When the Scaffold Stays On: AI, Practice Style, and Screening in Elite Skill Formation

Song Yao*

August 4, 2026

Abstract

Generative AI raises short-term productivity by completing tasks that learners would otherwise practice on their own. Whether this exchange erodes frontier skill depends on the mode of use: substitute-users let AI stand in for deliberate practice and fail to develop skill, while complement-users use it to accelerate skill development. The two modes look alike in AI-aided output, so organizations screening talent on that output cannot tell which mode produced it. We ask whether the AI-prohibited evaluation gates organizations already operate can separate the modes. In elite competitive programming, the International Collegiate Programming Contest (ICPC) and the International Olympiad in Informatics (IOI) prohibit AI under in-person proctoring, with qualification-round entry, whereas Codeforces (CF) practice is unproctored and open to all. From CF submission histories we build an AI-prompt signature, more first-attempt acceptances, fewer attempts, fewer debugging retries, consistent with AI-assisted practice. CF practice has shifted toward this signature across entry cohorts spanning two AI rollouts. In CF contests, a stronger signature predicts smaller rating gains for users with no ICPC-IOI affiliation, but not for those who qualified. Inside the AI-prohibited ICPC environment, a shift toward AI-style practice predicts higher non-AI-aided scores for AI-era entrants. The same signature carries opposite signs across the two environments, exactly the pattern a type-separating gate predicts. The message is constructive: AI-style practice is compatible with frontier skill; the erosion risk links to the substitute mode; and that mode is separable by gates standard at credential boundaries, from medical and legal boards to professional certification.

Keywords: generative AI; skill formation; human capital; expertise; technology and work; screening; credentialing; substitutes and complements; competitive programming.

* Olin Business School, Washington University in St. Louis. Email: songyao@wustl.edu. I thank Tommaso Bondi, Seth Carnahan, Xinlei (Jack) Chen, Kiara Kim, Lamar Pierce, and Dennis Zhang for insightful comments. All errors are my own. Song Yao holds concurrent appointments as a Professor of Marketing at Washington University in St. Louis and as an Amazon Scholar. This paper describes work performed at Washington University in St. Louis and is not associated with Amazon.

1 Introduction

Generative AI has become a routine part of skill formation: across domains, from software development to professional education, learners now use AI in the very activities that build skill. For the organizations that train and select talent, this creates an evaluation problem: the track records these organizations screen, such as individuals' practice histories, portfolios, and assessment results, are now often produced with AI tools. The same AI tools can be used in two modes during learning. Some learners use them as a *substitute*: the AI does the work that would otherwise serve as practice, and the underlying skill fails to develop. Others use them as a *complement*: a way to learn more efficiently while continuing to practice. The two modes of AI usage accrue different amounts of underlying skill in the long run, yet the scaffold holds either way: in contemporaneous AI-aided output the modes look alike, and a polished record reveals little about which mode produced it and hence how much skill accrued. The difference surfaces only when the scaffold comes off, at the moments when a person must perform unaided. The first-order question for these organizations is therefore not whether to allow AI during skill formation but how to govern the mode in which it is used, and whether the evaluation regimes they already operate, such as proctored examinations and staged qualification rounds, can tell the two modes apart.

Two levers could govern that mode. The first is the tool itself: organizations can design the tools to enforce the complement mode. In a classroom experiment, AI full-solution access damages later unaided performance, and an AI tutor that instead gives hints avoids the damage (Bastani et al. 2025). This paper studies the second, the one organizations already control at credential boundaries: the evaluation gate. Many elite credentials (e.g., medical boards, bar exams, professional certifications) are awarded through AI-prohibited, proctored evaluations. If such gates genuinely test unaided skill, substitute-users should fail and complement-users should pass, so the gate would do more than measuring skill; it would separate the two modes of AI use at the population level. This paper tests whether real gates accomplish this separation.

The stakes sit at the skill frontier. Calculators make many forms of skill loss look benign: if a tool is always available and reliable, the user's underlying capability matters less. That argument holds for routine skill but breaks at the frontier or when AI-assistance is unavailable. Current AI systems are weakest on genuinely novel, frontier problems, the problems farthest from the patterns in their training data. Catching AI errors requires unaided expertise; trusted institutions such as peer review, medical diagnosis, and judicial reasoning need people who can function without AI mediation; and the innovation pipeline that feeds future training data depends on humans pushing the frontier. If substitute-mode AI use displaces the deliberate practice that builds this frontier capability, the cost is invisible in contemporaneous, AI-aided output but arrives years later. The concern is not hypothetical: after AI-assisted colonoscopy was introduced, endoscopists' *unaided*

adenoma-detection rate fell from 28.4% to 22.4% in a multicentre observational study (Budzyń et al. 2025). The macro version of this concern is the knowledge-collapse dynamic of Acemoglu et al. (2026), in which AI substitutes for human effort and the stock of new knowledge decays. We study its individual-level counterpart with a skill-formation framework, in which AI exposure shifts practice effort differently for substitute-users and complement-users. The heterogeneity in AI use mode is the key: the same technology, under identical access, produces both modes, and the design of the evaluation regimes that certify talent, in the presence of that heterogeneity, is an organizational design problem with substantial socioeconomic stakes.

The question extends a fast-growing body of evidence on AI at work. Field experiments document a jagged frontier at the task level: the same AI raises a knowledge worker's performance on tasks inside its capability boundary and lowers it on tasks outside (Dell'Acqua et al. 2026). We study the person-level counterpart: under identical access, the mode of AI use varies across learners, and the consequences for unaided skill diverge. Organizations already govern how work is evaluated and recorded, increasingly through algorithmic means (Kellogg et al. 2020). The AI-prohibited gate is a different instrument in that same design space: an unambiguous, non-AI-aided evaluation rule that determines which capabilities remain visible when the training records themselves are AI-saturated.

Testing whether a gate separates the two modes requires three things that are rarely observed together: practice behavior detailed enough to measure how heavily it relies on AI, an evaluation that genuinely prohibits AI, and measured performance at that evaluation. Fortunately, elite competitive programming provides all three. Codeforces (CF) is an online programming contest platform *open to all*: its contests have no qualification filter, so the pool blends substitute-users and complement-users, and AI use during practice is unregulated. Individual CF practice submission histories provide a long panel of submission-level detail on how users practice; CF contest outcomes, on the other hand, provide continuous individual performance measurements. The International Collegiate Programming Contest (ICPC) and the International Olympiad in Informatics (IOI) sit at the other extreme: both prohibit AI with in-person, proctored finals, and run qualification rounds that screen entrants before the finals (ICPC through peer screening at team formation and team qualifiers, IOI through national olympiads), so the qualifying pool is selected on unaided performance. Our panel follows 10,419 CF users across two AI rollouts, the GitHub Copilot launch in 2021 and the ChatGPT release in 2022; a subset of these users are linked to ICPC and IOI rosters and contest scores.

Three patterns emerge from the analyses on these users, and they differ in how much identifying weight they carry. We present them in increasing order of that weight:

- **A shift in practice norms across entry cohorts (Prediction 1).** On the CF platform, *practice-mode* submissions of later entry cohorts lean toward a behavioral signature: more problems solved on the first attempt, fewer attempts per solved problem, and fewer rapid debugging

retries. We call this an *AI-prompt signature* because it matches what a prompt-and-fix workflow with an AI assistant would produce. To illustrate with one of the three components, the fraction of problems solved on the first attempt: the AI-era cohort's post-ChatGPT rate rose about six percentage points more than the pre-AI cohort's. The shift appears in both the ICPC-IOI-rostered and non-rostered subpopulations, so neither pool drives it alone. The other two components move in the same AI-prompt direction. We treat this cohort contrast as descriptive context: it concerns practice inputs, so it does not by itself separate substitution from complementarity. And because the shift is not cleanly timed to the AI releases (Section 5.1), we do not read it as AI-caused.

- **A cross-subpopulation asymmetry of performance in the open pool (Prediction 2).** On CF rating, which measures CF contest performance, a stronger AI-prompt signature during practice predicts smaller rating gains among non-rostered users, about twelve rating points per signature unit, but is uncorrelated with rating change among ICPC-IOI-rostered users, who passed the AI-prohibited gates of ICPC or IOI qualification rounds. The cross-subpopulation difference is statistically significant, though only marginally (Section 5.2). The sign is what the screening mechanism predicts: the signature tracks weaker rating gains where substitute-users remain in the pool but not where the gate has removed them.

- **A positive AI-intensity slope inside the screened pool (Prediction 3).** Inside the AI-prohibited ICPC environment, a within-user shift toward the AI-prompt practice style predicts *higher* non-AI-aided contest score for users who entered competitive programming in the AI era. Regressing ICPC score on this practice-style change interacted with cohort, the AI-era cohort interaction is positive and statistically significant, while the pre-AI cohort slope is a null. This within-contest association is the paper's central claim: it is estimated inside the AI-prohibited environment, is specific to the AI-era cohort, and survives career-stage and baseline-skill controls (Section 5.3). It implies that, once the gate has screened substitute-users out, the remaining AI-style practice tracks complement use associated with higher performance. Predictions 2 and 3 together are the sharpest test of the screening mechanism, the cross-environment contrast: the same practice signature shows opposite signs in the two environments.

The findings support three conclusions. First, what matters for frontier skill is the mode of AI use, not access to AI: the same practice signature predicts weaker rating gains in the open pool and higher unaided scores inside the screened pool. Second, the cross-environment evidence is consistent with AI-prohibited qualification gates separating the two modes at the population level. This sharpens the classical signaling and screening framework of credentials (Spence 1973, Stiglitz 1975) for a technology that can partially substitute for the very skill being certified: whether the

credential still separates types depends on how the gate filters. Third, and most constructively, AI-style practice is compatible with frontier skill. Within the screened pool, a within-user shift toward AI-style practice accompanies higher unaided performance, so the erosion risk links to the substitute mode, not to the technology itself. Our evidence supports the reading that the two modes are real and institutionally separable. For the organizations that train and select talent, now relying on AI-era records, the practical meaning is that regime design can restore the records' informativeness eroded by ubiquitous AI assistance: an evaluation built to test unaided skill keeps the credential's signal intact. The paper's contribution is to the emerging empirical literature on AI and skill formation, and to the organizational study of evaluation regimes: field evidence that the two modes leave footprints in practice records, and are separable at a screening gate organizations already operate. Whether organizations can also steer learners from one mode to the other is the open question; evidence suggests they can when the tool itself is redesigned (Bastani et al. 2025), and we discuss the gate-design and training-design levers in Section 6.

2 Related Literature

This paper draws on seven literatures. First, theoretical models of AI offer a human capital depreciation mechanism at the macro and individual levels; we provide the empirical counterpart, showing that skill formation varies across people along the AI substitute-complement margin and that an institution may sort substitute-users from complement-users. Second, randomized controlled trials (RCTs) on AI-assisted learning establish the short-term substitution mechanism; our question is whether the mechanism persists in long-term skill formation. Third, workplace field experiments and observational studies document how AI changes practice and production patterns at scale, and how over-reliance on a capable tool can erode the user's own judgment; we extend those patterns into an elite skill market with a multi-year time window. Fourth, organizational research on technology at work treats AI's effects as enacted in practice and studies how organizations evaluate and control workers through algorithmic systems; we bring that lens to skill formation, with a designed evaluation gate as the sorting device. Fifth, in the labor-economics literature on technological change and skill obsolescence, the value of a skill is endogenous to the technology that can replace it; we shift the object of study from a skill's wage return to its upstream formation. Sixth, pre-generative-AI evidence from chess training engines shows distributional compression (lower-tail player ratings rise while the world-champion rating stays flat) under complement-style tool use; the open question is whether the elite frontier erodes when the AI can substitute for the work itself, which the chess setting does not test. Finally, we frame AI-prohibited credentialing as a mechanism-design problem, drawing on signaling-and-screening theory in economics. The first six streams establish that AI use splits into two modes with real stakes at the skill frontier; the seventh

is the institutional lever this paper studies. We discuss each stream in turn.

2.1 Theoretical models of AI and human capital

Acemoglu et al. (2026) provide a dynamic theoretical model in which agentic AI substitutes for human effort, weakening the flow of new knowledge people could contribute to society's general knowledge stock, so general knowledge depreciates. Under certain conditions, if AI accuracy rises above a critical threshold, complete knowledge collapse becomes the unique stable outcome because people stop investing any effort and delegate all work to AI. The model is macro-level, theoretical, and assumes symmetric agents.

At the individual level, Bondi and Johnson (2026) independently develop a related model. They study a forward-looking agent who delegates tasks to AI while human capital accumulates through learning-by-doing. They use a single parameter, *pedagogical quality* μ, to govern how much learning survives delegation: when $\mu < 1$, AI substitutes for skill formation and unaided capability atrophies; when $\mu > 1$, AI augments learning and the effect reverses. Their $\mu < 1$ and $\mu > 1$ line up with our substitute and complement modes; we reach that split from the data, they from a model. Their model and Acemoglu et al. (2026) are both theoretical. Two features are unique to our setting: the substitute-complement mode varies across people under the same AI access, and an institution, the AI-prohibited gate of Section 2.7, sorts on that mode difference. We further provide the field test of the gate in an elite skill market.

2.2 Substitution mechanism evidence from RCTs on AI-assisted learning

Recent RCTs have tested AI's substitution-for-effort mechanism in short-horizon settings. Stadler et al. (2024) find that students using ChatGPT for a 20-minute research task expended about 35% less germane cognitive load (the deep-processing facet that builds understanding) than students using Google search, and produced worse output as a result. Melumad and Yun (2025) show across seven experiments that learning from LLM summaries yields shallower knowledge than learning from web search, because the model performs the synthesis the user would otherwise do themselves. Closest to our design, Shen and Tamkin (2026) randomized 52 software developers learning a new Python library: those with high-engagement patterns (asking conceptual questions, requesting explanations alongside generated code) matched or exceeded non-AI users on an unaided follow-up quiz, while those with low-engagement patterns (debug delegation) scored lower. Our *AI-prompt signature* on CF submissions is the multi-year, field-observational analog of their usage patterns, high- and low-engagement alike.

Bastani et al. (2025) provide the cleanest existing test of substitution-induced skill erosion: their field RCT randomizes AI access across 1000 students during mathematics practice and then

evaluates their performance on AI-free in-class exams immediately after each practice session. The students were randomized across three arms (Control with no AI access, "GPT Base" directly giving answers to questions, "GPT Tutor" giving hints for solving questions). GPT Base raised practice grades by 48% but reduced exam grades by 17%; GPT Tutor raised practice grades by 127% without harming exam grades. In short, because GPT Base students use AI as substitute for problem-solving during practice, their higher practice grades did not carry through to the AI-free exam.

These RCTs establish substitution at the session and task scale; whether it persists is our question. Four features separate our setting. Bastani et al. (2025) run four 90-minute sessions, whereas we follow contestants across multiple years (2018–2025); they estimate mean effects in high-school mathematics, whereas we estimate top-tail effects in elite competitive programming; their AI-prohibited test is a high school course exam, whereas ours are career-defining contests (ICPC, IOI) feeding university admissions and employer signals; and they study students working below the frontier, whereas we study the formative-to-elite pathway in a domain where the frontier is what AI is least able to substitute for. We test the multi-year, top-tail, real-stakes joint cell these dimensions define.

2.3 Workplace field experiments and observational AI studies

A growing body of field experiments and observational studies documents how AI tools change practice and production at scale. Noy and Zhang (2023) find in a writing-task experiment that ChatGPT raised output and compressed the productivity distribution, mostly by substituting for worker effort rather than complementing skill. Brynjolfsson et al. (2025) document the same compression in a call center: AI raises mean output, with lower-skilled workers gaining more than higher-skilled ones. In a field experiment with management consultants, Dell'Acqua et al. (2026) show that AI lifts performance on tasks inside its capability frontier but degrades it on tasks just beyond that frontier, a "jagged" boundary that workers struggle to perceive. Hausman et al. (2025) show that ChatGPT raises university grades, especially for lower-performing students, compressing the grade distribution and eroding grades' signal value for employers. At the market level, Hui et al. (2024) find that ChatGPT's release reduced employment and earnings for freelancers in writing-heavy occupations on a large online labor market, with top-rated freelancers affected no less than others. And in a field comparison, Boussioux et al. (2024) find that human-AI collaboration produces more valuable though less novel solutions than a human crowd on a real innovation challenge, shifting how organizations source creative work.

Within programming specifically, the practice workflow itself shifts. Peng et al. (2023) find developers randomized to GitHub Copilot completed a task 55.8% faster; Hoffmann et al. (2024) show that Copilot shifts developers' time toward core coding and away from project management,

with larger effects for lower-ability developers; and Shihab et al. (2025) document a workflow shift from `read` → `understand` → `implement` to `prompt` → `view response` → `implement`, with manual code entry dropping from 95.6% to 52.5%.

A distinct strand asks what AI does to the user rather than to the output. Dell'Acqua (2022) finds that recruiters given a high-quality AI tool disengaged (exerting less effort, deferring to the model, and failing to improve over time), whereas those given a weaker tool stayed critical and built their own skill. Shaw and Nave (2026) document the same blind spot at the moment of use: across preregistered experiments, people adopt AI suggestions with little scrutiny and report *higher* confidence even after the AI has erred, a pattern they call cognitive surrender. This "falling asleep at the wheel" is the in-the-moment analog of the substitute-versus-complement distinction (Section 2.2) whose multi-year accumulation we study.

Closest to our complement-versus-substitute distinction, Kim et al. (2026) conduct a survey of 1,943 U.S. workers and find that those who report viewing AI as a complement to their own effort raise their work effort after AI's introduction while those who report the opposite do not, and that self-reported career-relevant learning rises most among intensive AI users; the patterns hold with controls and an instrumental-variables strategy exploiting employer-provided AI access. We study the same margin through revealed behavior in an elite skill market, instead of self-reported belief in the general workforce, and add an institution that sorts complement versus substitute.

These studies show that AI changes how people work and learn. Their scale is session-level, within-firm, or platform-wide; the time window is months. We extend the inquiry to multi-year individual horizons and elite-frontier outcomes. The 2018–2025 CF panel follows individual contestants across two major AI rollouts, and the contestants who also appear in ICPC and IOI rosters link their AI-permitted practice to AI-prohibited contest performance.

2.4 AI and work in organizations

Organizational research treats a technology's effects as enacted in practice rather than fixed by its design (Orlikowski 2000): the same tool, in different hands, produces different work. A central strand of this research studies evaluation as an instrument of control. Kellogg et al. (2020) synthesize this strand: organizations increasingly direct, evaluate, and discipline workers through algorithms, with evaluation operating through comprehensive recording and rating. In that literature, evaluation extracts effort and compliance and predicts existing skill; whether the evaluation regime sorts on, or shapes, the *formation* of the capability it measures is not asked; the evidence is largely case-based, and the review itself notes the lack of counterfactuals. Rahman (2021) supplies the sharpest account of what evaluation design does to workers: when a freelancing platform replaced its transparent five-star rating with an opaque algorithmic score, freelancers could no longer tell what the evaluation

valued; most reported being unable to learn from it, and they responded by either experimenting on the algorithm or constraining their activity on the platform. That study closes by noting that the long-term impacts of such evaluations remain unclear. Our setting supplies the sorting half of what this strand is missing. The evaluation criterion is fully transparent, where the score in that study is deliberately opaque. The evidence is a within-person quantitative panel rather than a single-regime case study. And the outcome is capability itself: whether the skill the evaluation certifies actually formed.

A second strand focuses on how expertise forms, or fails to form, around analytical technologies. In a two-year ethnography of four investment-banking groups, Anthony (2021) finds that where juniors only assembled analyses with algorithmic tools while seniors interpreted them, juniors and seniors alike treated the tools as black boxes, never questioning the assumptions inside. The juniors acquired a deep command of the spreadsheets but not the professional expertise itself; the juniors' expertise formed when they had to defend their interpretations to skeptical seniors. The mechanism is an organizational division of labor, and the expertise outcomes are career proxies observed for a few dozen bankers over five years. Generative AI moves that division of labor inside a single person: the AI now does the assembling, and the person either interrogates the output or accepts it. Our AI-prompt signature picks up the footprint of that choice. Rather than treating AI as one more tool in an individual user's hands, Anthony et al. (2023) propose studying the full arrangement of work around the technology: who designs it, who deploys it, and who uses it. They call for grounded accounts of how AI intersects with work. We answer both papers on their own terms: the substitute and complement modes give that division of labor a person-level, multi-year footprint in an elite skill market, and the AI-prohibited gate widens the empirical frame beyond the user and the tool, to the organization that evaluates.

2.5 Technological change and skill obsolescence

A labor-economics literature studies how technological change revalues skills over a career. Deming and Noray (2020) show that the return to experience reflects a race between on-the-job learning and skill obsolescence: in fast-changing technical fields, early-career wage premia erode as the skill content of work shifts, and workers sort out of the faster-changing occupations as they gain experience. Horton and Tambe (2025) examine a discrete skill-death event, the deprecation of Adobe Flash, and find that wages did not collapse, because the affected developers reskilled rapidly toward fallback skills; the market absorbed the shock through adjustment. Acemoglu and Restrepo (2020) document how automation displaces labor and reshapes task demand across local labor markets. These studies treat the value of a skill as endogenous to the technology that can replace it, but they measure that value downstream, through wages, occupational sorting, and reskilling. We

move upstream, asking whether AI exposure during training erodes the formation of the deep skill itself, at the individual frontier and before any wage consequence. The rapid-reskilling result of Horton and Tambe (2025) also marks a boundary condition: adjustment is fast when a fallback skill exists, whereas our concern is the formation of the frontier skill for which, by construction, there is no fallback.

2.6 Pre-generative-AI evidence: chess engines and why they differ from AI in programming

Gaessler and Piezunka (2023) document that pre-generative-AI computer-assisted training in chess raised players' annual Elo ratings (a relative skill rating based on chess game outcomes) by approximately 11 points per year on average. The effect concentrated in lower-skilled players, with world-champion-level Elo being flat during the same period. The chess engines' impact is therefore a population-level distributional compression: the lower tail rises, the frontier holds steady.

This positive Elo result is the most natural objection to our framing. Computer engines have been part of chess training for decades, raising the lower tail while leaving the frontier intact; if the tool did not erode the chess frontier, why would AI-assisted training in programming threaten deep-skill formation at all? The answer rests on three concrete differences between the chess-engine training regime and current AI-programming practice. Each difference moves chess closer to a complement-style integration and AI-programming closer to a substitute-style integration.

1. **Complement versus substitute.** A chess engine serves as an artificial training partner, a scalable substitute for the scarce human sparring partners a player would otherwise need (Gaessler and Piezunka 2023). Crucially, what it substitutes for is the *opponent*: the player still calculates, plans, and internalizes *why* a move is best, and still plays their rated games unaided. AI for programming, by contrast, integrates into the work itself (prompt, generate, accept or modify) and can deliver functioning code without the user grasping the algorithm, instead substituting for the user's own design-time cognition. Stadler et al. (2024) measure this latter kind of substitution at the cognitive level: AI use suppresses germane cognitive load (the deep processing that builds skill) by 35%. Chess engines do not produce comparable suppression because they replace the training partner, not the learner's own processing.

2. **Training-test separation.** Competitive chess maintains a hard institutional boundary: engines are forbidden during competition, so players know precisely when they are using the tool and when they are competing unaided. Everyday programming has no equivalent boundary as AI is integrated continuously into one's workflow. Bastani et al. (2025) operationalize the distinction: GPT Tutor (scaffolded, no full solutions) did not harm unaided performance, while GPT Base

(continuous full-solution access) cut unaided exam scores by 17%. Same domain, same students; the difference was scaffolded versus unscaffolded use. AI-prohibited contests (Section 2.7) reinstate the chess-style separation that everyday programming lacks, and that separation is the feature our research design exploits.

3. **What the tool can substitute for.** A chess engine is a powerful study aid, but unaided games assess runtime cognition of the player, including pattern recognition, calculation, and strategic planning under time pressure. There is a gap between what the engine does (analysis between games) and what competition measures (live play). Such a gap does not exist in programming contests: AI in programming substitutes for design-time cognition (algorithm design, debugging, problem decomposition), which is precisely what programming contests measure.

Read together, these differences make the positive Elo result of Gaessler and Piezunka (2023) consistent with our framing, not against it: chess engines raised ratings through a complement-style mode of use, whereas current programming practice sits closer to the substitute side. Whether this substitute-style risk shows up in an elite skill market over multi-year horizons is the question this paper tests.

2.7 Selection into AI-prohibited credentialing

The seventh stream motivates our research design: can evaluation regimes screen productive from unproductive AI use when the technology partially substitutes for the skill being evaluated?

The distinction we draw is between substitute-users and complement-users of AI: contestants with the same AI access during practice differ in whether they use it as a substitute for deliberate practice or as a complement, and that difference accumulates into divergent non-AI-aided performance. This is the multi-year, field-scale analog of the engagement-pattern heterogeneity that Shen and Tamkin (2026) document in their learning RCT (Section 2.2): the same tool yields different skill outcomes depending on how the user engages it.

The evaluation gate filters through two channels at once: formal rules (no AI, no internet, in-person, supervised) and social selection (peer screening at team formation, national-olympiad reputation), which together determine who reaches the elite qualifying pool. The design question is not whether to permit AI but how to make the gate filter on the substitute-vs-complement margin. Classical signaling (Spence 1973) and screening (Stiglitz 1975) theory treat credentials as costly actions or tests that separate types; our setting differs in that the effort is genuinely skill-forming rather than purely type-sorting, and the organization actively shapes which separating equilibrium obtains by designing the gate. What is new with generative AI is that the deep skill is now partially substitutable by the tool itself, so whether the credential still separates types depends on how the

gate filters, not only on whether it exists. Two recent studies bracket the signaling question from two directions. Galdin and Silbert (2025) show that generative AI can *destroy* a costly signal: once AI makes tailored job applications cheap to write, they no longer separate strong applicants from weak ones, and hiring becomes less meritocratic. Jabarian and Reshidi (2025) show that a designed *choice* can *restore* the signal: when applicants choose whether an AI bot or a human evaluates them, the choice itself signals ability, because stronger and weaker applicants choose differently. Our AI-prohibited gate is a third design. By testing the unaided skill, it preserves signal informativeness precisely where AI may substitute for the underlying skill. We add to this stream observational evidence at multi-year individual scale that the substitute-vs-complement distinction is empirically real and accumulates into divergent non-AI-aided performance; identification of a selection mechanism, the screening of substitute-users out of the pool by ICPC and IOI, which we treat as a research-design feature rather than a sample-selection nuisance; and a sharpened policy claim that gate design should target the substitute-vs-complement margin rather than rely on generic gatekeeping. Together these position the contribution as a design principle for evaluation systems in the presence of partially substitutable AI.

3 Background and Empirical Setting

3.1 The AI shocks

GitHub Copilot entered technical preview on June 29, 2021 and reached general availability one year later. It targeted programming directly and was among the first AI tools widely adopted by programmers. ChatGPT followed on November 30, 2022, putting generative AI into broad public use across populations and tasks. Subsequent releases (e.g., GPT-4, Claude 2 and 3, DeepSeek-Coder) further extended AI's coding capability but did not introduce a comparably sharp adoption shock.

The two releases play different roles in our research design. Copilot was a programmer-specific tool and arrived more than a year before ChatGPT, so users who first appeared on CF after the Copilot release had already begun training with AI tools available to them. ChatGPT was a broader population-level event with a sharper date, observable across the entire panel; Copilot's reach was confined to its early adopters. We use the Copilot date to partition users by training vintage, and the ChatGPT date to define the within-user pre/post comparison; we elaborate on the motivation for this split in Section 5.

3.2 Competitive programming as the empirical setting

CF is the dominant online competitive programming platform. It has run contests several times a week since 2010. It issues individual ratings based on each user's performance in the contests compared to others. The CF rating is Elo-style, like chess, measuring each user's *relative* skill. Besides weekly contests, anyone can create an account and submit practice solutions. Practice submissions do not enter the rating system and are a good training sandbox. There is no qualification gate, no proctoring during practice, and nothing preventing a user from running an AI tool while solving problems. Both modes of AI use therefore mix in the practice pool. During rated contests AI use is again hard to police, but AI may be a weaker substitute there: under time limits and a point penalty for incorrect submissions, contestants who lean on AI without the underlying skill tend to lose rating (see Section 5).

The International Collegiate Programming Contest (ICPC) and the International Olympiad in Informatics (IOI) sit on the other end of the institutional spectrum. Both run multi-stage qualification rounds before the finals, and both prohibit AI tools at every qualification stage. ICPC builds teams of three from a single university and holds its AI-prohibited qualification rounds and world finals in person. IOI selects individual contestants through national olympiads and runs the international finals in person (online only during 2020 and 2021 due to COVID-19). Users whose CF performance leans heavily on AI are unlikely to pass these rounds, so the populations that reach the ICPC and IOI finals have been screened of the most AI-dependent practitioners. These are high-stakes contests whose medals are recruiting signals and are cited on university and job applications, so the non-AI-aided scores measure frontier skill under real stakes.

The contrast between the more AI-lenient CF pool and the screened ICPC and IOI pools is what lets us identify the substitution mechanism. The CF pool keeps both user types and shows how AI exposure correlates with rating change in an unscreened population. The qualified ICPC and IOI pools show how AI-intensive practice translates into non-AI-aided performance. Reading the two together gives cross-environment evidence on whether the same practice signature marks substitute or complement mode.

On the CF platform, one further feature matters for our empirical strategy. Each submission carries a tag indicating whether it is in practice or contest mode. The behavioral signature of AI-style practice is cleanest in practice mode and the data are dense, so we use the practice-mode signature as our main measure of what we term *AI-intensity*: a practice-style proxy for how far a user's practice has shifted toward an AI-style prompt-and-fix workflow, although not a verified measure of AI use. The same signature computed on contest-mode submissions, noisier because contest behavior also reflects contest skill and time pressure, provides a robustness check against a roster-versus-skill confound (Appendix A.1): the cross-subpopulation pattern it shows is consistent with the practice-mode measure.

3.3 Conceptual framework

We start with a simple framework that defines the two modes of AI use and formalizes why an AI-prohibited gate separates them. It organizes the empirical strategy: combined with the assumptions stated below, its primitives yield the three predictions we test.

Consider a learner who builds unaided frontier skill z through deliberate practice. Deliberate-practice effort e and AI use a produce that skill through a learning technology in the human-capital tradition of Ben-Porath (1967),

$$z = L(e, a), \tag{1}$$

$$L_e > 0, \qquad L_{ea} > 0. \tag{2}$$

Effort builds skill ($L_e > 0$), and AI, used alongside practice, strictly complements it ($L_{ea} > 0$, in the sense of Hicks 1970; the complementarity central to the technology of skill formation, Cunha et al. 2010): for a learner who exerts effort, more intensive AI use raises that practice's skill yield, so practicing more with AI builds strictly more skill, not less. Skill, however, comes only from the practice: AI can deliver a task's output in place of the learner's own effort, but that output is not skill. A learner who substitutes AI for practicing still produces output, but without building z; a learner who practices alongside AI gets the same help with output and builds z as well. More AI use therefore raises contemporaneous output in either mode, so the AI-prompt signature (proxying the magnitude of a) is a footprint of AI use, not a direct measure of skill. Because the signature appears under both pure delegation and intensive practice, a given signature can reflect either mode. The gate, which reveals z, is what distinguishes the modes.

What separates learners is whether they invest effort in deliberate-practice. We call practice that pairs AI with genuine effort a *complement mode* and reliance on AI without genuine effort a *substitute mode*. Under the same AI access, the two modes accrue different amounts of unaided skill z: the complement-mode learner practices and accrues $z = L(e, a)$ with $e > 0$, while the substitute-mode learner delegates and ends with $z = L(0, a)$, which stays low. This gap is what our empirical strategy exploits.

The deliberate-practice effort e that builds z is costly with a cost level κ. In the open, AI-permitted environment, leaning on AI substitutes for that effort and saves κ, which is why it is attractive to the learner. The AI-prohibited gate, the qualification rounds that select the elite pool, changes the incentive: the gate scores only on realized z by blocking AI, and it awards an elite credential value V to entrants whose unaided skill clears a threshold, $z \geq \bar{z}$. Accordingly, a learner would invest the effort when the credential payoff covers its cost,

$$V \cdot \Pr(z \geq \bar{z}) - \kappa > 0. \tag{3}$$

Complement-mode learners pay κ, build z, and clear the gate; substitute-mode learners save κ by leaning on AI, arrive with low z, and cannot pass. The CF contests have no such gate and are nearly free to enter, so both modes stay on the open CF platform.

The gate sorts learners by realized skill: it admits the high type, complement-mode learners with high z, and turns away the low type, substitute-mode learners with low z. What our data cannot tell us is how a learner came to be one type or the other, and the following two accounts leave the same cross-sectional footprint. Under *intrinsic traits*, the type is fixed and the gate sorts a distribution of types it did not shape. Under *anticipation*, the type forms in response to the gate: a learner who foresees an AI-prohibited evaluation invests in deliberate practice and becomes the high type, while one who does not foresee the evaluation substitutes AI for effort and stays the low type. The opposite-signs pattern is consistent with either, so the data identify the sorting but not how the types arise. Neither account is established, and the pessimistic one, in which types are fixed and the gate changes no one, is no better supported than the optimistic, improvable one; the two differ in policy implications, which we elaborate on in Section 6.

3.4 Testable predictions

The framework of Section 3.3 yields three predictions. All three hinge on the output/skill split drawn above: the AI-prompt signature tracks output, which AI raises in either mode, whereas the skill z the gate scores diverges between modes. The first concerns practice inputs across the platform; the second and third concern how the same practice signature maps to performance in the un-screened and screened environments: the difference in mode drives the second, and the gate's separation drives the third. We test these predictions in Section 5.

Prediction 1 (universal practice shift). After the AI releases, CF practice behavior shifts toward an AI-prompt signature (more first-attempt acceptances, fewer attempts per solved problem, and fewer rapid debugging retries) across the platform as a whole, and the shift grows with how much of a user's training postdates AI. The signature is a footprint of AI use, which raises output in either mode. Consequently, the shift appears for substitute- and complement-users alike. This prediction concerns practice inputs and does not by itself separate substitution from complementarity.

Prediction 2 (asymmetry in the un-screened pool). In the open CF pool, where both user types coexist, a stronger AI-prompt signature predicts smaller rating gains among un-screened (non-rostered) users, because substitute-users remain in this pool and their non-AI-aided skill erodes. Among roster-linked users, who have passed AI-prohibited gates, the same signature does not predict weaker gains, because the gate has already removed the substitute-users. The signature-to-rating slope is therefore more negative in the un-screened subpopulation than in the screened one.

Prediction 3 (positive slope inside the screened pool). Inside the AI-prohibited ICPC

environment, among users who entered competitive programming in the AI era, a larger shift toward AI-intensive practice predicts *higher* non-AI-aided contest score. Once the gate has screened substitute-users out, the remaining AI-style practice reflects complement use, which builds frontier skill rather than eroding it.

Predictions 2 and 3 together form the sharpest test. The same AI-style practice signature, operationalized as a post-ChatGPT level in the open pool and as a within-user shift inside ICPC, is predicted to carry *opposite* signs depending on whether the contest environment screens substitute-users out: negative in the un-screened CF pool (Prediction 2) and positive inside the screened ICPC pool (Prediction 3). A single mechanism operating uniformly, whether pure erosion or pure enhancement, does not produce this reversal, which arises from heterogeneous use combined with selective screening. The two tests also differ in construct, a cross-user level in the open pool and a within-user change among the screened, a difference dictated by where each mode is identifiable rather than an arbitrary choice.

4 Data

4.1 Data Sources

The analysis draws on three data sources, listed in Table 1. The two AI-prohibited contests, ICPC and IOI, have public rosters of finals that identify the credentialed competitors who anchor the CF panel. The Competitive Programming Hall of Fame (`cphof.org`) supplies ICPC rosters and team scores from 2018 through 2025; IOI's website (`stats.ioinformatics.org`) hosts IOI individual scores over the same time window. The CF individual panel is then assembled from the CF API in stages: we link CF accounts to the ICPC and IOI rosters and scrape those rostered competitors, then add a non-rostered expansion, resulting in the full CF rating and submission histories of 10,419 CF users. Next, we detail the scraping process and the linkage- and cohort-window exclusions that define this analysis panel.[1]

4.2 Panel Construction

The analysis uses two unbalanced panels: the panel from the AI-prohibited ICPC and IOI contests, and the CF within-individual panel. Table 2 reports summary statistics across four panels: the ICPC university-year panel (Panel A), the IOI country-year panel (Panel B), the CF within-individual

[1]We considered the larger public CF submissions dataset on Hugging Face (17.6 million submissions, 14,957 anonymized users, `https://huggingface.co/datasets/denkCF/UsersCodeforcesSubmissionsEnd2024`), but its users are anonymized and cannot be matched to ICPC or IOI rosters; we therefore scrape real, identifiable users directly from the CF API.

Table 1: Data sources and roles.

Source	Records	Outcome type	Role
cphof.org/standings/icpc/<year>	1,055 team-year records; 3,163 contestant-year records (2018–2025)	Team-level problems solved, ranks, and medals	ICPC AI-prohibited test events; CF user-level linkage via team rosters
stats.ioinformatics.org/results/<year>	2,779 contestant-year records (2018–2025)	Individual-level points, ranks, and medals	IOI AI-prohibited test events; CF user-level linkage via country delegation rosters
CF API	10,419 users; 793,920 user-month records; 2018–2026	CF rating + submission verdicts + submission timestamps + profile	Within-individual panel, assembled from the scrapes detailed in Section 4.2; user linkage to ICPC/IOI rosters

panel (Panel C), and the roster linkage connecting the contests to CF accounts (Panel D).

The AI-prohibited panel comprises 1,673 distinct ICPC contestants and 1,030 IOI contestants. Together they result in 2,494 distinct ICPC-IOI rostered contestants (209 competing in both) whose first rated CF contest falls between 2015 and 2024 (Panel D of Table 2 lists 3,117 roster-linked CF accounts in total; these 2,494 are the *cohort-assignable* ones out of the 3,117, i.e., their first-rated-contest dates allow a cohort assignment under the definitions of Table 3, while the remaining 623 fall outside the window and belong to no cohort). The Competitive Programming Hall of Fame (cphof.org) publishes the CF usernames for participants of ICPC finals if they have a CF account. For IOI contestants (stats.ioinformatics.org), we use name-and-country matching to find their CF accounts. We match these contestants to their respective CF account, forming the ICPC-IOI roster-linked subpopulation. These ICPC and IOI contestants span 354 universities and 90 countries over 2018–2025.

The CF within-individual panel contains 10,419 users, of whom 6,399 (61.4%) span both the pre-Copilot window (on or before 2021-06-28) and the post-ChatGPT window (on or after 2022-11-30) with at least 30 practice submissions on each side. These 10,419 users come from two groups. The first is 2,480 ICPC-IOI roster-linked users (from the 2,494 linked contestants above, less 14 with no practice-mode submissions in the analysis window; see footnote 2). The second and larger group is a non-rostered expansion of 7,939 users who are not anchored to either ICPC or IOI roster. These non-rostered users are scraped directly from the CF API among active accounts with a maximum rating of at least 1,400; these non-rostered users make up roughly three-quarters of the

panel. The rating floor applies only to the non-rostered expansion.[2]

Each user is assigned to a cohort by the date of their first rated CF contest (Table 3). The split captures training vintage relative to the GitHub Copilot launch: pre_AI users (2015 through 2018) trained before AI, transition users (2019 through Copilot launch) were partway through training when AI arrived, and AI_era users (Copilot launch through 2024) trained with AI available throughout. Table 4 reports each cohort's size, median CF rating, and median practice-submission counts, split by the pre- and post-ChatGPT windows. We observe the panel's CF practice and contest outcomes from 2018-01-01 to 2026-04-30. All estimation windows close at the end of 2025; the 2026 months enter only CF-career-total quantities, such as practice submissions per user and rated contests per user in Panel C of Table 2.

4.3 Outcome measurement: scoring system properties

Each contest has its own scoring system, so our outcome variable differs by contest; we discuss each choice in turn.

The CF rating is a modified, self-calibrating Elo-rating system inspired by chess (Mirzayanov 2015). After each contest, a competitor's rating moves up or down according to how they placed against the others in that round, and the ratings of those opponents enter the calculation directly. The rating therefore already reflects the strength of the field a competitor faced, and we do not need additional control for it. CF also applies an inflation correction that keeps the rating scale roughly stable from year to year, so a rating change of a given size means about the same thing across years. The rating is relative: it measures a contestant's standing in the pool, not an absolute skill level, so the rating does not move when the whole pool improves by the same amount. Our open-pool analysis (Section 5.2) therefore concerns relative performance, that is, whether a stronger AI signature predicts losing ground against one's peers. The absolute-performance measures come from the ICPC and IOI scores below.

In contrast, ICPC and IOI provide no comparable self-calibrating rating. ICPC and IOI do have relative performance measures such as the contest rank or medal, but they depend on the then-current contestant pool. Consequently, instead of using these relative measures without self-calibration,

[2]The analysis panel excludes three groups of scraped users. First, 1,374 users are matched to an ICPC or IOI roster only by a lower-confidence name-and-country match. Because the match is too uncertain to confirm their participation and attach a contest score, we exclude them from the analysis panel, and Appendix A shows the within-subpopulation slopes are insensitive to their handling (excluded, assigned to the non-rostered subpopulation, or assigned to the roster-linked subpopulation); only the pooled interaction softens, to $p = 0.110$, when these AI-era-heavy matches are forced into the roster-linked arm, as expected if unverifiable matches contaminate the screened arm. Second, 1,408 users cannot be assigned a cohort because their first rated CF contest falls outside the 2015–2024 cohort window: 623 rostered contestants (beyond the 2,494 rostered contestants above) and 785 non-rostered users. Third, 38 users have no practice-mode submissions in the analysis window of 2018-Q1 through 2025-Q4 (excluding the 2022-Q4 boundary quarter): 14 roster-linked and 24 non-rostered. Beyond these three exclusions, the panel is not filtered on any outcome variable.

Table 2: Summary statistics for the three analytical panels and the ICPC/IOI–CF linkage. Panels A and B are the ICPC institution-year and IOI country-year panels used for the non-AI-aided-performance tests; Panel C is the combined CF within-individual panel (contest-record roster-linked users and the non-rostered expansion); Panel D summarizes the linkage connecting them. Definitions are given in the table note.

Panel A. ICPC university-year panel	
University-year cells	1,055
Distinct universities	394
spanning pre-AI and AI-era cohorts	112
Cohort: pre-AI / transition / AI-era	272 / 249 / 534
Team CF rating: mean / median / SD	2,150 / 2,146 / 334
Teams with all members CF-linked	768 (72.8%)
Universities by contest year, 2018–2025	138, 134, 117, 132, 124, 130, 141, 139
Panel B. IOI country-year panel	
Country-year cells	708
Distinct countries	103
spanning pre-AI and AI-era cohorts	86
Cohort: pre-AI / transition / AI-era	353 / 89 / 266
Delegation CF rating: mean / median / SD	1,953 / 1,999 / 401
Delegations with all contestants CF-linked	251 (35.5%)
Country-years by contest year, 2018–2025	88, 88, 88, 89, 89, 88, 93, 85
Panel C. CF within-individual panel	
Unique users	10,419
of which spanning subset (≥30 subs both windows)	6,399 (61.4%)
User-month observations	793,920
Cohort: pre-AI / transition / AI-era	3,347 / 5,293 / 1,779
spanning subset, same order	2,061 / 4,006 / 332
Cohort CF rating (mean): pre-AI / trans / AI_era	1,818 / 1,610 / 1,531
Practice submissions per user (median)	1,489
Rated contests per user (median)	54
Source: roster-linked / non-rostered	2,480 / 7,939
Panel D. ICPC/IOI → CF linkage	
ICPC published-username linkage: accounts	2,189
IOI name-and-country linkage: accounts	1,164
Roster-linked CF users (ICPC ∪ IOI), all cohorts	3,117 (236 in both; 2,494 cohort-assignable)

Source: authors' calculations over the cleaned panels. Cohorts are assigned by first rated contest date: pre-AI (2015–2018), transition (2019 to the day before the Copilot release), and AI-era (Copilot release to 2024-12-31); users first rated outside 2015–2024 fall out of window and are excluded (Section 4.2). "Spanning" CF users have ≥30 practice submissions in both the pre-Copilot and post-ChatGPT windows; spanning universities and countries appear in both a pre-AI and an AI-era cohort cell. Cohort CF rating means are computed over all user-months within each cohort; practice-submission and rated-contest counts are medians over the 10,419-user analysis sample. The published-username linkage uses the CF username that the Hall of Fame lists on each ICPC contestant profile; Panel A and the within-ICPC test (Section 5.3) rely on it, while the name-and-country audit is the linkage route for IOI. Appendix A documents the candidate-matching confidence tiers and the sensitivity of the cross-subpopulation asymmetry to the 1,374 excluded name-only matches. Per-analysis estimation subsamples are documented in Section 4.2.

Table 3: Cohort definitions.

Cohort	First-rated-contest window	AI availability during training
pre_AI	2015-01-01 to 2018-12-31	Trained entirely pre-Copilot. Reference group.
transition	2019-01-01 to 2021-06-28	Partly pre-AI; AI tools arrived mid-career.
AI_era	2021-06-29 (Copilot launch) to 2024-12-31	Trained with AI tool availability.

Table 4: Cohort composition of the CF headline cohort × post DiD sample (Section 5.1). All counts are over the three cohorts (pre_AI, transition, AI_era). Spanning N marks numbers of users with ≥ 30 practice submissions in both the pre-Copilot and post-ChatGPT windows. Per-user medians are computed on the practice-pattern panel after the 2018–2025 and non-2022-Q4 filters.

Cohort	N	Spanning N	Median rating		Median practice sub counts	
			pre-ChatGPT	post-ChatGPT	pre-ChatGPT	post-ChatGPT
pre_AI	3,347	2,061	1,874	1,874	422	49
transition	5,293	4,006	1,582	1,665	462	142
AI_era	1,779	332	1,348†	1,674	163	408

† *Median over the 1,336 AI_era users with at least one pre-ChatGPT month of rated activity in the monthly rating panel; the remaining 443 AI_era users entered their first rated contest after ChatGPT and have no pre-ChatGPT rating to summarize. Cohort sample sizes are unbalanced: transition is the largest cohort (5,293) and AI_era is the smallest (1,779), and the spanning sub-sample collapses AI_era to 332 because most AI_era users cannot yet supply 30 pre-Copilot submissions. The DiD uses the full analysis sample (10,419 users; see Section 4.2), not the spanning sub-sample. Median rating is the per-user most recent rating in each window; median practice submissions are per-user totals on the practice-mode panel. Pre/post practice activity moves in opposite directions across cohorts (pre_AI's median drops from 422 to 49 as veterans exit; AI_era's median rises from 163 to 408 as new entrants accumulate post-ChatGPT activity); the* $\log(1 + n_{h,t})$ *control in the DiD absorbs this directly at the user-quarter level.*

we choose the absolute scores as the outcome measures: problems solved at the ICPC finals, and total points earned at the IOI. A score does not depend on the contestant pool at all; it reflects a competitor's own performance on a fixed problem set. It does, however, vary with the difficulty of each year's problem set. Accordingly, we use contest-year fixed effects to account for this variation.

4.4 Descriptive statistics

Table 5 reports descriptive statistics for the variables that enter the analyses of Section 5, one panel per analysis. Panel A reports the three AI-prompt practice proxies by cohort and pre/post-ChatGPT period: the gap between the AI_era cohort's pre-to-post change and the pre_AI cohort's is the raw difference-in-differences (DiD) that Equation 4 in Section 5.1 estimates. Figure 1 complements Panel A by tracing the three practice proxies by cohort over the full sample window. The large confidence intervals in the three proxies obscure any clear pattern change at the AI releases. The DiD therefore does the formal identification. Panel B reports the cross-subpopulation variables by subpopulation.

Specifically, the within-subpopulation spread of the AI-signature (its 10th-to-90th-percentile range) is the variation that identifies the slopes θ_s (Equation 5 in Section 5.2), and the level gap between the two subpopulations in baseline rating motivates the $\bar{r}_h^{\text{pre}}$ control. Panel C reports the within-ICPC variables by cohort; the coefficients of interest ψ_c are the difference across cohorts in how AI-intensity relates to score (Equation 7 in Section 5.3).

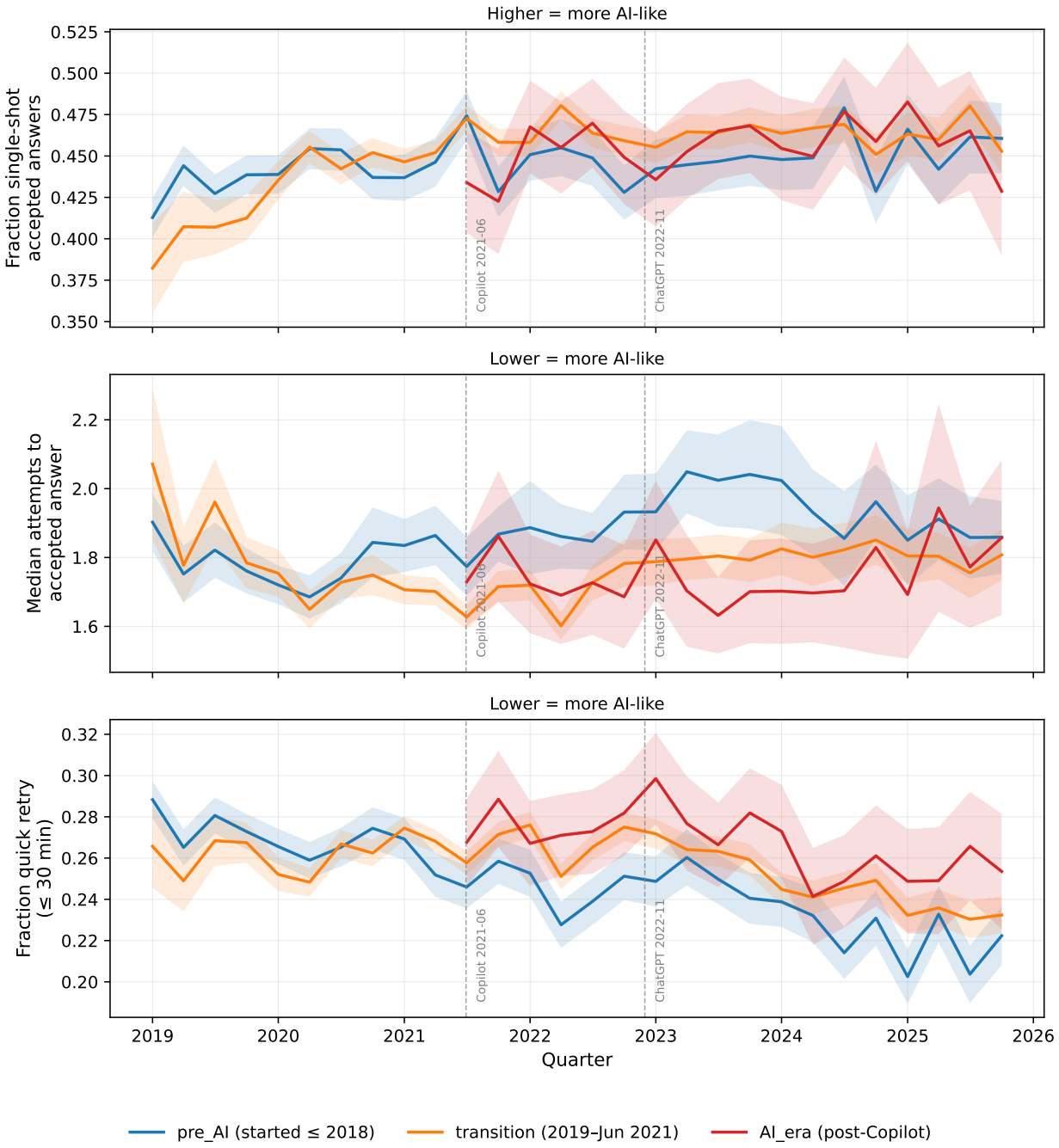


Figure 1: AI-prompt signature components by cohort and quarter. Each panel plots the cohort mean of one practice proxy by quarter with 95% confidence bands; the vertical axis names the metric and the panel title gives its AI-prompt direction. Vertical dashed lines mark the Copilot (2021-06) and ChatGPT (2022-11) releases, and the AI_era series begins after the Copilot launch by cohort construction. To hold the sample fixed across the time series, the figure uses spanning users (those with at least 30 practice submissions in both the pre-Copilot and post-ChatGPT windows; Table 2) in the three focal cohorts.

Table 5: Descriptive statistics for the Section 5 analyses. Each panel reports the variables of one analysis, sliced along the dimension that identifies its key coefficient: Panel A by cohort × pre/post-ChatGPT period (the DiD of Section 5.1), Panel B by subpopulation (the cross-subpopulation contrast of Section 5.2), and Panel C by cohort within the ICPC sample (the cohort contrast of Section 5.3). Cells are mean (standard deviation) unless noted.

Panel A. CF practice proxies by cohort × period (Section 5.1; user-quarter panel)

Cohort, period	Single-shot accept rate	Median attempts	Quick-retry fraction	Subs. / quarter	N user-qtrs
pre_AI, pre	0.429 (0.269)	1.85 (1.59)	0.264 (0.193)	61.6 (94.2)	38,627
pre_AI, post	0.450 (0.325)	1.95 (2.07)	0.228 (0.216)	33.4 (67.7)	13,387
transition, pre	0.449 (0.248)	1.73 (1.34)	0.259 (0.178)	88.6 (129.3)	43,360
transition, post	0.461 (0.290)	1.81 (1.78)	0.247 (0.203)	51.9 (87.5)	32,751
AI_era, pre	0.414 (0.270)	1.87 (1.45)	0.264 (0.191)	78.6 (129.4)	6,558
AI_era, post	0.456 (0.251)	1.74 (1.43)	0.272 (0.188)	79.4 (112.2)	14,780

Panel B. Cross-subpopulation rating asymmetry (Section 5.2; N = 6,690 users)

	Non-rostered (n = 5,283)	Roster-linked (n = 1,407)	All (n = 6,690)
Rating change Δr_h (post − pre)	286.9 (272.4)	350.2 (288.9)	300.2 (277.1)
AI-signature (post-ChatGPT PC1)	−0.003 (1.447)	−0.270 (1.567)	−0.059 (1.477)
p10 / p50 / p90	−1.72 / 0.11 / 1.67	−2.03 / −0.17 / 1.43	−1.79 / 0.05 / 1.62
Baseline rating $\bar{r}_h^{\text{pre}}$	1377 (381)	1766 (451)	1459 (427)

Panel C. Within-ICPC AI-intensity by cohort (Section 5.3; 579 obs, 444 users)

	pre_AI	transition	AI_era
AI-intensity (first-difference PC1)	−0.150 (1.324)	−0.014 (1.868)	0.161 (0.825)
p10 / p50 / p90	−1.55 / −0.09 / 1.28	−1.26 / 0.12 / 1.41	−0.91 / 0.19 / 1.10
Baseline rating $\bar{r}_h^{\text{pre}}$	2268 (372)	1946 (371)	1338 (337)
ICPC score (problems solved)	5.9 (1.9)	5.1 (2.0)	4.8 (1.5)
N observations	190	309	80
N users	144	233	67

Notes: Cohorts are defined in Table 3. Panel A cells are unadjusted cohort × period means over user-quarters; the estimated DiD (Table 6) adds user and quarter fixed effects and the $\log(1 + n_{h,t})$ *submission control. "pre" is 2018-Q1 through 2022-Q3 and "post" is 2023-Q1 through 2025-Q4 (the 2022-Q4 boundary quarter is dropped); N counts user-quarters in each cell on the single-shot support, and the estimation sample sizes (marginally smaller once fixed-effect singletons are dropped) are reported in Table 6. Panel B reports the cross-subpopulation sample of Section 5.2; "AI-signature" is* `ai_pc1_post`*, the post-ChatGPT first principal component of the three sign-aligned proxies. Panel C reports the within-ICPC sample of Section 5.3; "AI-intensity" is the within-user first-difference principal component* `ai_intensity_pc1`*. The IOI sister sample (Section 5.3; 345 obs, 243 contestants) has cohort AI-intensity means of* 0.16, −0.24, *and* −0.14 *and IOI-score means of* 263, 272, *and* 227 *for pre_AI, transition, and AI_era respectively.*

5 Estimation and Results

This section presents three sets of analyses that jointly distinguish the substitute and complement roles of generative AI in skill formation. They test Predictions 1–3 of Section 3.4, respectively. The first set (Section 5.1) documents a cross-cohort shift in CF practice patterns toward an AI-prompt signature. The magnitude of the shift grows with how much of a cohort's training window postdates AI. We read this shift as descriptive, because the cohort contrast is not cleanly timed to the AI releases. The second (Section 5.2) shows a cross-subpopulation asymmetry in CF users' ratings: this same AI-prompt signature predicts weaker rating gains among users in the open CF user pool, but not among rostered users who are linked to AI-prohibited ICPC and IOI contests. The third (Section 5.3) moves inside the AI-prohibited ICPC environment, where AI-intensive CF practice predicts *higher* non-AI-aided scores for AI_era entrants. This resolves the asymmetry: team formation and the qualifiers screen substitute-users out, and the AI use that remains in the pool is complement-style practice. Section 5.4 summarizes the robustness battery, which Appendix A reports in full.

All three analyses share the cohort structure of Section 4.2 (Table 3), with the pre_AI cohort as the reference. The cohort cutoff is anchored at the Copilot launch and the within-user pre/post comparisons at the later ChatGPT release; we elaborate on the reasons for this choice in Section 5.1.

5.1 The behavior shift toward an AI-prompt signature in CF practice

We first evaluate how CF practice behavior shifts after the AI releases, focusing on three patterns that together form an AI-prompt signature:

1. `frac_single_shot`: the fraction of problems solved on the first attempt;
2. `median_attempts`: the median number of attempts to solve a problem;
3. `frac_quick_retry`: the fraction of quick retries after a failed submission, where a quick retry is defined as a reattempt within thirty minutes of the incorrect submission.

Solving a problem without AI is iterative: the user submits, reads the verdict, locates the bug, and resubmits, often several times before the solution is accepted as correct ("AC," the platform's Accepted verdict). A user who relies on AI-assistance can instead submit a generated solution that often clears the problem on the first try, collapsing the submit-debug-resubmit cycle into a single submission, consistent with the keystroke-level workflow shift documented by Shihab et al. (2025). If users rely on AI during practice, three patterns are likely to surface:

1. more problems solved on the first attempt (greater `frac_single_shot`);
2. fewer total submissions per solved problem (lower `median_attempts`); and

3. fewer quick retries after a failed submission (lower `frac_quick_retry`).

Fewer quick retries are the subtlest of the three. Quick retries signal rapid human debugging, where the programmer submits, identifies the bug in code they wrote, and resubmits within a short time span. AI reliance reduces quick retries on two fronts: more first submissions succeed and leave nothing to retry, and, more importantly, when an AI-generated solution does fail, the user often cannot quickly debug code that they did not write themselves.

Throughout, this signature is a behavioral proxy inferred from submission timing. It is consistent with AI-style practice, but not a direct, verified measure of AI use. Appendix B reports construct-validation tests that bound the leading alternative: the signature merely proxies programmer skill because more skilled programmers may show similar behavior patterns even without AI. We show that, before AI was available, higher-rated (more skilled) users were, if anything, *less* AI-like on these behaviors, not more (the correlation between CF rating and each signature component is weak, $|r| < 0.20$, and runs opposite to the AI direction), so skill does not produce the signature.

To assess shifts in the three behavioral patterns after AI releases, we consider a DiD design on the user-quarter CF practice panel. For user h in quarter t:

$$\begin{aligned} y_{h,t} = \alpha_h + \tau_t + \beta_1 \cdot \text{post}_t \cdot \mathbb{1}[\text{cohort}_h = \text{transition}] \\ + \beta_2 \cdot \text{post}_t \cdot \mathbb{1}[\text{cohort}_h = \text{AI_era}] + \lambda \cdot \log(1 + n_{h,t}) + \varepsilon_{h,t}, \end{aligned} \tag{4}$$

where $y_{h,t}$ is one of the three patterns, $\text{post}_t = \mathbb{1}[t \geq \text{2023-Q1}]$ is the post-ChatGPT indicator (the boundary quarter 2022-Q4 is dropped from the panel), α_h are user fixed effects, τ_t are quarter fixed effects, $n_{h,t}$ is the user's submission count in the quarter, and standard errors are clustered at the user level. The coefficients of interest are β_1 on the transition cohort and β_2 on the AI_era cohort, both relative to the pre_AI reference cohort. Quarter fixed effects τ_t absorb the post main effect.

The sample is the 10,419-user CF panel of Section 4.2, restricted to user-quarter observations on practice-mode submissions from 2018-Q1 through 2025-Q4 (with 2022-Q4 dropped as the ChatGPT boundary quarter). The number of observations varies between 142,774 and 149,441 depending on the outcome variable (see Table 6). Per-cohort composition is shown in Table 4.

The two AI events anchor this design in different ways. The Copilot launch on 2021-06-29 defines the cross-section: a user's cohort assignment is set by where their first rated contest falls relative to this date. The ChatGPT release on 2022-11-30 defines the time series: the within-user pre/post indicator post_t flips from zero to one at this date. The two boundaries do not coincide for the following reasons. If both anchors were pinned to Copilot, AI_era users' pre period would consist only of whatever practice they did before their first rated contest, a period that spans the novice-to-competitor transition for this cohort. A within-user pre/post comparison anchored at Copilot would therefore conflate that career-stage transition with pre/post AI exposure.

Pinning the time series to ChatGPT instead gives AI_era users who entered before ChatGPT a clean pre-period of up to five quarters (2021-Q3 through 2022-Q3), all observed after they entered rated competition, while the cohort cutoff at Copilot still anchors the cross-section. AI_era users who first became active after ChatGPT have no pre-period practice and enter the DiD through their post-release observations only. The choice is also substantively defensible. ChatGPT was the broader shock: free, web-based, and immediately mainstream. In contrast, Copilot in 2021 was a programmer-specific tool with a much narrower user base. In the aggregate series, the practice-pattern signature shift is also larger around the ChatGPT date than around the Copilot date; the transition cohort viewed on its own is the exception, with the larger shift at Copilot (Appendix A.3). As a robustness check against the choice of date, we re-run the DiD with the pre/post split anchored at the Copilot launch (Appendix A.3). Only the pre_AI and transition cohorts enable the identification in this version, since AI_era users entered after Copilot and have no pre-Copilot practice. This check shows that the transition cohort's shift toward the AI-prompt signature is already present around the Copilot launch on all three proxies, so the headline does not depend on anchoring at ChatGPT.

The identification of β_1 and β_2 comes from the within-user pre-to-post change in the outcome, differenced across cohorts. User fixed effects absorb time-invariant user characteristics (e.g., skill baseline, education, country). Quarter fixed effects absorb calendar-time shocks common to all users, such as platform policy changes and seasonal effects. The specification does not identify the causal effect of AI use. Because cohort assignment is based on platform-entry timing rather than observed AI use, β_1 and β_2 measure the differential pre-to-post change in practice behavior of the transition and AI_era cohorts, respectively, relative to the pre_AI reference. We read these coefficients as a descriptive cohort contrast rather than as an AI effect: as the event study below shows, the gradient across cohorts (β_1 and β_2) does not coincide with the AI releases, so attributing it solely to AI exposure during training, rather than to a secular change in practice norms or to novice-to-competitor maturation, is not warranted by this design.

Results. The AI_era cohort × post interaction (β_2) is significant on all three outcomes (Table 6). Each β_2 is a DiD term: the AI_era cohort's pre-to-post change measured against the pre_AI cohort. All three margins move toward the AI-prompt signature: single-shot acceptances rise ($\beta_2 = 0.0586$, SE 0.0063), median attempts fall (-0.3271, SE 0.0374; about one-sixth of the 1.9-attempt baseline), and quick retries fall (-0.0190, SE 0.0041), all at $p < 0.001$.

The transition cohort lies between, but closer to the pre_AI cohort: $\beta_1 = 0.0139$ (SE 0.0046, $p = 0.003$) on single-shot submission and -0.0680 (SE 0.0297, $p = 0.022$) on median-attempts per problem, roughly a quarter of the corresponding AI_era effect. On these two outcomes the gradient pre_AI (reference, 0) → transition → AI_era is monotonic, consistent with the practice

Table 6: AI-Prompt Signature Shifts in CF Practice Patterns (Equation 4)*

Dependent variable	(1) `frac_single_shot`	(2) `median_attempts`	(3) `frac_quick_retry`
post × transition (β_1)	0.0139**	−0.0680*	0.0090**
	(0.0046)	(0.0297)	(0.0029)
post × AI_era (β_2)	**0.0586*****	**−0.3271*****	**−0.0190*****
	(0.0063)	(0.0374)	(0.0041)
$\log(1 + n_{h,t})$ (λ)	−0.0045***	0.0120**	0.0587***
	(0.0007)	(0.0037)	(0.0005)
User FE	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes
Observations**	149,398	142,774	149,441
Users	10,419	10,401	10,419
Quarters	31	31	31

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

*Notes: *Sample is the CF user-quarter panel, 2018-Q1 through 2025-Q4 excluding 2022-Q4. Reference cohort: pre_AI; the post main effect is absorbed by quarter fixed effects. Standard errors clustered at the user level in parentheses.*

***Sample sizes and user counts differ across columns because the three outcomes have different definitional supports: `median_attempts` is undefined for user-quarters with no solved problems (dropping ~6,600 user-quarters and 18 users); the small gap between `frac_single_shot` and `frac_quick_retry` reflects different denominators (problems attempted vs total submissions). Each regression uses the natural sample for its outcome.*

signature strengthening with how much of a user's training occurred post-AI. The quick-retry margin is the exception: the transition coefficient is small and runs the opposite way (0.0090, SE 0.0029, $p = 0.002$) while the AI_era coefficient is negative, so the monotone gradient holds on the other two outcomes but not on quick-retry. One plausible explanation: transition users adopted AI assistance mid-career but retained the iterative debugging habits they formed before AI. Generating code with AI and then debugging by hand produces *more* quick retries than the pre_AI baseline (who use AI less overall) and more than the AI_era cohort, who lack the pre-AI debugging fluency needed to repair AI-generated bugs. On this conjecture, the positive transition coefficient on quick-retry reflects this hybrid workflow, which single-shot and median-attempts do not capture.

This cross-cohort shift is descriptive, and it is evident for both ICPC-IOI-rostered and non-rostered users: re-estimating the `frac_single_shot` DiD separately by subpopulation, the AI_era coefficient is 0.0356 ($p = 0.004$) among the rostered users and 0.0617 ($p < 0.001$) among the non-rostered users.

The cohort contrast is not cleanly timed to AI. We do not read the cohort gradient (β_1 or β_2) as a causally identified AI effect because of three reasons. First, the parallel-trends assumption behind the DiD fails. We re-estimate Equation 4 with the post×cohort interactions replaced by a full set of

quarter×cohort interactions (each quarter dummy interacted with the transition and AI_era indicators, with the pre_AI cohort and the 2022-Q3 quarter as the omitted references), and plot the coefficients in Figure 2: neither the transition nor the AI_era cohort moves discontinuously at the ChatGPT release, and for the transition cohort (the testable contrast, since AI_era users barely appear in the panel before 2023) a joint test of the pre-release coefficients rejects parallel pre-trends on all three practice patterns ($\chi^2(18) = 96.7$, 76.8, and 84.6; all $p < 0.001$). Second, the cohort differences do not line up with AI exposure. If AI were the driver, the heavily exposed AI_era cohort should look more AI-like than the transition cohort; instead, at matched career stage (each user-quarter observation re-indexed from calendar time to time since the user's own first rated contest, with fixed effects on the new index absorbing the career maturation trajectory) the two cohorts are statistically indistinguishable (single-shot difference 0.0002, $p = 0.95$). The gap is between the pre_AI cohort and the later cohorts and does not widen from transition to AI_era, despite AI_era's far greater AI exposure. Moreover, the transition cohort entered before the Copilot launch, predating both AI releases. Thus the difference could be a secular, cohort-level shift in practice norms, not one that grows with AI exposure or begins at the AI rollout. Third, the cohort difference in the signature is largest on the problems AI handles worst, not the ones it handles best. Because the signature is first-attempt success, and AI is more likely to supply a correct first attempt on problems it can solve (easy, templated ones, not hard ones), a cohort leaning on AI should show the signature concentrated on easy problems. Instead, holding career stage fixed and splitting by Codeforces problem-difficulty rating, the AI_era-versus-pre_AI single-shot difference is null on the easiest problems (+0.002, $p = 0.63$) and rises with difficulty, peaking at +0.039 ($p < 0.001$) on the 2000–2399 band (harder problems) before attenuating slightly on the hardest problems. A secular shift in practice norms together with novice-to-competitor maturation could fit these three facts besides a clean AI shock, so we carry the practice shift forward as descriptive context that AI alone does not explain. Full event-study, career-time, and difficulty-gradient specifications are available from the author.

This cross-cohort shift is descriptive, but it is the premise for the analyses that follow. The AI-prompt signature is now a common feature of practice across this population. Practice patterns alone do not tell us whether shifting toward this style erodes or builds frontier skill. We examine this question in two settings: the cross-subpopulation asymmetry on CF rating (Section 5.2) in the open pool, and the within-ICPC contest performance (Section 5.3) inside the AI-prohibited environment.

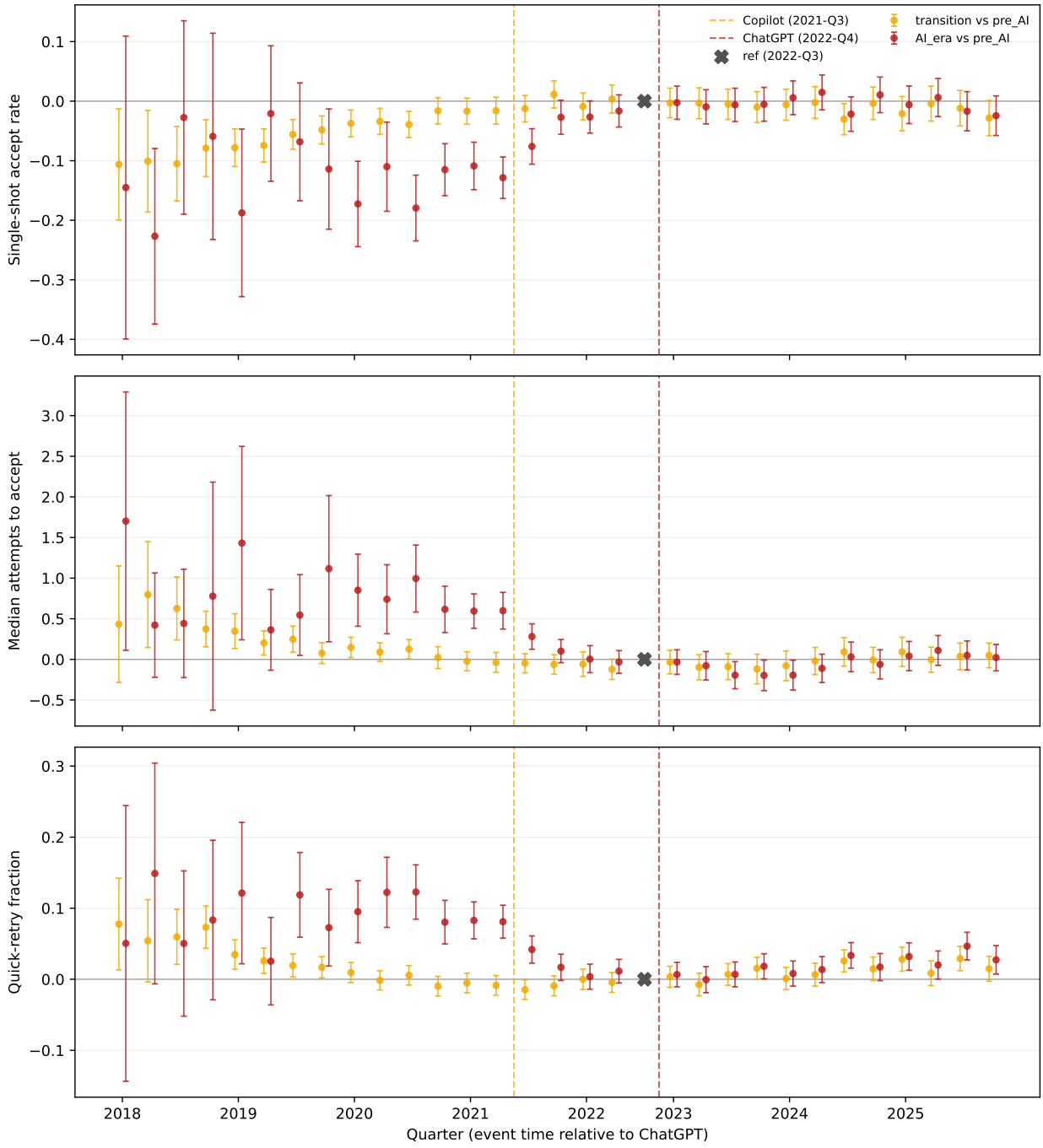


Figure 2: Cohort event study of the three AI-prompt signature proxies (Equation 4 with quarterly event-time interactions). Each panel plots event-time coefficients for the transition (orange) and AI_era (red) cohorts relative to the pre_AI cohort; the reference period is 2022-Q3 (gray X, normalized to zero) and the vertical dashed lines mark the Copilot (2021-Q3) and ChatGPT (2022-Q4) releases. Bands are 95% confidence intervals (user-clustered standard errors). Neither cohort jumps at either release. Cohorts are defined by the date of a user's first *rated* contest, but the event study runs on the practice panel, whose submissions can predate that contest debut, so a small number of AI_era users appear before 2021; these few-user leads are uninformative, and the testable pre-trend contrast is therefore the transition cohort, which rejects parallel trends on all three proxies. A pure experience-accumulation account, rather than AI exposure, is addressed separately by the matched-career-stage test in Section 5.1, which finds the AI_era and transition cohorts indistinguishable at the same career stage.

5.2 A Cross-subpopulation CF Rating Asymmetry

In this section, we assess CF rating changes under the AI-prompt shift discussed above, comparing two subpopulations of CF users: roster-linked users who participate in AI-prohibited ICPC and IOI contests, and non-rostered users who do not. The substitute-vs-complement mechanism predicts an asymmetry across these pools: the AI-prohibited contests filter substitute-users out of the roster-linked pool, since their non-AI-aided skill is too weak to qualify. The AI signature should therefore predict weaker rating gains in the non-rostered pool (where substitute-users remain) but be neutralized in the roster-linked pool (where they have been screened out).

We implement this analysis both ways, two subpopulation regressions and one pooled regression.

- The subpopulation version fits two parallel cross-sectional regressions, one per subpopulation $s \in \{\text{non-rostered}, \text{rostered}\}$:

$$\Delta r_h = \alpha_s + \theta_s \cdot \text{ai_pc1_post}_h + \gamma_s \cdot \bar{r}_h^{\text{pre}} + \varepsilon_h, \tag{5}$$

  with robust standard errors. The outcome Δr_h is the user's mean CF rating after ChatGPT minus their mean rating before, computed from the monthly rating panel as $\bar{r}_h^{\text{post}} - \bar{r}_h^{\text{pre}}$, where the post-ChatGPT window covers months $t \in$ [2022-12, 2025-12] and the pre window covers $t \in$ [2018-01, 2022-12).[3] The covariate $\bar{r}_h^{\text{pre}}$ controls for the user's baseline skill level.

  The coefficient θ_s of `ai_pc1_post` is the coefficient of interest, measuring how AI-signature predicts the subpopulation's rating change. The covariate `ai_pc1_post` is the user's post-ChatGPT AI-signature score, defined as the first principal component of the three sign-aligned behavioral AI proxies introduced in Section 5.1 (`frac_single_shot`, the negation of `median_attempts`, and the negation of `frac_quick_retry`), so that all three components contribute positively to the AI-prompt direction. The bootstrap-validated principal component analysis (PCA) loadings of the three AI-prompt patterns are 0.614, 0.535, and 0.580 respectively.[4] The *level* is the construct the cross-sectional question calls for: conditioning on how AI-style a user's practice is, we ask whether this between-user difference predicts the user's rating change and whether screening severs the link. This differs from the within-user *change* used in the within-ICPC test (Section 5.3), which asks a within-person

[3] ChatGPT released on 2022-11-30, so December 2022 falls in the post window. These monthly rating windows split at the release month without excising a buffer; this differs from the quarterly DiD, which drops the 2022-Q4 boundary quarter because its coarser quarterly resolution cannot separate the pre- and post-release weeks within that quarter.

[4] Because the three proxies share a single submission stream, we check that the composite is not one timing pattern restated three times. After sign alignment, the proxies are positively but only moderately correlated (post-shock levels, n = 8,597: pairwise Pearson r between 0.45 and 0.68, mean 0.56), so the average pair shares roughly a third of its variance, while the first principal component explains 71% of the total. The composite thus aggregates three related but non-redundant indicators rather than a single repackaged variable.

skill-formation question.

- The pooled fully-stratified version runs both subpopulations through a single regression with an indicator NR_h for non-rostered users:

$$\begin{aligned}\Delta r_h = \alpha_0 + \alpha_1 \cdot \mathrm{NR}_h + \theta_0 \cdot \mathrm{ai_pc1_post}_h + \theta_1 \cdot \mathrm{ai_pc1_post}_h \cdot \mathrm{NR}_h \\ + \gamma_0 \cdot \bar{r}_h^{\mathrm{pre}} + \gamma_1 \cdot \bar{r}_h^{\mathrm{pre}} \cdot \mathrm{NR}_h + \varepsilon_h.\end{aligned} \tag{6}$$

  The interaction term coefficient θ_1 is the coefficient of interest, measuring the cross-subpopulation asymmetry in rating changes.

The sample consists of users who remain active in the post-ChatGPT window: a user enters if they have at least two quarters between 2023-Q1 and 2025-Q4 with at least five practice submissions in each. After dropping users with missing Δr_h or $\bar{r}_h^{\mathrm{pre}}$, the analysis sample contains 6,690 users (5,283 non-rostered and 1,407 roster-linked).

Identification works cross-user within a subpopulation for θ_s (Equation 5) and cross-subpopulation for θ_1 (Equation 6). The within-subpopulation comparison holds the screening environment fixed and asks how AI-style behavior co-moves with rating change among users facing the same gate (or no gate). The cross-subpopulation comparison holds the AI signature fixed and varies the screening environment. The asymmetry test relies on the testable assumption that the roster-linked pool has been screened of substitute-users by the qualifying rounds; so it does not require roster status to be randomly assigned.

The baseline-skill control $\bar{r}_h^{\mathrm{pre}}$ blocks omitted-variable bias. The AI signature `ai_pc1_post` correlates negatively with $\bar{r}_h^{\mathrm{pre}}$, since high-AI users are on average lower-skill, and $\bar{r}_h^{\mathrm{pre}}$ independently predicts Δr_h (its coefficient is -0.32 in non-rostered and -0.34 in roster-linked, both with $p < 0.0001$). Without the control, the AI coefficients $\theta_\cdot$ would be biased.

This specification, however, does not identify the *causal* effect of AI on rating change. The AI signature is observational, and high-signature users may differ from low-signature users in unobservables. We control for baseline skill and exploit the cross-subpopulation screening difference, not random assignment of AI use; the asymmetry test is interpretable as evidence of a screening filter under the maintained assumption that the ICPC-IOI qualifier screens substitute-users out of the roster-linked pool.

Results. The within-subpopulation slopes on `ai_pc1_post` are starkly asymmetric (Table 7). The non-rostered slope is -12.24 rating points (SE 2.45, $p < 0.0001$, 95% CI $[-17.04, -7.44]$). The roster-linked slope is statistically null: -2.41 (SE 4.33, $p = 0.58$). The interaction coefficient θ_1 of the pooled regression is -9.82 (SE 4.98, $p = 0.048$, 95% CI $[-19.58, -0.07]$): the AI signature predicts around ten points weaker rating gains per signature unit in the un-screened pool than in the

screened roster-linked pool. A one-unit change in `ai_pc1_post` is hard to interpret because it is a principal-component score. To gauge the magnitude of the estimates, we scale by `ai_pc1_post`'s standard deviation (SD), which is 1.48 on the estimation sample (see Table 7). Accordingly, a one-SD-stronger AI signature is associated with an 18-point lower rating gain among non-rostered users (-12.24×1.48), roughly 6% of the mean rating gain of 300 points post-ChatGPT across all users (see Panel B, Table 5). The same scaling applies to the cross-subpopulation gap θ_1: a one-SD-stronger signature is associated with a 14.5-point lower rating gain in the unscreened non-rostered pool than in the screened roster-linked pool (-9.82×1.48).

Table 7: Cross-subpopulation CF rating asymmetry on `ai_pc1_post`. Sample: N = 6,690 users active in the post-ChatGPT window (5,283 non-rostered + 1,407 roster-linked). Robust standard errors. The fully-stratified pooled spec tests $\theta_{\text{non_rostered}}$ and θ_{rostered} being statistically different. `ai_pc1_post` is a principal-component score with standard deviation 1.48 on this sample.

Coefficient	θ	SE	p	95% CI
Subpopulation specifications (Equation 5)				
Non-rostered slope $\theta_{\text{non_rostered}}$	**−12.24**	2.45	< 0.0001	$[-17.04, -7.44]$
Roster-linked slope θ_{rostered}	−2.41	4.33	0.58	$[-10.91, 6.08]$
Pooled fully-stratified spec (Equation 6)				
`ai_pc1_post×non_rostered` (θ_1)	**−9.82**	4.98	0.048	$[-19.58, -0.07]$

The asymmetry is in the direction the screening-filter mechanism predicts. The AI signature predicts weaker rating gains in the non-rostered pool, where substitute-users still appear. The same signature is uncorrelated with rating change in the roster-linked pool, consistent with the AI-prohibited ICPC and IOI qualifiers having filtered substitute-users out. This roster-linked null is a failure to reject rather than a precise zero (95% confidence interval $[-10.91, 6.08]$), so the slightly negative point estimate should not be read as skill erosion among the screened. The differential-attrition check below bounds how much of the null could reflect sample selection rather than screening.

However, the asymmetry on CF rating, an outcome where in-contest AI use is hard to police, raises a reasonable objection: if roster-linked users use AI more proficiently during CF contests than non-rostered users, the gap could be measuring who uses AI less effectively during CF contests rather than who developed weaker non-AI-aided skill. The within-ICPC result (Section 5.3) addresses this objection by moving to the AI-prohibited ICPC environment where in-contest AI use is procedurally blocked. Together with the positive within-ICPC slope (Section 5.3), this negative slope yields the same-training-inputs, opposite-signs pattern that operationally distinguishes substitute use from complement use of AI.

Selection robustness. The post-ChatGPT activity requirement (≥ 2 quarters with ≥ 5 practice submissions each) is needed for `ai_pc1_post` to be reliably measured at the user level, but it cuts the sample from 10,419 users (the Section 5.1 sample) to 6,690, raising a sample-selection concern that we address in three ways (Appendix A.1 reports the full results). First, the Section 5.1 practice-pattern shift survives on the Section 5.2 sub-sample with 8 to 13% attenuation across all three outcomes (all $p < 0.0001$), so the sub-sample is not behaviorally distinct from the full panel. Second, the cross-subpopulation interaction θ_1 is stable under filter relaxation: weaker filter thresholds (fewer required quarters and submissions) yield -11.92 ($p = 0.017$) and -7.14 ($p = 0.081$), bracketing the headline -9.82. Third, roster-linked users drop out of the activity filter more often than non-rostered users (43.3% vs 33.5%), likely because ICPC and IOI contestants step back from open CF practice after their qualifying year. The remaining roster-linked sample is therefore more selected than the non-rostered sample, and the asymmetry magnitude varies non-monotonically across filter thresholds (the $[-11.92, -7.14]$ range above brackets -9.82). Appendix A.1 reports the full balance decomposition by subpopulation.

Selection bounds. Next, we bound the cross-subpopulation interaction θ_1 against selection on unobservables and against the differential roster-linked attrition just noted (Table 8). Against selection that mimics the observables, the asymmetry is robust: adding the baseline-rating control barely moves θ_1 (from -10.04 to -9.82), so by the coefficient-stability logic of Oster (2019) an unobservable would have to be roughly twelve times as important as all observables combined to drive θ_1 to zero, and the asymmetry holds within baseline-rating-by-cohort matched cells (coarsened exact matching: $\theta_1 = -9.73$, $p = 0.050$). It is not, however, robust to the worst case of the differential attrition between the two subpopulations. Since the non-rostered pool is retained more often than the roster-linked pool, the Lee trimming bound (Lee 2009) brackets θ_1 in $[-11.95, -0.35]$: benign attrition (-11.95) leaves the asymmetry intact, while the worst-case trim shrinks it to near zero (-0.35). Relatedly, given the marginal significance of θ_1 ($p = 0.048$), its robustness value (Cinelli and Hazlett 2020) is small: a confounder explaining only 2.4% of the residual variance in both the treatment and the outcome would be enough to zero θ_1. As a benchmark, the baseline rating control explains 15.7% of that residual variance, so a confounder far weaker than baseline rating would suffice to overturn the result. We therefore read the cross-subpopulation asymmetry as suggestive and attrition-bounded. The mechanism evidence rests on the within-ICPC test (Section 5.3), which is measured in an environment where AI use is procedurally blocked. In particular, we rely on the sign of its estimate rather than the precise magnitude, and the sign is consistently positive across various robustness checks.

Table 8: Selection bounds on the cross-subpopulation interaction θ_1 (headline -9.82, $p = 0.048$). The asymmetry is robust to selection on observables (Oster; matching) but bounded by the differential roster-linked attrition (Lee) and fragile at the significance margin (Cinelli–Hazlett).

Method	Result for θ_1	Reading
Oster (2019) δ for $\theta_1 = 0$	12.6 to 147	Robust to observable-like selection
Coarsened exact matching	-9.73 ($p = 0.050$)	Holds on baseline-rating × cohort support
Lee (2009) attrition bound	$[-11.95, -0.35]$	Near zero in the worst case
Cinelli and Hazlett (2020) $\text{RV}_{q=1}$	0.024 (vs. 0.157, baseline rating)	Fragile at the $p = 0.048$ margin

The Oster δ range spans $R^2_{\max} = 1$ to $R^2_{\max} = 1.3\tilde{R}^2$; $\delta > 1$ means unobservable selection must exceed observable selection to overturn the estimate. The Lee bound trims the over-retained non-rostered pool to the roster-linked retention rate. $\text{RV}_{q=1}$ is the share of residual variance a confounder must explain in both treatment and outcome to zero the estimate. Full results available from the author.

5.3 AI-complement practice and higher non-AI-aided scores

The third set of analyses asks whether AI-intensity during CF practice predicts AI-prohibited ICPC performance differently across cohorts. We estimate a pooled cross-sectional ordinary least squares (OLS) regression with university fixed effects and team-clustered standard errors:

$$s_{h,y} = \alpha + \psi_0 \cdot \text{ai_intensity_pc1}_h + \sum_{c \in \{\text{trans,AI}\}} \psi_c \cdot \text{ai_intensity_pc1}_h \cdot \mathbb{1}[\text{cohort}_h = c] + \gamma \cdot \bar{r}_h^{\text{pre}} + \delta_y + \nu_u + \varepsilon_{h,y}, \tag{7}$$

where each observation is a user h appearing at an ICPC contest in year y, δ_y is a contest-year fixed effect, and ν_u is a university fixed effect.

The outcome $s_{h,y}$ is the ICPC score of user h's team in year y (the number of problems the team solved at the contest). ICPC is a team competition: three contestants compete as one university team, so the same team score outcome is shared within team. This shared-outcome structure motivates the team-level clustering of standard errors discussed below. The treatment `ai_intensity_pc1` measures each user's within-user shift toward AI-style CF practice. Each of the three AI-prompt proxies of Section 5.1 is computed over a pre window (2021-Q3 through 2022-Q3, the five quarters between Copilot and ChatGPT releases) and a contest-year-conditional post window (2023-Q1 through the quarter immediately preceding the contest); the user-level change on each proxy (the first difference) is the post-window mean minus the pre-window mean. `ai_intensity_pc1` is the first principal component of the three sign-aligned changes. The PCA loadings are fit on a broad CF reference group ($n = 4{,}959$), comprising CF practice-panel users of any cohort, rostered or not, with sufficient practice activity in both the pre- and post-windows, and

then frozen and applied to the within-ICPC sub-sample.[5] The PCA loadings on `frac_single_shot`, the negation of `median_attempts`, and the negation of `frac_quick_retry` are 0.630, 0.530, and 0.568 respectively. The first-difference construction nets out each user's pre-shock baseline practice style, so the treatment captures the within-user *change* in signature rather than the cross-user level. We assign each user a single value, anchored at their first qualifying year, so that the treatment is a time-invariant, pre-determined user characteristic rather than one per user-by-contest-year.[6] This choice is deliberate on two grounds. First, anchoring at the user's entry into the qualifying pool measures the practice shift before any contest outcome could feed back into how the user practices. Second, it avoids a mechanical artifact for users with multiple 2023+ ICPC contests: a contest-year-conditional value would drift with the length of the post window (one or two quarters for a 2023 contest versus eight or more for a 2025 contest) rather than with any genuine change in practice. The baseline-skill control $\bar{r}_h^{\text{pre}}$ (defined as in Section 5.2) is included for the same omitted-variable reason. Cohort assignment follows Section 4.2 (Table 3), with pre_AI as the reference. Contest-year fixed effects δ_y absorb year-specific score-distribution variation (different problem sets, different scoring rubrics across years), and university fixed effects ν_u absorb the time-invariant university-selection layer described below.

University fixed effects and team-clustered standard errors together address two identification issues. The first is selection into university. Top universities recruit more homogeneous high-skill students, and subsequently their teams may contain both higher scores and more AI_era high-AI-intensity members. With the fixed effects, the AI_era × AI-intensity coefficient identifies off within-university variation only, comparing team members at the same university across years. The second issue is the shared team score within team members described above: residuals are mechanically correlated within team, and team-level clustering captures that correlation. With university fixed effects already absorbing the time-invariant university-selection component, the residual structure across teams within a university is plausibly small, and team-level clustering is enough. A more conservative alternative would cluster at the university level ($n = 196$); this inflates the standard error on the headline coefficient from 0.087 to 0.104 and shifts the p-value from 0.028 to 0.068, leaving the headline marginal but the point estimate unchanged. Dropping the university fixed effects entirely (and clustering at the team level) returns a larger $\psi_{\text{AI}} = 0.323$ (SE

[5] We fit the loadings on a broad reference group rather than on the 444 within-ICPC users so that the axis is estimated from population-level co-variation and is not defined by the same sample on which it is then tested. The reference group is every user in the CF practice panel (all cohorts, not restricted to the ICPC or IOI rosters) with ≥ 3 active quarters in the pre-window (2021-Q3 through 2022-Q3) and ≥ 2 active quarters in the full post-window (2023-Q1 through 2025-Q4), where a quarter is active if it has ≥ 5 practice submissions. About 91% of these users are in the analysis panel; the remainder are out-of-window-cohort or name-only-matched users that the panel excludes. The frozen loadings are then applied to each within-ICPC user's contest-year-conditional first-differences to produce that user's `ai_intensity_pc1`.

[6] The anchor is the user's earliest year in which both the pre- and post-windows contain sufficient practice activity; for users with multiple 2023+ ICPC contests this can differ from the year of the literal earliest contest when that earliest year lacked qualifying activity.

0.157, $p = 0.040$), because cross-university confounding is no longer absorbed; we therefore treat the within-university, team-clustered version as the headline.

The sample for this analysis is the intersection of the three cohorts and the ICPC roster, restricted to contest years on or after 2023 where AI-intensity is meaningfully defined. Table 9 reports the chain that produces the sample.

Table 9: Sample filter chain for the within-ICPC regression (Equation 7). The within-ICPC regression and the practice-pattern DiD (Equation 4) share the same analysis sample of 10,419 users; the within-ICPC regression retains 444 users after the ICPC-specific filters, yielding 579 user-by-contest-year observations across 321 teams in 196 universities.

Step	Filter	Users	Drop	Retention
0	The three cohorts in the CF panel (shared with Equation 4)	10,419	—	—
1	Linked to any ICPC team roster entry	1,673	−8,746	16.1%
2	ICPC participation in contest year ≥ 2023	1,026	−647	61.3%
3	`ai_intensity_pc1` defined (qualifying pre- and post-window practice activity)	447	−579	43.6%
4	$\bar{r}_h^{\text{pre}}$ also defined (rated CF history before ChatGPT)	444	−3	99.3%

The coefficient of interest, ψ_{AI} on the interaction between `ai_intensity_pc1` and the AI_era cohort indicator, identifies how the AI-intensity–score relationship differs between the AI_era and pre_AI cohorts (within a university and a contest year, and holding baseline skill fixed), rather than the causal effect of AI use on score. The pre_AI cohort serves as a built-in placebo: its members had little AI exposure during training, so its slope ψ_0 should be near zero under any reading of the mechanism, and the substantive test is whether the AI_era slope is distinguishable from it. The estimate is observational rather than experimental: the within-user change in AI signature is not randomly assigned, so AI_era users who shifted more toward AI-style practice may differ on unobservables from those who shifted less. The cross-cohort comparison conditions on baseline skill and contest-year fixed effects but cannot rule such differences out; we therefore read the coefficient through the screening-filter lens discussed with the results below, not as a causal effect of AI use.

Results. The AI_era cohort interaction is positive and significant: $\psi_{\text{AI}} = 0.190$ (SE 0.087, $p = 0.028$). For pre_AI users (the reference cohort and the de facto placebo), the slope is small and null ($\psi_0 = -0.048$, SE 0.047, $p = 0.303$). For transition users the interaction is also null ($\psi_{\text{trans}} = 0.060$, SE 0.068, $p = 0.373$, implying a cohort slope of +0.012). Combining the pre_AI reference slope and the interaction, the AI_era cohort's own AI-intensity slope is +0.142 (SE 0.074, $p = 0.055$, 95% CI $[-0.003, 0.287]$): positive and of the predicted sign, though only marginally significant on its own. The design's substantive test is the cohort *contrast* with the pre_AI placebo

(the interaction $\psi_{\text{AI}} = 0.190$, $p = 0.028$), which is also the quantity that completes the opposite-signs comparison with Section 5.2. The slope on AI-intensity therefore differs across cohorts only for the AI_era group, with the sign in the predicted direction: among AI_era users, who started their competitive programming careers with AI tools, a larger within-user shift toward AI-style practice predicts *higher* ICPC score in 2023–2025, controlling for baseline skill and university (Table 10).

Table 10: Within-ICPC AI-intensity by cohort. Headline specification with university fixed effects and team-clustered standard errors. Sample: 579 user-by-contest-year observations across 444 users, 321 teams, and 196 universities. Contest years 2023–2025. Reference cohort: pre_AI. The AI_era cohort's marginal AI-intensity slope (pre_AI reference slope plus the AI_era interaction) is +0.142 (SE 0.074, $p = 0.055$).

Term	ψ	SE	p	95% CI
`ai_intensity_pc1` (pre_AI reference)	−0.048	0.047	0.303	[−0.140, 0.043]
`ai_intensity_pc1`:transition	0.060	0.068	0.373	[−0.072, 0.193]
`ai_intensity_pc1`:AI_era	**0.190**	**0.087**	**0.028**	[**0.020**, **0.360**]
$\bar{r}_h^{\text{pre}}$	0.001	0.000	< 0.001	[+0.000, 0.001]
University FE	✓			
Contest-year FE	✓			
SE clustering	team			

This positive sign inside the AI-prohibited environment is the cross-environment counterpart to the negative slope in the un-screened CF pool (Section 5.2): the same signature, opposite signs. Under the screening-filter reading, the ICPC qualifier has already removed users whose AI-style practice reflects substitution; the within-user shift toward AI among the qualified pool then reflects complement-style adaptation that raises score, not substitution that erodes it.

A natural question is why the roster-linked slope of Section 5.2 is a (negative) null while the slope here is positive. The two estimates answer different questions, and two differences line them up. First, Prediction 2 called for a null there: on CF rating, screening removes the erosion signal, and rating is a relative measure that complement-style skill gains need not move among elite peers. Second, the populations differ. The positive slope is specific to the AI_era cohort, while the roster-linked pool is dominated by the pre_AI and transition cohorts, whose within-ICPC slopes are null (−0.048 and +0.012).[7] A pooled, level-based regression on a mostly pre-AI elite pool should

[7]Weighting the three within-ICPC cohort slopes (−0.048, +0.012, and +0.142) by their user counts (144, 233, and 67; Table 5, Panel C) gives a pooled slope of about +0.01, indistinguishable from zero. That is what the roster-linked regression of Section 5.2 finds in its own units: −2.41 with a standard error of 4.33 ($p = 0.58$). Restricting the Section 5.2 rating regression to exactly this 444-user ICPC sample makes the reconciliation direct rather than weighted. The open-pool slope on `ai_pc1_post` is null (+4.31, SE 5.88, $p = 0.46$) on the very users whose within-contest slope is +0.190, while the cross-subpopulation asymmetry estimated against this screened group is −16.55 ($p = 0.009$), if anything sharper than the headline −9.82. The 444-user slope is noisy and its positive sign is not distinguishable from zero, and part of the sharper gap reflects this group being younger and more AI_era than the full roster pool.

find nothing, and it does.

Construct and selection robustness. Three caveats apply to the headline estimate $\psi_{\text{AI}} = 0.190$.

- First, the construct is a within-user first difference, aggregated to a single per-user value anchored at the user's first qualifying year, with principal-component loadings fit on a broad reference group, and frozen before they touch the within-ICPC sample. We read the within-ICPC evidence as establishing the *sign* of the effect, the direction that completes the opposite-signs test with Section 5.2, rather than the precise magnitude 0.190.

- Second, the pre-window of the first-difference (2021-Q3 through 2022-Q3) partly predates the rated competition careers of AI_era users (whose first rated contest is post-Copilot by construction): these contestants are the novices in the 2023–2025 sample (mean tenure 3.9 years since their first rated contest, against 7.4 years for the pre_AI cohort), so the AI_era change could in part capture novice-to-competitor maturation rather than an AI-related shift. To test this, we add exogenous career-stage controls to Equation 7: tenure, tenure interacted with cohort, and the user's total count of rated contests. The AI_era slope does not move: ψ_{AI} goes from 0.190 to between 0.195 and 0.196 (p between 0.025 and 0.027) across these specifications. It even survives adding the user's CF rating growth over the pre-to-post span as a control (0.190, $p = 0.038$). That check deliberately over-controls: rating growth is itself partly an outcome of AI use, so conditioning on it removes part of the effect being tested. Were the slope mainly driven by elite-novice maturation, holding career stage fixed would attenuate it. It does not, so the novice-growth confound is bounded.

- Third, the within-ICPC sample is reached through the filter chain in Table 9, and two steps raise selection concerns: the Step-3 activity filter drops 579 (56%) of ICPC participants who lack the practice activity needed to measure a first-difference, and the Step-2 participation cutoff drops 647 users with pre-2023 ICPC history but no recent participation. The Step-2 exit is ambiguous between graduation, loss of interest, and AI-driven dropout. If the AI-driven-dropout channel is present, the cutoff itself would remove substitute-users, doing part of the screening we attribute to the qualification gate. Appendix A.2 reports the full filter sensitivity, balance check, and Section 5.1 cross-check.

IOI sister analysis. The IOI counterpart to the within-ICPC AI-intensity test is weaker than ICPC's in two ways. First, precision: the IOI is a smaller pool and the test runs on only 345 contestant-year observations from 243 CF-linked contestants over the 2023–2025 window, in comparison with the 579 observations from 444 users for ICPC. So the estimate is far less precise. Second, age and timing: IOI participants are typically high-schoolers aged 17–19, while ICPC participants are

college students aged 18–23, so a higher share of an IOI contestant's skill-formation window is post-AI, leaving less pre-AI training to anchor the cohort contrast.

Nevertheless, we estimate the same AI-intensity-by-cohort specification on the IOI individual contestants linked to CF accounts, clustered at the contestant level. The AI_era cohort interaction on IOI score is 18.1 (SE 18.0, $p = 0.31$): directionally consistent with the ICPC result but far noisier, with a 95% confidence interval including zero. We anchor the non-AI-performance headline on ICPC and treat IOI as a directional check only.

5.4 Robustness

This subsection summarizes the full robustness battery in Appendix A. The first three items address result-specific threats, ordered by severity; the last two checks apply to all three analyses.

- **The cross-subpopulation asymmetry (Section 5.2).** This is the most fragile result. The activity filter cuts the sample from 10,419 to 6,690 users, and roster-linked users exit more often (43.3% versus 33.5%). Consequently, differential selection could mimic screening. Table 8 bounds the asymmetry against selection on observables and against this attrition. Appendix A.1 adds the filter-threshold sweep, the dropout balance check, the Section 5.1 cross-check, and the contest-mode signature.

- **The within-ICPC AI-intensity slope (Section 5.3).** The within-user treatment is measurable only for users with sufficient practice in both pre and post windows, a requirement that drops 56% of ICPC participants, so the activity filter itself could select the sample. Appendix A.2 varies the filter thresholds, checks balance on the dropped users, and confirms the Section 5.1 practice shift on the retained sub-sample.

- **The cross-cohort practice shift (Section 5.1).** This result is descriptive. Its checks therefore concern timing. The event study presented there tests pre-trends and release-timed jumps. Appendix A.3 re-anchors the DiD to the Copilot release and restricts the post window to the five quarters before ChatGPT: the transition cohort's signature shift is already present around the first AI shock, so the headline is not a ChatGPT-specific artifact.

- **The COVID-19 disruption (Appendix A.4).** The pandemic falls inside the empirical window and just precedes the Copilot release, so pandemic-era disruptions to practice behavior and to the contests could be mistaken for AI-timed changes. The practice DiD absorbs common pandemic shocks through its quarter fixed effects, the rating-change windows average over multiple years on each side of ChatGPT, and the within-ICPC test uses contest years that post-date the disruption.

- **The roster-to-CF linkage (Appendix A.5).** The subpopulation split behind the Section 5.2 asymmetry rests on the roster-to-CF linkage, so ambiguous matches, if mishandled, could shift users across the two arms and bias the estimate. The 1,374 excluded lower-confidence matches are users linked to a contest roster by name and country alone rather than a published CF username. The within-subpopulation slopes are stable under all three handlings of these users: excluded (the headline), kept as non-rostered, or included in the roster-linked arm.

6 Discussion

The findings combine into one internally consistent picture: a cross-cohort shift in practice patterns toward an AI prompt-and-fix style; a cross-subpopulation asymmetry in which the AI-prompt signature predicts weaker CF rating gains in the un-screened, non-rostered pool but not in the screened, roster-linked pool; and a positive within-ICPC association between a shift toward AI-style practice and non-AI-aided contest score among AI-era entrants who survived the qualification gate. These pieces fit together under the selection-into-AI-prohibited-contests structure, which is the central contribution of this paper.

6.1 Contributions

This paper makes three contributions to the emerging empirical literature on AI and skill formation. First, the paper introduces the selection-into-AI-prohibited-contests structure as a central feature of empirical designs in this literature. Competitive credentialing systems with AI-prohibited gates filter substitute-dependent agents out of the elite pool. This filtering changes which signals appear in which observational windows: the substitution-driven rating signal from substitute-users is observable only on AI-allowed platforms, while AI-prohibited contests, having screened those users out, reveal only the complement-user association among the deep-skill subpopulation that remains. Recognizing this selection structure reconciles the cross-subpopulation and within-ICPC findings that would otherwise appear contradictory, and it has implications for an empirical design that uses credentialing-gate variation to identify AI substitution effects.

This selection structure also yields a design principle for the organizations that evaluate talent. The AI-prohibited gate is a designed evaluation regime: its rules (no AI, in person, proctored) and its qualification structure are choices that determine whether differential AI use is rendered visible or screened out. The AI-prompt signature, in turn, tracks how AI is enacted in routine practice, not whether a user has access to it, a distinction developed in research on technology-in-practice (Orlikowski 2000). Our core result is that the same practice carries opposite performance signals depending on the screening design of the environment in which skill is later tested. The policy

lever therefore sits in the design of the evaluation regime, which is the margin directly controlled by organizations. Evidence from organizational economics points the same way: the returns to a capable algorithm depend on the design of the decision-authority structure around it, not on access alone (Kim et al. 2024).

These results feed back to the organizational literatures of Section 2.4. To the algorithmic-control conversation (Kellogg et al. 2020, Rahman 2021), they add the transparent case: where opaque ratings leave workers unable to learn what the evaluation values, the gate's criterion is explicit, and our estimates speak to what such a design renders visible. To the accounts of expertise forming or failing around analytical tools (Anthony 2021, Anthony et al. 2023), they supply a quantitative, multi-year footprint of the same dynamics in an elite skill market. For the organizations that recruit from AI-era talent pools (R&D labs, engineering firms, professional partnerships alike), the design question our evidence sharpens is which stage of the evaluation pipeline supplies the unaided signal.

Second, the paper operationalizes the mechanism of human capital depreciation in Acemoglu et al. (2026) at the individual frontier scale, providing a multi-year, field-observational counterpart to the macro-theoretical argument and the laboratory evidence in Bastani et al. (2025), Stadler et al. (2024), and Shen and Tamkin (2026). Our opposite-signs result shares its structure with the independent work of Bondi and Johnson (2026): the same AI input reads differently depending on the design that measures it. Their single-agent model derives this across evaluation designs: as workers keep using AI, a within-worker panel makes the AI gain look larger while a long-horizon experiment makes it look smaller. They call this divergence a "scissors." We find it across populations: the same practice signature carries a negative sign in the un-screened pool and a positive sign inside the screened pool, because the AI-prohibited gate separates substitute-users ($\mu < 1$, pedagogical quality parameter in their model) from complement-users ($\mu > 1$). Our setting therefore adds what their single-agent model does not contain, heterogeneity in μ across people and an institution that sorts on it. We do not reproduce their over-time scissors, which would require a long-horizon experiment our design lacks. The cross-cohort practice-pattern shift we document, toward an AI prompt-and-fix submission style at multi-year scale, is consistent with the debug-skipping mechanism that Shen and Tamkin (2026) identify in an RCT.

Third, the paper provides field evidence that the substitute-versus-complement distinction is empirically real: at multi-year scale, the two modes of AI use are measurable from submission behavior alone, and they accumulate into divergent unaided performance. This substitute-versus-complement distinction is the premise on which both the measurement and the policy readings of our results hinge. That premise carries substantial socio-economic stakes because unaided frontier capability still matters in an AI-saturated world: catching AI errors requires unaided expertise, schemas not built during formative years may never form (Stadler et al. 2024), capability that depends on AI access disappears when that access is disrupted, and knowledge/human capital

frontier creation feeds the training data on which the advance of AI depends (Acemoglu et al. 2026).

6.2 Practical implications and policy levers

The findings suggest two complementary policy levers for preserving non-AI-aided frontier capability in elite skill markets. They operate at different scales and through different mechanisms, and they reinforce rather than substitute for each other. Both rest on the same premise: that the mode of AI use is a behavior, not a fixed trait.

Substitute use is a mode, not a fixed trait. A reader could take our results to mean that two pre-existing segments simply sort, weaker users into substitute mode and stronger users into complement mode, and that the only forecast is a widening gap between the two segments in their skill levels. However, three facts resist that reading. First, the signature is not skill: before the AI releases, more skilled users were, if anything, *less* AI-like on these behaviors (see Appendix B), and the cross-subpopulation asymmetry holds conditional on baseline rating. Second, the underlying behavior moves within person: the within-ICPC treatment is a within-user change in practice style, so the practice patterns that carry the two modes are changeable at the individual level, not fixed at entry. Third, when the mode has been manipulated experimentally, it responds to how AI is designed and used: hint-based GPT tutoring removes the harm that full-solution GPT access does to unaided performance (Bastani et al. 2025), and engagement patterns separate outcomes under the same AI tool (Shen and Tamkin 2026).

Our gate evidence concerns the detection and separation of modes, not the reassignment of individuals between them. Whether a user can be steered from one mode to the other is an experimental question our design cannot answer. Nevertheless, the distributional concern, a widening skill gap between weaker and stronger users, remains real, and it is why governing the mode of AI use matters: if substitute-style practice concentrates among weaker users, ungated environments will widen gaps that gated environments can check. Two levers follow, one that shapes the mode within training and one that screens it at the population level:

- **Lever 1: within-training, complement-style AI integration.** Our within-gate result makes this lever more than a conjecture. Among survivors of the qualification gate, a stronger shift toward AI-style practice is associated with *higher* non-AI-aided ICPC performance (Section 5.3): the complement channel, AI that supports rather than supplants skill formation, is visible among elite performers. What our design cannot identify is whether the *mode* of integration can be steered to put more users on that channel; Bastani et al. (2025)'s classroom RCT supplies the causal step, showing that scaffolded AI (problem-specific hints rather than full solutions) eliminates the substitution effect on non-AI-aided performance. Read together, the field pattern and the RCT motivate a training principle for elite settings: integrate AI as a

scaffolded complement rather than a full-solution substitute.

- **Lever 2: AI-prohibited evaluation gates at the population level.** The selection-into-AI-prohibited-contests structure functions as a population-level filtering mechanism. Programmers whose competitive performance relies heavily on AI cannot pass AI-prohibited qualification rounds; the elite pipeline retains deep-skill investors while substitute-users remain visible only on AI-allowed platforms. The cross-subpopulation asymmetry (Section 5.2) is consistent with the filter screening substitute-users out before the contest. The within-ICPC test (Section 5.3) is consistent with the same filter: the positive within-gate association noted under Lever 1 reflects complement-user selection, the substitute-users having been screened out upstream. The lever is the existence and design of AI-prohibited evaluation checkpoints at the population scale, not the elimination of AI from training.

Sorting vs. anticipation: what we identify and what we do not. The mechanism we identify for Lever 2 is *sorting*: AI-prohibited gates filter substitute-dependent programmers out of the elite pipeline. A separate, complementary mechanism, *anticipation*, would operate behaviorally: programmers who know the AI scaffold will be removed at critical evaluation moments invest more in deep skill during training and use AI as a complement rather than a substitute. The two mechanisms have different population-level implications. Under sorting alone, AI-prohibited gates winnow the pool but do not change individual behavior; under anticipation, the gates' existence shifts the training distribution itself. Our data are consistent with both channels (the sorting and anticipation cases of the framework in Section 3.3) but cannot distinguish them. The symmetry is worth stating: the purely pessimistic reading, in which gates only sort and never change behavior, is equally unproven. Identifying the anticipation channel would require a natural experiment in gate-introduction (or removal), or a longitudinal design tracking training-behavior changes after credible AI-prohibition signals. We flag this as promising future work and rest the policy argument here on the sorting channel that we identify.

Both levers rest on the screening we document, not on a proven erosion effect. The opposite-signs pattern shows what each side of the gate looks like: any long-run cost of substitute-style practice concentrates among the users the gate screens out, while among the complement-users who pass, AI-style practice accompanies higher non-AI-aided performance (Lever 1). The gate is therefore the relevant lever whether or not the erosion materializes. The argument is population-level, not within-individual: AI access is privately valuable to most users, and the externality, if any, sits in the elite pool from which frontier capability is drawn.

6.3 Limitations

Identification. Observational identification has bounded reach. Selection-on-entry into the broad CF population is partially handled by within-individual fixed effects, but selection on continued participation, on AI tool use, and on the choice to enter qualification rounds is not fully eliminated. The robustness battery of Appendix A defends the primary identification but does not resolve all of these concerns.

Non-AI alternatives. Forces unrelated to AI could also contribute to the cross-subpopulation asymmetry (Prediction 2). A turbulent tech labor market during 2022–2025 may have distracted elite programmers from contest participation. Universities have been redirecting top students toward applied AI/ML coursework at the expense of competitive programming. Both could result in weaker average CF rating gains among non-rostered users. Also, for non-rostered users whose pre-period ratings were unusually strong, part of their weaker post-2022 performance could be regression to the mean. These alternative channels, however, move participation and rating levels, not the Section 5.2 estimand, namely a slope on the AI signature conditional on baseline rating. The labor-market and curricular stories could lead to a similar slope only with an added assumption: they would have to both push practice toward the AI-style signature and lower ratings among non-rostered users but not rostered ones. Regression to the mean cannot mimic it at all, because the baseline-rating control (and the matched-cells robustness check) blocks its channel. Furthermore, none of the three channels predicts the positive within-ICPC slope of Prediction 3. The joint pattern is the sharpest test: weaker rating gains that track the AI signature in the open pool ($\theta_{\mathrm{NR}} < 0$), no such shortfall among the screened ($\theta_{\mathrm{RL}} \approx 0$), and *higher* non-AI-aided scores with the same signature inside the AI-prohibited contest ($\psi_{\mathrm{AI}} > 0$). Screening at the AI-prohibited gate generates all three signs with one mechanism, whereas each non-AI alternative would have to explain them separately. Nevertheless, we acknowledge that we cannot fully rule these alternatives out.

Generalizability. Our setting is competitive programming, and extension to other elite skill markets (e.g., mathematics, theoretical computer science, professional research) requires domain-specific natural experiments or cross-domain meta-evidence. The urgency of the AI risk to frontier skill also differs across fields. AI is unusually strong at code generation and debugging, the very skills competitive programming trains, so the substitution-erosion risk is most immediate in our setting. Fields where AI is currently weaker at the focal task face a smaller near-term risk, though model capabilities are expanding rapidly and may erase that difference quickly.

Measurement. We do not directly observe which CF users adopt which AI tools. The cohort × post DiD identifies practice-pattern shifts at the cohort level; the subpopulation comparison computes a

within-user proxy from observed submission-pattern changes. Neither directly measures AI use. The construct-validity case for the AI-prompt signature (discriminant validity against pre-shock skill, and within-user stability around AI release among established users) is developed in Appendix B.

Filter strength. The selection-into-AI-prohibited-contests framing rests on the empirical claim that qualification rounds effectively screen out substitute-users, but we do not observe filter strength directly. Imperfect proctoring at earlier qualification stages would weaken the framework. The within-ICPC positive slope partially probes filter strength: the positive sign is consistent with effective filtering, whereas a strong negative sign would have indicated that substitute-users survive in the qualifier pool. A platform-level or survey-based measurement of user-level AI use would close both this gap and the measurement gap above.

6.4 Conclusion

We ask whether AI-prohibited evaluation gates can separate two coexisting types of AI practice in elite competitive programming, and we assemble evidence in three patterns. The practice-pattern shift is a cross-cohort move toward the AI-prompt signature on the CF practice panel. The cross-subpopulation asymmetry appears in CF rating gains: among non-rostered CF users (no ICPC/IOI selection), a stronger AI-prompt signature during practice predicts lower contest performance; among roster-linked users (selected through AI-prohibited gates), the relationship becomes null. The within-ICPC result shows that a shift toward AI-style practice predicts *higher* non-AI-aided performance for AI_era entrants. Where the gate screens, AI-style practice predicts gains; where it does not, the same practice predicts losses. The unified reading is that AI-prohibited qualification gates separate substitute-style from complement-style practice, the gate removing substitute-users before they reach elite contests. That selection structure is the pivotal feature of the design, and it is what makes the cross-environment evidence coherent. Our evidence speaks to the gatekeeping organization that decides whether any long-run cost of AI-era training reaches the elite skill frontier; it does not establish the existence or the size of that cost. The constructive message is that erosion is not a property of the technology. Within the screened pool, AI-style practice coexists with frontier skill; the risk attaches to the substitute mode of use. Two design levers follow: integrate AI into training so that it complements deliberate practice, and design AI-prohibited evaluation gates so that the two modes remain separable at credential boundaries. Our evidence supports the second lever's premise, that the modes are real and institutionally separable.

Several extensions follow naturally from our research. First, distinguishing the sorting channel from the anticipation channel for AI-prohibited evaluation gates is crucial, but requires a natural experiment in gate introduction or removal, or a longitudinal design tracking training-behavior

changes after credible AI-prohibition signals. Second, extension to other elite skill markets where AI-prohibited gates exist (e.g., medical board and legal bar exams, scientific peer review) could test whether the selection mechanism generalizes. Third, user-level AI-use measurement through platform telemetry or developer surveys would close the identification gap on individual AI use and sharpen the substitution test at the individual level. Fourth, a future update incorporating ICPC 2026–2027 and IOI 2026–2027 outcomes would extend the post-shock observation window for the AI-era cohort whose effects are most informative for long-run skill formation. Beyond these, the deepest open question is how to steer learners between modes rather than merely sort them, both for research and for the organizations that train the next generation of frontier talent.

References

Acemoglu, D., D. Kong, and A. Ozdaglar (2026, February). AI, Human Cognition and Knowledge Collapse. NBER Working Paper 34910, National Bureau of Economic Research.

Acemoglu, D. and P. Restrepo (2020). Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy 128*(6), 2188–2244.

Anthony, C. (2021). When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies. *Administrative Science Quarterly 66*(4), 1173–1212.

Anthony, C., B. A. Bechky, and A.-L. Fayard (2023). "Collaborating" with AI: Taking a System View to Explore the Future of Work. *Organization Science 34*(5), 1672–1694.

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman (2025). Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences 122*(26), e2422633122.

Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Earnings. *Journal of Political Economy 75*(4, Part 1), 352–365.

Bondi, T. and G. Johnson (2026). Skill Atrophy and AI Productivity Measurement. SSRN Working Paper 6169671, Social Science Research Network.

Boussioux, L., J. N. Lane, M. Zhang, V. Jacimovic, and K. R. Lakhani (2024). The Crowdless Future? Generative AI and Creative Problem-Solving. *Organization Science 35*(5), 1589–1607.

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at work. *Quarterly Journal of Economics 140*(2), 889–942.

Budzyń, K. et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*.

Cinelli, C. and C. Hazlett (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B 82*(1), 39–67.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica 78*(3), 883–931.

Dell'Acqua, F. (2022). Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters. Working paper, laboratory for innovation science at harvard, Harvard Business School.

Dell'Acqua, F., E. McFowland, E. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. Lakhani (2026). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Organization Science 37*(2). Earlier circulated as Harvard Business School Working Paper 24-013 (2023).

Deming, D. J. and K. Noray (2020). Earnings Dynamics, Changing Job Skills, and STEM Careers. *Quarterly Journal of Economics 135*(4), 1965–2005.

Gaessler, F. and H. Piezunka (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal 44*(11), 2724–2750.

Galdin, A. and J. Silbert (2025). Making Talk Cheap: Generative AI and Labor Market Signaling. Working Paper 2511.08785, arXiv.

Hausman, N., O. Rigbi, and S. Weisburd (2025, April). Generative AI's Impact on Student Achievement and Implications for Worker Productivity. CESifo Working Paper 11843, CESifo. Also CEPR Discussion Paper 20206; SSRN 5224465.

Hicks, J. (1970). Elasticity of Substitution Again: Substitutes and Complements. *Oxford Economic Papers 22*(3), 289–296.

Hoffmann, M., S. Boysel, F. Nagle, S. Peng, and K. Xu (2024). Generative AI and the Nature of Work. Working Paper 25-021, Harvard Business School. SSRN 5007084.

Horton, J. J. and P. Tambe (2025). The Death of a Technical Skill. *Information Systems Research 36*(3), 1799–1820.

Hui, X., O. Reshef, and L. Zhou (2024). The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market. *Organization Science 35*(6), 1977–1989.

Jabarian, B. and P. Reshidi (2025). Choice as Signal: Designing AI Adoption in Labor Market Screening. SSRN Working Paper 5856382, Social Science Research Network.

Kellogg, K. C., M. A. Valentine, and A. Christin (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals 14*(1), 366–410.

Kim, H., E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca (2024). Decision Authority and the Returns to Algorithms. *Strategic Management Journal 45*(4), 619–648.

Kim, K., N. Mester, and G. Sun (2026). AI and the Labor Market: A Worker's-Eye View. SSRN Working Paper 6240619, Social Science Research Network.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies 76*(3), 1071–1102.

Melumad, S. and J. H. Yun (2025). Experimental evidence of the effects of large language models versus web search on depth of learning. *PNAS Nexus 4*(10), pgaf316.

Mirzayanov, M. (2015, October). Open Codeforces Rating System. Codeforces blog entry 20762.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*(6654), 187–192.

Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science 11*(4), 404–428.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *arXiv preprint*.

Rahman, H. A. (2021). The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. *Administrative Science Quarterly 66*(4), 945–988.

Shaw, S. D. and G. Nave (2026). Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. Working paper, The Wharton School, University of Pennsylvania.

Shen, J. H. and A. Tamkin (2026). How AI Impacts Skill Formation. *arXiv preprint*.

Shihab, M. I. H., C. Hundhausen, A. Tariq, S. Haque, Y. Qiao, and B. Mulanda (2025). The Effects of GitHub Copilot on Computing Students' Programming Effectiveness, Efficiency, and Processes in Brownfield Programming Tasks. In *Proceedings of the 2025 ACM Conference on International Computing Education Research (ICER 2025)*.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics 87*(3), 355–374.

Stadler, M., M. Bannert, and M. Sailer (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior 160*, 108386.

Stiglitz, J. E. (1975). The Theory of "Screening," Education, and the Distribution of Income. *American Economic Review 65*(3), 283–300.

Appendix

A Robustness

This appendix reports the robustness checks referenced in Section 5.4: the selection-bias robustness for the cross-subpopulation and within-ICPC tests, the Copilot-anchored DiD, the COVID-19 disruption, and the ICPC/IOI–CF record linkage with the sensitivity of the cross-subpopulation asymmetry to the excluded name-only matches.

A.1 Selection-bias robustness for the cross-subpopulation asymmetry test

The cross-subpopulation asymmetry of Section 5.2 (Equations 5 and 6) is estimated on cohort-assigned users active in the post-ChatGPT window: users with at least two post-ChatGPT quarters in 2023-Q1 through 2025-Q4, each containing at least five practice-mode submissions. This filter is required so that `ai_pc1_post` is reliably measured at the user level: a single quarter with one or two submissions produces a wildly noisy first principal component. The cost is sample restriction: 3,729 of the 10,419 cohort-assigned users who pass the Section 5.1 practice-window filter do not pass this activity requirement or lack the rating variables, dropping the Section 5.2 sample to 6,690 (5,283 non-rostered + 1,407 roster-linked). This appendix documents three robustness checks against selection bias from this restriction: a filter-sensitivity check on the asymmetry test, a cross-section consistency check with the Section 5.1 practice-pattern shift, and a balance check that decomposes the dropped users by observables and by subpopulation, together with two further checks: skill-control sensitivity and the roster-versus-skill confound.

Filter sensitivity

We re-estimate Equation 5 (per-subpopulation) and Equation 6 (pooled fully-stratified) under three activity thresholds: the headline ($\geq$ 2 quarters, $\geq$ 5 submissions each), an intermediate variant ($\geq$ 2 quarters, $\geq$ 1 submission each), and a permissive variant ($\geq$ 1 quarter, $\geq$ 1 submission). The first principal component is fit once on the headline sample; the resulting PCA loadings and standardization parameters are frozen and applied to the relaxed samples so that `ai_pc1_post` is on the same scale across all three specifications.

The pooled interaction θ_1 is significant at HEADLINE and INTERMEDIATE (with INTERMEDIATE actually larger than HEADLINE), and attenuates to marginal significance at PERMISSIVE. The PERMISSIVE attenuation is consistent with classical errors-in-variables: at $\geq$ 1 quarter with $\geq$ 1 submission, the per-user signature is computed from very few observations and the resulting measurement noise pulls θ_1 toward zero. Within the non-rostered subpopulation, θ_{NR}

Table 11: Cross-subpopulation asymmetry (Equations 5 and 6) under three activity thresholds. PCA loadings frozen on the HEADLINE sample and applied to the relaxed samples. Heteroskedasticity-consistent (HC1) standard errors. The HEADLINE row reproduces Table 7's θ_1 within numerical noise; small differences arise because the PCA is refit on the Section 5.2 reconstruction sample rather than reusing the broad-panel `ai_pc1_post` scores directly.

Filter	N	θ_{NR}	θ_{RL}	θ_1
PERMISSIVE	8,478	-14.61^{***}	-7.47^{*}	-7.14
($\geq$ 1 q, $\geq$ 1 sub)		(1.87)	(3.64)	($p = 0.081$)
INTERMEDIATE	7,753	-16.38^{***}	-4.46	$\mathbf{-11.92}^{*}$
($\geq$ 2 q, $\geq$ 1 sub)		(2.32)	(4.45)	($p = 0.017$)
HEADLINE	6,690	-12.42^{***}	-2.54	$\mathbf{-9.88}^{*}$
($\geq$ 2 q, $\geq$ 5 subs)		(2.45)	(4.37)	($p = 0.049$)

θ_{NR} is the non-rostered slope on `ai_pc1_post`, θ_{RL} is the roster-linked slope, θ_1 is the pooled cross-subpopulation interaction. Standard errors below each estimate for θ_{NR} and θ_{RL}; pooled-interaction p-value below θ_1. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

is significantly negative across all three filters and ranges from -12.42 to -16.38. Within the roster-linked subpopulation, θ_{RL} is null at the two stricter thresholds but reaches significance at PERMISSIVE (-7.47, $p = 0.040$), a pattern we revisit alongside the balance check below.

Cross-section consistency with Section 5.1

If the Section 5.2 sub-sample is materially different from the full Section 5.1 sample in ways that affect the practice-pattern shift, the two analyses are not characterizing the same population, and reading them together is harder to justify. We re-estimate Equation 4 on the 6,690 Section 5.2 users' user-quarter practice observations in 2018-Q1 through 2025-Q4 (excluding 2022-Q4) and compare to the Section 5.1 headline.

Table 12: Equation 4 AI_era × post coefficient β_2 on the full Section 5.1 sample (10,419 users) versus the Section 5.2 sub-sample (6,690 users). Cluster-robust (user-level CRV1) standard errors in parentheses.

Outcome	Section 5.1 (N = 10,419)	Section 5.2 sub-sample (N = 6,690)	Attenuation
`frac_single_shot`	0.0586^{***} (0.0063)	0.0509^{***} (0.0067)	13%
`median_attempts`	-0.3271^{***} (0.0374)	-0.2981^{***} (0.0400)	9%
`frac_quick_retry`	-0.0190^{***} (0.0041)	-0.0175^{***} (0.0043)	8%

*** $p < 0.001$ in all six estimates.

All three practice-pattern outcomes survive on the Section 5.2 sub-sample with modest attenuation (between 8% and 13%); statistical significance is preserved in all three. The Section 5.1 and Section 5.2 findings are about the same underlying population: the sub-sample restriction does not

select on the practice-pattern signal in a way that would explain the cross-subpopulation asymmetry by sample composition alone.

Balance check: in-sample versus dropped users

The 3,729 users dropped at the activity filter are compared to the 6,690 retained users on pre-ChatGPT mean CF rating, Δ rating (post − pre), the spanning indicator (whether the user is active on both sides of the ChatGPT cutoff), the pre-window single-shot AC rate (a proxy for pre-window AI signature among the subset of dropped users with at least one pre-window practice submission), and cohort composition.

Table 13: Continuous observables: Section 5.2 in-sample users versus dropped users. Mean ± standard error of the mean; Welch t-test p-values.

Observable	In-sample	Dropped	Diff	p
Pre-ChatGPT mean CF rating	1,459 ± 5 (n = 6,690)	1,642 ± 7 (n = 3,286)	−183	< 0.0001
Δ rating (post − pre)	300 ± 3 (n = 6,690)	92 ± 3 (n = 3,286)	208	< 0.0001
Spanning indicator	0.79 ± 0.00 (n = 6,690)	0.30 ± 0.01 (n = 3,729)	0.50	< 0.0001
Pre-window single-shot AC rate	0.45 ± 0.00 (n = 5,952)	0.42 ± 0.01 (n = 1,894)	0.03	< 0.0001

Table 14: Cohort composition: Section 5.2 in-sample users versus dropped users. Chi-square test of independence $p < 0.0001$.

Cohort	In-sample %	Dropped %	In count	Drop count
pre_AI	24.2%	46.4%	1,616	1,731
transition	57.7%	38.4%	3,861	1,432
AI_era	18.1%	15.2%	1,213	566

The dropped users are not low-skill quitters; they are higher-skilled veterans. Pre-ChatGPT mean CF rating averages 1,642 for the dropped group versus 1,459 for the in-sample group (a 183-point gap), with much smaller post-shock rating gains (92 versus 300) consistent with mean reversion at higher rating levels. They are predominantly non-spanning (30% versus 79%), reflecting that many are users who reduced activity after ChatGPT rather than users who exited entirely. The pre-window single-shot AC rate is only marginally different across groups (0.42 dropped versus 0.45 in-sample), suggesting that dropout is not strongly correlated with pre-window AI signature. The cohort composition is sharply different: the pre_AI cohort is over-represented among dropped users (46.4% versus 24.2%), consistent with older veterans aging out of CF practice after ChatGPT.

The most consequential finding from the balance check is the asymmetric dropout between the two subpopulations. Roster-linked users (those with ICPC or IOI affiliation) drop out at the activity filter more often (43.3%) than non-rostered users (33.5%). This is the opposite direction from the

Table 15: Differential dropout at the activity filter, by subpopulation.

Subpopulation	N at Section 5.1 sample	N at Section 5.2 sample	Dropout %
roster-linked	2,480	1,407	43.3%
non-rostered	7,939	5,283	33.5%

naive hypothesis that roster-linked users are more committed and therefore more consistently active. The likely substantive cause is that ICPC and IOI contestants compete in time-bound events and tend to step back from open Codeforces practice after their qualifying year ends; non-rostered users, who chose CF as their primary competitive outlet, are more consistently active across years.

The implication for the asymmetry test: the roster-linked sample is the more-selected of the two subpopulations, and the null roster-linked slope ($\theta_{\text{RL}} = -2.41$ in Table 7) could be partially produced by differential censoring of high-signature roster-linked users rather than purely by the AI-prohibited-qualifier screening mechanism. The filter-sensitivity result in Table 11 above is consistent with this reading: when the PERMISSIVE filter brings additional low-density roster-linked users into the sample, θ_{RL} becomes significantly negative (-7.47, $p = 0.040$). The asymmetry test θ_1 should be interpreted as bounded by this differential censoring; the cross-environment opposite-signs pattern with the within-ICPC test (Section 5.3) remains the more robust evidence of the screening-filter mechanism.

Skill-control sensitivity

The headline specification includes the baseline-skill control $\bar{r}_h^{\text{pre}}$ to block omitted-variable bias from the negative AI-signature-skill correlation (Section 5.2). Dropping the control shifts the within-subpopulation slopes substantially: non-rostered moves from -12.24 to -5.15 ($p = 0.050$), roster-linked from -2.41 to 4.88 ($p = 0.29$). The pooled slope-difference θ_1 is essentially preserved at -10.04 (SE 5.29, $p = 0.058$). Because the AI-signature-skill correlation and the mean-reversion rate are similar in both subpopulations, the omitted-variable bias shifts within-subpopulation levels by similar amounts and largely cancels in the cross-subpopulation difference. The cross-subpopulation asymmetry test does not depend on the skill-control specification, even though the within-subpopulation levels do.

Roster-skill confound check

The cross-subpopulation asymmetry could in principle reflect a roster-versus-skill confound: roster-linked users are systematically higher-skill, having qualified for ICPC and IOI rosters through competitive selection, so a story in which higher-skill users have systematically different post-shock rating dynamics could in principle produce the asymmetry without any AI-specific channel. To

rule out the simplest version of this confound, we recompute the three AI-signature components on contest submissions instead of practice submissions and re-run the within-mode roster comparison. The bootstrap-validated contest-mode PCA loadings are within roughly 0.05 of the practice-mode loadings (both computed on the within-user-change panels) (0.653, 0.551, 0.520 against the practice-mode 0.602, 0.570, 0.560), so the composite construct is stable across modes.

If the practice-mode roster gap were entirely a roster-skill artifact, the contest-mode gap should attenuate with the skill control in the same direction the practice-mode gap does. The empirical pattern is the opposite. Without $\bar{r}_h^{\text{pre}}$, the practice-mode roster gap is -0.256 ($p < 0.001$, $N = 7{,}585$) and the contest-mode roster gap is -0.226 ($p < 0.001$, $N = 7{,}639$). Adding $\bar{r}_h^{\text{pre}}$ attenuates the practice-mode gap to -0.168 ($p < 0.001$, $N = 7{,}151$), as a roster-skill story would predict, but it *widens* the contest-mode gap to -0.262 ($p < 0.001$, $N = 7{,}191$). The roster-skill story predicts both gaps shrink together; the data show one shrinks and the other grows. The roster gap on the contest-mode signature, conditional on baseline skill, is therefore robust to the simplest version of the confound.

The contest-mode and practice-mode signatures are correlated with $r \approx 0.32$ at the user level (identical across roster-linked and non-rostered subpopulations to three decimal places). The two measures share roughly ten percent of variance: enough that the contest-mode test is not orthogonal to the practice-mode test, but not so much that contest mode is a tautological re-test. The contest-mode evidence is therefore partial corroboration of the skill-confound rule-out, not direct measurement of in-contest AI use. A non-rostered user can show high contest signature by actually using AI during CF contests (which is hard to police), by carrying learned reflexes from AI-style practice into contest mode, or by producing faster but lower-quality submissions because deliberative debugging skill has atrophied. The contest-mode test cannot separate these channels.

A.2 Robustness for the within-ICPC AI-intensity test

The within-ICPC AI-intensity test of Section 5.3 (Equation 7) is estimated on the 444-user, 579-observation sample described in Table 9. The headline $\psi_{\text{AI}} = 0.190$ ($p = 0.028$) depends on the activity filter that defines the qualifying practice window. This appendix documents three robustness exhibits: filter sensitivity, a balance check on filter dropout by cohort, and a Section 5.1 cross-check on the within-ICPC sub-sample.

Filter sensitivity

We re-estimate Equation 7 under three activity-filter thresholds for measuring within-user changes in the three behavioral proxies. The treatment is built per (user, contest year), matching the contest-year-conditional window logic of the main spec but with a faithfully per-(user, year) aggregation

(the manuscript headline instead collapses each user to a single per-user value). PCA loadings are fit on the headline-threshold sample and frozen across the moderate and least-strict thresholds so that the treatment is on a comparable scale across thresholds.

Table 16: Within-ICPC AI-intensity test under three activity-filter thresholds, with the treatment built per (user, contest year) and aggregated as within-user changes. AI_era × ai_intensity coefficient with team-clustered standard errors. PCA loadings are fit on the headline-threshold sample and frozen.

Filter	N_{obs}	N_{users}	ψ_{AI}	SE	p
LOOSE (≥ 1 quarter, ≥ 1 sub)	541	465	**0.434**	0.257	0.091
MEDIUM (≥ 2 quarters, ≥ 1 sub)	405	355	0.162	0.246	0.512
HEADLINE (≥ 2 quarters, ≥ 5 subs each)	386	336	0.108	0.213	0.611

Direction is positive across all three thresholds and grows as the filter relaxes (0.434 at the loosest, 0.108 at the headline threshold), so the Step-3 activity filter that drops 56% of ICPC participants is not manufacturing the positive sign. None of the relaxed cells reaches conventional significance given the smaller per-user-year samples (smallest $p = 0.091$). This table builds the treatment per user-year to vary the filter cleanly, whereas the manuscript headline collapses each user to a single per-user value. The per-user-year magnitudes here are therefore not directly comparable to the headline 0.190.

Balance check on filter dropout by cohort

Step 3 of Table 9 drops 579 of the 1,026 ICPC participants at Step 2, retaining 447. The first-difference construct is stricter than a post-window-levels filter because it requires both a pre-window (2021-Q3 to 2022-Q3) and a post-window (2023+) with the qualifying activity thresholds. Dropout rates differ substantially across cohorts:

Table 17: Within-ICPC sample dropout at Step 3 of Table 9 (`ai_intensity_pc1` defined under the within-user first-difference construct), by cohort.

Cohort	N at Step 2	N at Step 3	Dropped	Dropout %
pre_AI	431	144	287	66.6%
transition	412	233	179	43.4%
AI_era	183	70	113	61.7%

Transition users drop out the least (43.4%), pre_AI users the most (66.6%); AI_era (61.7%) sits between. The pattern is consistent with the first-difference construct's both-windows requirement: pre_AI users are more likely to have stopped active CF practice by 2023, AI_era users are more likely to lack pre-2022 practice history (their rated career started post-Copilot by construction), and

transition users span both periods most fully. Dropped users have higher pre-ChatGPT mean rating (1,932 vs 1,802, Welch $p < 0.0001$) and lower spanning rate (0.74 vs 0.85, $p < 0.0001$). ICPC score itself does not differ between retained and dropped users ($p = 0.73$), so the dropout is not directly outcome-correlated; pre-window single-shot AC rate is modestly higher among retained users (0.46 vs 0.42, $p < 0.001$). The cohort interaction in the headline is estimated on a sample whose AI_era cell is the second-most-attrited; whether this biases the AI_era × ai_intensity slope upward or downward depends on the within-cohort relationship between the within-user change in AI signature and dropout, which we do not directly observe.

Cross-check with the practice-pattern shift

If the within-ICPC sub-sample participates in the broader practice-pattern shift documented in Section 5.1, the AI_era × post DiD coefficients on the three behavioral proxies should survive on these users' user-quarter observations. We re-run Equation 4 on the 446 users that the reconstruction retains at Step 4 (within ±2 of the headline 444-user sample, an acceptable reconstruction tolerance).

Table 18: Equation 4 AI_era × post coefficient β_2 on the full Section 5.1 sample (10,419 users) versus the within-ICPC sub-sample (446 users). Cluster-robust (user-level) standard errors in parentheses.

Outcome	Full sample (10,419 users)	Within-ICPC sub-sample (446 users)
`frac_single_shot`	0.0586 (0.0063, $p < 0.001$)	0.0736 (0.0195, $p < 0.001$)
`median_attempts`	−0.3271 (0.0374, $p < 0.001$)	−0.2705 (0.1130, $p = 0.017$)
`frac_quick_retry`	−0.0190 (0.0041, $p < 0.001$)	−0.0193 (0.0134, $p = 0.15$)

The single-shot effect is *amplified* on the within-ICPC sub-sample to 126% of its full-panel magnitude (0.0736 vs 0.0586, $p < 0.001$). The median-attempts effect retains 83% of full (−0.2705 vs −0.3271, $p = 0.017$). The quick-retry effect has essentially the same magnitude as the full panel (−0.0193 vs −0.0190) but loses statistical significance on the smaller sample ($p = 0.15$). The amplified pattern is consistent with the within-ICPC sub-sample being more-active by construction (the first-difference construct requires qualifying practice in both the 2021-Q3 to 2022-Q3 pre-window and the 2023+ post-window), so the practice-pattern shift is more precisely estimated than on the full panel.

A.3 Copilot-anchored DiD

The Section 5.1 headline pins post_t to ChatGPT. A natural question is whether the practice-pattern shift is detectable around the first AI shock (Copilot, 2021-06-29) before ChatGPT was released. We re-run Equation 4 with post_t pinned to the Copilot date and drop the AI_era cohort, whose first rated contest is post-Copilot by construction so the cohort × post-Copilot interaction is unidentified for

them. The Copilot-isolation specification restricts the post window to 2021-Q3 through 2022-Q3, five quarters during which Copilot is the only AI shock (N = 76,035 observations across 8,586 users). The transition × post-Copilot coefficient on `frac_single_shot` is 0.0398 (SE 0.0053, $p < 0.001$), on `median_attempts` is −0.1768 (SE 0.0306, $p < 0.001$), and on `frac_quick_retry` is −0.0100 (SE 0.0033, $p = 0.002$), all three significant in the AI-prompt direction. The transition cohort's signature shift is therefore already present around the first AI shock, before ChatGPT, so the headline DiD is not a ChatGPT-specific artifact.

The Copilot-isolation quick-retry coefficient (−0.0100, $p = 0.002$) is also notable on its own: it is negative here (the AI-prompt direction), in contrast to the wrong-signed positive coefficient (0.0090, $p = 0.002$) for the transition cohort in the ChatGPT-anchored headline. This is consistent with the conjecture in Section 5.1: Copilot's integrated development environment (IDE) autocomplete-style integration may have reduced human retries cleanly, while ChatGPT's generate-and-debug workflow reintroduced human-debugging cycles among transition users. If anything the shift is more concentrated around Copilot than around ChatGPT for this mid-career cohort: the ChatGPT-anchored transition coefficient on `frac_single_shot` (0.0139 in Table 6) is smaller because its pre-period baseline already absorbs the post-Copilot shift. A second specification that extends the post window through 2025-Q4 (N = 127,124) recovers similar magnitudes on single-shot and median-attempts (0.0339 and −0.1569), with the quick-retry coefficient attenuating to null (−0.0027, SE 0.0028, $p = 0.333$) as the ChatGPT-era observations rejoin the post window.

A.4 COVID-19 disruption (2020–2021)

The empirical window straddles the COVID-19 pandemic. ICPC 2020 was postponed and held in 2021 with reduced attendance (117 teams in our rosters, against 134–138 in 2018–2019); IOI 2020 was held virtually (participation in our rosters did not fall); multiple national qualification pipelines were disrupted across 2020 and 2021. The CF practice panel is the most COVID-resistant of the three samples because it is continuous, while the contest panels run on an annual schedule; the cohort × post DiD on practice-pattern proxies (Section 5.1) absorbs year-level shocks through its quarter fixed effects, and within-user rating dynamics include 2020 and 2021 observations as part of the natural trajectory rather than as anomalies to be dummied out. The cross-subpopulation CF rating asymmetry (Section 5.2) is similarly COVID-resistant: the rating-change outcome Δr_h is constructed as a pre-ChatGPT-versus-post-ChatGPT difference where both windows span multiple years, washing out year-specific shocks. The within-ICPC AI-intensity test (Section 5.3) restricts to contest years ≥ 2023, which entirely post-dates COVID, so the COVID years cannot directly contaminate the AI-intensity coefficient. The quarterly cohort event study (Figure 2) does span 2020–2021. Its quarter fixed effects absorb the common COVID shock, and any cohort-differential

COVID response is one more reason Section 5.1 reads the cohort contrast as descriptive rather than causal.

A.5 ICPC/IOI–CF linkage and sensitivity to the excluded name-only matches

Candidate competitors from the ICPC and IOI rosters (2018–2025) are linked to CF accounts in two ways. ICPC contestants carry a CF username published on their Competitive Programming Hall of Fame profile, a direct link; the analysis ICPC roster is the 2,189 distinct accounts (2,784 contestant-year records, 88.0% of ICPC contestant-years) that carry such a username. IOI contestants are linked by a name-and-country match to the IOI participant lists, yielding 1,164 accounts. The two sets overlap in 236 accounts, for 3,117 roster-linked CF users.

The name-and-country matcher screens 4,219 candidate persons and returns 1,592 matched to at least one CF username (1,036 ICPC, 467 IOI, 89 both). Each match carries a best-status confidence tier (verified ≥ 0.85, likely 0.70–0.85, uncertain 0.55–0.70, rejected below 0.55), with counts 441 / 327 / 267 / 557 by person (441 / 338 / 330 / 1,106 across 2,215 match pairs). The matched set spans the competitive frontier: 15 World Champions and 86 gold, 114 silver, and 154 bronze medalists.

The 1,374 users matched to a roster only by a lower-confidence name-and-country match, for which we cannot confirm a contest record or attach a contest score, are excluded from the analysis panel (Section 4.2). The cross-subpopulation asymmetry of Section 5.2 is insensitive to their handling: excluding them (the headline) gives a non-rostered slope of -12.24 and a roster-linked null of -2.41 (pooled interaction -9.82, $p = 0.048$); leaving them in the non-rostered subpopulation gives -12.16 and -2.41 (interaction -9.75, $p = 0.048$); and assigning them to the roster-linked subpopulation gives -12.24 and -4.88 (interaction -7.36, $p = 0.110$). The within-subpopulation slopes are stable across all three handlings; only the pooled-interaction p-value softens when the AI-era-heavy name-only matches are forced into the roster-linked arm, as expected if unverifiable matches contaminate the screened arm. Full results are available from the author.

B Construct Validation of the AI-Prompt Signature

The cohort practice-pattern results (Section 5.1), the cross-subpopulation asymmetry test (Section 5.2), and the within-ICPC AI-intensity test (Section 5.3) rest on the AI-prompt signature: the behavioral composite of higher single-shot accept rate, lower median attempts-to-AC, and lower quick-retry rate that we attribute to an AI prompt-and-fix workflow. A skeptical reading is that this signature captures programmer skill or an analytic work style rather than AI use directly. This appendix bounds that concern with three tests on the Codeforces panel restricted to submissions where `author.participantType == "PRACTICE"` (the canonical practice filter on the combined re-scraped panel; each test below reports its own sample).

The tests address two distinct alternatives to the AI-practice interpretation. Subsection B.1 tests whether the signature is a proxy for skill, by examining its relationship to CF rating in the strictly pre-shock window (2018-Q1 to 2022-Q3). Subsection B.2 examines whether established practitioners exhibit a within-user discontinuity in the signature at AI release dates, which would distinguish a discrete AI-availability effect from a smooth skill-acquisition trend. Subsection B.3 synthesizes the two tests against the experimental causal anchor from Shen and Tamkin (2026).

B.1 Discriminant validity against pre-shock skill

If the signature were a tight proxy for CF rating, pre-2022 high-rated users would already exhibit the AI-prompt signature: higher single-shot accept rates, lower median attempts to AC, and lower quick-retry rates than lower-rated users. We test this directly using the pre-shock window 2018-Q1 through 2022-Q3 (strictly before ChatGPT, which released 2022-11-30).

Cross-sectional test. For each user with at least four active pre-shock quarters (n = 8,535), we compute the user-level mean of each signature component and the user-level mean pre-shock CF rating. Pearson and Spearman correlations between the signature components and mean rating are reported in Table 19.

Table 19: Pre-shock cross-sectional correlation: signature components vs. CF rating

Signature component	n users	Pearson r	SE	p	Spearman r
Single-shot AC rate	8,535	−0.091	0.011	< 0.001	−0.095
Median attempts to AC	8,535	+0.157	0.011	< 0.001	+0.175
Quick-retry rate	8,535	+0.147	0.011	< 0.001	+0.161

Notes. Pre-shock window: 2018-Q1 to 2022-Q3. User-level means; users included if active in ≥ 4 pre-shock quarters with ≥ 5 practice submissions per quarter. Filter: `author.participantType == "PRACTICE"` on the combined re-scraped panel.

All three correlations have signs *opposite* to the AI-prompt prediction. In the pre-shock window, higher-rated users had *lower* single-shot AC rates, *higher* median attempts to AC, and higher

quick-retry rates than lower-rated users. Skill in the pre-shock era was associated with attempting harder problems that required more iteration, not with the prompt-and-fix pattern. Magnitudes are weak ($|r| < 0.20$) but the directional contrast with the AI-prompt prediction is unambiguous.

Figure 3 plots the user-level scatter and binned mean for each component against pre-shock rating. The binned-mean lines make the directional contrast visible.

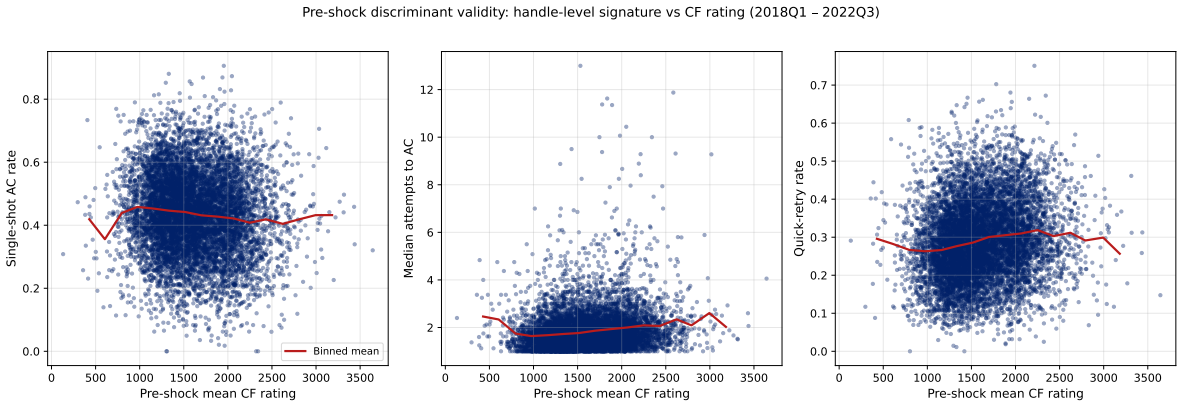


Figure 3: Pre-shock discriminant validity: user-level mean signature components vs. pre-shock mean CF rating (n = 8,535 users, 2018-Q1 to 2022-Q3). Red lines are binned means. The cross-sectional relationships have signs opposite to the AI-prompt prediction: higher pre-shock rating is associated with lower single-shot AC, higher median attempts, and higher quick-retry rates — inconsistent with a skill-proxy interpretation of the signature.

Within-user dynamics. The cross-sectional contrast could in principle mask a within-user skill-on-signature effect. We test this with the specification

$$\text{signature}_{ht} = \alpha_h + \gamma_t + \delta \cdot \frac{\text{rating}_{ht}}{100} + \varepsilon_{ht}, \tag{8}$$

where α_h is a user fixed effect, γ_t is a quarter fixed effect, and δ is the within-user elasticity per 100 Elo points. The sample is the pre-shock quarterly panel; standard errors are clustered at the user level. Table 20 reports the results.

Table 20: Pre-shock within-user elasticity: signature vs. rating

Signature component	n obs	δ per 100 Elo	SE	p
Single-shot AC rate	80,184	+0.00434	0.00045	< 0.001
Median attempts to AC	79,514	−0.03335	0.00367	< 0.001
Quick-retry rate	80,203	+0.00110	0.00036	0.002

Notes. Spec (8): user and quarter FE, cluster-robust SE at user. Pre-shock window: 2018-Q1 to 2022-Q3.

Within-user, a user's signature does shift modestly with its own rating, but the magnitudes are small. The single-shot AC elasticity is $\delta = 0.0043$ per 100 Elo, meaning that a user would need to

gain roughly 1,350 Elo points to produce a single-shot AC shift of magnitude comparable to the AI-era cohort × post effect of +0.0586 documented in Section 5.1. Median attempts has an elasticity of −0.033 per 100 Elo against the cohort effect of −0.327, roughly a 10:1 ratio of cohort-effect to skill-elasticity. Skill movements at the magnitudes plausibly achievable within the 2022-Q4 to 2025 post-shock window (typically tens to low hundreds of Elo for established users, not 1,000+) cannot reproduce the cohort effect.

Taken together, the cross-sectional and within-user tests establish that the signature is not a tight proxy for skill: the cross-sectional sign runs opposite to the AI-prompt prediction, and the within-user skill-on-signature elasticity is an order of magnitude too small to explain the cohort effect.

B.2 Within-user stability around AI release

The remaining alternative for the cohort effect is a smooth-trend story: the signature has been drifting toward the AI-prompt signature over time for reasons unrelated to AI release, and we mistakenly attribute the cohort contrast to AI exposure. A direct test is whether established users exhibit a discrete change in their signature at known AI release dates.

We estimate a within-user event study around the ChatGPT release (2022-11-30) on the spanning panel of users active both pre- and post-shock (n = 1,712 users, 25,395 user-month observations):

$$\text{signature}_{ht} = \alpha_h + \sum_{k=-18,\, k\notin\{-1,0\}}^{+18} \beta_k \cdot \mathbf{1}[\text{event}_{ht} = k] + \varepsilon_{ht}, \tag{9}$$

where event time is months from 2022-11. User FE absorbs user-level heterogeneity; the reference period is one month before release, and the release month itself ($k = 0$) is dropped from estimation. Standard errors are clustered at the user level. Figure 4 plots the coefficient time series.

Because the release month is dropped, the profile is read from the adjacent coefficients, and what movement exists after the release is directed *away* from the AI-prompt signature: single-shot acceptance falls and median attempts rise at $k = +1$, with quick retries rising as well. Established users show no shift *toward* the signature at the release. For the single-shot rate, coefficients on the pre-shock lead months are individually indistinguishable from zero; the other two components have scattered individually significant lead months (details available from the author).

The Copilot release (2021-06-29) yields a similar conclusion with additional pre-shock parallel-trends concerns (coefficients on the pre-shock lead months are scattered significantly positive for the single-shot rate). We treat the Copilot event study as inconclusive due to the parallel-trends violation and report it for completeness; full results are available from the author.

Cohort-stratified event studies on the ChatGPT shock confirm the absence of within-user adoption

among existing practitioners. Within the pre-AI cohort (n = 733 users, 8,791 user-month obs in the event window), post-release movement again runs away from the signature or is flat. Within the transition cohort (n = 775 users, 14,460 obs), the post-shock pattern matches the pre-AI cohort's direction, with no movement toward the signature at the release.

Interpretation. Established practitioners did not shift their practice patterns at ChatGPT release. Two implications follow. First, the smooth-trend alternative is bounded: if a smooth time trend were driving the cohort contrast, established users would also drift toward the AI-prompt signature over time, and they do not. Second, the cohort effect must originate primarily in cohort *composition*: users who entered the platform after AI's broad availability arrived with practice patterns that differ from pre-AI cohorts'. This is consistent with the paper's skill-formation framework: AI exposure during the formation period shapes practice habits; AI exposure to experienced practitioners does not change theirs.

B.3 Synthesis with experimental evidence

The two subsections' tests bound two distinct alternatives to the AI-practice interpretation. The discriminant test (Subsection B.1) rules out the signature as a proxy for *static* skill: the cross-sectional sign runs opposite to the AI-prompt prediction and the within-user skill-on-signature elasticity is an order of magnitude too small to explain the cohort effect. The event-study test (Subsection B.2) rules out adoption by established practitioners: there is no within-user drift in the signature around the AI releases, which locates the cohort effect in cohort composition rather than in existing-user adoption.

What these two tests establish is narrower than AI causation: the cross-cohort signature difference is a genuine, cohort-compositional shift in practice style, not an artifact of static skill or of established users adopting AI. They do not establish that the cohort difference is *caused* by AI. As Section 5.1 shows, the cohort gradient does not coincide with the AI releases (the cohort event study rejects parallel trends and shows no discontinuity), the AI-era and transition cohorts are indistinguishable at matched career stage, and the maturation-clean difficulty gradient is skill-like rather than AI-amenable. The signature is therefore best read as a practice-style proxy whose cross-cohort variation carries a secular and maturation component; the AI interpretation is the leading reading, not an identified effect.

Shen and Tamkin (2026)'s randomized experiment on 52 software developers provides the complementary causal anchor. Their RCT randomly assigns AI use during a learning task and finds that AI users score substantially lower on a subsequent non-AI-aided assessment (a 17% reduction, Cohen's d = 0.738, p = 0.010), with no significant gain in completion time; the deficit is concentrated among low-engagement, prompt-and-fix usage patterns rather than the cognitively engaged ones. That low-engagement, prompt-and-fix way of working with AI is one of the usage

patterns our signature tracks; because AI use is randomly assigned in their design, its link to weaker non-AI-aided performance is causally identified in that setting. Our signature is the multi-year, field-observational analog of their experimentally induced pattern.

Taken together, the field tests and the experimental anchor establish that the AI-prompt signature is a behavioral practice-style proxy consistent with AI-assisted practice: the discriminant and event-study tests rule out static skill and existing-user adoption, and Shen and Tamkin (2026)'s RCT shows that AI use produces this very pattern. What they do not establish is that the cross-cohort variation in the signature is AI-caused; the maturation-clean evidence in Section 5.1 attributes part of it to secular practice-norm change and novice maturation. We therefore use the signature as a proxy for AI-style practice throughout, and treat its AI interpretation as the leading but unverified reading rather than a measured fact.

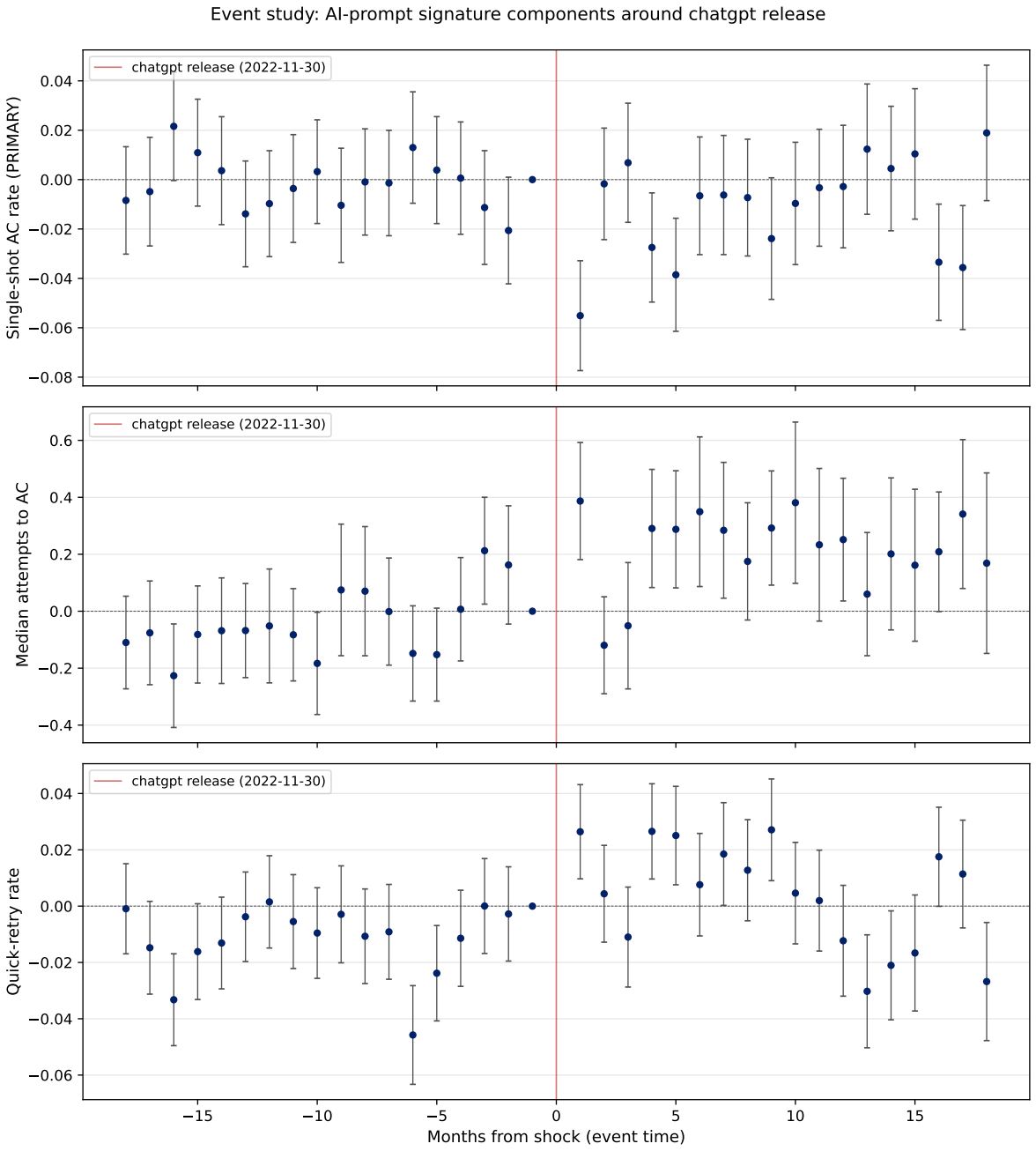


Figure 4: Within-user event study at the ChatGPT release. Spanning panel of 1,712 users active both pre- and post-shock. Reference period: $k = -1$ (one month before release, October 2022). Vertical red line marks release. The release month is dropped, so no $k = 0$ coefficient is estimated. The post-release movement that exists runs *away* from the AI-prompt signature (fewer single-shot accepts, more attempts); for the single-shot rate, coefficients on the pre-shock lead months are individually indistinguishable from zero.

C Data and code availability

CF rating and submission data: scraped from the CF API (real usernames; sample construction in Section 4.2). ICPC standings: scraped from cphof.org. IOI results: scraped from stats.ioinformatics.org. The analysis code and cleaned panels, together with the scoring-system specifications (CF Elo, ICPC team scoring, IOI individual scoring), are available from the author on request. A de-identified replication package will accompany the published article.