HUMAN-AI COLLABORATION AND THE FUTURE OF THE WORKFORCE

Dr. Muhammed Yusuf Kocyigit
Research Scientist, Google

Attribution: Kocyigit, M. Y. (2026). Human-AI collaboration and the future of the workforce. In H. Arslan, A. Taşdelen, E. Kuzucu Hıdır, & O. A. Çetin (Eds.), *Artificial intelligence and national technology reflections* (pp. 503–526). Turkish Academy of Sciences. https://doi.org/10.53478/TUBA.978-625-6110-86-1.ch21

10.53478/TUBA.978-625-6110-86-1.ch21

HUMAN-AI COLLABORATION AND THE FUTURE OF THE WORKFORCE

Dr. Muhammed Yusuf Kocyigit
Research Scientist, Google

Abstract

This chapter argues that artificial intelligence is becoming a general-purpose tool for converting energy into intellectual labor, analogous to how the electric motor converted electricity into physical force. The underlying capabilities already exist, and the conditions for extending them to new domains are improving rapidly. Drawing on evidence from software engineering, scientific research, and workforce studies, we show that AI systems are most powerful when supported by high-quality domain data and robust verification signals, and that their economic impact follows a recognizable cycle of data curation, reward design, iterative refinement, and human-AI workflow integration. The workforce implications are significant: the consistent finding across coding, customer support, consulting, and medical screening is that AI augments human capability before it displaces workers, with the largest gains going to less experienced practitioners. At the same time, early signals of hiring suppression in entry-level roles suggest that the transition will not be frictionless. The central shift is from execution to orchestration, where human comparative advantage moves toward judgment, verification, and system design rather than task completion. Whether this transition leads to broad-based gains or concentrated benefits will depend on institutional choices around compute access, education, career architecture, and gain-sharing mechanisms. The chapter examines macro constraints including energy demand and the productivity J-curve, and concludes that the window for shaping equitable outcomes is measured in years, not decades.

Keywords

Human-AI Collaboration, Future of Work, Large Language Models, AI Productivity

Introduction

In the span of three years, the world has gone from debating whether AI can write a passable essay to watching autonomous AI agents resolve real software engineering issues, predict complex biomolecular structures and interactions (Abramson et al., 2024), and achieve gold-medal performance at the International Mathematical Olympiad (DeepMind, 2025). The pace of change is not merely fast; it is accelerating in ways that confound traditional economic forecasting frameworks.

The central argument of this chapter is that AI will become a ubiquitous tool for turning energy into intellectual labor. The core capabilities are already in place, and the conditions for adapting new domains to this framework should improve over time. The workforce will shift from a model where humans execute every step to one where humans orchestrate autonomous systems, setting objectives, verifying outputs, and intervening when things go wrong. Whether this change converges toward augmentation or displacement will likely depend on institutional positioning and policy choices. The prevailing economic literature projects modest GDP gains from AI adoption over the coming decade (Acemoglu, 2025), but such estimates capture a snapshot of where capability stands today and miss the trajectory of capability improvement. This trajectory is visible in the empirical record of capability growth, deployment patterns, and organizational adaptation.

We present this argument by first examining the current state of model capabilities and limits, then looking at software engineering as the best-measured case. From there we offer a generalization framework for other domains, review workforce evidence on augmentation and displacement, and finally discuss the macro constraints around compute and energy. The future remains uncertain, but the direction of travel is becoming easier to see.

1. Gains and Limits of Current AI Models

1.1. On Large Language Models

Early public interpretations often treated large language models as text-prediction engines with fundamentally limited capabilities. The web-scale data that made their behavior impressive was also a source of fragility. Because these models were trained to represent broad distributions of text, they could generate highly fluent passages yet fail on seemingly simple tasks (McCoy et al., 2024). This pattern became known as the jagged frontier (Dell'Acqua et al., 2023). Some of this behavior may have been more predictable to people involved in model training, but there were still surprising effects, including elicitation studies in which cash incentives or emotional pressure changed model performance (Li et al., 2024). As the field learned more about how models acquire trajectories during pre-training and how post-training shapes their behavior (Yue et al., 2025), it became possible to treat models more concretely, focusing on their strengths without assuming they are uniformly capable.

Thus frontier systems in 2024-2026 show clear gains on reasoning and coding benchmarks through more task-focused data, reinforcement learning, and test-time compute across multiple labs (OpenAI, 2024; DeepSeek-AI, 2025; AI, 2026; Team, 2025; Hassabis & Kavukcuoglu, 2025). This improvement reflects better training recipes as well as better data and training signals for the target domains. Current frontier models are especially strong on in-distribution tasks, and for well-represented distributions it is often possible to push performance to levels that previously seemed out of reach for AI systems. Scaling and progress are not confined to one frontier; new frontiers are continually being discovered that improve how efficiently and effectively these models can advance (Kaplan et al., 2020; Hoffmann et al., 2022; DeepSeek-AI, 2024; Team, 2024; AI, 2025; Team, 2025).

Beginning in 2024, a new scaling axis emerged in test-time compute. Rather than only making larger models, researchers discovered that allocating more computation at inference time (step-by-step reasoning, candidate search, self-verification, and even test-time training) can yield performance improvements rivaling or exceeding training-time scaling (Snell et al., 2025; Akyurek et al., 2025). This opened a new paradigm in which models evolved from simple input-output systems into components in a scaffolding of other models and tools that execute specific roles in what are now called agentic systems. Although benchmark scores should be interpreted cautiously, models have surpassed expert-level performance on many challenging benchmarks. This pattern now appears across multiple independent labs, with Kimi K2.5, Qwen3, and Gemini 3 demonstrating similar gains (AI, 2026; Team, 2025; Hassabis & Kavukcuoglu, 2025). The lesson is that current models can excel at a task when they receive enough relevant data and training signal. For any domain where AI is expected to deliver economically useful work, those ingredients are still central. If the right data are curated, the right reward signals are designed, and the system is iterated, resulting performance tends to be far stronger than initial impressions of these models would suggest.

1.2. How Reward Signals Shape What Models Can Do

Language models begin by representing human-generated text according to patterns in the training distribution. But some generated trajectories are more useful, accurate, or valuable than others. Alignment improves this by increasing the probability of preferred trajectories and reducing the probability of unwanted ones. Recent reasoning-model work including OpenAI's o-series, DeepSeek-R1, Kimi K2.5, and Qwen3 all emphasize reinforcement-learning stages that improve hard reasoning and coding outcomes (OpenAI, 2024; OpenAI, 2025; DeepSeek-AI, 2025; AI, 2026; Team, 2025). In coding and mathematics, the loop is especially efficient because correctness can be checked by compilers, tests, proof systems, and execution traces. These were therefore among the first domains where AI systems began to rival or exceed human performance. Coding also appears to be the first domain where industry-scale impact and rapid workforce adoption are visible in daily work.

2. Coding as the Best-Measured Case

Software engineering is one of the earliest domains where the necessary conditions for a full AI development cycle are simultaneously visible, making it a uniquely informative case study. Two properties make software engineering the ideal proving ground. First, data abundance: public code ecosystems provide a massive corpus of observable expert work (issue reports, pull requests, code reviews, test outcomes) including the full debugging process, not just finished artifacts. Second, automated verifiability: a model can write code, execute test suites, and revise based on results without needing a human to judge correctness. These properties, precisely what the scaling and alignment literature identifies as necessary for strong in-distribution performance, create a tight feedback loop that can enable AI models to improve rapidly.

The empirical evidence from AI coding assistants is substantial and growing rapidly. The earliest controlled experiment by Peng et al. (2023) found that developers using GitHub Copilot completed a coding task 55.8% faster than a control group. The largest study to date, conducted across Microsoft, Accenture, and a Fortune 100 company with 4,867 developers, found a 26% increase in weekly pull request completions, with less experienced developers seeing gains of 35–39% (Cui et al., 2024).

The picture is more nuanced, however. The METR study (METR, 2025), an RCT with experienced open-source developers on their own repositories, found AI tools slowed them by 19%, even as developers believed they were faster. AI tools excel with abundant pattern-matched context but may introduce friction where developers possess deep tacit knowledge the model lacks. A difference-in-differences study of Cursor adoption across 807 repositories found transient velocity gains alongside persistent increases in technical debt (He et al., 2025).

The macro-level picture complements these task-level studies. PwC's 2025 Global AI Jobs Barometer found that productivity growth in AI-exposed industries (financial services, software publishing) nearly quadrupled from 7% to 27% since 2022, with AI-exposed industries generating 3× higher revenue per employee and AI-skilled roles commanding a 56% wage premium (PricewaterhouseCoopers, 2025).

Beyond controlled experiments, broader adoption data tells a consistent story. The Stack Overflow 2025 Developer Survey reports that 84% of respondents are using or plan to use AI coding tools, with 51% of professional developers using them daily. Speed is not the only dimension improving either. Among teams that integrate AI into code review workflows, 81% report measurable quality gains, suggesting that the tools are learning to assist not just with generation but with verification (Overflow, 2025; Qodo, 2025). METR's "50% task-completion time horizon" (the length of task an AI agent can complete autonomously with 50% reliability) has been doubling approximately every seven months, reaching roughly one hour by early 2025, and this metric is measured on tasks not in public training data (METR, 2025).

Anthropic's study of millions of real-world Claude Code interactions confirms that coding accounts for nearly 50% of all agentic API tool calls (Anthropic, 2026). The 99.9th percentile of uninterrupted session duration nearly doubled between October 2025 and January 2026 (from under 25 to over 45 minutes). Part of the slower aggregate growth reflects user adaptation to the technology. Early movers and domain experts describe a sharper shift, reporting that coding has become less about typing each implementation detail and more about oversight, task definition, and selective intervention while agents perform much of the execution (Karpathy, 2025). This pattern is likely to spread across knowledge-work domains as more tasks are integrated into AI-mediated workflows.

The coding domain reveals the economic logic of AI capability: it is not that AI "understands" code in a human sense, but that it has absorbed vast in-distribution data and can be refined through verifiable feedback. Where these conditions are met, AI performs economically useful work. This is a structural feature, and the current path to AI impact in any domain runs through data curation and verification infrastructure. Breakthroughs in general reasoning could accelerate this process, but the path no longer depends on them.

3. From Coding to Other Domains

3.1. A Repeating Pattern Across Domains

If we abstract the pattern observed in software engineering, a recognizable development cycle emerges for making AI economically productive in any domain. It begins with data curation: assembling a corpus of expert work, from web-crawled data to expert-generated reasoning traces and records of complete real-world work units, ideally at scale. The next requirement is a reward signal: some mechanism for evaluating outputs without human judgment on every instance, whether that is test execution in coding, proof verification in mathematics, or compliance checks and multi-agent debate in other domains. With data and signal in hand, iterative refinement and reasoning-oriented RL becomes possible (Bai et al., 2022; OpenAI, 2024; DeepSeek-AI, 2025; AI, 2026; Team, 2025), with each iteration generating higher-quality training data for the next. Finally, the system is deployed in collaboration with human workers who use it in real workflows through orchestrated agentic loops, providing feedback and generating even more signal. This turns task-specific models into more capable systems that can collaborate with and assist humans in their work. The cycle is technically feasible for domains where data curation and reward design are both workable, and each successive cycle should become faster, cheaper, and more effective as transfer learning improves, reward design becomes systematic, and integration patterns mature.

The coding pattern does not transfer automatically to every knowledge domain. Where output quality cannot be cheaply and reliably verified, the optimization loop runs slower and produces less trustworthy results. In medicine, law, and safety-critical operations, even small error rates carry unacceptable liability and social cost. While these issues limit the pace of full automation, AI's impact as a human collaborator can still grow significantly over time. Domains with limited digital traces of expert process (sparse workflow telemetry) provide weaker training signal, delaying capability transfer.

Software engineering was the first major domain where the full development cycle was proven out, but scientific research is where the pattern becomes most vivid. In structural biology, AlphaFold predicted protein structures from amino-acid sequence alone (Jumper et al., 2021), work later recognized with the 2024 Nobel Prize in Chemistry (Outreach, 2024); AlphaFold 3 then extended structure prediction to biomolecular complexes involving proteins, nucleic acids, small molecules, ions, and modified residues, with improved accuracy over many specialized tools (Abramson et al., 2024). In formal mathematics, AlphaProof solved four of six 2024 IMO problems through RL and formal proof verification (DeepMind, 2025); by 2025, Gemini with "Deep Think" achieved a gold-medal score on five of six problems (DeepMind, 2025). In materials science, GNoME identified 2.2 million new crystal structures (equivalent to roughly 800 years of human discovery) including 381,000 previously unknown stable materials (Merchant & others, 2023). Weather forecasting saw a similar transformation: GenCast now outperforms ECMWF's ENS forecast on 97.2% of evaluation targets and produces a 15-day probabilistic ensemble in about eight minutes on a single TPU (Price et al., 2024). In medical screening, the MASAI trial (a randomized study of more than 100,000 women in Sweden) found that AI-supported mammography reduced interval cancer diagnoses and increased the share of cancers detected during screening without increasing false positives or overdiagnosis (Gommers et al., 2026). It is one of the largest randomized trials of AI in cancer screening to date, demonstrating that the data-plus-verification cycle can produce clinically meaningful results even in high-stakes domains.

3.2. Why Each Cycle Gets Easier

A critical feature of this cycle is that it compounds, driven by both better research methods and better AI models. As stronger models are trained, they can help create better training data for the next generation. An AI system that writes code can build data pipelines for other domains; one that does literature review can curate training data for scientific reasoning models. This does not imply a discontinuous jump; the current evidence points to staged, compounding gains rather than unconstrained recursive improvement (Anthropic, 2026; OpenAI, 2025; AI, 2025; Team, 2025; Anthropic, 2026). Each domain where AI becomes productive expands the toolkit for

making AI productive in the next. Researchers are also improving the adaptation, generalization, and efficiency of learning methods, so expansion to new domains and tasks should become easier over time, accelerating adoption and economic impact.

4. Workforce Evidence So Far

4.1. How Much of the Workforce Is Affected

The workforce transformation ahead is not a distant prospect, but exposure should not be conflated with displacement. Global evidence is now more granular than the first-wave 2023 analyses. The ILO's 2025 refined global index estimates that roughly one in four jobs has meaningful generative-AI exposure, while the share at highest automation risk remains much smaller (around 3.3%) (Organization, 2025). Complementary usage evidence from millions of real model interactions shows that AI work today is concentrated in software and writing-heavy occupations, and that augmentation still outweighs full automation in observed task execution (Handa et al., 2025; Anthropic, 2025). The World Economic Forum's 2025 Future of Jobs Report projects that 39% of core job skills will transform by 2030 (Forum, 2025).

4.2. Augmentation Before Replacement

The consistent finding across domains is that AI augments human capability before it replaces human workers, and that gains are largest for less experienced workers. AI systems trained on expert-generated data effectively disseminate the tacit knowledge of high performers to everyone. In customer support, Brynjolfsson et al. (2023) found that AI assistance raised average productivity by 14%, with a 34% improvement for the least experienced agents. In professional writing, Noy & Zhang (2023) found that ChatGPT reduced task completion time by 40% and raised output quality by 18%, with the largest gains for lower-ability workers. In consulting, a pre-registered RCT with 758 BCG consultants found that access to GPT-4 increased task completion by 12.2%, improved speed by 25.1%, and raised output quality by over 40% for tasks within the AI's capability frontier, while below-average consultants improved by 43% compared with 17% for top performers. On tasks outside the frontier, however, AI users performed 19 percentage points worse than the control group, demonstrating that augmentation gains are bounded by the in-distribution limits we described in Section 1 (Dell'Acqua et al., 2023). In medicine, the MASAI mammography trial (Section 3) demonstrated that AI-supported screening could improve clinical detection patterns without increasing false positives or overdiagnosis (Gommers et al., 2026).

This leveling pattern is powerful, but not guaranteed to persist, and early displacement signals deserve serious attention. A Federal Reserve Bank of St. Louis analysis found a 0.47 correlation between occupational AI exposure and rising unemployment rates over 2022–2025, climbing to 0.57 for occupations with highest

actual AI adoption (Ozkan & Sullivan, 2025). A Stanford Digital Economy Lab study found that employment among 22–25 year-olds in highly AI-exposed occupations fell 13% relative to less-exposed occupations from late 2022 to mid-2025, suggesting that hiring suppression in entry-level positions may precede outright displacement (Brynjolfsson et al., 2025). These effects are concentrated in specific demographics and occupations, and are not yet visible in aggregate employment data: the Yale Budget Lab found no clear economy-wide disruption attributable to AI (Lab, 2025). The divergence between aggregate stability and occupational-level stress is itself informative; it suggests a transitional phase in which AI is powerful enough to reshape hiring patterns at the margin, but not yet broad enough to register as macroeconomic displacement. Whether augmentation gains or displacement pressures dominate in the medium term will depend on market structure, bargaining institutions, and deployment design (Fund, 2025).

4.3. New Work Creation

An important counterbalance to the displacement narrative is the historical evidence that technological change systematically creates new types of work. Autor et al. (2024) analyzed 80 years of U.S. census data and found that roughly 60% of employment in 2018 was in job titles that did not exist in 1940. The task-based framework of Acemoglu and Restrepo (2019) emphasizes that automation displaces workers from existing tasks while simultaneously creating new tasks where humans have a comparative advantage.

In the AI context, the early signal is visible less as a clean set of new titles than as changing skill bundles inside existing occupations. PwC's 2025 barometer finds that employer-requested skills are changing 66% faster in occupations most exposed to AI, while the WEF projects rising demand for AI and big data, technological literacy, analytical thinking, and creative thinking (PricewaterhouseCoopers, 2025; Forum, 2025). As AI handles routine cognitive work, human comparative advantage shifts toward judgment under uncertainty, ethical reasoning, creative synthesis, and contextual adaptation.

The broader implication is that automation and new-task creation operate simultaneously, as the task-based framework itself predicts (Acemoglu & Restrepo, 2019). The empirical question is whether the rate of new-task creation keeps pace with automation, a question whose answer depends on deployment design and institutional support.

5. Energy, Compute, and Diffusion Frictions

5.1. Adoption Frictions and the Productivity J-Curve

Historical evidence suggests that transformative general-purpose technologies initially show declining measured productivity before producing large gains, because firms must invest in complementary intangibles (reorganization, training, new processes) before benefits materialize (Brynjolfsson et al., 2021). We may be in the early phase of this J-curve now. BCG's 2025 global survey of 1,400 C-suite executives found that 62% cite talent and skills shortages as the biggest barrier to AI integration, yet only 6% have meaningfully upskilled their workforce (Group, 2025). The WEF finds that the skills gap, not the technology itself, is the top barrier to AI adoption (Forum, 2025). This gap between capability availability and organizational readiness suggests that the technology is ahead of the institutions that must absorb it.

A historical parallel sharpens the point. The electric motor converted electricity into physical force and was subsequently deployed across virtually every context in which human or animal muscle power had previously been required (factories, homes, transportation, agriculture, and various electrical tools). The range of applications was as broad as the space of physical work itself (David, 1990). Large language models and their successors may represent a similar phenomenon, converting electricity into something resembling intellectual capacity (reasoning, analysis, pattern recognition, language production). The range of eventual applications is open-ended for this general-purpose technology still early in its diffusion. As it matures, working without it will resemble a carpenter using manual tools rather than any electrical ones. It might be a process of artistry or exercise, but rarely will it be practical or efficient for most purposes.

5.2. When Energy and Compute Become Dominant Constraints

However, another aspect of our metaphor is that a carpenter using electrical tools will now need to pay the manufacturers of those tools and the electricity companies, creating new economic constraints and dependencies. These dependencies will shift the balance and distribution of economic production. The largest component in intellectual productivity so far has been human capital, but the weights are being readjusted to give more importance to energy and compute infrastructure than before. Under high-adoption scenarios with sustained capability growth, compute and energy become increasingly important marginal constraints on output growth. The IEA reports data-center electricity demand at approximately 415 TWh in 2024, with projections to roughly 945 TWh by 2030 (growing at 15% per year, four times faster than total electricity demand) (Agency, 2025). Goldman Sachs Research projects a 165% increase in data center power demand by 2030 (Research, 2024). The IEA also estimates that data-center capital investment reached roughly half a

trillion dollars in 2024, while Amazon, Microsoft, Meta, and Alphabet announced more than $300 billion in planned 2025 capital expenditures, much of it directed toward AI and data centers (Agency, 2025). The IMF confirms that AI-producing sectors in the United States have grown at nearly triple the rate of the broader economy (Fund, 2025).

The scale of demand has pushed data-center operators toward long-term power procurement. The IEA reports that renewable power purchase agreements already cover about one-fifth of estimated data-center electricity demand, that technology companies accounted for more than 30% of corporate PPA volumes from 2010 to 2023, and that another 60 GW of contracted capacity is under development (Agency, 2025).

At the same time, chip-level efficiency improvements partially offset demand growth. NVIDIA markets Blackwell as supporting up to 30× faster real-time inference for large models, and Blackwell Ultra as delivering up to 50× higher performance and 35× lower cost for agentic AI (NVIDIA, 2026). This creates a dynamic analogous to Jevons' paradox (Jevons, 1865), where cheaper compute enables new applications that absorb saved capacity. The race between efficiency gains and workload expansion means energy projections should be treated as conditional scenarios, not deterministic endpoints (Agency, 2025; Fund, 2025; Shehabi et al., 2024; Administration, 2026).

Access to cheap, abundant electricity and compute infrastructure is becoming a defining factor of production, a conclusion supported not only by industry projections but by the revealed preferences of firms committing hundreds of billions to infrastructure that would be worthless absent expectations of transformative demand (Agency, 2025; Fund, 2025).

6. Human-AI Collaboration in the Agentic Era

6.1. A Thought Experiment on Friction

To understand the next phase of collaboration, consider a mobility metaphor alongside the power-tools metaphor above. Today's work often resembles a walking world: movement is effortful, local, and constrained by the time and energy required to cover distance. Infrastructure, habits, and planning all reflect those constraints. Now imagine a teleportation world where someone can instantly travel as far as they can see. Routes, homes, cities, and investments would be organized around different assumptions. A similar change occurs when building a useful artifact becomes constrained less by manual execution and more by the ability to specify the desired outcome. In a walking world, a common route justifies a train; in software, a common workflow justifies a reusable product sold to many users. If custom software could be

produced from a short request, the value of the train would change. The metaphor is imperfect, since real AI systems still require electricity, compute, and human oversight, but it helps show how lower execution friction can reorganize work. Science fiction has explored how near-instant mobility reorganizes society, both as liberation (Bester, 1956) and as a warning about dependency when human skill atrophies (Forster, 1909).

The transition will not be seamless. As the workforce evidence in Section 4 shows, aggregate labor market disruption has not yet materialized even as firm-level differentiation accelerates. The key variable, as the MIT Task Force on the Work of the Future emphasized, is institutional response: labor market policies, education systems, and corporate practices determine whether the friction collapse is broadly shared or concentrated (Future, 2020; Institute, 2024).

6.2. How Collaboration Actually Works

Human-AI collaboration is not one thing; it operates across a spectrum of modes, each placing different cognitive demands on the human participant. At one end, the human retains full control and invokes the AI for bounded assistance (autocomplete, lookup, single-step generation). The human decides what to do, when to do it, and how to evaluate the result. This is the mode most current productivity studies measure, and it is where gains are easiest to capture but also smallest in scope.

Further along the spectrum, human and AI begin to iterate jointly where the AI drafts, the human critiques; the human specifies, the AI stress-tests. The locus of value shifts from raw output speed to the quality of the exchange: how well the human frames the problem, how precisely they diagnose model errors, and how effectively they steer successive refinements toward a goal the model cannot fully represent on its own. Evidence from writing, consulting, and product-innovation teamwork experiments suggests that this kind of collaboration can improve outputs, but only when the human actively evaluates rather than passively accepting suggestions (Noy & Zhang, 2023; Dell'Acqua et al., 2023; Dell'Acqua et al., 2025).

At the far end lies delegation. The human assigns a scoped objective (resolve this bug, draft this regulatory filing, analyze this dataset) and the AI executes an autonomous chain of actions: reading files, running code, searching documents, assembling outputs. The human's role becomes supervisory, monitoring progress, intervening on exceptions, verifying final results. This is the mode captured by Anthropic's agent-autonomy data, where experienced users grant broad pre-authorization and step in selectively rather than approving each action (Anthropic, 2026).

The progression along this spectrum represents a fundamental shift in the cognitive division of labor. As delegation increases, oversight becomes management-by-exception and the human must develop reliable intuitions about when to intervene

(Parasuraman et al., 2000; Bainbridge, 1983; Lee & See, 2004; Hoff & Bashir, 2015). Pushing workers into full delegation without building the supervisory judgment that iterative collaboration develops risks the automation-complacency failures documented for decades in aviation and other safety-critical domains (Parasuraman & Riley, 1997). The goal is not to remove the human from the loop but to place them at the right level of it, where comparative advantage shifts toward orchestration, decomposing problems, setting evaluation criteria, auditing outputs, and deciding when to override.

Making this work in practice requires something like a contract between the human and the agent, defined before execution begins. Current agents are mostly optimized to complete tasks, not to collaborate gracefully with human supervision at every stage. As models improve, it should become easier for users and models to align on what the agent is trying to achieve and what is out of scope, which actions it may take autonomously and which require approval, under what conditions it must pause and hand control back, and what kind of audit trail it leaves behind. Once these conditions are explicit rather than implicit, both over-reliance and under-utilization decrease, and the human can focus supervisory attention where assumptions are most likely to break down (Parasuraman et al., 2000; Lee & See, 2004; Hoff & Bashir, 2015; Bainbridge, 1983).

It is worth noting that speed alone is a poor measure of whether any of this is working. A short-run speed gain accompanied by rising static-analysis warnings, cyclomatic complexity, or other downstream quality debt is not a productivity improvement; it is debt accumulating off the balance sheet, a pattern the Cursor adoption study documents directly (He et al., 2025). What matters is whether cycle time and rework move together in the right direction, whether genuinely ambiguous cases get routed to human reviewers, whether supervisory effort per task declines with organizational experience, and whether outcome quality holds up where it actually matters (defect rates, compliance, downstream consequences) not just initial completion speed (Brynjolfsson et al., 2021; Anthropic, 2026).

6.3. Open Challenges for Trustworthy Collaboration

The operating model described above does not arrive ready-made. Before delegation and joint iteration can deliver reliable value, a cluster of unresolved issues has to be designed into the collaboration loop rather than treated as after-the-fact governance. The first is the division of responsibility. When a human executes a task directly, accountability is comparatively legible; when an autonomous agent reads files, runs code, invokes tools, and assembles an output, responsibility can blur into what has been called a responsibility gap, where neither the operator who delegated the task nor the designer who built the system fully controls the behavior in question (Matthias, 2004). A human-agent contract helps by specifying what the agent may do

unsupervised, which actions require sign-off, and what audit trail it must leave (Parasuraman et al., 2000). But this should be understood as an input to accountability, not a complete solution. Traceability makes responsibility easier to allocate; it does not by itself settle legal liability, managerial responsibility, or vendor obligations.

Closely related is the problem of trust, which is easy to miscalibrate in either direction. Too little trust can erase the productivity gain that motivated the collaboration, though in high-stakes or immature deployments extensive re-checking may be exactly the rational posture. Too much trust is more dangerous: humans may wave through errors precisely on the out-of-distribution tasks where the model is least reliable, the failure mode the jagged-frontier experiment documented when consultants followed AI suggestions past its capability frontier (Dell'Acqua et al., 2023). The goal is therefore calibrated reliance, trust scaled to the system's demonstrated competence on the task at hand rather than to its fluency or apparent confidence (Lee & See, 2004; Hoff & Bashir, 2015). Calibration is domain-specific. A drafting suggestion in a low-stakes document and a dosage recommendation in a clinical workflow warrant different default levels of delegation, verification, and sign-off. Trust, in other words, is not a global setting but a function of consequence and evidence.

The same point applies to explainability. Humans need enough visibility into a system's behavior to know when to accept, question, or override it, but useful visibility is not always the same thing as a verbal explanation from the model. In some settings, especially structured high-stakes prediction tasks, there is a strong argument for inherently interpretable models over post-hoc explanations of opaque ones (Rudin, 2019). In agentic workflows, however, the more important oversight object may be the action trace: what files were read, what tools were invoked, what assumptions were made, what tests were run, and where independent verification succeeded or failed. Explanations still matter, but they must improve the joint outcome rather than merely increase confidence; poorly designed explanations can make teams more comfortable without making them more accurate (Doshi-Velez & Kim, 2017; Bansal et al., 2021). This connects directly to the chapter's earlier emphasis on verification signals: oversight works best when explanations, logs, and tests constrain each other.

A fourth challenge is bias in decision-making, especially when the reward signal or proxy target is mis-specified. Because models inherit the statistical structure of their training data, they can reproduce and amplify historical inequities even when aggregate accuracy looks strong (Mehrabi et al., 2021). The canonical illustration is a widely deployed healthcare algorithm that used prior spending as a proxy for need and consequently underestimated the illness of Black patients, a flaw invisible in

headline performance metrics but consequential at the level of individual care (Obermeyer et al., 2019). This is where the data-plus-verification cycle of Section 3 cuts both ways: the same infrastructure that drives capability can also optimize the wrong target at scale. Fairness auditing, subgroup evaluation, and bias detection therefore have to be part of the reward and deployment design, not external checks added after the system is already in use. High-stakes deployments such as AI-supported mammography make the point concrete: average performance gains are not sufficient if errors or benefits are distributed unevenly across populations (Lang & others, 2026).

Finally, effective collaboration requires co-adaptation. Humans have to learn the contours of a system's competence, where it is reliable, where it fabricates, and how to specify objectives so an agent acts on the intended goal. Systems, in turn, improve only when deployment generates usable corrections, preferences, and failure cases that can be incorporated into evaluation and refinement pipelines, rather than disappearing as unstructured user frustration (Anthropic, 2026). The productivity studies bear this out: gains are largest when humans actively evaluate and steer rather than passively accept, and the value of the pairing comes from genuine complementarity, each party covering the other's characteristic weaknesses, rather than from the model alone (Bansal et al., 2021; Brynjolfsson et al., 2023). The skills and habits that make human-AI teams work are themselves intangible investments, which is one more reason the J-curve is real (Brynjolfsson et al., 2021).

7. Skills, Education, and Career Architecture

7.1. From Speed to Foresight

In a walking world, value often comes from doing more steps faster. In a teleportation-like work environment, speed at low-level execution matters less than foresight: selecting better objectives, anticipating downstream consequences, and aligning multi-step systems with real human needs. This does not reduce the need for expertise; it shifts where expertise matters most.

What emerges is a workforce organized less around tasks executed and more around the type of judgment contributed to a human-AI system. Some people will provide deep contextual expertise: the physician who catches what an AI differential diagnosis misses, the lawyer who knows a clause is unenforceable in a specific jurisdiction. Others will design the agentic pipelines themselves: decomposing objectives, sequencing workstreams, inserting human checkpoints where they matter. Still others will focus on guardrails, audit protocols, and failure-recovery procedures, work that becomes especially critical in regulated industries. A growing number will improve the underlying infrastructure (training data, evaluation benchmarks, deployment pipelines). Most professionals will blend elements of several of these

roles. The overall direction is clear: human comparative advantage is shifting away from execution throughput and toward judgment, design, governance, infrastructure, and the ability to recognize quality when no specification fully captures it.

7.2. Education and the Apprenticeship Problem

If the frontier moves from raw execution toward orchestration, education must follow. The highest-leverage capabilities become problem formulation, epistemic rigor, evaluation design, collaboration with autonomous tools, and continuous reskilling. Current labor evidence already points to rapid skill transformation and uneven readiness across institutions (Forum, 2025; Organization, 2025; OECD, 2025).

What this means in practice is that curricula need to teach people how to convert ambiguous objectives into specifications that agents can act on, how to check model outputs against independent evidence rather than accepting fluent text at face value, how to calibrate when to delegate and when to intervene, and how to treat both tools and domain knowledge as continuously evolving rather than fixed. Evidence confirms that firms gain more from AI when deployment and skill upgrading happen together rather than sequentially (Institute, 2024; OECD, 2025; Forum, 2025).

A deeper concern is what might be called ladder erosion: if AI absorbs routine tasks, entry-level workers may lose the repetitive practice that historically built expert judgment (Autor et al., 2024; Brynjolfsson et al., 2025). Progression systems will need to evaluate people on orchestration capability (clean delegation boundaries, effective quality checks, minimal rework) not just output volume. Error detection and prudent intervention need to be recognized as valuable contributions rather than treated as friction. Early-career professionals will periodically need to work through problems manually, building the first-principles understanding that makes later supervisory judgment reliable. Medical residencies and legal clerkships already embody this principle; the challenge is extending it to domains where AI adoption is eroding the natural apprenticeship that routine work once provided. None of this works without gain-sharing mechanisms that connect productivity improvements to compensation and job quality. Workers who see AI as enriching employers while compressing their own wages will resist adoption or disengage from the oversight responsibilities that make agentic workflows safe (Acemoglu & Restrepo, 2019; Fund, 2025; Organization, 2025).

Conclusion

We used two metaphors in this chapter, and both point in the same direction. The electric motor converted electricity into physical force and reshaped every industry that relied on muscle power. Large language models and their successors are doing something analogous, converting electricity into a form of intellectual capacity. Like

the carpenter who once worked with hand tools and now reaches for powered ones, knowledge workers are beginning to reach for AI not as a curiosity but as a basic instrument of their craft (David, 1990; Karpathy, 2025).

The teleportation metaphor captures a different dimension of the same shift. When execution friction collapses, the scarce resource is no longer the ability to do the work but the ability to decide what work is worth doing, to verify that it was done well, and to intervene when something goes wrong (Anthropic, 2026; Parasuraman et al., 2000). The evidence from coding, customer support, consulting, and scientific research all suggests we are already moving in this direction (Cui et al., 2024; Brynjolfsson et al., 2023; Dell'Acqua et al., 2023).

This does not make humans obsolete. It makes human responsibility more strategic. The workforce is reorganizing around judgment, oversight, design, and quality recognition. The question going forward is less about whether AI will transform knowledge work and more about the terms on which that transformation unfolds. Energy and compute are becoming defining factors of production (Agency, 2025; Fund, 2025). Entry-level career ladders are already under pressure (Brynjolfsson et al., 2025). The skills gap, not the technology itself, remains the top barrier to adoption (Forum, 2025).

Whether the gains from this transformation are broadly shared or narrowly captured will depend on institutional choices made in the next few years, not the next few decades. Societies that invest in public access to compute, meaningful worker transition systems, and education that teaches orchestration rather than execution will be better positioned than those that do not (Fund, 2025; Programme, 2025; Organization, 2025). The technology is moving fast. The institutions that absorb it need to move faster.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., & others. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature, 630*, 493--500. https://doi.org/10.1038/s41586-024-07487-w

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3--30.

Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy, 40*(121), 13--58. https://www.nber.org/papers/w32487

Administration, U. E. I. (2026). *Electricity demand growth is driven by data centers and manufacturing in our latest forecast.* https://www.eia.gov/todayinenergy/detail.php?id=64304

Agency, I. E. (2025). *Energy and AI.* https://www.iea.org/reports/energy-and-ai

AI, M. (2025). *Kimi K2.* https://doi.org/10.48550/arXiv.2507.20534

AI, M. (2026). *Kimi K2.5.* https://github.com/MoonshotAI/Kimi-K2.5

Akyurek, E., Damani, M., Zweiger, A., Qiu, L., Guo, H., Pari, J., Kim, Y., & Andreas, J. (2025). *The surprising effectiveness of test-time training for few-shot learning.* https://arxiv.org/abs/2411.07279

Anthropic. (2025). *Uneven adoption and underuse from the 2025 Anthropic Economic Index.* https://www.anthropic.com/research/economic-index-geography

Anthropic. (2026). *Anthropic Economic Index, January 2026 report: Economic primitives.* https://www.anthropic.com/research/economic-primitives

Anthropic. (2026). *Measuring AI.* https://www.anthropic.com/research/measuring-agent-autonomy

Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940--2018. *Quarterly Journal of Economics.* https://economics.mit.edu/people/faculty/david-autor

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., & others. (2022). *Constitutional AI.* https://arxiv.org/abs/2212.08073

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775--779. https://doi.org/10.1016/0005-1098(83)90046-8

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445717

Bester, A. (1956). *The stars my destination.* Signet.

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve. *American Economic Journal: Macroeconomics, 13*(1), 333--372. https://www.aeaweb.org/articles?id=10.1257/mac.20180386

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI.* https://www.nber.org/papers/w31161

Brynjolfsson, E., Chandar, B., & Chen, R. (2025). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence.* https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

Cui, Z. (., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2024). *The effects of generative AI.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4945566

David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review, 80*(2), 355--361.

DeepMind, G. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature.* https://www.nature.com/articles/s41586-025-09833-y

DeepMind, G. (2025). *Advanced version of Gemini.* https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

DeepSeek-AI. (2024). *DeepSeek-V3.* https://doi.org/10.48550/arXiv.2412.19437

DeepSeek-AI. (2025). *DeepSeek-R1.* https://doi.org/10.48550/arXiv.2501.12948

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI.* https://doi.org/10.2139/ssrn.4573321

Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., & Lakhani, K. R. (2025). *The cybernetic teammate: A field experiment on generative AI.* https://doi.org/10.3386/w33641

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* https://arxiv.org/abs/1702.08608

Forster, E. M. (1909). *The machine stops. Oxford and Cambridge Review.*

Forum, W. E. (2025). *The Future of Jobs Report 2025.* https://www.weforum.org/publications/the-future-of-jobs-report-2025/

Fund, I. M. (2025). *The global impact of AI.* https://www.imf.org/en/publications/wp/issues/2025/04/11/the-global-impact-of-ai-mind-the-gap-566129

Fund, I. M. (2025). *Power hungry: How AI.* https://www.imf.org/en/publications/wp/issues/2025/04/21/power-hungry-how-ai-will-drive-energy-demand-566304

Fund, I. M. (2025). *AI.* https://www.imf.org/en/publications/wp/issues/2025/04/04/ai-adoption-and-inequality-565729

Future, M. T. F. o. t. W. o. t. (2020). *The work of the future: Building better jobs in an age of intelligent machines.* https://workofthefuture-taskforce.mit.edu/

Group, B. C. (2025). *The AI.* https://www.bcg.com/publications/2025/ai-adoption-puzzle-why-usage-up-impact-not

Handa, A., Bengio, Y., Klyman, K., Ouyang, C., Richard, A., Sato, A., Tsotsos, A., Weller, A., Weidinger, L., Whitney, W. F., Kinnebrew, J., Voss, C., Rosenberg, C., & Tamkin, A. (2025). *Which economic tasks are performed with AI.* https://doi.org/10.48550/arXiv.2503.04761

Hassabis, D., & Kavukcuoglu, K. (2025). *Gemini 3: Introducing the latest Gemini AI.* https://blog.google/products/gemini/gemini-3/

He, H., Miller, C., Agarwal, S., Kaestner, C., & Vasilescu, B. (2025). *Speed at the cost of quality: A difference-in-differences study of the causal effect of AI.* https://arxiv.org/abs/2511.04427

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407--434. https://doi.org/10.1177/0018720814547570

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., & others. (2022). Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS).* https://arxiv.org/abs/2203.15556

Institute, M. G. (2024). *A new future of work: The race to deploy AI.* https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond

Jevons, W. S. (1865). *The coal question: An inquiry concerning the progress of the nation, and the probable exhaustion of our coal-mines.* Macmillan.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., & \vZ. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583--589. https://doi.org/10.1038/s41586-021-03819-2

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models.* https://arxiv.org/abs/2001.08361

Karpathy, A. (2025). *2025 LLM.* https://karpathy.bearblog.dev/year-in-review-2025/

Lab, Y. B. (2025). *Evaluating the impact of AI.* https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs

Lang, K., & others. (2026). AI. *The Lancet.* https://doi.org/10.1016/S0140-6736(25)02464-X

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50--80. https://doi.org/10.1518/hfes.46.1.50_30392

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., & Xie, X. (2024). Large language models understand and can be enhanced by emotional stimuli. In *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2307.11760

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175--183. https://doi.org/10.1007/s10676-004-3422-1

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences, 121*(41). https://doi.org/10.1073/pnas.2322420121

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1--35. https://doi.org/10.1145/3457607

Merchant, A., & others. (2023). Scaling deep learning for materials discovery. *Nature, 624*, 80--85. https://doi.org/10.1038/s41586-023-06735-9

METR. (2025). *Measuring AI.* https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

METR. (2025). *Measuring the impact of early-2025 AI.* https://arxiv.org/abs/2507.09089

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187--192. https://doi.org/10.1126/science.adh2586

NVIDIA. (2026). *NVIDIA Blackwell.* https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447--453. https://doi.org/10.1126/science.aax2342

OECD. (2025). *Generative AI. OECD Artificial Intelligence Papers*(36). https://doi.org/10.1787/84c9f81f-en

OpenAI. (2024). *Learning to reason with LLMs.* https://openai.com/index/learning-to-reason-with-llms/

OpenAI. (2025). *OpenAI o3 and o4-mini system card.* https://openai.com/index/o3-o4-mini-system-card/

Organization, I. L. (2025). *Generative AI.* https://doi.org/10.54394/HETP0387

Outreach, N. P. (2024). *The Nobel Prize.* https://www.nobelprize.org/prizes/chemistry/2024/summary/

Overflow, S. (2025). *2025 Developer Survey.* https://survey.stackoverflow.co/2025/

Ozkan, S., & Sullivan, P. (2025). *Is AI.* https://www.stlouisfed.org/on-the-economy/2025/aug/is-ai-contributing-unemployment-evidence-occupational-variation

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230--253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE, 30*(3), 286--297. https://doi.org/10.1109/3468.844354

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The impact of AI.* https://arxiv.org/abs/2302.06590

Price, I., & others. (2024). Probabilistic weather forecasting with machine learning (GenCast). *Nature.* https://doi.org/10.1038/s41586-024-08252-9

PricewaterhouseCoopers. (2025). *Fearless future: PwC.* https://www.pwc.com/gx/en/issues/artificial-intelligence/job-barometer/2025/report.pdf

Programme, U. N. D. (2025). *The next great divergence: Why AI.* https://www.undp.org/asia-pacific/press-releases/ai-risks-sparking-new-era-divergence-development-gaps-between-countries-widen-undp-report-finds

Qodo. (2025). *State of AI.* https://www.qodo.ai/reports/state-of-ai-code-quality/

Research, G. S. (2024). *AI.* https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206--215. https://doi.org/10.1038/s42256-019-0048-x

Shehabi, A., Smith, S., Horner, N., Aghajanzadeh, A., Koomey, J., Sartor, D., Brown, R., Herrlin, M., & Lintner, W. (2024). *2024 United States data center energy usage report.* https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2001675_rev20241218.pdf

Snell, C., Lee, J., Xu, K., & Kumar, A. (2025). Scaling LLM. In *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2408.03314

Team, Q. (2024). *Qwen2.5.* https://doi.org/10.48550/arXiv.2412.15115

Team, Q. (2025). *Qwen3.* https://doi.org/10.48550/arXiv.2505.09388

Team, G. (2025). *Gemma 3 technical report.* https://doi.org/10.48550/arXiv.2503.19786

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., & Huang, G. (2025). Does reinforcement learning really incentivize reasoning capacity in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS).* https://arxiv.org/abs/2504.13837

About the Author

Dr. Muhammed Yusuf KOÇYİĞİT
Google | kocyigit[at]bu.edu
ORCID: 0009-0000-2904-4021

Dr. Muhammed Yusuf Kocyigit is a Research Scientist at Google, where he works on training efficient language models for Google Search and contributes to the Gemini Large Scale Pretraining team. He received his PhD in Computer Science from Boston University in 2025, where he was advised by Prof. Derry Wijaya and conducted research on natural language evaluation, with a particular focus on data contamination and its effects on model performance. Prior to his doctoral studies, he earned BS degrees in Electrical and Electronics Engineering from Bogazici University. Before joining Google as a Research Scientist, he held research internships at Google, Meta (FAIR) and Amazon, and co-founded TRIA AI, a machine learning startup where he led projects for major Turkish enterprises. His research interests include language model evaluation, data quality, and the relationship between training data and model generalization.