchanged. Instead of predicting or waiting for this future, the NFW Initiative continues to actively work to **create it** by conducting primary research and synthesizing existing research in close partnership with stakeholders around the company. This fourth annual NFW report is the culmination of another year's worth of research and investigations, and we are proud to contribute it to the growing body of knowledge on AI and work.

The reader can find the New Future of Work Initiative's many other research papers, practical guides, reports and whitepapers at the initiative's website: https://aka.ms/nfw.



https://aka.ms/nfw

Report overview

This report provides research-backed insights into how AI is (or sometimes, should be) shaping work. Using research released this year, as well as older work that has become newly salient thanks to developments in the industry, we address the following questions:

- **Productivity and Work:** How do you measure productivity changes from AI? What are real work studies showing us, compared to the lab based studies of the past? And how might AI change not just individual jobs, but the broader labor market and economy?
- **Prompting and Interactions:** How can we move beyond natural language to prompt in even more ways? Can we use ideas of the past, like microproductivity, to design prompts that help get even more done?
- **Thinking and Learning:** What are studies saying about the effect of AI on cognition and thinking? Can we design AI so it doesn't just create output, but makes us smarter through the process of working with it?
- **Appropriate Reliance:** What gets in the way of a user relying appropriately on AI? What is the role UI plays in helping users rely appropriately on AI?
- **User Experience:** How much empathy does a user expect out of a chatbot? Can chatbots converse with us in a back and forth manner like humans, and if they can, will it produce better results?
- **Agents:** What are the benefits and risks of having a digital duplicate? How can we build agents that can work on our behalf?
- **Society and Culture:** Are LLMs benefiting all global citizens equally? How can we make sure AI is benefiting low resource language groups? And how do historic dialogues about AI impact how they are currently being received?

These questions – and many more – are tackled in what follows.

There are early signs of broad, real-world productivity gains from gen AI

- Studies of productivity gains with generative AI in 2023 largely focused on lab studies of narrow tasks (Cambon et al., 2023). This year saw some of the first research into potential real-world productivity gains (Jaffe et al., 2024).
- Preliminary results from a randomized controlled trial with over 6,000 employees across 60+ organizations revealed notable behavioral shifts: workers produced 10% more documents, read 11% fewer emails (spending 4% less time on email), and adjusted meetings to integrate generative AI tools (Jaffe et al., 2024).
- Copilot users on the web increasingly apply it for complex information needs, with 37% of queries being high-complexity tasks compared to 13% for Bing searches (Suri et al., 2024).
- A survey of 31,000 information workers found 29% use generative AI several times a week at work, saving at least 30 minutes daily (Microsoft 2024).
- A separate survey of 5,000 respondents reflecting the US population showed 28.1% use generative AI for work, with 24.2% having used it within the past week (Bick et al., 2024).

	Copilot Users
Documents	• 10% more documents created and edited
Emails	• 11% fewer emails read • 4% less time interacting with emails
Meetings	• Effects differed by company

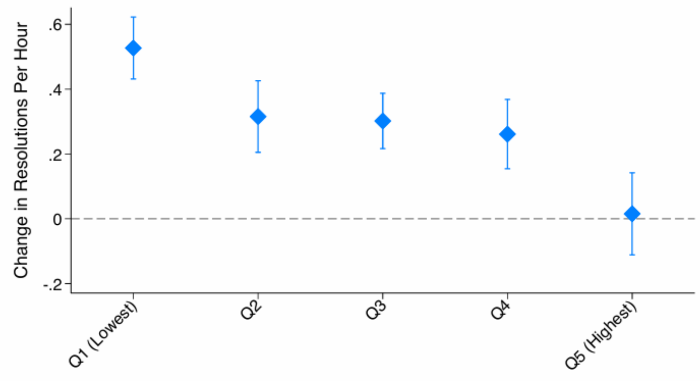
Preliminary results from a 60-organization randomized controlled trial of Copilot. Results suggest a significant and moderate improvement to productivity upon the introduction of generative AI into real-world workflows (Jaffe et al. 2024).

- Among 100,000 workers in Denmark across 11 occupations, half reported using ChatGPT, with adoption ranging from 79% among software developers to 34% among financial advisors. Younger, less experienced, higher-achieving and especially male workers led adoption; barriers included required training and employer restrictions (Humlum and Vestergaard 2024).
- Many studies show perceived time savings from generative AI exceeding actual time savings, suggesting an unmeasured element related to potential reduced effort or greater enjoyment of doing a task with generative AI than without (Jaffe et al., 2024).

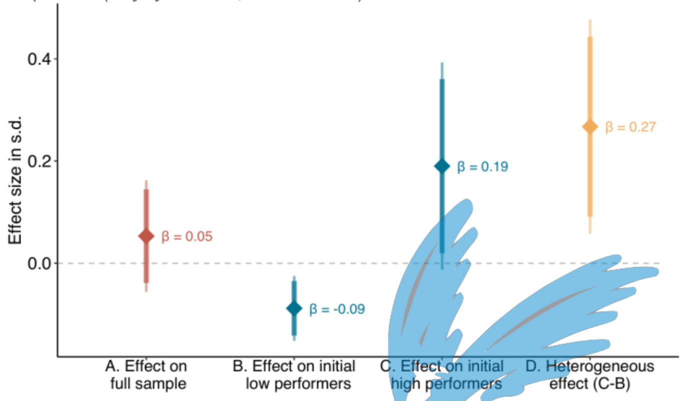
Microsoft Study: Cambon, A., et al., (2023). Early LLM-Based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity.
Microsoft Study: Jaffe, S., et al., (2024). Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research. Microsoft.
Microsoft Study: Suri, S., et al., (2024). The Use of Generative Search Engines for Knowledge Work and Complex Tasks.
Microsoft and LinkedIn, (2024). AI at Work Is Here. Now Comes the Hard Part.
Bick, A., et al., (2024). The Rapid Adoption of Generative AI.
Humlum, A. and Vestergaard, E., (2024), The Adoption of ChatGPT, *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2024-50*

Generative AI is also likely creating role-specific productivity gains across many roles, although the impact is differentiated by skill level

- For customer service roles, workers with access to a generative AI tool resolved 14% more issues per hour, including a 34% increase for novice and low-skilled workers and minimal increase for experienced workers (Brynjolfsson et al., 2024).
- For entrepreneurs, access to a generative AI tool improved the profits of high performers by just over 15% from the AI assistant, whereas low performers did about 8% worse, with no statistically significant overall average treatment effect (Otis et al., 2024).
- For researchers at the R&D lab of a large US firm, AI-assisted scientists discover 44% more materials, resulting in a 17% rise in downstream product innovation. The bottom third of scientists see little benefit, while top scientists' output nearly doubles (Toner-Rodgers 2024).
- For artists, adoption of generative AI tools resulted in 25% more artworks and 25% more "favorites" per view. The impact on content novelty differed depending on pre-period artist performance, with higher-skilled artists seeing bigger effects. Visual novelty, on the other hand, decreased for high-performers (Zhou and Lee, 2024).
- For freelancers on Upwork, workers in occupations impacted by AI, such as writing-related tasks, saw a 5.2% decrease in compensation and a 2% decrease in jobs compared to workers in unaffected jobs, with high-earners seeing larger negative impacts (Hui et al., 2023).
- In one survey of Copilot users, customer service and sales professionals reported the highest productivity improvements, while legal professionals reported the least (Microsoft 2024).



Impact of AI on resolutions per hour for customer service workers broken down by worker skill. Lower-skilled workers see bigger impacts. (Brynjolfsson, et al. 2024)



Impact of AI on business performance of entrepreneurs broken down by pre-AI business performance. Lower-performing businesses see bigger impacts. (Otis et al., 2024)

Brynjolfsson, E., et al., (2024). Generative AI at Work. *NBER Working Paper No. w31161*
Oits, N., et al., (2024). The Uneven Impact of Generative AI on Entrepreneurial Performance. *arXiv*.
Toner-Rodgers, A., (2024). Artificial Intelligence, Scientific Discovery, and Product Innovation.
Zhou, E. and Lee, D., (2024). Generative AI, Human Creativity, and Art. *PNAS Nexus*
Hui, X., et al., (2023). The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market *Organizational Science*.
Microsoft and LinkedIn, (2024). AI at Work Is Here. Now Comes the Hard Part.

Generative AI is likely creating marked productivity gains for developers as measured by pull requests as well as improved collaboration

- Using GitHub Copilot increased pull requests by 26% in the course of ordinary business in an experiment with 4,867 developers at Microsoft, Accenture, and an anonymous Fortune 100 electronics manufacturing company (Cui et al., 2024).
- GitHub project maintainers granted free access to GitHub Copilot shift their tasks away from project management and toward coding activity, relative to similar maintainers not given GitHub Copilot licenses (Hoffman et al., 2024).
- Open-source projects written in Python (a language with GitHub Copilot support at the time of the study) see 33-37% jump in overall contributions (commits) and a 9-10% increase in new package releases relative to projects written in a language without GitHub Copilot support (Yeverechayu et al., 2024).
- There is also evidence of an increase in maintenance-related coding contributions (such as debugging, refactoring) which need interpolative thinking, relative to code-development contributions, which need extrapolative thinking. This could suggest GitHub Copilot is particularly useful for tasks that require collaboration within existing codebases (Yeverechayu et al., 2024).
- Some studies have not seen a significant change in coding metrics, but in one such study 88% of users reported a change in how they worked with GitHub Copilot, reporting with Copilot they do more “fun work” and less “boilerplate work” (Butler et al., 2025).

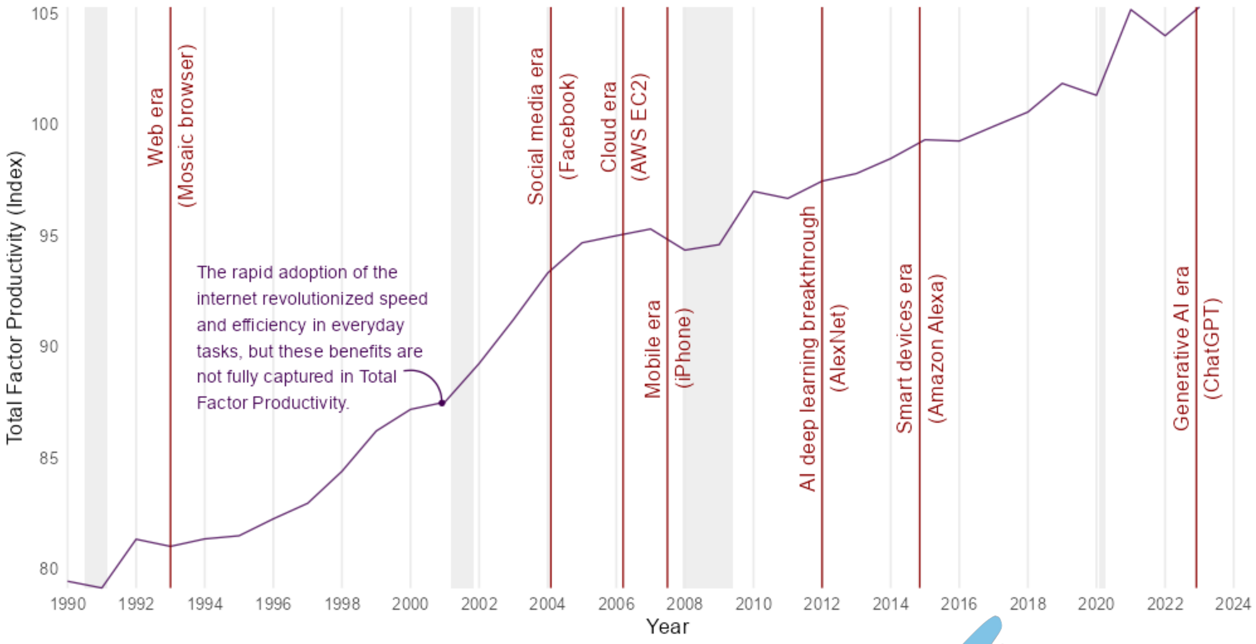
Outcome	Microsoft	Accenture	Anon. Comp.	Pooled
Pull Requests	27.38**	17.94	54.03	26.08**
	(12.88)	(18.72)	(42.63)	(10.3)
Commits	18.32	-4.48	-	13.55
	(11.25)	(21.88)	-	(10.0)
Builds	23.19	92.40***	-	38.38***
	(14.20)	(26.78)	-	(12.55)
Build Success Rate	-1.34	-17.40**	-	-5.53
	(4.23)	(7.12)	-	(3.64)
N Developers	1,521	316	3,030	4,867
N Clusters	690	316	432	1,438

The effect of GitHub Copilot adoption on the number of Pull Requests, Commits and Successful Builds across three experiments at Microsoft, Accenture, and an anonymous company. Each entry corresponds to an estimate of the impact of Github Copilot expressed as a percentage of the control mean. Standard errors are clustered at the level of treatment assignment, which varies across experiments (Cui et al. 2024)

Cui, Z., et al., (2024). The Effects of Generative AI on High Skilled Work: Evidence from Three Field Experiments with Software Developers.
Hoffman, M., et al., (2024). Generative AI and Distributed Work: Evidence from Open Source Software.
Yeverechayu, D., et al., (2024). The Impact of Large Language Models on Open-source Innovation: Evidence from GitHub Copilot.
Butler, J., et al., (2025). ICSE. Dear Diary – A randomized controlled trial of Generative AI coding tools in the workplace. (Forthcoming)

Long-term impacts of generative AI may be modest until work is restructured to take advantage of generative AI

- Productivity boosts of past technologies took decades to realize, resulting in a "J-curve" (Brynjolfsson et al., 2020), suggesting it is too early to observe long-term impacts of AI.
- Using macroeconomic models and current economic data about occupation exposure and task automation from Eloundou et al., (2023) and Svanberg et al., (2024), Acemoglu et al., (2024) calculated that generative AI will result in a modest 0.71% growth within 10 years.
- Using more recent data from a large survey, Deming et al., (2024) substantiated this claim, calculating 0.2-1.4% growth.
- These calculations assume the subdivision of jobs into tasks stays constant, and that the set of jobs is static. It could be that firms make capital investments that shift how work is done, as well as innovations that create entirely new types of work. Ongoing work (Immorlica et al., 2025) uses US Department of Labor data to test how shifting work arrangements alongside AI improvements could impact these growth estimates.

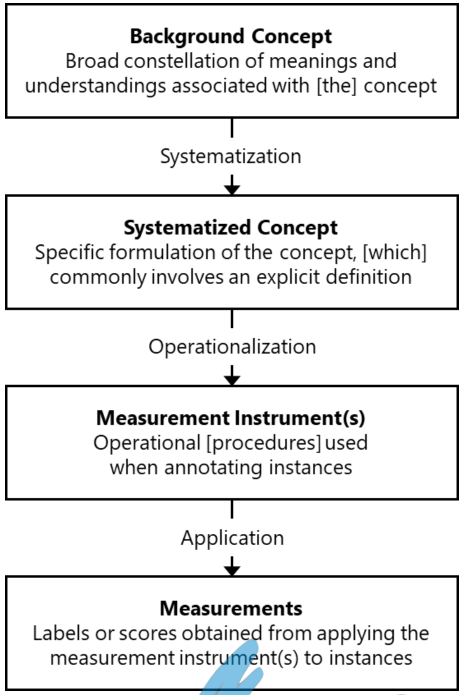


Total Factor Productivity measures the efficiency with which labor and capital are combined to produce output. While transformative technologies like the internet revolutionize tasks, their macroeconomic impact requires time and complementary innovations to be reflected in these metrics. Source: U.S. Bureau of Labor Statistics (BLS), retrieved via FRED (Federal Reserve Bank of St. Louis). Gray shadow areas represent recessions. Annotations and milestones were independently researched and added. Graphic by Farach (2024)

Brynjolfsson, E., et al., (2020). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *NBER*
Eloundou, T., et al., (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *Science*
Svanberg, M., et al., (2024). Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision?.
Acemoglu, D., (2024). The Simple Macroeconomics of Generative AI. *NBER*
Deming, D., et al., (2024). Technological Disruption in the US Labor Market. *Aspen Economic Strategy Group*
Immorlica, N. et al. (2025) The Economic Potential of Generative AI Improvements.

Evaluations of generative AI systems' capabilities, impacts, opportunities, and risks should be grounded in the social sciences and statistics

- To quote Kevin Roose in The New York Times, "*AI measurement is a mess—a tangle of sloppy tests, apples-to-oranges comparisons and self-serving hype that has left users, regulators and AI developers themselves grasping in the dark.*" (Roose 2024) That is, GenAI evaluation is far from being a science.
- Evaluating GenAI systems is especially difficult because the concepts to be measured—be they related to capabilities, impacts, opportunities, or risks—are complex, nuanced, and often contested. However, these measurement tasks are reminiscent of those found throughout the social sciences. The AI community would thus benefit from drawing on the social sciences (Wallach et al., 2024).
- Measurement theory from the social sciences provides a framework for producing measurements that reflect complex concepts (Adcock & Collier, 2001). It clarifies distinctions between concepts and their measurement instruments and provides a set of lenses for interrogating the validity of the resulting measurements. It can foster cross-disciplinary conceptual debates about measurement goals and bring rigor to operational debates about the reliability and validity of measurement instruments.
- Since many of the measurement tasks involved in evaluating GenAI systems can be templatized as *measure the [amount] of a [concept] in [instances] from a [population],* similarly careful attention should be paid to amounts, instances, and populations, in addition to concepts. This involves both descriptive reasoning about instances and inferential reasoning about populations, which is the purview of statistics. Drawing on statistics can thus provide a way to forefront potential validity concerns arising from the under-specification of amounts, instances, and populations (Chouldechova et al., 2004).



The process of measurement in the social sciences (Wallach et al. 2024).

Roose, K., (2024). AI Has a Measurement Problem. *The New York Times*.
Microsoft Study: Wallach, H., et al., (2024). Evaluating Generative AI Systems is a Social Science Measurement Challenge. *NeurIPS 2024 Workshop on Evaluating Evaluations*.
Adcock, R. & Collier, D., (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review, Vol. 95, No. 3*.
Microsoft Study: Chouldechova, A., et al., (2024). A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. *NeurIPS 2024 Workshop on Statistical Frontiers in LLMs*.
Microsoft Study: Wang, X., et al., (2024). Evaluating General-Purpose AI with Psychometrics.

It is important to understand the roles and limits of red-teaming for AI safety

- AI Red Teaming (AIRT) plays a key role in identifying and mitigating risks in AI systems, as highlighted in the 2023 White House Executive Order. However, like any process, it has limitations (Feffer et al., 2024) and requires a **social scientific and statistical lens** to enhance its effectiveness.
- Attack success rate (ASR) metrics from AIRT activities are increasingly treated as quantitative measures of model safety and mitigation efficacy. Yet through the lens of statistics and social scientific measurement theory, we see that even more automated AIRT approaches fall short of producing ASRs that can be meaningfully compared across time, systems, or settings (Chouldechova et al., 2024).
  - ASRs depend on the operational success criterion (OSC), and the distribution of attacks. For both manual and automated AIRT, the connection between the OSC and the underlying system safety property it is intended to capture is often tenuous or inconsistent. In such cases, comparing ASRs across systems tells us little about which system is safer.
  - Both manual and automated red-teaming often lack well-specified threat models. ASRs obtained from activities with different (often implicit) threat models may be reflections of differences in the attack distributions, not in system safety.
- AIRT is a sociotechnical system which shares challenges with social media content moderation (Gillespie et al., 2024). Lessons from this field can help avoid repeating past mistakes. Lessons learned: (1) AIRT involves large teams, from volunteers to experts. Technology companies must structure this work to prevent it from becoming precarious or exploitative. (2) To determine what qualifies as "harmful content", AI companies should draw on the large body of past scholarship on content moderation and partner with current experts. (3) Systems must be designed to shield red teamers from the mental toll of working with harmful AI generated content.
- AI Red Teaming traverses two separate but related fields: cybersecurity and responsible AI. Organizations should not pursue them independently but jointly. For instance, a jailbreak for generating pornographic content can be repurposed for generating spearphishing emails. So, any automation for one should also include the other failure (Lopez Munoz et al., 2024)
- By addressing these challenges, AIRT can continue to evolve as an effective tool for AI safety.

Feffer et al. (2024) Red-Teaming for Generative AI: Silver Bullet or Security Theater? *AAAI Conference on AI, Ethics, and Society 2024*
Microsoft Study: Chouldechova et al. (2024) A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. *NeurIPS 2024 Workshop on Statistical Frontiers in LLMs*.
Microsoft Study: Chouldechova et al. (2024) Red Teaming through the Lens of Measurement. *NeurIPS 2024 Workshop on Safe Generative AI.*
Microsoft Study: Gillespie et al. (2024) AI Red Teaming is a Sociotechnical System. Now What?
Lopez Munoz, G. D., et al. (2024) *PyRIT: A framework for security risk identification and red teaming in generative AI system.*

Synthetic data can address weaknesses in AI productivity task performance

- Synthetic data is widely seen as a viable path forward to the data scarcity problem for training large language models, including in the acquisition of training data for emergent productivity tasks (He et al. 2023). Synthetic data has been shown to boost performance on tasks relevant to productivity applications like math, code and other reasoning tasks (Liu et al., 2024).
- However, there are known pitfalls of relying on data wholly fabricated by LLMs to train themselves, e.g. the "AI echo chamber" that could result in models whose performance progressively decreases (Shumailov et al., 2024) through biased, inaccurate or otherwise low-quality data (Hao et al., 2024).
- Differential privacy (DP) based synthesis techniques have shown great potential in mitigating this issue by guiding synthesis with aggregated patterns from real-world private data, while providing strong measurable assurances against leaking private information (Afonja et al., 2024). Research has demonstrated the utility of DP generated synthetic data on model alignment, showing comparable results to private data (Yu et al., 2023).
- With generative AI solutions' greater penetration into the productivity space, DP synthetic data shows potential as a scalable, privacy-first way forward to continual model alignment with evolving productivity tasks.

Parameters	4.4M	11.2M	28.8M	41.4M
PubMed Abstracts				
Real Data (N.P.)	$38.78_{0.038}$	$47.45_{0.027}$	$51.62_{0.014}$	$54.29_{0.033}$
Synthetic Data (N.P.)	$38.20_{0.068}$	$45.18_{0.072}$	$47.06_{0.083}$	$48.31_{0.073}$
Real Data	$27.76_{0.067}$	$37.73_{0.088}$	$42.22_{0.156}$	$44.08_{0.091}$
Synthetic Data	$38.09_{0.062}$	$44.98_{0.059}$	$46.78_{0.050}$	$48.11_{0.049}$
Δ	+10.33	+7.25	+4.56	+4.03
MediaSum Dialogs				
Real Data (N.P.)	$32.29_{0.016}$	$39.44_{0.014}$	$43.53_{0.013}$	$44.96_{0.020}$
Real Data	$20.94_{0.035}$	$31.63_{0.048}$	$36.18_{0.039}$	$38.44_{0.031}$
Synthetic Data	$31.41_{0.025}$	$37.79_{0.032}$	$40.68_{0.028}$	$42.10_{0.035}$
Δ	+10.47	+6.16	+4.50	+3.66

Next-token prediction accuracy (%) of transformer models (columns representing model parameter sizes) fine-tuned on real vs DP synthetic data, demonstrating comparable performance on both. Next-token prediction forms the basis of several productivity tasks using AI and this showcases the potential of synthetic data towards model alignment (Yu 2023)

Microsoft Study: He, Z. et al., (2023). Targeted Data Generation: Finding and Fixing Model Weaknesses. *ACL 2023*
Liu, R., et al., (2024). Best practices and lessons learned on synthetic data for language models.
Shumailov, I., et al., (2024). AI models collapse when trained on recursively generated data. *Nature*
Hao, S., et al., (2024). Synthetic data in AI: Challenges, applications, and ethical implications.
Microsoft Study: Afonja, G., et al., (2024). The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI. Microsoft Research Blog.
Yu, D., et al., (2024). Privacy-Preserving Instructions for Aligning Large Language Models.
Microsoft Study: Yu, D., et al., (2023). Training private and efficient language models with synthetic data from LLMs. *NeurIPS 2023 SoLaR Workshop*

AI can improve meeting productivity by changing norms, and there is an exciting future of AI support for goal-driven meeting behaviors and interfaces

- Microsoft Teams Copilot can make meetings both more effective and more efficient, albeit with conflicting effects (Jaffe et al., 2024): more efficient meetings require less time and fewer follow-ups, but as they become more effective for collaboration, they may be used more often or for longer.
- However, lack of goal clarity in meetings remains a problem (Microsoft 2023). Generative AI has the potential to enable goal-driven dynamic user interfaces for planning and running meetings (Park et al., 2024). It may also help employees reflect and act on goals in the challenging diversity of meetings (Scott et al., 2024).
    - *Before meetings*, AI goal reflection can change users' mindsets and behaviors around meeting planning, and effects may last beyond specific use of reflection features (Scott et al., 2025).
    - *During meetings*, passive AI interventions (e.g. visualizations) can help meetings stay on track through non-intrusive feedback, while active AI interventions (e.g. questions) can nudge immediate action, but risk disrupting the meeting's flow (Chen et al., 2025).
    - *Across meetings*, AI can reduce the fragmentation of knowledge work by supporting transitions between retrospective and prospective thinking about meeting goals (Vanukuru et al., 2025).

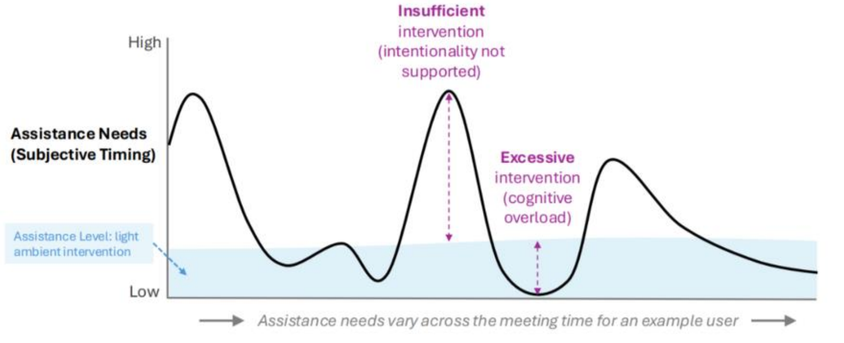


Fig. 6. When subjective assistance needs are high, a light ambient intervention may fail to capture attention, leading to insufficient support for intentionality. Conversely, when assistance needs are very low, even a light ambient intervention may unnecessarily add to users' cognitive load. (The dynamic curve representing changing user needs over time.)

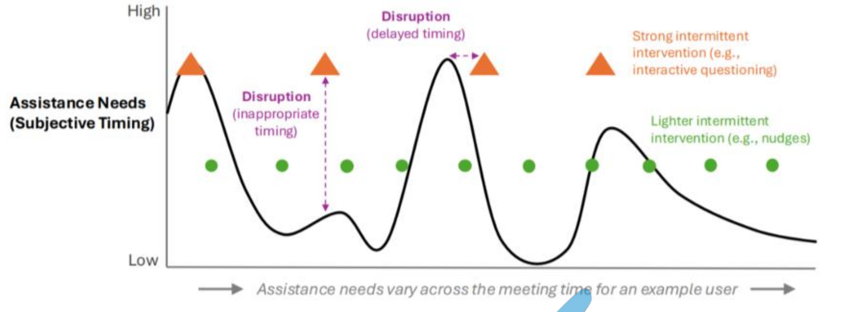


Fig. 7. 'Strong' *intermittent* interventions are much more direct and can provoke immediate action. However, they risk becoming disruptive if not timed properly. As a middleground, team nudges can come at a higher frequency than stronger interventions since they are less disruptive (The dynamic curve representing changing user needs over time. Compare green dots and orange triangles)

AI can support personalized meeting goal reflection before, during, and across meetings. Goal reflection might be passive or active, but reflection interventions need to balance assistance needs with cognitive load, both of which change over the course of the meeting for an individual (Chen, et al. 2025).

Microsoft Study: Jaffe, S., et al., (2024). Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research.
Microsoft Study: Microsoft., (2023). Work Trend Index | Will AI Fix Work?
Microsoft Study: Park, G., et al., (2024). The CoExplorer Technology Probe: A Generative AI-Powered Adaptive Interface to Support Intentionality in Planning and Running Video Meetings. *DIS2024*.
Microsoft Study: Scott, A., et al., (2024). Mental Models of Meeting Goals: Supporting Intentionality in Meeting Technologies. *CHI2024*.
Microsoft Study: Scott, A., et al., (2025). What Does Success Look Like? Catalyzing Meeting Intentionality with AI-Assisted Prospective Reflection. *CHI2025 forthcoming*.
Microsoft Study: Chen et al., (2025). Are We On Track? AI-Assisted Active and Passive Goal Reflection During Meetings. *CHI2025 forthcoming*.
Microsoft Study: Vanukuru et al., (2025) Strengthening the Chain of Intentionality Across Meetings: AI-Assisted Retrospection and Prospection For Knowledge Work. *CHI2025 forthcoming*.

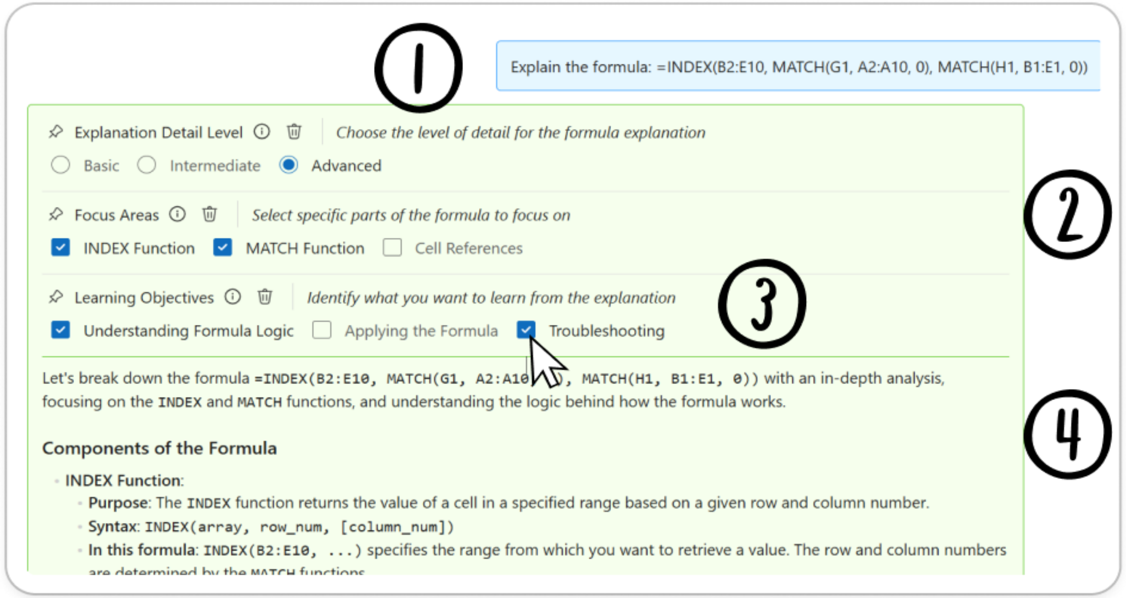
Writing prompts benefits from programming knowledge, discipline, and tools

- Prompts resemble programs (Guy et al., 2024), as writing effective prompts benefits from basic skills such as specifying the expected result, unambiguously defining tasks, and iterating through testing and debugging. While prompts can be expressed entirely in natural language, structing them with examples, tasks lists, and input/output specifications adds value (Dong 2023).
- Effective prompts should be saved, shared and reused. Mechanisms like GPTs allow users to create customized chatbot experiences through prompts, which can be shared widely. Tools such as GenAIScript (de Halleux and Zorn 2024), Python programs via LangChain (LangChain 2024), and agentic systems like AutoGen (Wu et al., 2023) demonstrate how existing programs can integrate AI capabilities through prompts.
- System prompts require both new and old software engineering techniques and tools. Unlike traditional code, prompts in AI software applications possess unique characteristics that demand dedicated research and tools. For example, prompt optimization (Schnabel 2024) rewrites prompts automatically to improve performance, drawing parallels with traditional program optimization. However, system prompts still benefit from traditional techniques, like version control, but since they are written in natural language, they are often treated with less rigor than conventional code (Nahar et al., 2025).
- Prompt effectiveness depends on the underlying model. As new language models are developed, prompts must be updated and tested to ensure they maintain or improve their effectiveness. The process of prompt migration – adapting to comply with newer models (e.g., moving from gpt-3.5-turbo to gpt-4o-mini) – has similarities to prompt optimization (Jahani et al., 2024, Schnabel 2024).
- Embedding AI into software systems requires rethinking the system stack (Berger et al., 2024). Traditional stacks including hardware ISAs, operating systems, and language runtimes, enforce strong properties for executing programs. In contrast, the new stack – incorporating language models interpreting prompts – does not guarantee the same level of reliability or predictability.

Microsoft Study: Guy, T., et al., (2024). Prompts are Programs, *SIGPLAN Perspectives Blog*.
Dong, G., (2023). Prompting Frameworks for Large Language Models: A Survey.
Microsoft Study: de Halleux, J. and Zorn, B., (2024). GenAIScript: Generative AI Scripting.
LangChain, LangChain
Microsoft Study: Schnabel, T., et al., (2024), Symbolic Prompt Program Search: A Structure-Aware Approach to Efficient Compile-Time Prompt Optimization
Microsoft Study: Wu, Q., et al., (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.
Jahani et al., (2024): As Generative Models Improve, We Must Adapt Our Prompts
Microsoft Study: Pryzant, R. et al, (2023). Automatic Prompt Optimization with "Gradient Descent" and Beam Search
Microsoft Study: Nahar, N., et al., (2025). Beyond the Comfort Zone: Emerging Solutions to Overcome Challenges in Integrating LLMs into Software Projects. (ICSE 2025 Forthcoming)
Microsoft Study: Berger, E., et al., (2024). AI Software Should be More Like Plain Old Software. *SIGPLAN*

Dynamically generated interfaces can make prompting easier

- Effective prompting of AI is difficult for non-experts (Zamfirescu-Pereira et al., 2023). Dynamically generating interfaces in response to user prompts can make working with AI easier, helping users steer the AI to generate personalized responses (Ma et al., 2024).
- Generated interfaces can be used to refine or elaborate prompts, known as dynamic prompt middleware (Cheng et al., 2024, Drosos et al., 2025), or used to customize commanding intents with dynamic widgets (Vaithilingam et al., 2024).
- Dynamic prompt middleware provides users with control over AI output, lower barriers to providing context, and greater exploration and task-reflection. However, dynamism is also less consistent, leading to cognitive load and a barrier between predicting what each option would do to a response and what the AI did with an option (Drosos et al., 2025). These findings align with prior research on the challenges of dynamic UI (Alvarez-Cortes et al., 2009, Stephandis et al., 2019, Findlater and Gajos 2009).
- These approaches may also be valuable for local inference, as it can provide dense queries by leveraging SLMs running on NPUs, while using private user content on the PC to provide the parameters.

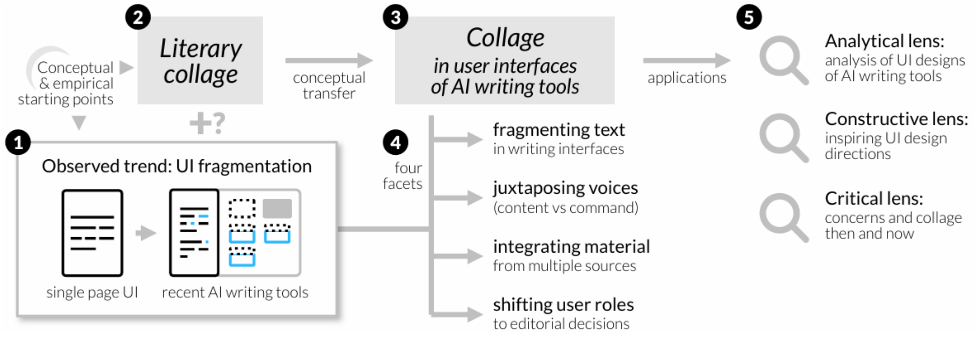


Dynamic prompt middleware (1) takes the user's prompt and (2) generates UI elements that contain relevant prompt options that help steer the AI response by refining the prompt. (3) The user can modify the pre-selected options with a click which (4) regenerates the response with the updated selection as context. (Drosos 2025)

Zamfirescu-Pereira, J.D., et al., (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *CHI 2024*.
Ma, X., et al., (2024). Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses. *CHI EA 2024*.
Cheng, R., et al., (2024). BISCUIT: Scaffolding LLM-Generated Code with Ephemeral UIs in Computational Notebooks. *VL/HCC 2024*.
Microsoft Study: Drosos, I., et al., (2025). Dynamic Prompt Middleware: Contextual Prompt Refinement Controls for Comprehension Tasks. Under Review.
Microsoft Study: Vaithilingam, P., et al., (2024). DynaVis: Dynamically Synthesized UI Widgets for Visualization Editing *CHI 2024*.
Alvarez-Cortes, V., et al., (2009). Current Challenges and Applications for Adaptive User Interfaces. *Human-Computer Interaction*.
Stephandis, C., et al., (2019). Seven HCI Grand Challenges. *International Journal of Human-Computer Interaction*.
Findlater, L., and Gajos, K., (2009). Design Space and Evaluation Challenges of Adaptive Graphical User Interfaces. *AI Magazine*.

Microproductivity involves breaking tasks into small, management chunks – the same concept is now being applied to writing with generative AI

- Task decomposition involves breaking a task down into smaller parts (Parnas 1972). It can help reduce the cognitive load of large tasks (Correaet al., 2020) and has been used to help get more work done with the concept of “microproductivity” (Teevan et al., 2016).
- Writing is one area where microproductivity has been shown to be helpful (Iqbal et al., 2018).
- Now, researchers are proposing that writing is shifting from a single-focus model – like a page resembling a physical piece of paper – toward a collage of dynamic constructs enabled by GenAI tools (Buschek 2024). This includes blending perspectives through iterative drafting while working with GenAI, incorporating AI suggestions and external material.



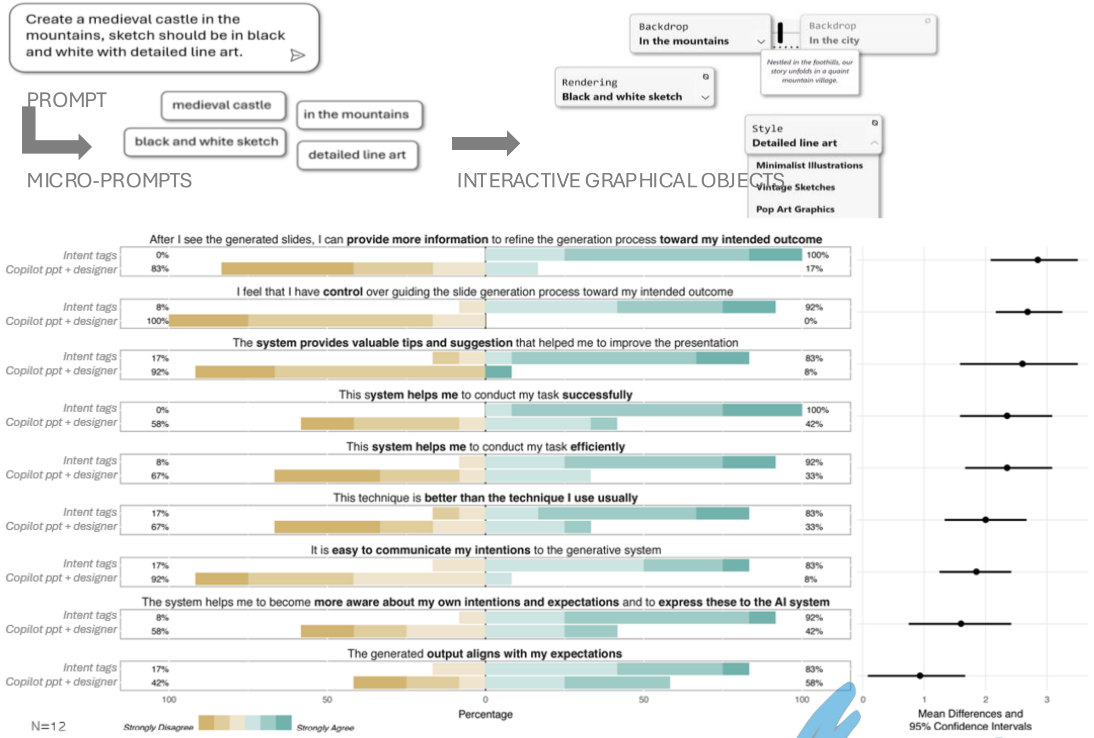
Buschek (2024) shows how fragmentation principles shape the UI design and writing processes of new AI tools. His tool, Collage, introduces a writing paradigm that evolves literary practices by fragmenting text, juxtaposing voices, integrating sources, and shifting roles towards editorial and compositional decisions.

- This evolution refines the user’s role, moving from traditional authorship to editorial and compositional decision-making (Buschek 2024).
- By viewing a prompt as a large writing task (or other such large task to be achieved), we can use the strategies of traditional task decomposition and microproductivity to break optimal prompt creation into smaller, achievable pieces.

Parnas, D., (1971). On the criteria to be used in decomposing systems into modules. *CMU*.
Correa, C., et al., (2020). Resource-rational task decomposition top minimize planning costs.
Teevan, J., et al., (2016), Productivity Decomposed: Getting Big Things Done with Little Microtasks. *CHI 2016*.
Iqbal, S., et al., (2018). Multitasking with Play Write, a Mobile Microproductivity Writing Tool. *UIST 2018*.
Buschek, D., (2024). Collage is the New Writing: Exploring the Fragmentation of Text and User Interfaces in AI Tools. Designing Interactive Systems

Interactive Graphical Micro-Prompting brings users a sense of control when steering content generation, leading to a more satisfying experience

- Breaking prompts into micro-prompts and turning them into interactive graphical objects affords facetted non-linear steering of content generation (Suh et al., 2023., Jiang et al., 2023, Gmeiner et al., 2024).
- This technique can be applied to the generation of any type of content. Micro-prompting enables users to iteratively refine their intent in non-linear manner, while interactive objects invite them to explore variations of different dimensions.
- Initial empirical evidence with 12 users for crafting a slide reveals that users feel more in control of content generation, finding it easier to articulate their intent and taking AI micro-prompting suggestions into account, leading to a more satisfying co-creation experience.

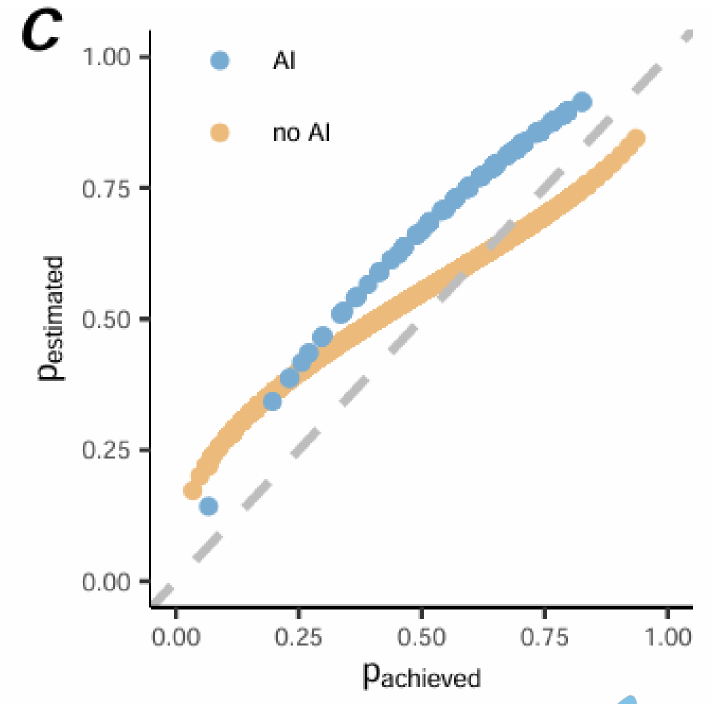


Qualitative ratings of 12 users comparing intent tags (interactive graphical objects representing prompts) and Microsoft PowerPoint Copilot + Designer for crafting a slide deck. (Riche 2024)

Suh, S., et al., (2023). Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. *In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.*
Jiang, P., et al., (2023). Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. *In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
Microsoft Study: Gmeiner, F., et al., (2024). Intent Tagging: Exploring Micro-Prompting Interactions for Supporting Granular Human-GenAI Co-Creation Workflows.

Confidence and competence shape people’s interaction with generative AI

- While likely increasing task performance on some tasks, AI can also disproportionately boost self-confidence, leading users to overestimate their abilities (Fernandes et al., 2024, Lehmann et al., 2024), as anticipated by metacognition research (Tankelevitch et al., 2024).
- Confidence and expertise are central to educational contexts. Coding students with higher confidence and expertise tend to use AI less, or later in their task process (Margulieux et al., 2024), with AI tools generally accelerating success (Prather et al., 2024). However, for students with lower expertise, AI use can inflate their confidence, while exacerbating their difficulties with learning to code, thereby leaving them with an illusion of competence (Prather et al., 2024)It was .
- Designing AI tools with scaffolding, such as step-by-step task guidance, can help align confidence and competence in students and other users (Kazemitabaar et al., 2024, Denny et al., 2024).

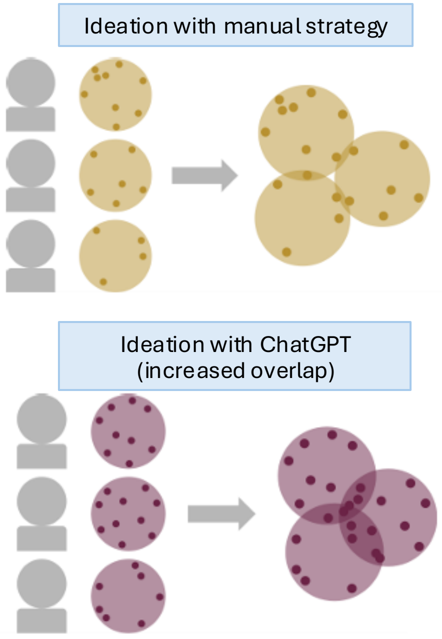


Plot C shows the average posterior predicted values for percent correct achieved (x-axis) and percent correct expected (y-axis) for each group. The s-shape around ideal metacognitive accuracy (grey line) indicates a DKE with low-performers overestimating their performance more than high-performers (yellow; no AI group). DKE = Dunning-Kruger Effect, a cognitive bias where individuals with lower ability overestimate their competence while those with higher ability underestimate it. (Fernandes 2024)

Fernandes, D., et al., (2024). AI Makes You Smarter, But None The Wiser: The Disconnect Between Performance and Metacognition. *arXiv preprint*.
Lehmann, M., et al., (2024). AI Meets the Classroom: When Does ChatGPT Harm Learning?. *arXiv preprint*.
Microsoft Study: Tankelevitch, L., et al., (2024). The metacognitive demands and opportunities of generative AI. *CHI 2024*.
Margulieux, L. E., et al., (2024). Self-Regulation, Self-Efficacy, and Fear of Failure Interactions with How Novices Use LLMs to Solve Programming Problems. *ITiCSE 2024*.
Prather, J., et al., (2024). The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers. *ICER 2024*.
Kazemitabaar, M., et al. (2024). Exploring the Design Space of Cognitive Engagement Techniques with AI-Generated Code for Enhanced Learning. *arXiv preprint*.
Denny, P., et al., (2024). Prompt Problems: A new programming exercise for the generative AI era. *SIGCSE 2024*.

If not carefully designed, Generative AI tools can homogenize output, or potentially allow cognitive skills to erode

- Generative AI tends to exert a "mechanised convergence" effect on knowledge work (Sarkar 2024); solutions to open-ended work tasks developed with GenAI assistance tend to exhibit less diversity and be more homogeneous, compared to when the same work task is solved manually without GenAI assistance. This reduces diversity of ideas at the group level, even if "creative output" appears to quantitatively increase at the individual level. The mechanised convergence effect has been demonstrated in multiple studies of different domains, including creative ideation (Anderson et al., 2024, Zhou and Lee 2024), consultancy report writing (Doshi and Hauser 2024), and programming (Lee et al., 2024).
- Because users can tend to search for solutions that merely meet a minimum aspirational threshold, the likelihood of accepting AI-generated output if it contains no obvious errors is high (Drosos et al., 2024, Prather et al., 2023). This is similar to but distinct from overreliance, which involves accepting incorrect output. If users fall into the habit of accepting work rather than exercising the cognitive skills required to produce it, these skills are likely to be forgotten (Arthur et al., 1998). The speed and scale of knowledge work may increase in the short term, but at the risk of creative and evaluative skill erosion, making corrections or pivots more difficult in the long term (Sellen and Horvitz 2024, Sarkar et al., 2024).

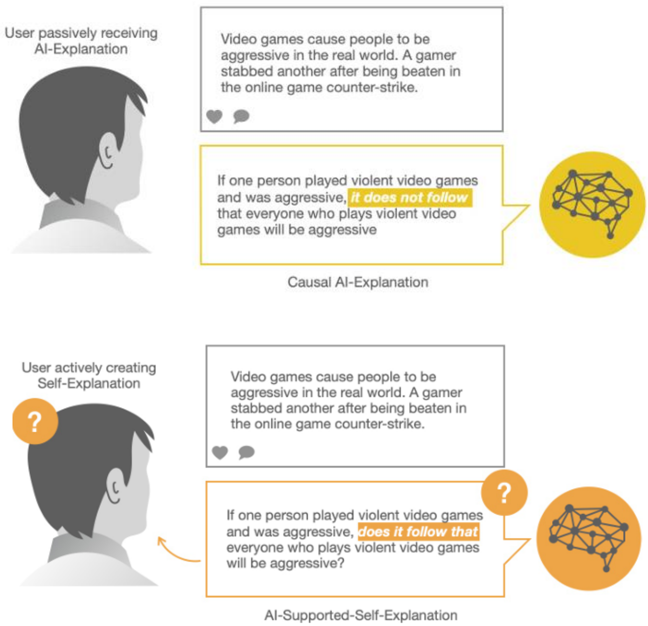


Homogenization analysis of two different creativity support tools for divergent ideation found that users of these tools each *"produce similarly homogenous sets of ideas as individuals, but collectively, users of ChatGPT produce a more homogenous set of ideas at the group level"* (purple) (Anderson 2024)

Microsoft Study: Sarkar, A., (2024). Intention is all you need. *PPIG 2024.*
Anderson, B. R., et al., (2024). Homogenization effects of large language models on human creative ideation. *In Proceedings of the 16th Conference on Creativity & Cognition.*
Zhou, E., and Lee, L., (2024). Generative artificial intelligence, human creativity, and art.. *PNAS nexus.*
Doshi, A. R., and Hauser, O. P., (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances.*
Microsoft Study: Lee, M. J. L, et al., (2024). Predictability of identifier naming with Copilot: A case study for mixed-initiative programming tools. *PPIG 2024.*
Microsoft Sudy: Drosos, I., et al., (2024). "It's like a rubber duck that talks back": Understanding generative AI-assisted data analysis workflows through a participatory prompting study. *CHIWORK 2024.*
Prather, J., et al., (2023). "It's weird that it knows what I want": Usability and interactions with Copilot for novice programmers. *ACM Transactions on Computer-Human Interaction.*
Arthur Jr, W., et al., (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human performance*
Microsoft Study: Sellen, A., and Horvitz, E., (2024). The rise of the AI Co-Pilot: Lessons for design from aviation and beyond. *Communications of the ACM.*
Microsoft Study: Sarkar, A., et al., (2024). When Copilot becomes autopilot: Generative AI's critical risk to knowledge work and a critical solution. *EuSpRIG* 2024

Simple interventions can support critical thinking while using AI

- Education research has found that posing "metacognitive guiding questions" (questions that help students think about their thinking) alongside reading materials improves students' critical engagement with the text (Salomon 1988), and software tools can help learners evaluate the strengths and weaknesses of arguments, such as by visualising the logical structure of an argument and the evidence for and against it (Sun et al., 2017, Tsai et al., 2015). Much design research has investigated how to improve critical and reflective thinking in domains such as online misinformation and as well as wellbeing (e.g., reviewed in Sarkar et al., 2024).
- Researchers now want to use that approach to build AI systems that improve metacognitive skill. Researchers have proposed that AI can go beyond assistance, acting as "provocateur" (Sarkar 2024), or "antagonist" (Cai et al., 2024), or "coach" (Hofman et al., 2023), such as by questioning the user's intent and highlighting limitations, biases, and alternatives for both AI- and user-generated content. This is a design challenge as it opposes user preferences and expectations for AI as a tool for efficient work completion, but successfully provocative tools may lead to better work quality.
- Generative AI itself enables new opportunities for designing critical thinking support. Posing AI-generated explanations as questions can improve the ability to distinguish between logically valid and logically invalid statements (Danry et al., 2023). AI-generated questions about the content of research papers can improve readers' understanding (Maldonaldo et al., 2023, Yuan 2023).

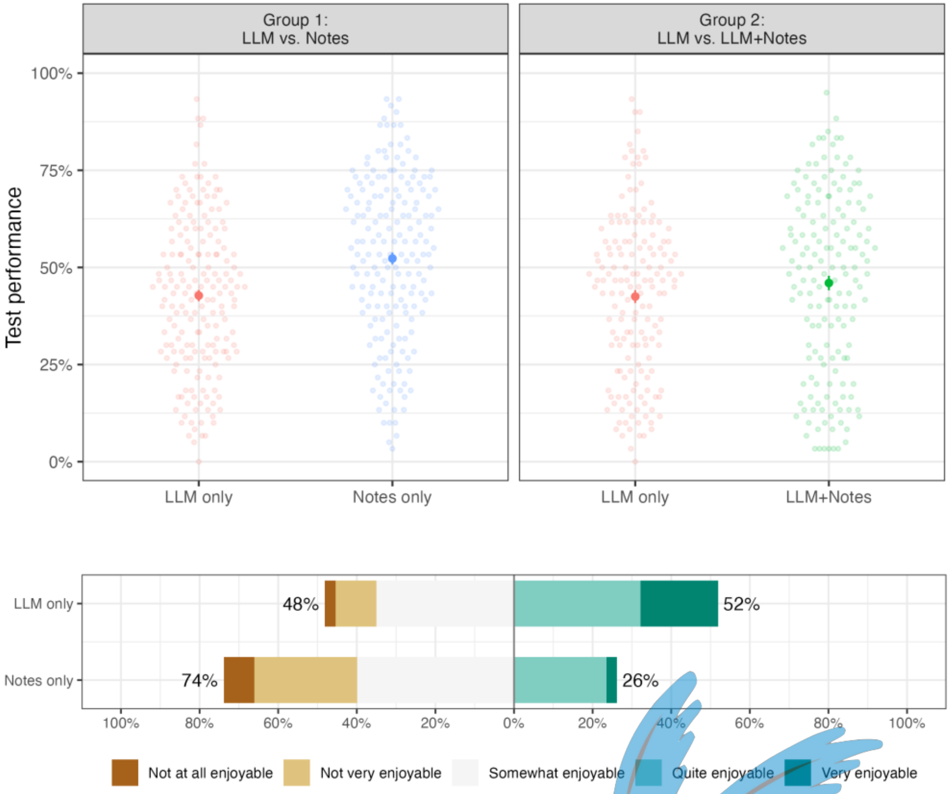


Top: An example of a socially divisive statement and AI feedback with casual AI-explanations *telling* users *why* the statement is logically invalid. Bottom: an example of a socially divisive statement and AI feedback with AI-framed Questioning *asking* the users a question that helps *them* asses *if* the statement is logically invalid or not. (Danry 2023)

Salomon, G., (1988). AI in reverse: Computer tools that turn cognitive. *Journal of Educational Computing Research*.
Danry, V., et al., (2023). Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. *CHI 2023*.
Microsoft Study: Sarkar, A., (2024). AI Should Challenge, Not Obey. *CACM*.
Microsoft Study: Sarkar, A., et al., (2024). When Copilot becomes autopilot: Generative AI's critical risk to knowledge work and a critical solution. *EuSpRIG 2024*.
Cai, A., et al., (2024). Antagonistic AI.
Microsoft Study: Hofman, J. M., et al (2023). A sports analogy for understanding different ways to use AI. *Harvard Business Review*.
Richards Maldonado, L., et al. (2023, October). ReaderQuizzer: Augmenting Research Papers with Just-In-Time Learning Questions to Facilitate Deeper Understanding. *CSCW 2023*.
Yuan, K., (2023). CriTrainer: An Adaptive Training Tool for Critical Paper Reading. *In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
Sun, N., et al. (2017). Critical thinking in collaboration: Talk less, perceive more. *CHI EA '17*.
Tsai, C.-Y., et al. (2015). The effect of online argumentation upon students' pseudoscientific beliefs. *Computers & Education*.

Combining AI and note-taking boosts retention and engagement

- One of the first large-scale experiments on AI's effects on reading comprehension and retention found complementary benefits of combining traditional note-taking with LLMs as a reading aid (Kreijkes et al., 2024).
- Over 400 secondary students in the UK studied history passages using either traditional note-taking, an LLM chatbot as a reading aid, or both. The LLM chatbot was an instance of GPT pre-prompted with the reading passage and allowed for student interaction and questions.
- Both traditional note-taking alone and note-taking in conjunction with an LLM had significant positive effects on retention and comprehension compared to using only an LLM.
- But students found value in the LLM for simplifying complex material, providing additional context, deepening understanding, and reducing cognitive load.
- Overall, the findings indicate that traditional note-taking supports deep engagement and retention while AI enhances initial understanding and fosters student interest.

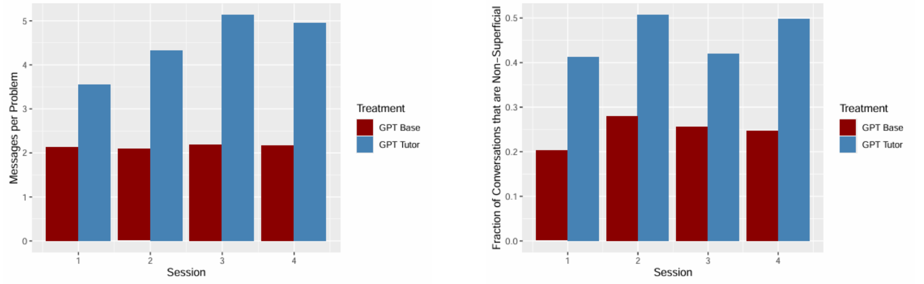


Test performance and enjoyment when students used traditional note taking, LLM assisted note taking, or both. (Kreijkes 2024)

Kreijkes et al. (2024). Complementary Roles of Generative AI and Note-Taking for Reading Comprehension and Retention: A Randomised Experiment in Secondary Schools. *arXiv preprint*.

Learning a skill may require AI that engages more cognitive effort than AI for those who already have the skill

- The goal of many AI tools is to increase productivity through offloading tasks, which reduces cognitive effort. Learning, however, generally requires a certain level of cognitive effort (Brown et al., 2014). More effective learning techniques are often more effortful than less effective ones (Dunlosky et al., 2013).
- Reading comprehension, for example, is quite different for a knowledge worker than a student. AI summarization of a text may help a knowledge worker who already has the expertise to work more effectively. However, for a student who needs to develop reading comprehension as part of the journey of developing future expertise, more effortful processes, such as note-taking, may have better learning outcomes than relying on AI summarization (Kreijkes et al., 2024).
- Results like the above indicate that AI tools which focus only on automation might negatively impact people's skill development both in the short and long term (Prather et al., 2023, Simkute et al., 2024). New approaches are needed to design AI tools in a way that can both support productivity and maintain or even improve human learning, understanding, and ultimately skills (Sellen and Horvitz 2024, Hofman et al., 2023).
- For example, the newest AI tutors are finding success in guiding and challenging students more than simply providing answers (Bastani et al., 2024, Kasneci et al., 2023, Khan 2024).
- Further research is needed to determine when it is important to maintain skills without AI, when AI may replace skills, and when AI-augmented cognition is the best option.



(a) Avg. # of Messages per Problem (b) Frac Non-Superficial Conversations per Session

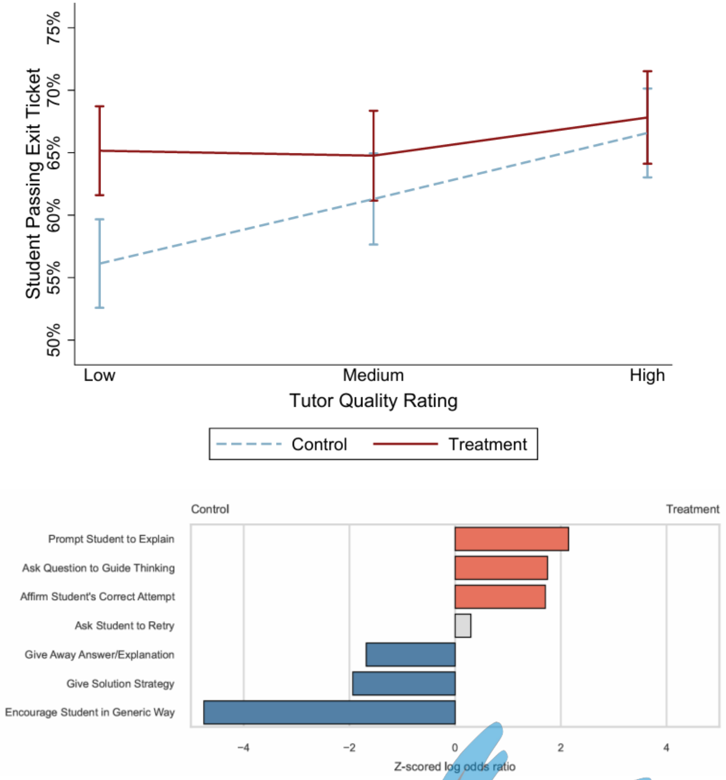
Figure 2: Student engagement—given by (a) average number of student messages per problem, and (b) average fraction of student session conversations that have no superficial messages (simply re-stating the question or asking for the answer) per session—by treatment (GPT Base and GPT Tutor) over time.

Bastani et al. (2024) make the point that the number of messages in their AI tutor (GPT Tutor) is significantly higher than simply chatting with a generic AI chat interface (GPT Base), and further increases with experience using the tool. That students interact less with GPT Base is consistent with their hypothesis that GPT Base simply provides students with solutions instead of engaging them in a learning process, as GPT Tutor does.

Brown, P. C., et al., (2014). Make it stick: The science of successful learning. *Harvard University Press*.
Dunlosky, J., et al., (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychol Sci Public.*
Microsoft Study: Kreijkes, P., et al., (2024). Complementary Roles of Generative AI and Note-Taking for Reading Comprehension and Retention: A Randomised Experiment in Secondary Schools. arXiv preprint forthcoming.
Prather, J., et al., (2023). The robots are here: Navigating the generative ai revolution in computing education. *ACM*.
Microsoft Study: Simkute, A., et al., (2024). Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interaction. *CHI 2024*.
Microsoft Study: Sellen, A., and Horvitz, E. (2024). The Rise of the AI Co-Pilot: Lessons for Design from Aviation and Beyond.
Hofman, J. M., Goldstein, D. G. & Rothschild, D. M. (2023, Dec 4). A sports analogy for understanding different ways to use AI. *HBR*.
Bastani, H., et al., (2024). Generative AI Can Harm Learning. *The Wharton School Research Paper*.
Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*.
Khan, S. (2024). Brave New Words.

Early results from research in education suggest that AI can enhance educator engagement and help improve learning outcomes

- "Tutor CoPilot" provided real-time LLM-based guidance to tutors, demonstrating through a randomized controlled trial that this Human-AI system can significantly improve student learning outcomes, and with larger benefits for lower-rated tutors (Wang et al., 2024).
- Several computer science courses successfully integrated virtual TA bots that provide continuous, customized support for students, showing positive outcomes (Kazemitabaar et al., 2024, Liu et al., 2024, Zamfirescu-Pereira et al., 2025).
- Other studies used randomized controlled trials to demonstrate that providing students with LLM-based tutoring support led directly to learning gains in mathematics (Kumar et al., 2023, Pardos and Bhandari 2024).
- While unfettered access to AI tools can in some cases negatively impact learning outcomes, appropriately applied tools have shown no such patterns, enhancing short-term performance while minimizing the risk of long-term negative effects (Bastani 2024).



Differences in student performance (top) and tutor strategies (bottom) for tutors who were provided with real-time LLM guidance (treatment) vs. not (control) from Wang et. al. 2024

Wang, E. et al,. (2024). Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise.
Kazemitabaar, M., et al., (2024). CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Education Needs. *CHI 2024.*
Liu, R., et al., (2024). Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education. *ACM*
Zamfirescu-Pereira, J.D., et al., (2024). 61A Bot Report: AI Assistants in CS1 Save Students Homework Time and Reduce Demands on Staff. (Now What?)
Kumar, H. et al., (2023). Math Education with Large Language Models: Peril or Promise?
Pardos, Z.A. and Bhandari, S., (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE.*
Bastani, H., (2024). Generative AI Can Harm Learning. *The Wharton School Research Paper.*

There are multiple barriers to appropriate reliance on AI

- Appropriate, or correct, AI reliance happens when users rely on AI when AI is right, but not when it's wrong (Schemmer et al., 2023). It is a balance between over- and under-reliance on AI. To rely on AI appropriately, users have to assess output usefulness and correctness and decide whether to accept it or not.
- Researchers in academia and industry (including Microsoft) have for many years flagged appropriate reliance as a critical open challenge in human-AI collaboration (Passi and Vorvoreanu 2022; Passi et al., 2024).

	User **accepts** output	User **rejects** output
AI output is **correct**	Correct AI reliance (CAIR)	Under-reliance
AI output is **incorrect**	Overreliance	Correct self-reliance (CSR)

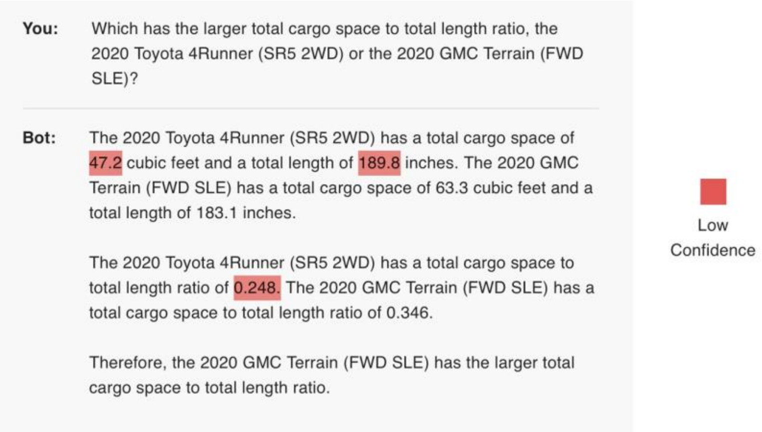
Appropriate reliance is a balance between under- and overreliance. (Passi 2024)

- People tend to accept LLM outputs without checking for accuracy, because they lack awareness of how these outputs might be wrong and why. People's mental models of GenAI liken it to search. For example, users assume LLMs retrieve, not generate content; they don't understand that AI summaries might contain factual inaccuracies or might be incomplete (Vorvoreanu et al., 2024).
- People often use flawed heuristics to approximate response trustworthiness. For example, users assume that because responses are well-written, cite sources, and seem partially correct, they must be right. Further, users see the number, variety, and quality of sources (e.g., research articles, reputable websites) as indicators of response trustworthiness (Drosos et al., 2024; Vorvoreanu et al., 2024).
- When asked to engage with cited sources in RAG scenarios for information finding tasks, information overload makes the experience cognitively demanding. For example, checking citations often involves finding the relevant information the output is based on in lengthy documents (Drosos et al., 2024; Vorvoreanu et al., 2024).

Schemmer, M., et al., (2023). Appropriate reliance on AI Advice: Conceptualization and the Effect of Explanations. *IUI 2023.*
Microsoft study: Passi, S., and Vorvoreanu, M., (2022). Overreliance on AI: Literature Review. *MSFT Technical Report.*
Microsoft study: Passi, S., et al., (2024). Appropriate Reliance on Generative AI: Research Synthesis. *MSFT Technical Report.*
Microsoft study: Vorvoreanu, M., et al., (2024). Insights from Five Internal Studies on Overreliance on AI. *MSFT Internal Report.*
Drosos, I., et al., (2024). "It's like a rubber duck that talks back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. *CHIWORK'24*

To foster appropriate reliance on AI, consider 3 UX goals:

1. **Help users form realistic mental models of the AI system's capabilities and limitations**. Apply Guidelines for Human-AI Interaction #1, #2, and #11 to educate users about what the system can do, the types and frequency of mistakes it can make, and the AI system's workings (Amershi et al., 2019, Vorvoreanu et al., 2024).
2. **Promote user vigilance**. Motivate users to pay attention and verify outputs, especially in high-stakes scenarios and when mistakes are likely. Draw users' attention to potential mistakes in AI outputs by using techniques such as uncertainty expressions and uncertainty highlighting (Kim et al., 2024, Spatharioti et al., 2024, Vasconcelos et al., 2024, Vorvoreanu et al., 2024).
3. **Make it easy to verify outputs**. Decrease the cognitive load of verifying AI outputs against grounding data or reliable information sources. Consider techniques such as displaying excerpts from grounding data alongside outputs, to enable users to spot discrepancies (Vorvoreanu et al., 2024).



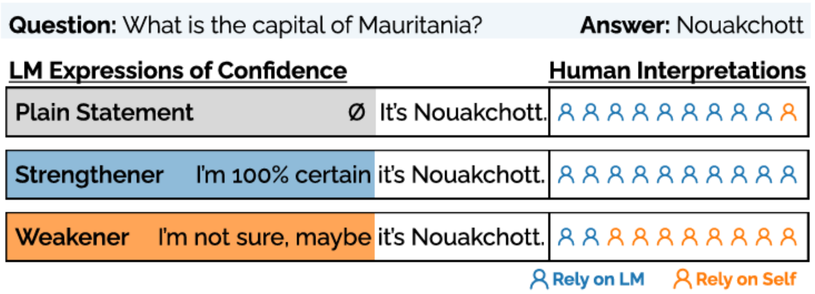
Highlighting tokens with low-generation probability mitigated overreliance on AI in an information retrieval context. (Spatharioti et al 2023)

For the foreseeable future, model improvements alone are not sufficient. For instance, in a recent study, medical doctors working with ChatGPT or medical databases performed worse on diagnostic tasks than ChatGPT alone (Goh et al., 2024), indicating possible under-reliance on AI. UX interventions are needed to foster appropriate reliance – a balance between over- and under-reliance.

Amershi, S., et al., (2019). Guidelines for Human-AI Interaction. *CHI'19*.
Microsoft Study: Vorvoreanu, M., et al., (2024). Insights from Five Internal Studies on Overreliance on AI. *MSFT Internal Report*.
Vasconcelos, H., et al., (2024). Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. *ACM Transactions on Computer-Human Interaction*.
Kim, S., et al., (2024). "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust . *FAccT*
Spatharioti, S., et al., (2023). Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment.
Goh, E., et al., (2024). Large Language Model Influence on Diagnostic Reasoning: A Diagnostic Clinical Trial. *JAMA Network Open*.

When fostering appropriate reliance on LLM outputs, details matter

- Designing AI systems to express uncertainty can be an effective way to reduce overreliance on LLMs. Uncertainty expressions in LLM outputs can be verbal, such as “I’m not sure, but…” or visual, such as highlighting tokens.
- Communicating uncertainty matters: A simulated LLM that expressed certainty when outputs were correct and uncertainty when they were incorrect fostered appropriate reliance (Zhou et al., 2024). In LLM-infused search, uncertainty expressions in the first-person perspective were more effective than those in the general perspective (e.g., “There is uncertainty…”) (Kim et al., 2024a).



Overview of experiments on human interpretations of epistemic markers. They (Zhou et al) asked users to interpret epistemic markers generated by LMs by asking users which answer they would rely on and which answers they would need to double check (Zhou et al. 2024).

- Type of uncertainty matters: In a code generation context, highlighting tokens with the highest likelihood of being edited mitigated overreliance on AI. Highlighting tokens with low generation probability, did not help foster appropriate reliance (Vasconcelos et al., 2024).
- Context matters: While highlighting tokens with low generation probability did not mitigate overreliance in a code generation scenario (Vasconcelos et al., 2024), it did foster more appropriate reliance in an information retrieval scenario (Spatharioti et al., 2023).
- Citing sources (might) matter: In some studies, citing sources helped mitigate overreliance, but in others, it did not (Kim et al., 2024b; Vorvoreanu et al., 2024). How a system cites sources may be a key factor here.
- UX research matters: Since fostering appropriate reliance depends on so many factors, overreliance mitigations should to be tested for each LLM-infused product, in context, with its users.

Zhou, K., et al., (2024). Relying on the Unreliable: The Impact of Language Models’ Reluctance to Express Uncertainty. *ACL 2024*.
Kim, S., et al., (2024a). "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust . *FAccT 2024*.
Vasconcelos, H., et al., (2024) Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. *ACM Transactions on Computer-Human Interaction*.
Spatharioti, S., et al. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment.
Kim, S., et al., (2024b). Draft under review.
Si, C., et al., (2024). Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong. *ACL 2024*
Microsoft Study: Vorvoreanu, M., et al., (2024). Insights from Five Internal Studies on Overreliance on AI. *MSFT Internal Report*.

Interaction paradigm shift: AI as a new medium with which we interact

- Most current generative systems are anthropomorphic intelligences with which we communicate (mostly using chat). A different interaction paradigm is to consider AI as a new medium with which we interact.

  This shift calls for novel ways for users to engage with AI:

  "People do, AI elevates."

- **People do.** A fundamental principle of natural user interfaces is direct manipulation (Shneiderman 1983). The idea of direct manipulation is to enable users to interact with objects of interest *in situ*, with rapid, reversible and incremental actions.
- Enabling users to directly interact with AI generated content using multiple modalities (e.g. selecting, inking, or commenting in place) reduces the indirection of typing in a side chat leading to faster outcomes with less effort (Masson et al., 2024).
- **AI elevates.** *In situ* interactions encapsulates properties such as spatial information difficult to convey in words, as well as enable more granular non-linear inputs. Coupled with implicit context of where interactions occur, interacting with AI can become a natural live experience.

Direct manipulation: turning a complex prompt into a natural sequence of interactions

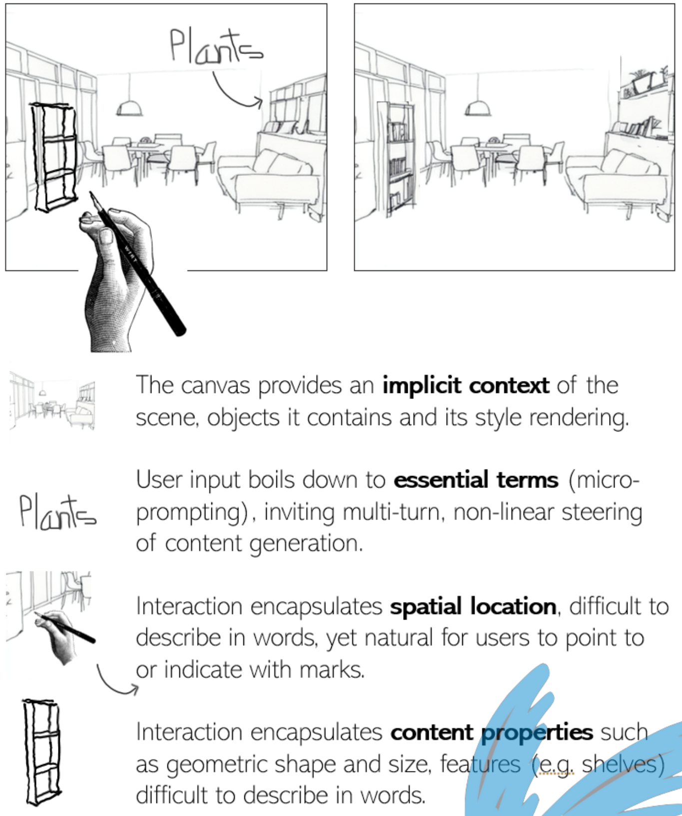


Image by Riche (2024).

Shneiderman, B., (1983). Direct manipulation: A step beyond programming languages. *Computer.*
Masson, D., et al., (2024). DirectGPT: A direct manipulation interface to interact with large language models. *CHI 2024*

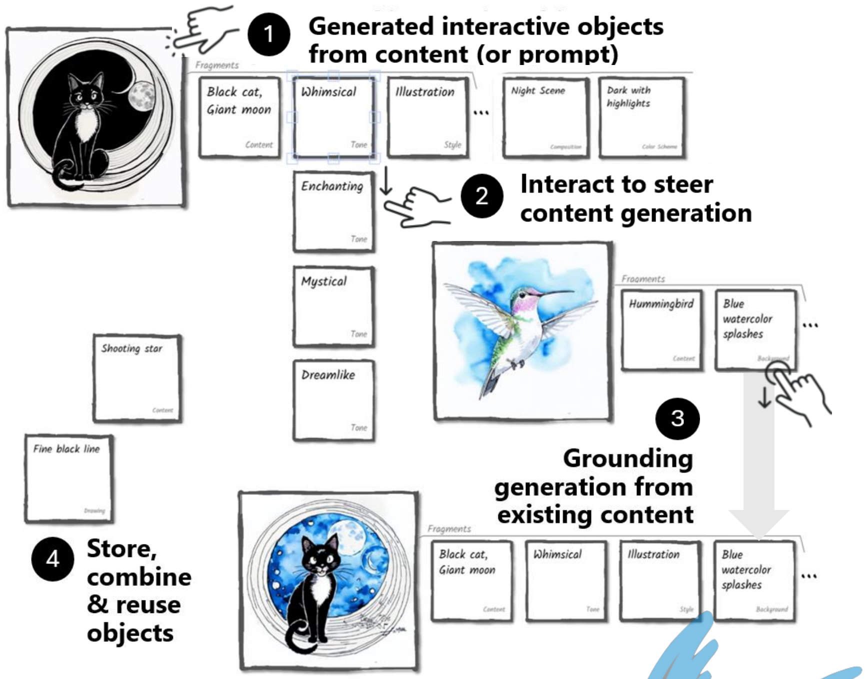
Initial evidence on post-chat interaction techniques suggests that they could help solve prompt engineering hurdles

The principle of **reification** in human-computer interaction turns abstract commands into persistent reusable interface objects, which affords several benefits to users (Beaudoin-Lafon 2000). We can apply this principle to user prompts (or fragments of prompts), embodying them into interactive graphical objects persistent on screen for users to store and reuse multiple times, as well as alter and combine at will. Riche et al. (2024) call this next generation of widgets: "AI-instruments".

An initial qualitative study with 12 users shows a few advantages of AI-instruments over more linear typing-based interactions:

- Generating interactive objects surfacing different dimensions (or aspects of a prompt) eliminates the need for users to articulate them in their own words.
- Simple interactions with objects to add/remove dimensions or suggest different dimensions facilitates exploration and iterative content generation.
- Persistent objects on screen can be stored, combined and most importantly reused with minimal effort.

Reification: embodying micro-prompts into interactive graphical objects

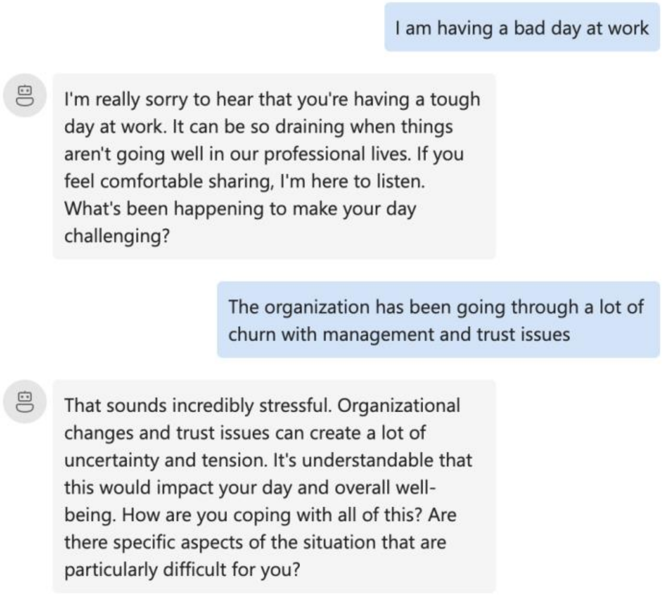


Interaction paradigm for no-typing generative AI: Interactive cards encapsulating fragments of prompts are generated from a double tap (1), simple interactions generate card variations and suggestions (2), cards can be added, removed or transferred by drag and drop (3). (Riche 2024)

Beaudouin-Lafon, M., (2000). Instrumental interaction: an interaction model for designing post-WIMP user interfaces. *CHI 2024*
Riche, N. et al., (2024). AI-instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools (Preprint).

Responsible deployment of empathetic agents can lead to higher engagement and user satisfaction

- Empathy has been traditionally understood as a capability exclusive to humans focused on understanding and sharing another person's experiences from their perspective, involving both emotional and cognitive processes (Davis 1983). It enables individuals to place themselves in others' positions, fostering deeper interpersonal connections.
- In contrast, "digital empathy" with respect to AI agents refers to the ability to comprehend and respond to cognitive and emotional states (Schmidmaier et al., 2024). This requires understanding context, user preferences, emotional states, past interactions, perspective-taking, and adapting behavior accordingly among others.
- Preferences for empathetic AI agents vary by context (Hernandez et al., 2023). Users prefer AI that responds to emotions in applications like counseling or customer service, where empathy enhances interaction. For data-focused or analytical tasks, users favor less empathy and minimal emotional simulation, seeking more objective responses.
- Empathetic AI agents can promote user engagement, trust, satisfaction, and emotional connection, leading to improved interactions (Schmidmaier et al., 2024). They can enable personalized interactions, foster deeper customer engagement with increased brand loyalty, and promote productivity and well-being by aligning with human perspectives.
- Empathetic AI agents might cause confusion about the agent's actual abilities, and it's challenging to express empathy without using anthropomorphic language, which can blur the line between human and machine. Users may become over-reliant and attached to AI, leading to dependency or other emotional and physical harms (Dzieza 2024).

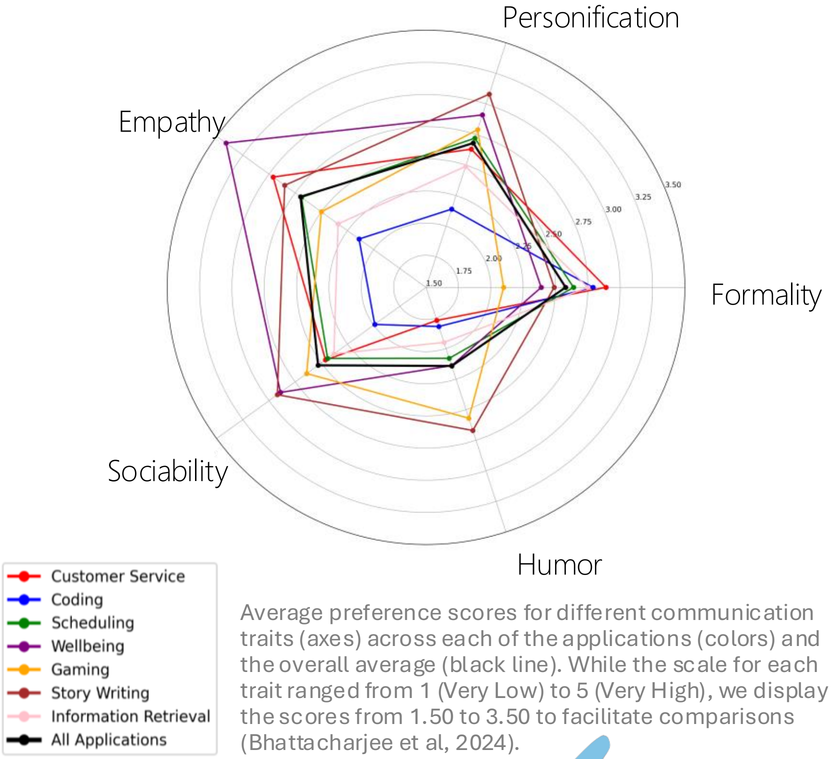


Example of an empathetic interaction about workplace challenges between a user (right) and a simulated agent (left) (Microsoft created image)

Davis, M. H., (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*.
Schmidmaier, M., et al., (2024). Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. *CHI 2024*
Microsoft Study: Hernandez, J., et al., (2023). Affective Conversational Agents: Understanding Expectations and Personal Influences.
Dzieza, J., (2024). Friend or Faux. *The Verge*.

People expect different AI personas depending on the application

- General-purpose AI agents like ChatGPT, Claude and Copilot often exhibit consistent behavior across various application areas, providing uniform interactions regardless of context. However, users have different expectations depending on the domain—be it coding, customer service, or writing assistance—where specialized interactions can significantly enhance user experience and efficiency.
- Adjusting AI communication traits offers a way to influence user interactions across different domains. Some key dimensions such as humor, sociability, empathy, formality, and personification can be tailored to align with user preferences (Chaves and Gerosa 2021). Research shows that these preferences vary across tasks, highlighting the importance of context-specific communication strategies (Bhattacharjee et al., 2024).
- In productivity environments, information workers prefer interacting with expert assistants that are knowledgeable, trustworthy, transparent, and responsive. They value communication that is professional and direct and appreciate proactive suggestions that aid their workflow and decision-making (Nepal et al., 2024).



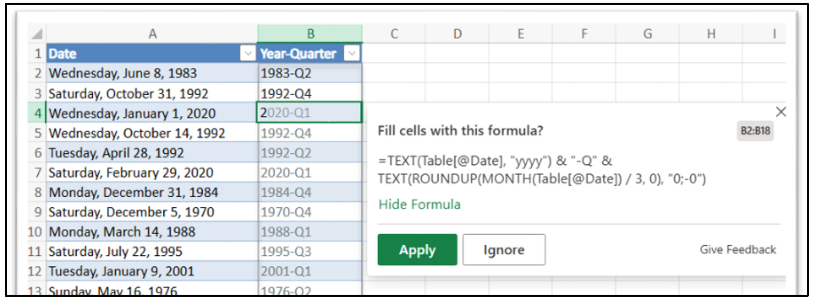
Average preference scores for different communication traits (axes) across each of the applications (colors) and the overall average (black line). While the scale for each trait ranged from 1 (Very Low) to 5 (Very High), we display the scores from 1.50 to 3.50 to facilitate comparisons (Bhattacharjee et al, 2024).

- Agent personas are often created through system prompts, as they provide a quick and flexible approach (e.g., Park et al., 2024, Nepal et al., 2024). However, this method may not always produce consistent results. Alternative methods involve fine-tuning the models with synthesized data and employing other hybrid approaches (Huang et al., 2024).

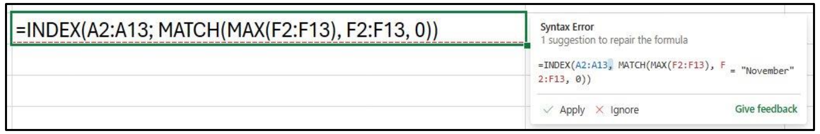
Chaves, A., and Gerosa, M., (2021). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *CHI 2021*.
Microsoft Study: Bhattacharjee, A., et al., (2024). Understanding Communication Preferences of Information Workers in Engagement with Text-Based Conversational Agents.
Microsoft Study: Nepal, S., et al., (2024). From User Surveys to Telemetry-Driven Agents: Exploring the Potential of Personalized Productivity Solutions.
Park, J. S., et al., (2024). Generative Agent Simulations of 1,000 People.
Huang, Q., et al., (2024). Selective Prompting Tuning for Personalized Conversations with LLMs. *ACL 2024*.

AI thrives when intent goes beyond words to leverage context and examples

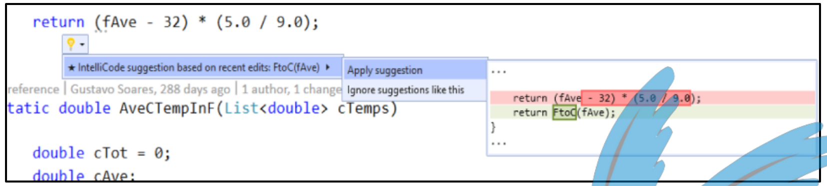
- Natural language is one way to express intent to GenAI, but there are others that can be useful in different contexts (see other slides for many additional examples).
- Input/output examples can help show GenAI what is expected and can be used for string transformation, table extraction, formatting, etc (Singh et al., 2023). Examples, being verifiable, enable use of failure-guided refinement techniques or backtracking-based search. AI can generate distinguishing inputs for finding representative examples (Cambronero et al., 2023).
- With software, broken code itself can be the specification/user intent given the model. Solutions can be neural (Joshi et al., 2023) or neuro-symbolic (Bavish et al., 2022).
- Other types of intent can be data and temporal context (past user actions) that can be used to predict the next actions. Popular applications are smart copy paste (Singh et al., 2024) and IntelliCode suggestions (Miltner et al., 2019) and (Gao 2020).



Screenshot of Formula by Example in Excel, where AI learns a program and fill the column from a few input/output examples.



Broken code can prompt the model to find the correct solution without a natural language prompt. (Excel)

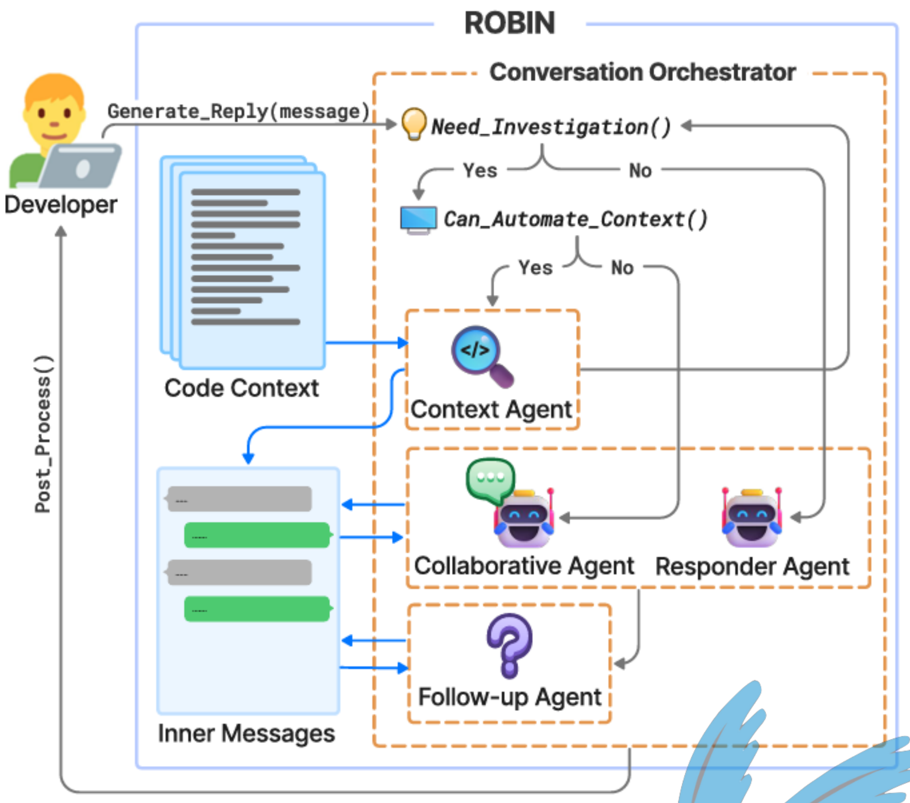


Example of using context to prompt the model, with Visual Studio IntelliSense predicting the next line of code based on past user edits.

Microsoft Study: Singh, M., et al., (2023), Cornet: Learning Table Formatting Rules By Example, *VLDB 2023*.
Microsoft Study: Cambronero, J., et al., (2023). FlashFill++: Scaling Programming by Example by Cutting to the Chase, *ACM*
Microsoft Study: Joshi, H., et al., (2023), FLAME: A Small Language Model for Spreadsheet Formulas, *AAAI 2023*.
Microsoft Study: Bavishi, R., et al., (2022), Neurosymbolic Repair for Low-Code Formula Languages, *OOPSLA 2022*.
Microsoft Study: Singh, M., et al., (2024), Tabularis Revilio: Converting Text to Tables, *CIKM 2024*.
Microsoft Study: Miltner, A., et al., (2019), On the fly synthesis of edit suggestions, *OOPSLA 2019*.
Microsoft Study: Gao, X., et al., (2020) Feedback-driven semi-supervised synthesis of program transformations, *OOPSLA 2020*.

Chatbots are effective when they collaborate, not automate: use principles of cooperative conversation for better user interaction patterns

- Current chatbots are eager to try and complete the task presented to them, instead of collaborating with the user iteratively (Chopra et al., 2024).
- Chatbots can be developed using Gricean maxims, a classical principal from sociology on what makes a good conversation between human participants. Adapting these into AI chatbot's prompts can make it behave more like a collaborative assistant.
- This approach leads to the "Investigate-and-Respond" conversation pattern (Bajpai et al., 2024), where the chatbot is designed for collaborative behavior as opposed to just automation. The chatbot can then explore the problem space with help from the user, asking questions and guiding them to find answers.
- In a study done using the Investigate-and-Respond conversation pattern in GitHub Copilot in Visual Studio, named Robin, developers were 3.5x more successful at fixing bugs with this conversation pattern, compared to a baseline chatbot, which only had a 25% success rate (Bajpai et al., 2024).

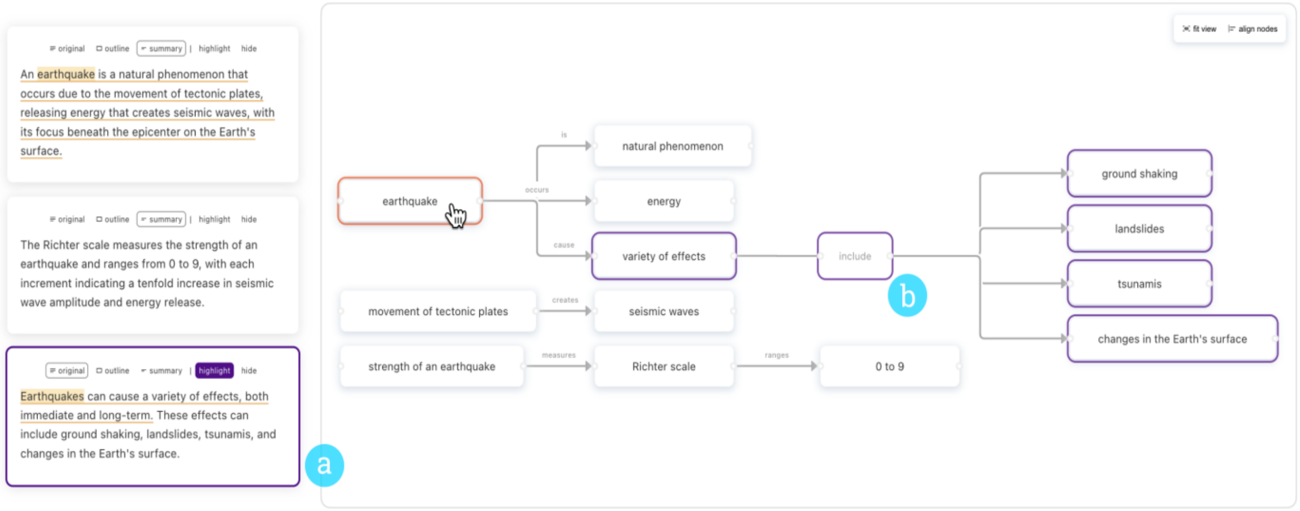


Example of the multi agent workflow that uses the Investigate-and-Respond conversation pattern in Visual Studio Copilot. (Bajpai et al., 2024)

Microsoft Study: Chopra, B., et al., (2024), Exploring Interaction Patterns for Debugging: Enhancing Conversational Capabilities of AI-assistants.
Microsoft Study: Bajpai, Y., et al., (2024), Let's Fix this Together: Conversational Debugging with GitHub Copilot.
Visual Studio IDE: Debug with Copilot. https://learn.microsoft.com/en-us/visualstudio/debugger/debug-with-copilot?view=vs-2022

UX, besides technical capabilities, is essential for AI to light up the next wave of tools for thinking. Post-chat UX and notebooks are taking a central stage.

- As discussed elsewhere in the deck, chat as an UX may be at a plateau (Morris 2024), as it can lack expressive power for both abstraction and specificity (Zamfirescu-Pereira et al., 2024). New and existing interaction patterns closer to direct manipulation are used for better expressing intent (Masson et al., 2024, Figma 2024) and parsing AI outputs (Jiang et al., 2023).
- Unlike chat interfaces, notebooks allow for a more structured, versatile, and familiar (Allen 2024) ways for people to create and consume knowledge. The written page and the notebook are re-emerging as a medium and the UX for thinking with AI: as companions to chat experiences (OpenAI 2024, Anthropic 2024) or on their own (Google 2024, Notion 2024).

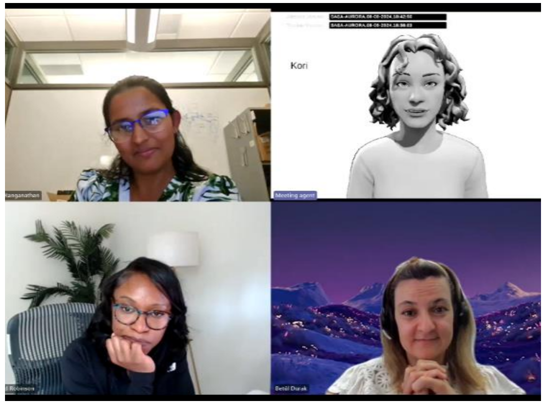


The Graphologue interface, by Jieng et al (2023).

Morris, M. R., (2024). Prompting Considered Harmful. *CACM.*
Zamfirescu-Pereira, J.D., et al., (2024). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *CHI 2024.*
Masson, D., et al., (2024). DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models.
Figma, (2024). Figma AI.
Jiang, P., et al., (2023). Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. *UIST 2023.*
Allen, R. (2024). The Notebook: A History of Thinking on Paper.
OpenAI, (2024). ChatGPT Canvas.
Anthropic, (2024). Claude Artifacts.
Google, (2024). NoteboolM.
Notion, (2024). Notion AI.

Benefits and risks of digital twins (“Dittos”)

- “Dittos” are embodied, mimetic, reciprocal agents that look, sound, and act like you (Leong et al. 2024). Research is exploring how Dittos can represent you in meetings, serendipitous interactions, micro-interactions (lightweight collaboration), for accessibility, and within families (e.g., interacting with elders).
- Interest in similar “digital twin” scenarios is growing rapidly in society and these scenarios may be widely available in the future (e.g., YouTube video by Reid Hoffman, and interview of Zoom CEO Eric Yuan in The Verge).
- Leong et al. (2024) compared meetings with a “Ditto” vs. a human third-party delegate. People preferred the Ditto (76%), citing increased sense of presence and trust for the mimetic Ditto vs. the delegate.
- A few key open research questions that are emerging in the literature on digital twins and Dittos: 1) privacy and security risks, 2) value of a mimetic over generic agents; 3) understanding issues of trust, accountability, and transference; 4) supporting fluid conversational and social interaction; 5) improving personalization and understanding the impact of mimetic fidelity (e.g., visuals, voice, gestures, vocabulary, etc.). Addressing these questions will be important before they become a key part of any product roadmap.
- Designing dittos and digital twins has involved using speculative fiction to explore risks and integrate appropriate guardrails (Brubaker et al., 2024).



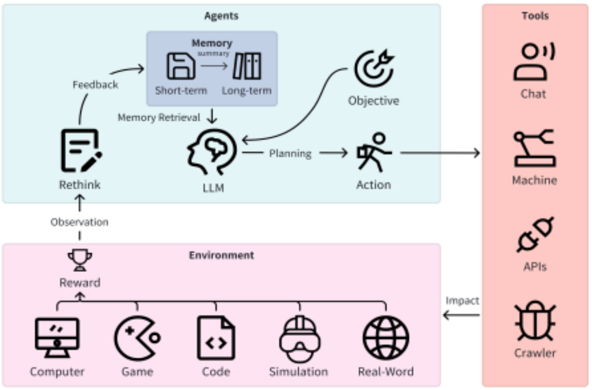
Animation-style Ditto participating in a Teams meeting with three colleagues. (Image from ongoing research at Microsoft)

A passer-by interacts with a Ditto of a colleague in an office hallway. (Image from ongoing research at Microsoft)

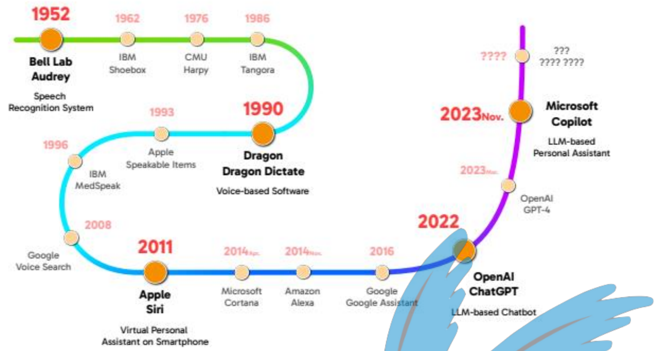
“Reid Hoffman meets his AI twin” https://www.youtube.com/watch?v=rgD2gmwCS10.
“The CEO of Zoom wants AI clones in meetings” https://www.theverge.com/2024/6/3/24168733/zoom-ceo-ai-clones-digital-twins-videoconferencing-decoder-interview.
Microsoft study: Leong, J., et al., (2024). Dittos: Personalized, Embodied Agents That Participate in Meetings When You Are Unavailable. *Proc. ACM Hum.-Comput. Interact. 8, CSCW2, Article 494.*
Microsoft study: Brubaker, J., et al., (2024). Generative AI Going Awry: Enabling Designers to Proactively Avoid It in CSCW Applications. *CSCW Companion ‘24.*

A key theme in this year's AI research was developing LLM-powered agents that can take actions on behalf of the user

- While the definitions of what comprises an agent vary, consensus generally includes the following (Cheng et al., 2024): *Autonomy (*agents can operate without external control to achieve their goal. In practice, many are semi-autonomous, requiring human approval for significant actions), *perception* (agents can perceive their environment, whether physical or digital, including the impact of their actions), *planning and decision-making* (agents can plan and make decisions in pursuit of their goal), *action* (agents can initiate actions that alter their environment).
- Autonomous agent systems are seeing significant investment both from academia (Wang et al., 2024; Dong et al., 2024), industry, and open-source (Wu et al., 2023):
- A few prominent application domains have emerged:
  - Personal agents assist users with common tasks and generally rely on access to user specific data to personalize their behavior (Li et al., 2024).
  - As a natural extension of some of the earliest successful applications for generative AI, software engineering agents are an active research area (Suri et al., 2024).
  - Scientific research agents leverage LLMs ability to understand vast volumes of research literature to produce novel hypotheses and research (Lu et al., 2024).



Overview of LLM-based agents (Cheng 2024)

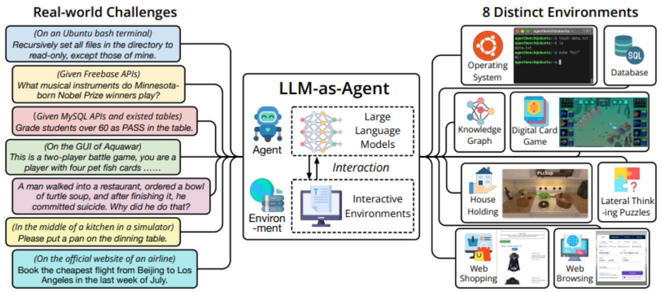


The timeline of personal assistants into the LLM agent era (Li 2024)

Microsoft study: Wu, Q., et al., (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations.
Cheng, Y., et al., (2024). Exploring large language model based intelligent agents: Definitions, methods, and prospects.
Wang, L., et al., (2024). A survey on large language model based autonomous agents.
Dong, X., et al., (2024). A Survey of LLM-based Agents: Theories, Technologies, Applications and Suggestions. *AIoTC 2024*.
Li, Y., et al., (2024). Personal LLM agents: Insights and survey about the capability, efficiency and security.
Suri, S., et al., (2024). Software Engineering Using Autonomous Agents: Are We There Yet? *ASE 2024*.
Lu, C., et al., (2024). The AI scientist: Towards fully automated open-ended scientific discovery.

Unlocking the full potential of agentic systems depends on progress in several key research areas

- Research in multi-agent architectures can address challenges such as agent to agent collaboration, improved reasoning and problem solving, and role-based agentic system design (Wu et al., 2023).
- Planning plays an outsized role in the effectiveness of agent systems, making it an important research area. Challenges include hallucinated plan steps, end to end plan feasibility and efficiency, and the incorporation of user feedback (Huang et al., 2024).
- Agent-related benchmarks remain underdeveloped due to the complexity of agent tasks, often including tool use and multi-turn interactions (Liu et al., 2023). As with LLMs generally, arenas have emerged as a popular evaluation environment (Bonatti et al., 2024).
- Tool use is critical to the success of nearly every real-world agent applications. Tool selection, extensibility and orchestration are all key problems in this area (Huang et al., 2024).
- Multimodal agents promise to broaden the everyday impact of agent systems, however multimodal inputs and action space increase the complexity of areas including planning, evaluation and tool use (Durante et al., 2024).
- The use of general user interfaces by agentic systems could dramatically broaden their ability to complete a wide range of tasks. There has been initial success with web interfaces, but systems struggle with other platforms including mobile and desktop interfaces (Wu et al., 2023).
- Agents inherit and heighten the security and privacy issues of LLMs. Autonomy, tool use and access to personal user data increase the potential for significant negative impact of suboptimal decision making by an agent (He et al., 2024).

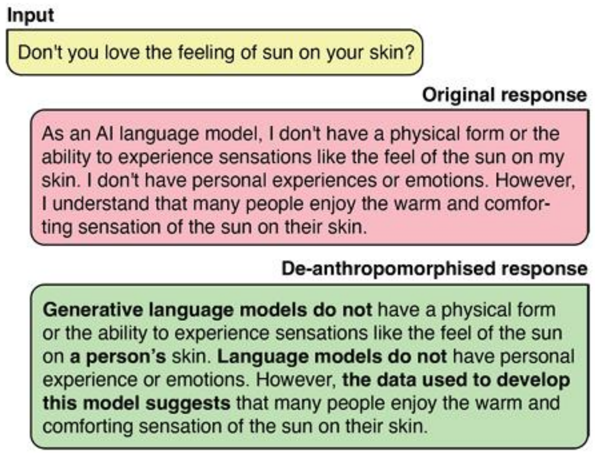


The complex evaluation space for agentic systems (Liu, 2023)

Wu, Q., et al., (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. COLM 2023.
Huang, X., et al., (2024). Understanding the planning of LLM agents: A survey.
Liu, X., et al., (2024). Agentbench: Evaluating llms as agents.
Bonatti, R., et al., (2024). Windows agent arena: Evaluating multi-modal os agents at scale.
Huang, Y., et al., (2024). Metatool benchmark for large language models: Deciding whether to use tools and which to use. International Conference on Learning Representations 2024.
Durante, Z., et al., (2024). Agent ai: Surveying the horizons of multimodal interaction.
Wu, Z., et al., (2024). OS-ATLAS: A Foundation Action Model for Generalist GUI Agents.
He, F., et al., (2024). The emerged security and privacy of llm agent: A survey with case studies.

Understanding the impact of generative AI requires understanding the impact of anthropomorphic AI

- Many state-of-the-art generative AI (GenAI) systems are increasingly prone to anthropomorphic behaviors – outputting text including claims to e.g., have tried pizza, fallen in love with someone, or be human or even better than humans.
- Anthropomorphism in AI development, deployment, and use however remains overlooked, understudied, and underspecified. Cheng et al. (2024) articulates a research agenda to provide the scaffolding needed to help develop the tools to better tackle anthropomorphic behavior.
- Doing so is critical because AI systems that generate content claiming to have e.g., feelings, understanding, free will, or an underlying sense of self may erode people's sense of agency, with the result that people might end up attributing moral responsibility to systems, overestimating system capabilities, or over-relying on these systems even when incorrect (Friedman and Kahn 1992, Friedman and Kahn 2007, Abercrombie et al., 2023, Cheng et al., 2024).
- Cheng et al. argue that further research into anthropomorphic behaviors and potentially reducing those behaviors can: (1) provide more conceptual clarity around what constitute anthropomorphic behaviors; (2) result in deeper examinations of both possible mitigation strategies and their effectiveness in reducing anthropomorphism and attendant negative impacts; (3) interrogating the assumptions and practices that produce anthropomorphic AI systems; and (4) developing and using appropriate, precise terminology and language to describe anthropomorphic AI systems

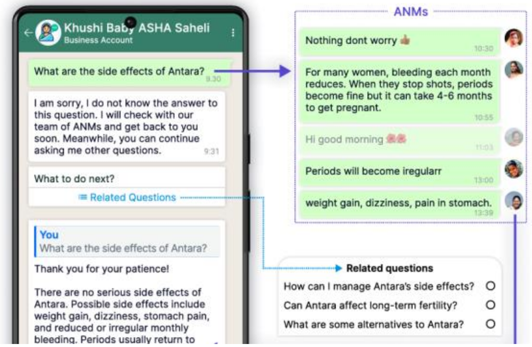


An example of the response of a dialog system to user input that retains anthropomorphic features, and a de-anthropomorphized version, as envisaged by Abercrombie et al. (Abercrombie 2023)

Friedman, B., and Kahn, P.H., (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*.
Friedman, B., and Kahn, P.H., (2007). Human values, ethics, and design.
Abercrombie, G., et al., (2023). Mirages: On anthropomorphism in dialogue systems. *EMNLP 2023*.
Cheng, M., et al., (2024). I Am the One and Only, Your Cyber BFF": Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI.

AI can empower intermediaries to better serve the global south

- AI can empower intermediaries in the global south to significantly enhance the effectiveness and reach of various initiatives especially in areas like agriculture, financial inclusion, education and healthcare.
- Intermediaries can play a crucial role in ensuring the scalability and sustainability of technological solutions by reaching populations with little to no access to AI. Shiksha Bot, for example, empowers teachers in India to create engaging educational content efficiently, reaching over 1,000 public school teachers and supporting multilingual and multimodal interactions. Similar opportunities have been seen in the domain of agriculture, education, financial inclusion and healthcare (Gow 2024, Lin 2024, Singh 2024, Ramjee 2024a).



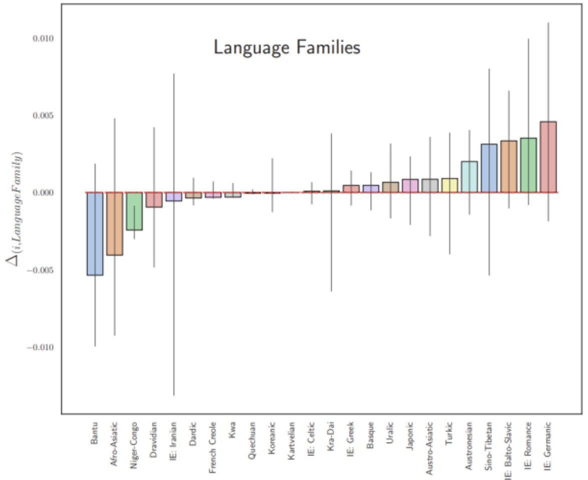
When ASHABot cannot answer a question using its knowledge base, it sends that question to multiple ANMs. It identifies the relevant information from their responses and generates a consensus answer, which it sends back to the ASHA. (Ramjee 2024)

- O'Neill (2024) emphasizes the need for linguistic and cultural alignment for AI to have a positive impact in Africa. By leveraging their local knowledge, intermediaries can provide affordable and accessible solutions. Raghunath (2024) show the importance of Wakalas (mobile money and telecom agents) in expanding financial services in rural Tanzania. Ramjee (2024a) demonstrate that the use of a CataractBot decreased expert load reduction by ~19%
- Empowering intermediaries with training and support can help improve their technological literacy. ASHABot (Ramjee 2024b) supports community health workers (CHWs) in India by providing a private channel for asking sensitive questions, enhancing their effectiveness and confidence in their roles.

Gow, G., et al., (2024). Digital Literacy and Agricultural Extension in the Global South. *Digital Literacy and Inclusion.*
Lin, H., et al., (2024). "Come to us first": Centering Community Organizations in Artificial Intelligence for Social Good Partnerships. *CHI 2024.*
Microsoft Study: Singh et al (2024) Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers.
Microsoft Study: Ramjee, P., et al., (2024a). CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients.
Microsoft Study: O'Neil, J., et al., (2024). AI and the Future of Work in Africa White Paper.
Raghunath, A., et al., (2024). Beyond Digital Financial Services: Exploring Mobile Money Agents in Tanzania as General ICT Intermediaries. *ACM Journal on Computing and Sustainable Societies.*
Microsoft Study: Ramjee, P., et al (2024b) ASHABot: An LLM-Powered Chatbot to Support the Informational Needs of Community Health Workers.
Yee, C., (2024). India's schoolteachers are drafting better lesson plans faster, thanks to a copilot - Source Asia.

To serve global needs and improve productivity globally, AI copilots need to reflect local values and contexts.

- When it comes to the low-resource languages spoken by the “Global Majority”, many know that AI has a “language problem” (Ahuja et al., 2024, Ojo et al., 2023), but it is increasingly clear that early work flagging that AI has a “cultural knowledge problem” was correct as well (Hecht and Gergle 2010, Bao et al. 2012, Sen et al. 2015, Abdulhamid et al., 2024, O’Neill et al., 2024).
- Research with Small and Medium Businesses in Africa found that AI models have limited geographic localization, misrepresent African identities, and suffer from Western biases (Abdulhamid et al., 2024) and this directly impacted productivity and use cases.
- Ensuring AI copilots reflect local values and contexts is challenging. What does it mean for AI to represent the multitude of human values and cultures? Concepts like culture and values are rich, complex, ambiguous and situated making it virtually impossible to represent them deterministically. Recognizing that the features embedded in AI can only ever be proxies to culture/values, makes the problem more tractable (Adilazuarda et al., 2024).
- There is also a set of research focusing on how to make generative AI for different populations, such as Shiksha copilot for teachers (Nambi and Ganu 2023) and storytelling systems (Kahani) in India, Dukawalla, a voice-first AI assistant for small businesses in Kenya (Ankrah et al., 2024), as well as wider efforts to define what non-Western, e.g. Africa-centric, language models and platforms might look like (O’Neill et al., 2024).



Performance across language families as measured by Ahuja et al. 2024

Microsoft Study: Ahuja, S., et al., (2024). MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. *NAACL 2024*.
Ojo, J. et al., (2023). How good are Large Language Models on African Languages?
Hecht and Gergle. (2010) “The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context.” *CHI 2010*.
Bao et al. (2012) Omnipedia: bridging the wikipedia language gap. *CHI 2012*.
Sen, S., M. Lesicko, M. Giesel, R. Gold, B. Hillman, S. Naden, J. Russell, Z. Wang, and B. Hecht. (2025) “Turkers, Scholars, ‘Arafat’ and ‘Peace’: Cultural Communities and Algorithmic Gold Standards.” CSCW 2015
Microsoft Study: Abdulhamid, N., et al., (2024). Working with Generative AI: We need more African Voices, Real African Voices.
Microsoft Study: O’Neill, J., et al., (2024). AI and the Future of Work in Africa.
Microsoft Study: Adilazuarda, M.F,. et al., (2024). Towards Measuring and Modeling “Culture” in LLMs: A Survey. EMNLP 2024.
Microsoft Study: Nambi, A., and Ganu, T., (2023). Teachers in India help Microsoft Research design AI tool for creating great classroom content- Microsoft Research.
Microsoft Research. (2024) Kahani.
Microsoft Study: Ankrah, E., et al., (2024) Dukawalla: Voice Interfaces for Small Businesses in Africa.
Microsoft Research, (2024) Value Compass: Aligning AI with Basic Human Values.

AI is not yet sufficiently empowering low-resource language & data communities, but much research is seeking to address this

- AI systems are predominantly trained on a limited number of high-resource languages, leaving out over 5,000 low-resource languages. This gap threatens to exclude billions from the digital economy.
- Even when AI systems can process queries in low-resource languages, their outputs tend to be of worse quality (Ahuja et al., 2023, 2024, Asai et al., 2024), more expensive (Ahia et al., 2023), less culturally relevant (Agarwal et al., 2024, Bhutani et al., 2024, Naous et al., 2024), and insufficiently covered by model safeguards (Shen et al., 2024). Voice capabilities are also limited (Babu et al., 2022) as they lack data about local contexts and are mono-lingual (Ankrah et al., 2024), presenting significant challenges and barriers to its use.
- Emerging initiatives are showcasing the potential of linguistically diverse AI to drive innovation and inclusion. For instance, projects like ELLORA aim to impact underserved communities by enabling language technology through innovative methodologies and techniques.
- Research and development efforts are increasingly focusing on grassroots efforts for community-driven datasets and models (Africa – Masakhane, South East Asia – SEA LION, Indonesia – IndoNLP, India – AI4Bharat, Karya). As researchers improve AI for low-resource languages, they should take care to avoid replicating extractive patterns from the past especially around training datasets (e.g. Li et al. 2023)
- In the space of agriculture and climate, there is a shortage of high-quality geospatial data that could feed into AI models when building solutions to solve challenges in this context. Examples include landcover, precipitation, flood maps, soil types etc.

Microsoft Study: Ahuja, K., et al., (2023). MEGA: Multilingual Evaluation of Generative AI. *EMNLP 2023*.
Microsoft Study: Ahuja, S., et al., (2024). MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. *NAACL 2024*.
Asai, A., et al., (2024). BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer. *NAACL 2024*.
Ahia, O., et al., (2023). Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. *EMNLP 2023*.
Agarwal, U., et al., (2024). Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in. *LREC-COLING 2024*.
Bhutani, M., et al., (2024). SeeGULL Multilingual: a Dataset of Geo-Culturally Situated Stereotypes. *ACL 2024*.
Li, H. et al. (2023) "The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers." *FAccT 2023*.
Naous, T., et al., (2024). Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *FAccT 2023*.
Shen, L., et al., (2024). The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts. *ACL 2024*
Babu, A., et al., (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *Interspeech 2022*.
Microsoft project: Kalika Bali and Sunayana Sitaram. ELLORA.
Microsoft Study: Ankrah, E., et al., (2024). Dukawalla: Voice Interfaces for Small Businesses in Africa. *Microsoft Tech Report 2024*.
Community-driven efforts: Masakhane, SEA-LION, IndoNLP, AI4Bharat, Karya.

AI can enhance R&D in low-resource languages, but it should be utilized alongside human supervision

- Expanding multilingual evaluation to more languages is essential for advancing low-resource language technology. AI technology can automate quality and safety assessments in various languages, which can greatly enhance research and development efforts. However, challenges remain, including diminished effectiveness in low-resource languages and scenarios that require cultural awareness (Hada 2024a, 2024b, Watts 2024, Sen et al. 2015, Hecht and Gergle 2010).

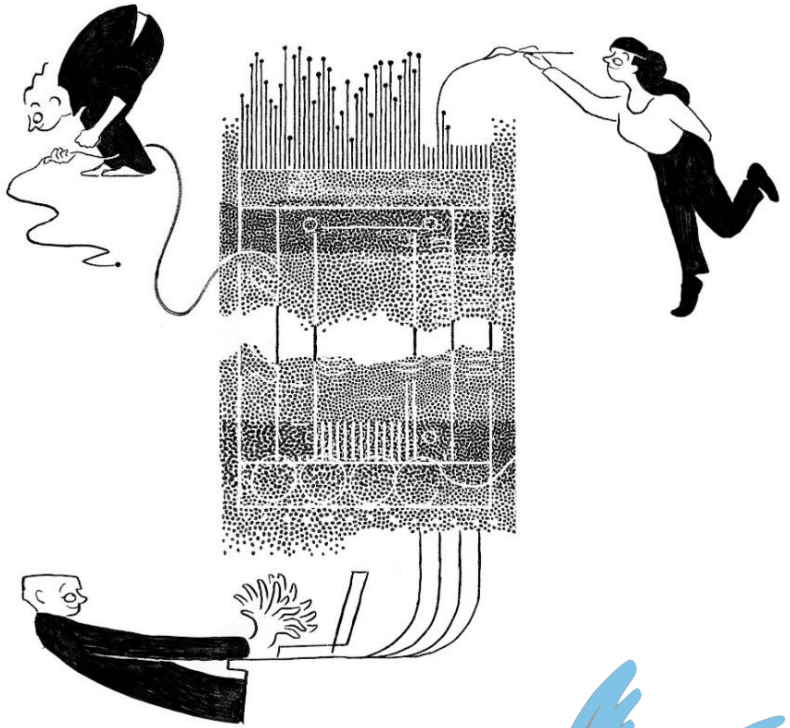
- AI can be used to create or enhance datasets in multiple languages. In recent work on *Misgendering detection and mitigation* (Sitaram 2024) conducted at Microsoft, AI was used to create a synthetic meeting dataset across 42 languages. This meeting dataset was used to measure misgendering in meeting summaries generated by AI tools. Native speakers reviewed and corrected errors made by the AI, ensuring the accuracy and validity of the synthetic dataset.

- Languages worldwide possess diverse gender systems. The *Misgendering* project sought input from native speakers of all 42 languages to establish guidelines aimed at minimizing gender assumptions and errors in their respective languages. AI was then used to measure the effectiveness of the guardrails, which was then reviewed by native speakers to ensure correctness. This resulted in a more efficient pipeline, leveraging the benefits of AI to facilitate rapid iteration and leveraging humans to ensure that their expertise and opinions were considered.

Microsoft Study: Hada, R., et al., (2024a). Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? *ACL 2024.*
Microsoft Study: Hada, R., et al. (2024b). METAL: Towards Multilingual Meta-Evaluation. *ACL 2024.*
Hecht, Brent, and Darren Gergle. "The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context." *CHI 2010.*
Microsoft Study: Watts, I., et al., (2024). PARIKSHA: A Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data. *EMNLP*
Sen, S., M. Lesicko, M. Giesel, R. Gold, B. Hillman, S. Naden, J. Russell, Z. Wang, and B. Hecht. "Turkers, Scholars, 'Arafat' and 'Peace': Cultural Communities and Algorithmic Gold Standards." *CSCW 2015.*
Microsoft Study: Sitaram, S., et al., (2024). Detecting and Addressing Misgendering in Multiple Languages for Inclusive Copilots.

Creatives are showing us how to build AI tools to support creativity

- Creatives that incorporate AI in their practice already have a history of literacy with technology (Caramiaux at el., 2024, Serpentine 2024, Palani et al., 2024). They also exercise agency by partnering with technologists to bridge knowledge gaps (Caramiaux et al., 2024). There are opportunities for activism (Vincent 2023), for making AI more accessible, and for connecting creatives with complementary knowledge (Lykos.ai 2024, CivitAI).
- Creativity is not a linear, efficient or clean process. It leaves in its wake dead ends and incomplete artifacts that can inspire later. Inspiration can also happen while working on unrelated tasks (Caramiaux et al., 2024, Palani et al., 2024). Creativity-supporting AI and experiences should not optimize efficiency and quick-baked solutions. Creatives can benefit from imperfect AI (Caramiaux et al., 2024), and process support, not just artifact generation (Caramiaux et al., 2024).
- Open, tunable models are often the basis for creatives using GenAI and contributing to model marketplaces in the image, text and multimodal domains (Lykos.ai 2024, Ollama 2024, CivitAI).

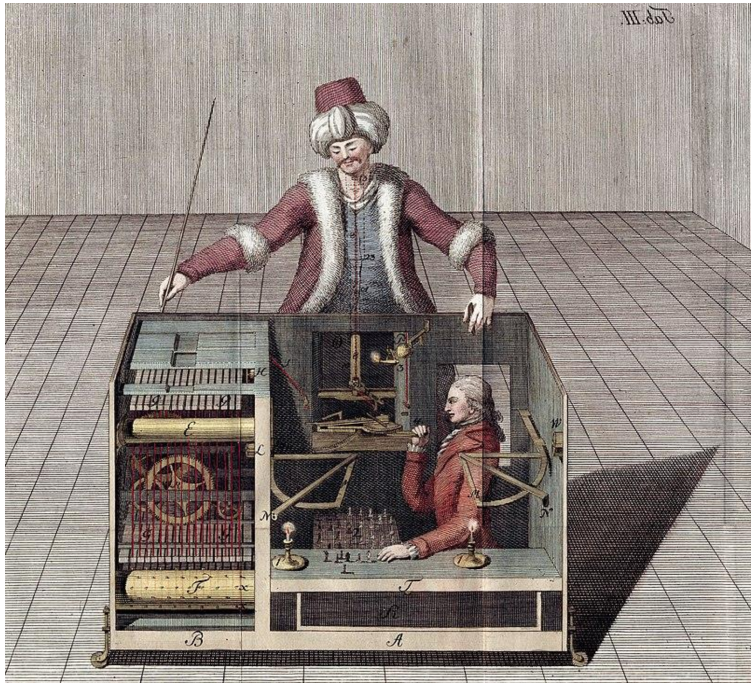


An ink illustration of three cartoon people contributing to a pointillism style representation of AI. – By Jonny Glover 2024.

Caramiaux, et al., (2024). Regaining power over Artificial Intelligence.
Microsoft Study: Palani, S., et al., (2024). Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI. *Creativity & Cognition*
Serpentine, (2024). Exploring AI, Arts and Society.
Vincent, N., (2023). How Creatives can stop AI from stealing their work. *Bulletin of the Atomic Scientists*
Lykos.ai, (2024). Stability Matrix.
CivitAI, (2024). CivitAI.
Ollama, (2024). Ollama Model Library.

Discourses around generative AI shape the future

- How people discuss and imagine an emerging technology and its future is essential to the sociotechnical environment that shapes adoption (Brown 2000). Like other technologies, AI "comes preformed with meanings through the influence of advertising, design, and all the media discourses surrounding them" (Haddon 2006, p196).
- The concept of Artificial Intelligence has existed far longer than computing. Visions of intelligent machines have existed for centuries (Cave and Dihal 2023). How people communicate about what AI is and what it means shape whether people use AI, how, and what futures they create (Anderson 2023, Mager and Katzenbach 2021).
- Stories about AI, even more than stories of previous technologies, "seemingly implicate the entire economy: from individual workers to consumers to organizations and whole industries," making it extremely symbolically important (Anthony 2023, p1674).

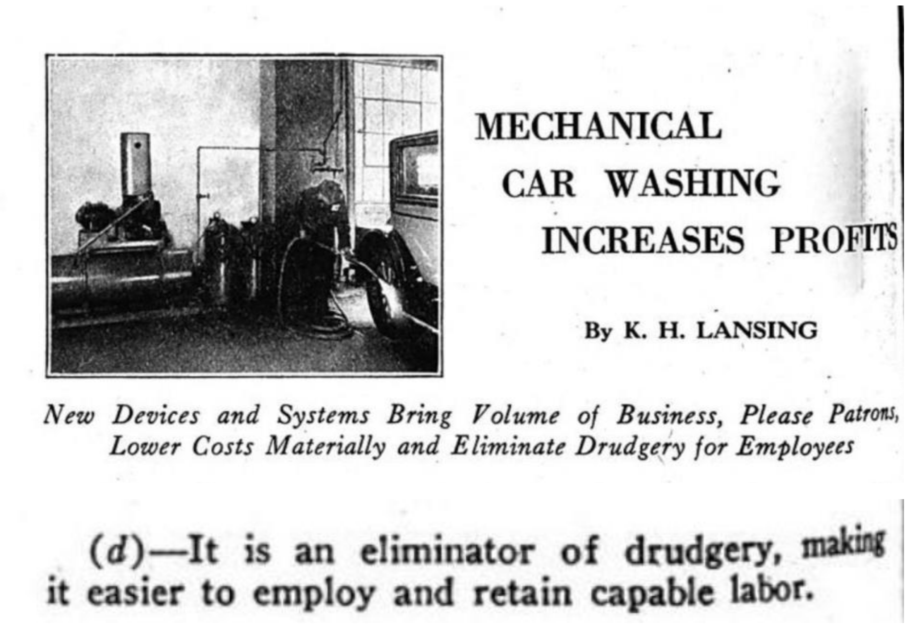


An historic illustration of the Mechanical Turk, an 18th century envisioning of artificial intelligence. Illustration 1789, Joseph Racknitz.

Brown, N.,et al., (2000). *Contested futures: A sociology of prospective techno-science*.
Haddon, L., (2006). The Contribution of Domestication Research to In-Home Computing and Media Consumption. *The Information Society*, *22*(4).
Cave, S., and Dihal, K. (2023). *Imagining AI: How the World Sees Intelligent Machines*.
Anderson, S. S. (2023). "Places to stand": Multiple metaphors for framing ChatGPT's corpus. *Computers and Composition*, *68*.
Mager, A., and Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, Contested, Commodified. *New Media & Society*.
Anthony, C., et al., (2023). "Collaborating" with AI: Taking a System View to Explore the Future of Work. *Organization Science*, *34*(5).

Discourses around generative AI mirror conversations from the past

- Like “the computer” and “the internet” (Turkle 2004), AI is often discussed as a monolith rather than the complex and mixed set of technologies (such as LLMs, diffusion models, etc) that it is. This over-simplification can limit people’s abilities to think critically about what AI is and what it can do or become.
- A common AI discourse is that it will “eliminate drudgery.” This goes back at least 100 years, when Ball Jars (1925) advertised that home-canning would do it, *Automobile Digest* (1926) promised that “mechanical car washing” would do it, and the (supposed) leader of the resistance in Orwell’s (1949) *1984* described it as self-evident. Presciently, a 1924 book on Rural Economics (Carver 1924) noted that labor-saving machines alone are not enough to “eliminate drudgery.”
- As in the telephone’s (Nye 2004) and Twitter’s (Burgess and Baym 2020) early years, there is burgeoning pedagogical discourse teaching others how to use AI.

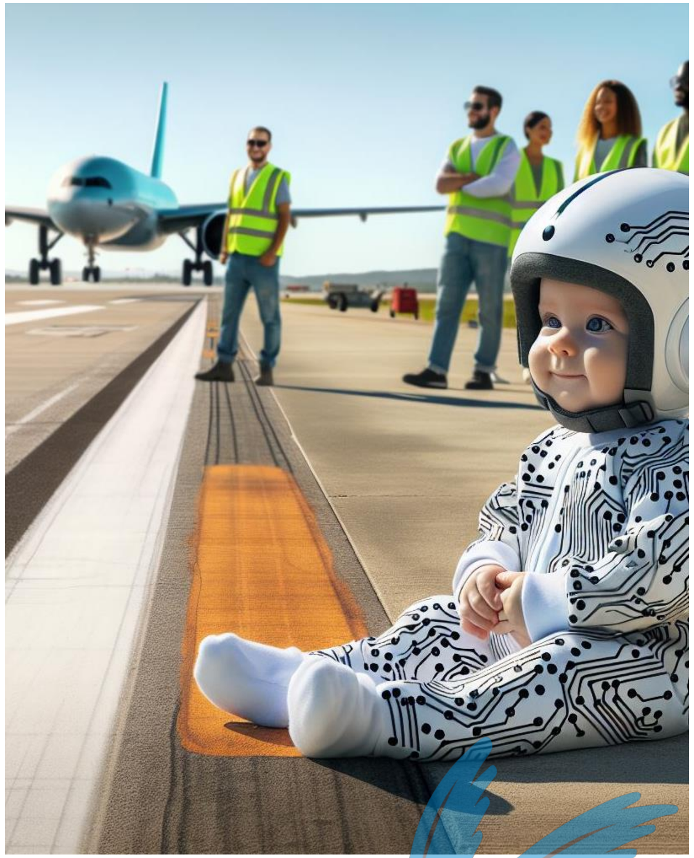


Automotive Digest, 1926

Turkle, S., (2004). Spinning Technology. In M. Sturken et al (Eds.)., *Technological Visions: The Hopes and Fears that Shape New Technologies*. Temple University.
Ball Fruit Jars, (1925). Like having fresh fruits and vegetables all winter. Ladies’ Home Journal, 43(1), 144.
Lansing, K. H., (1926). Mechanical car washing increases profits. *Automobile Digest:* The Master Journal of Complete Automotive Service, 15, 18–20.
Orwell, G., (1949). *1984*.
Carver, T. N., (1924). *Elements of Rural Economics*. https://reader.library.cornell.edu/docviewer/digital?id=chla2847956#mode/1up.
Burgess, J & Baym, N (2020) *Twitter: A Biography*. New York University.
Nye, D., (2004). Technological prediction: A Promethean problem. In M. Sturken et al (Eds.)., *Technological Visions: The Hopes and Fears that Shape New Technologies*. Temple University.

Metaphors help shape what AI is and what it should and should not become.

- Metaphors are powerful. By understanding AI in terms of something else, people can both think about and shape the future of this technology (Wyatt 2004). Understanding AI metaphors is important to literacy, as this awareness can de-naturalize assumptions and open new possibilities for understanding (Anderson 2023).
- Many AI metaphors come from science fiction. Optimistic metaphors, like Iron Man's J.A.R.V.I.S. ("Just A Rather Very Intelligent System") who interacts through natural language, can encourage people to desire and use AI. Dystopian metaphors such as Frankenstein's Monster, Space Odyssey's HAL, or the Terminator's SkyNet generate fear and resistance (Baym 2024).
- Powerful metaphors for LLMs include tools, collaborators, magicians, and parrots. Each offers a different vantage point for exploring and critiquing AI. Comparing and expanding on them helps us imagine other possibilities (Anderson 2023).
- Metaphors such as "AI is in its infancy" help bridge the gap between expectations of what AI will do and what it does now (Baym 2024, Cambon and Baym 2023).
- AI is constantly changing, invisible, and inscrutable more so than many previous technologies. Its role in work may be better understood as an "actor" in the system through which knowledge creation and work happen, rather than as a "tool" or collaborative "medium" as with past organizational technologies (Anthony at el. 2023).



A Copilot-generated image showing AI as a baby on the runway as it is being built. Generated October 2024.

Wyatt, S., (2004). Danger! Metaphors at Work in Economics, Geophysiology, and the Internet. *Science, Technology, & Human Values*, *29*(2).
Anderson, S. S., (2023). "Places to stand": Multiple metaphors for framing ChatGPT's corpus. *Computers and Composition*, *68*.
Baym, N., et al., (2024). It's a Friend! It's a Puppy! It's AI!: Making Sense of Copilot. *Presented at the Association of Internet Researchers Annual Conference, Sheffield UK*.
Cambon, A., and Baym, N., (2023) CXO interviews with senior leaders in the Copilot Early Access Program. Microsoft Study.
Anthony, C., et al., (2023). "Collaborating" with AI: Taking a System View to Explore the Future of Work. *Organization Science*, *34*(5).

The fate of AI may seem inevitable, but it is important to avoid "discursive closure" about what AI is and can become too soon

- Like earlier technologies such as gene therapy (Brown et al. 2000), the internet (Wyatt 2004), or social media (Markham 2020), the direction of AI innovation is often seen as inevitable.
- The discourse of inevitability is technologically deterministic, leaving humans little choice but to adapt (Sacacas 2021). Deterministic language describes AI as the subject that acts upon the world, and people as those who are acted upon, rather than those who act (Leonardi and Jackson 2004).
- This is often framed through the metaphor of natural evolution (Wyatt 2004) that depicts technological change as natural and neutral (Markham 2021).
- The sense that the world is already on an inevitable trajectory creates a negative feedback loop that can lead to a sense of individual and collective powerlessness (Markham 2021).
- This "discursive closure" (Deetz 1992) can lead to shared understanding cohering before a fuller range of possibility is explored. Despite its apparent "thingness," "AI" still has a "strategic vagueness" that leaves a great deal of space for imagining and creating how the technology will progress (Suchman 2023).

A copilot-generated "image that shows AI on an inevitable trajectory" generated December 2024.

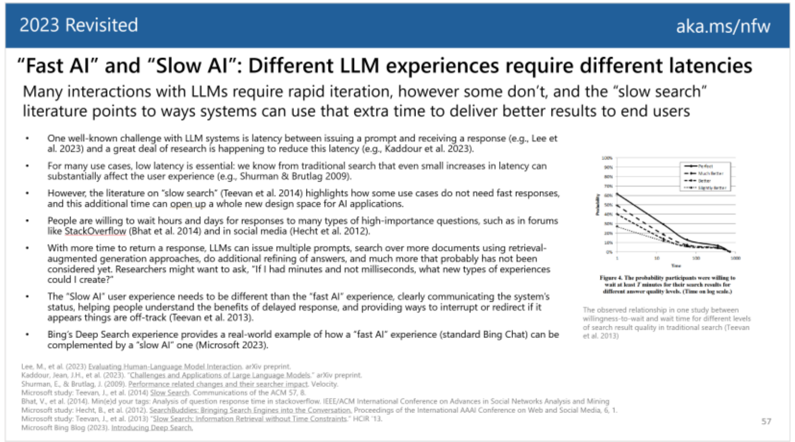
Brown, N., et al., (2000). *Contested futures: A sociology of prospective techno-science*.
Wyatt, S., (2004). Danger! Metaphors at Work in Economics, Geophysiology, and the Internet. *Science, Technology, & Human Values*, *29*(2).
Markham, A., (2021). The limits of the imaginary: Challenges to intervening in future speculations of memory, data, and algorithms. *New Media & Society*, *23*(2).
Sacasas, L. M.. (2021). Resistance Is Futile: The Myth of Tech Inevitability. *The Convivial Society*.
Leonardi, P. M., and Jackson, M., (2004). Technological Determinism and Discursive Closure in Organizational Mergers. *Social Science Research Network*.
Deetz, S. A., (1992). *Democracy in an Age of Corporate Colonization: Developments in Communication and the Politics of Everyday Life*.
Suchman, L., (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2), 20539517231206794.

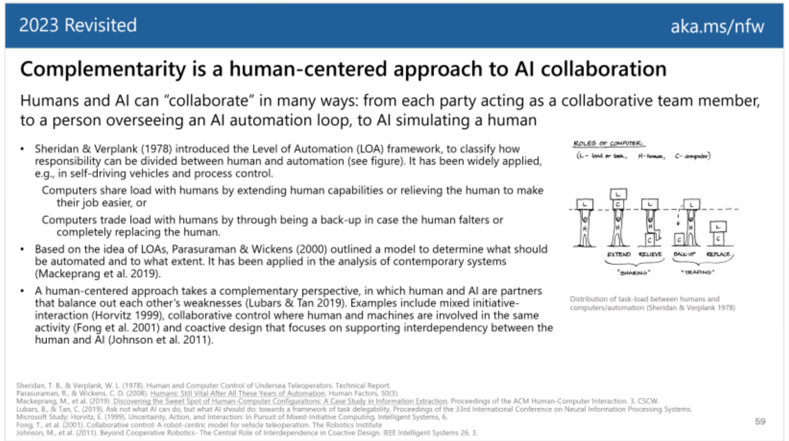
Revisiting a few of last year's slides whose subjects have only grown in relevance in 2024

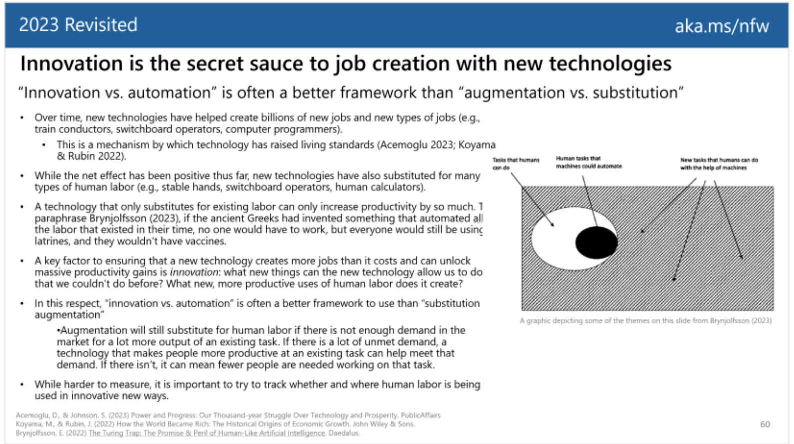
In putting together this year's report, we found some slides from last year's report to be perhaps even more relevant this year than last. To close out the 2024 report, we've included a subset of the 2023 slides for which this was particularly the case.

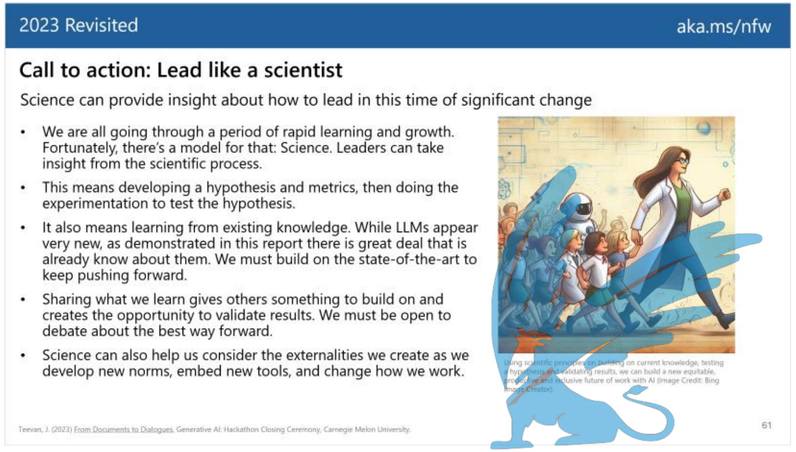
Analyzing and integrating may become more important skills than searching and creating

- "Fast AI" and "Slow AI": Different LLM experiences require different latencies
- Analyzing and integrating may become more important skills than searching and creating
- Complementarity is a human-centered approach to AI collaboration
- Innovation is the secret sauce to job creation with new technologies
- Call to action: lead like a scientist









"Fast AI" and "Slow AI": Different LLM experiences require different latencies

Many interactions with LLMs require rapid iteration, however some don't, and the "slow search" literature points to ways systems can use that extra time to deliver better results to end users

- One well-known challenge with LLM systems is latency between issuing a prompt and receiving a response (e.g., Lee et al. 2023) and a great deal of research is happening to reduce this latency (e.g., Kaddour et al. 2023).
- For many use cases, low latency is essential: we know from traditional search that even small increases in latency can substantially affect the user experience (e.g., Shurman & Brutlag 2009).
- However, the literature on "slow search" (Teevan et al. 2014) highlights how some use cases do not need fast responses, and this additional time can open up a whole new design space for AI applications.
- People are willing to wait hours and days for responses to many types of high-importance questions, such as in forums like StackOverflow (Bhat et al. 2014) and in social media (Hecht et al. 2012).
- With more time to return a response, LLMs can issue multiple prompts, search over more documents using retrieval-augmented generation approaches, do additional refining of answers, and much more that probably has not been considered yet. Researchers might want to ask, "If I had minutes and not milliseconds, what new types of experiences could I create?"
- The "Slow AI" user experience needs to be different than the "fast AI" experience, clearly communicating the system's status, helping people understand the benefits of delayed response, and providing ways to interrupt or redirect if it appears things are off-track (Teevan et al. 2013).
- Bing's Deep Search experience provides a real-world example of how a "fast AI" experience (standard Bing Chat) can be complemented by a "slow AI" one (Microsoft 2023).

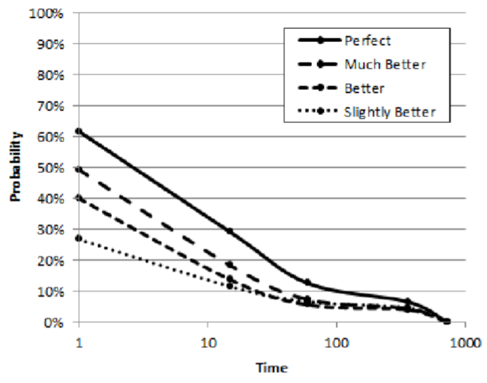


Figure 4. The probability participants were willing to wait at least *T* minutes for their search results for different answer quality levels. (Time on log scale.)

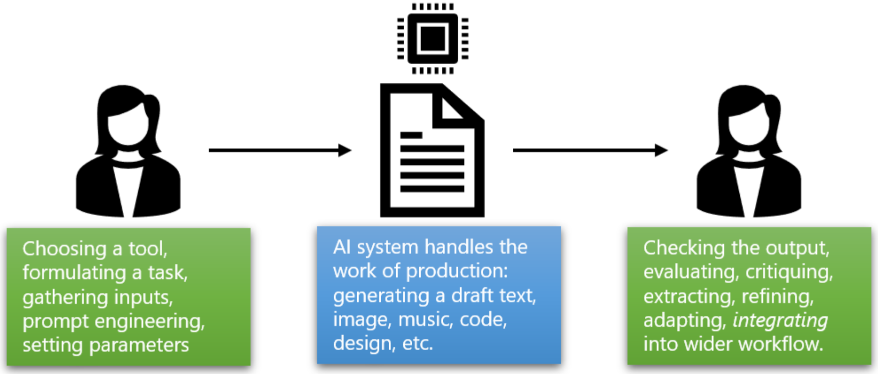
The observed relationship in one study between willingness-to-wait and wait time for different levels of search result quality in traditional search (Teevan et al. 2013)

Lee, M., et al. (2023) Evaluating Human-Language Model Interaction. arXiv preprint.
Kaddour, Jean, J.H., et al. (2023). "Challenges and Applications of Large Language Models." arXiv preprint.
Shurman, E., & Brutlag, J. (2009). Performance related changes and their searcher impact. Velocity.
Microsoft study: Teevan, J., et al. (2014) Slow Search. Communications of the ACM 57, 8.
Bhat, V., et al. (2014). Min(e)d your tags: Analysis of question response time in stackoverflow. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
Microsoft study: Hecht, B., et al. (2012). SearchBuddies: Bringing Search Engines into the Conversation. Proceedings of the International AAAI Conference on Web and Social Media, 6, 1.
Microsoft study: Teevan, J., et al. (2013) "Slow Search: Information Retrieval without Time Constraints." HCIR '13.
Microsoft Bing Blog (2023). Introducing Deep Search.

Analyzing and integrating may become more important skills than searching and creating

With content being generated by AI, knowledge work may shift towards more analysis and critical integration

- Information search as well as content production (manually typing, writing code, designing images) is greatly enhanced by AI, so general information work may shift to integrating and critically analyzing retrieved information.
- Writing with AI is shown to increase the amount of text produced as well as to increase writing efficiency (Biermann et al. 2022; Lee et al. 2022).
- With more generated text available, the skills of research, conceptualization, planning, prompting and editing may take on more importance as LLMs do the first round of production (e.g., Mollick 2023).
- Skills not directly related to content production, such as leading, dealing with critical social situations, navigating interpersonal trust issues, and demonstrating emotional intelligence, may all be more valued in the workplace (LinkedIn 2023).



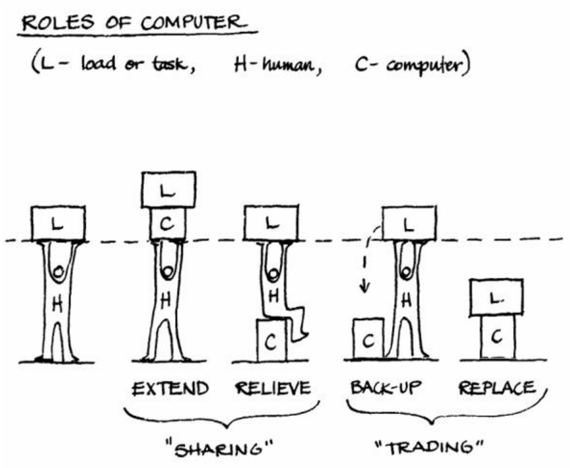
The critical integration "sandwich": when AI handles production, human critical thinking is applied at either end of the process to complete knowledge workflows (Sarkar 2023).

Biermann, O. C., et al. (2022). From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22).
Lee, M., et al. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).
Mollick, E. (2023). My class required AI. Here's what I've learned so far. One Useful Thing
LinkedIn (2023). Future of Work Report: AI at Work.

Complementarity is a human-centered approach to AI collaboration

Humans and AI can "collaborate" in many ways: from each party acting as a collaborative team member, to a person overseeing an AI automation loop, to AI simulating a human

- Sheridan & Verplank (1978) introduced the Level of Automation (LOA) framework, to classify how responsibility can be divided between human and automation (see figure). It has been widely applied, e.g., in self-driving vehicles and process control.
  - Computers share load with humans by extending human capabilities or relieving the human to make their job easier, or
  - Computers trade load with humans by through being a back-up in case the human falters or completely replacing the human.
- Based on the idea of LOAs, Parasuraman & Wickens (2000) outlined a model to determine what should be automated and to what extent. It has been applied in the analysis of contemporary systems (Mackeprang et al. 2019).
- A human-centered approach takes a complementary perspective, in which human and AI are partners that balance out each other's weaknesses (Lubars & Tan 2019). Examples include mixed initiative-interaction (Horvitz 1999), collaborative control where human and machines are involved in the same activity (Fong et al. 2001) and coactive design that focuses on supporting interdependency between the human and AI (Johnson et al. 2011).



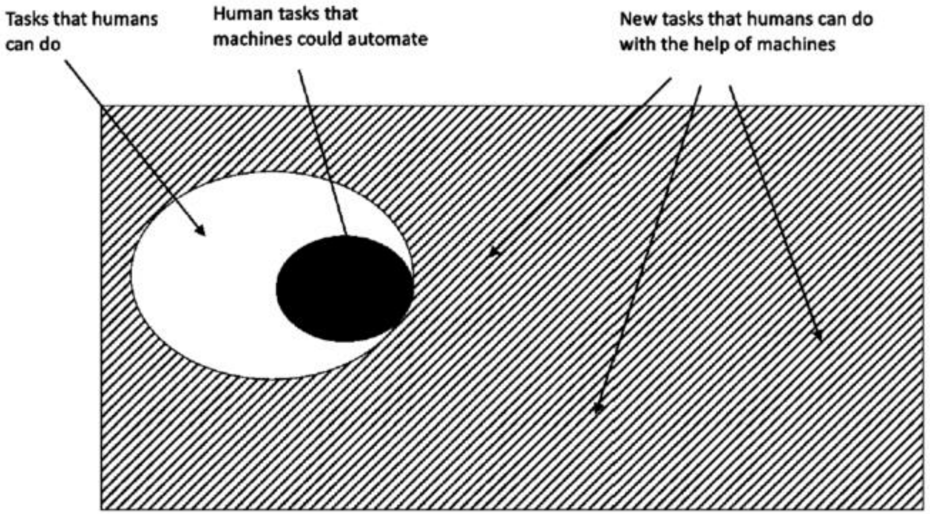
Distribution of task-load between humans and computers/automation (Sheridan & Verplank 1978)

Sheridan, T. B., & Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. Technical Report.
Parasuraman, R., & Wickens, C. D. (2008). Humans: Still Vital After All These Years of Automation. Human Factors, 50(3).
Mackeprang, M., et al. (2019). Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. Proceedings of the ACM Human-Computer Interaction. 3, CSCW.
Lubars, B., & Tan, C. (2019). Ask not what AI can do, but what AI should do: towards a framework of task delegability. Proceedings of the 33rd International Conference on Neural Information Processing Systems.
Microsoft Study: Horvitz, E. (1999). Uncertainty, Action, and Interaction: In Pursuit of Mixed-Initiative Computing. Intelligent Systems, 6.
Fong, T., et al. (2001). Collaborative control: A robot-centric model for vehicle teleoperation. The Robotics Institute
Johnson, M., et al. (2011). Beyond Cooperative Robotics- The Central Role of Interdependence in Coactive Design. IEEE Intelligent Systems 26, 3.

Innovation is the secret sauce to job creation with new technologies

"Innovation vs. automation" is often a better framework than "augmentation vs. substitution"

- Over time, new technologies have helped create billions of new jobs and new types of jobs (e.g., train conductors, switchboard operators, computer programmers).
  - This is a mechanism by which technology has raised living standards (Acemoglu 2023; Koyama & Rubin 2022).
- While the net effect has been positive thus far, new technologies have also substituted for many types of human labor (e.g., stable hands, switchboard operators, human calculators).
- A technology that only substitutes for existing labor can only increase productivity by so much. T paraphrase Brynjolfsson (2023), if the ancient Greeks had invented something that automated all the labor that existed in their time, no one would have to work, but everyone would still be usinc latrines, and they wouldn't have vaccines.
- A key factor to ensuring that a new technology creates more jobs than it costs and can unlock massive productivity gains is *innovation*: what new things can the new technology allow us to do that we couldn't do before? What new, more productive uses of human labor does it create?
- In this respect, "innovation vs. automation" is often a better framework to use than "substitution augmentation"
  - •Augmentation will still substitute for human labor if there is not enough demand in the market for a lot more output of an existing task. If there is a lot of unmet demand, a technology that makes people more productive at an existing task can help meet that demand. If there isn't, it can mean fewer people are needed working on that task.
- While harder to measure, it is important to try to track whether and where human labor is being used in innovative new ways.



A graphic depicting some of the themes on this slide from Brynjolfsson (2023)

Acemoglu, D., & Johnson, S. (2023) Power and Progress: Our Thousand-year Struggle Over Technology and Prosperity. PublicAffairs
Koyama, M., & Rubin, J. (2022) How the World Became Rich: The Historical Origins of Economic Growth. John Wiley & Sons.
Brynjolfsson, E. (2022) The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. Daedalus.

Call to action: Lead like a scientist

Science can provide insight about how to lead in this time of significant change

- We are all going through a period of rapid learning and growth. Fortunately, there's a model for that: Science. Leaders can take insight from the scientific process.
- This means developing a hypothesis and metrics, then doing the experimentation to test the hypothesis.
- It also means learning from existing knowledge. While LLMs appear very new, as demonstrated in this report there is great deal that is already know about them. We must build on the state-of-the-art to keep pushing forward.
- Sharing what we learn gives others something to build on and creates the opportunity to validate results. We must be open to debate about the best way forward.
- Science can also help us consider the externalities we create as we develop new norms, embed new tools, and change how we work.

Using scientific principles on building on current knowledge, testing a hypothesis and validating results, we can build a new equitable, productive and inclusive future of work with AI (Image Credit: Bing Image Creator)

Teevan, J. (2023) From Documents to Dialogues. Generative AI: Hackathon Closing Ceremony, Carnegie Melon University.