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

Song Yao*

October 18, 2026

Abstract

Generative AI can perform many of the tasks through which programmers build skill. When it substitutes for practice, skill development suffers; when it complements practice, learning may accelerate. Therefore, one programmer in the substitute mode and another in the complement mode can produce the same AI-assisted output yet differ in the skill they have acquired. We ask what digital practice records reveal about skill formation and whether AI-prohibited evaluations help distinguish the two modes of AI use. We follow 10,419 users of Codeforces (CF), an open, unproctored programming practice and contest platform with no enforced restriction on AI assistance, from 2018 to 2025. We link them to the rosters of the International Collegiate Programming Contest (ICPC) and the International Olympiad in Informatics (IOI), which admit entrants through qualification rounds and prohibit AI under in-person proctoring. From CF practice telemetry, we construct an AI-prompt signature (higher first-attempt acceptance rates, fewer attempts per solved problem, and fewer rapid debugging retries), consistent with AI-assisted practice. Three patterns emerge. Practice shifted toward the signature across entry cohorts spanning the releases of GitHub Copilot and ChatGPT. On CF, a stronger signature is associated with slower skill growth among users not linked to ICPC or IOI rosters, but not among roster-linked users. Among AI-era entrants who reached ICPC, larger within-user shifts toward the signature are positively associated with non-AI performance. These patterns are consistent with qualification screening on acquired skill, which admits users who complement practice with AI and excludes those who substitute AI for practice.

Keywords: generative AI; technology-mediated learning; delegation to AI; skill formation; human capital; expertise; digital trace data; screening; credentialing; substitutes and complements; competitive programming.

1 Introduction

Generative artificial intelligence (AI) now performs many of the tasks through which people build skill. In software development, professional education, and technical training, learners use AI in the very activities that

* 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, Dennis Zhang, and participants at the 13th Triennial Invitational Choice Symposium for insightful comments. AI coding assistants were used for analyses and drafting; I have verified all results against the underlying data. All remaining errors are my own. I hold 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.

once served as practice, and the records that organizations later evaluate, such as practice histories, portfolios, and assessment results, are produced with AI tools. Educators state the resulting problem plainly: a 2026 MIT committee on AI in teaching asks, "When AI makes it possible to offload the cognitive work of learning, how can we assess what students actually know and understand?" (MIT Ad Hoc Committee 2026). Two streams of information systems (IS) research speak to this problem: how technology mediates learning and what happens when people delegate work to intelligent systems. This paper joins the two streams to address this question. We ask what digital practice records reveal about skill built with AI and whether AI-prohibited evaluations can measure what those records hide.

Those records can hide skill because the same AI tool can be used in two modes during learning. In the first mode, the learner delegates the task to the AI: the tool produces the code, the proof, or the analysis, and the learner's own practice effort is replaced. In the second, the learner practices with the AI: the tool explains, hints, and checks, and the effort is sustained. We call them the *substitute* and *complement* modes. The distinction draws on IS research on delegation to agentic artifacts (Baird and Maruping 2021, Fügener et al. 2022) and on automation and augmentation as forms of task allocation (Fügener et al. 2026), but it is a distinction about learning rather than task performance: delegating a task can still achieve learning through review and critique, and what matters here is whether the practice effort survives. Both modes produce polished output while the tool is present. That is the scaffold of our title: it holds up the output either way, and only when it comes off, at the moments when a person must perform without AI, does the difference in accumulated skill appear.

The mechanism has a precedent in a technology that was not generative. When robotic surgery arrived, procedures came under the control of the expert surgeon at the console, trainees' participation became less necessary, the approved ways of learning surgery stopped working, and some trainees developed alternative learning practices of their own (Beane 2019). Generative AI moves the same displacement inside a single learner's practice session: the tool can do the task, and whether the learner still practices is a choice the output does not record.

We therefore ask two questions. First, in an open practice pool where both modes mix, how is a measure of AI-style practice associated with skill growth, and does that association differ between people who do and do not appear on the rosters of AI-prohibited evaluations? Second, among people on those rosters, how does a shift toward AI-style practice beforehand relate to their performance in the AI-prohibited evaluation?

These questions build on and contribute to the two IS research streams. The first is technology-mediated learning, the study of how information technology changes learning processes and outcomes (Alavi and Leidner 2001). It has shown that unguided access to a technology can lower learning outcomes (Belo et al. 2014) while teacher-guided use raises them (Deng et al. 2026), a finding our two modes build on. Like earlier studies in this stream (Singh et al. 2011, Zhang et al. 2019), we work from platform traces: multi-year practice histories with AI access, which we relate to non-AI-aided performance. The second stream studies delegation to AI and its consequences for the human, such as how well people delegate and how AI advice changes their judgments (Fügener et al. 2021, 2022, Jussupow et al. 2021). These studies measure AI-assisted performance and the impact of AI on human knowledge while the tool is present, in short experimental sessions. In contrast, we measure performance where the tool is available and where it is prohibited, after years of practice.

Performance without the tool matters most at the skill frontier, the level of skill that only the top performers in a field reach. If AI is always available and reliable, routine skill loss looks benign; the argument breaks where the tool is weakest and where it is absent. On tasks beyond its capabilities, AI makes errors (Dell'Acqua et al. 2026), and catching the errors requires human expertise. Future training data also depend on whether we keep pushing the knowledge frontier (Acemoglu et al. 2026). If substitute-mode use displaces the practice that builds this capability, the cost is invisible in AI-aided output but arrives years later. Such a cost has been measured in medicine: after AI-assisted colonoscopy was introduced, endoscopists' adenoma-detection rate in procedures without AI fell from 28.4% to 22.4% (Budzyń et al. 2025).

At the skill frontier, elite competitive programming contests let us answer our two questions with data on the same people. Codeforces (CF) is an online programming practice and contest platform open to all. It has no qualification screen and no proctoring, and AI use is unrestricted in practice, so learners in both modes share the pool. The International Collegiate Programming Contest (ICPC) and the International Olympiad in Informatics (IOI) prohibit AI at in-person, proctored finals and admit entrants through qualification rounds that also prohibit AI. Our panel follows 10,419 CF users from 2018 through 2025, across the GitHub Copilot launch in 2021 and the ChatGPT release in 2022, and links a subset of them to ICPC and IOI rosters and scores. While we do not observe AI use directly, from practice submissions we construct an *AI-prompt signature*: more problems solved on the first attempt, fewer attempts per solved problem, and fewer rapid debugging retries, the pattern a prompt-and-fix workflow with an AI assistant would leave (see Shihab et al. 2025). We relate the signature to two outcomes that measure skill. On CF, results in rated contests set each

user's rating, which measures skill relative to other users. Rating gains therefore track skill growth. At ICPC, the outcome is a team's score at the AI-prohibited finals.

Three patterns emerge. First, although not cleanly timed to the release of either GitHub Copilot or ChatGPT, a shift in practice toward the AI-prompt signature occurs across the entry cohorts spanning both AI tool releases. Second, among users in the open CF pool who are not linked to ICPC or IOI rosters, a one-standard-deviation stronger signature level after ChatGPT's release is associated with smaller rating gains of about eighteen points (roughly losing 6% of their average gain during that period). Among roster-linked users, the association is smaller and statistically insignificant. This difference between the two groups grows with each later entry cohort. It is near zero among pre-AI entrants and largest among AI-era entrants. Third, within AI-prohibited ICPC, a larger shift toward the signature before the contest is associated with higher team scores among AI-era entrants. This association is positive on its own ($p = 0.055$) and larger than among pre-AI entrants ($p = 0.028$ for the difference), and the difference stays positive across robustness checks. One reading reconciles both results. Where the two modes mix, a stronger signature tracks the substitute mode. Among users who passed AI-prohibited qualification rounds, by contrast, a strengthening signature tracks the complement mode.

The results carry an implication for evaluation design. Where AI-assisted output is less diagnostic of skill (Cowgill et al. 2026), an AI-prohibited evaluation shows what practice records cannot: performance without the tool. The results also bear on two decisions that training providers and employers face. The first is how to integrate AI into learners' practice. IS evidence favors designs that induce the learner's own effort (Santhanam et al. 2008) or give feedback rather than answers (Wambsganß et al. 2025). The second is where to place AI-prohibited evaluations in training and hiring. For training, the MIT committee suggests unrestricted AI use on assignments in courses whose primary assessment prohibits AI (MIT Ad Hoc Committee 2026). Our results are consistent with the same principle: where AI is allowed during practice, an AI-prohibited evaluation distinguishes the two modes of AI use.

Section 2 reviews the two IS research streams. Section 3 describes the setting and develops a framework with three predictions. Section 4 describes the data, Section 5 tests the predictions, and Section 6 discusses the contributions, implications, and limitations.

2 Theoretical background and related literature

This section reviews the two IS research streams that we build on and contribute to, followed by two related literatures. The first stream, technology-mediated learning (Section 2.1), supplies our outcome: the skill that forms through practice. The second, delegation to AI (Section 2.2), supplies the decision the learner faces: whether to delegate the task to AI or to practice with it. We add the setting in which the two modes can be compared: an AI-prohibited evaluation. Section 2.3 reviews the evidence on generative AI and productivity, which forms the backdrop to both streams. Section 2.4 reviews research on evaluation and screening under AI, which informs the implications of our research.

2.1 Technology-mediated learning and skill formation

Research on technology-mediated learning asks how technology shapes what people learn. We ask the same question about generative AI. Alavi and Leidner (2001) defined technology-mediated learning as an IS research area and called for deeper study of this question. In an early study of basic IT skills training, students in a web-based learning environment performed about as well as classroom students (Piccoli et al. 2001). A later study found that instruction encouraging self-regulated learning improved outcomes in e-learning-based IT training (Santhanam et al. 2008). Research on the IT workforce shows why IT skill matters: it is an investment (Bapna et al. 2013, Tambe et al. 2020) that loses value as technology changes (Bapna et al. 2013, Horton and Tambe 2025, Deming and Noray 2020). These studies take for granted that training and on-the-job practice build the skill, and ask what the skill is worth. We ask a more basic question: whether the skill forms at all when AI can do the practice for the learner.

Like this paper, many studies in this stream measure learning from the digital traces that work leaves. Archival records of software work show learning curves. Specialized experience in a system raises individual productivity the most (Boh et al. 2007), and task variety helps only up to a point (Narayanan et al. 2009). Six years of open-source project data reveal that developers learn most from peers (Singh et al. 2011), and a 50-month panel of software-design contests shows that participants improve more when they compete against a superstar (Zhang et al. 2019). Coding contests and their traces are thus a familiar setting in IS research, and our signature comes from the same kind of trace: CF practice submissions. We extend this work by using such traces to ask whether the way people practice when AI is available predicts their non-AI-aided performance.

Studies in this stream show that a technology can help or hurt learning, depending on how it is used. Our two modes build on this finding. Unstructured access hurts. Across Portuguese middle schools, high broadband use lowered national exam grades (Belo et al. 2014), and in two large randomized trials in Chinese classrooms, unstructured smartphone use reduced performance (Deng et al. 2026). Technology that supports the learner's own work helps. In the same trials, teacher-guided smartphone use raised performance (Deng et al. 2026). A system that gives dynamic feedback rather than answers improved students' skills more than fixed scripted support or adaptive support did (Wambsganß et al. 2025), and AI trainers for a strategic board game improved learning outcomes and narrowed a gender gap (Bao et al. 2026). Unlike generative AI, none of these technologies can do the practice for the learner. In our terms, AI in the substitute mode replaces the learner's practice effort, and AI in the complement mode sustains it.

Randomized experiments provide the causal evidence for the two modes: AI access and the design of the assistance change what learners can later do without AI. When the tool does the cognitive work, learning is shallower. Students who used ChatGPT for a research task reported about 35% less of the mental effort that builds understanding than students who used web search, and they produced worse work (Stadler et al. 2024). People who learned from large language model (LLM) summaries gained shallower knowledge than those who used web search, because the model did the synthesis for them (Melumad and Yun 2025). How much is lost depends on the design of the assistance and on how the learner engages with it. In high school mathematics, access to a standard ChatGPT interface raised practice grades by 48% but lowered grades on AI-free exams by 17% (Bastani et al. 2025). In the same study, an AI tutor that gave hints raised practice grades without that loss in exam grades. Among developers learning a new Python library, AI lowered quiz scores on average (Shen and Tamkin 2026). However, within the AI group, those who asked AI for explanations still learned the library, while those who delegated code generation or debugging scored far lower. Participants chose these patterns themselves, so this split is descriptive. In a three-month experiment with practicing patent lawyers, AI raised the quality of assisted drafting, but in a later task without AI only senior lawyers outperformed the control group (Autor et al. 2026). Junior lawyers showed no average gain, but compared with the control group their scores were more often either poor or good, which the authors read as AI accelerating skill acquisition among some junior lawyers and substituting for it among others. These experiments randomize AI access over sessions to months. We instead follow an elite population with career stakes through years of naturally occurring practice and link that practice to AI-prohibited evaluations.

The closest precedent to ours is chess. Training with chess computers raised players' Elo ratings, a measure of skill relative to other players, by roughly 11 points a year. The rating gains were mostly among lower-rated players, while the world-champion-level players' ratings stayed flat (Gaessler and Piezunka 2023). Chess computers work like our complement mode: the computer replaces the opponent rather than the player's own training, and the ban on computers in contests keeps a hard boundary between training and evaluation. A coding assistant, by contrast, can deliver working code without the user grasping the algorithm. In our setting, the AI-prohibited contests restore the boundary between training and evaluation.

What we add to this stream is field evidence from practice records at platform scale. In the open CF pool, we ask whether a user's AI-prompt signature is associated with rating gains after ChatGPT's release. At ICPC, we ask whether a user's shift toward the signature is associated with the score in the AI-prohibited contest.

2.2 Delegation to AI and its consequences for human capability

In this stream's terms, a learner who lets AI solve a practice problem delegates the task to AI. Baird and Maruping (2021) note that IS research has treated technology as a passive tool that people use. They argue that newer systems can take on tasks themselves, so people delegate tasks to them rather than simply using them. People often delegate poorly. AI judges what it knows and delegates to humans efficiently, whereas humans misjudge what they know and delegate knowledge work to AI poorly (Fügener et al. 2022). Fügener et al. (2026) model two roles AI can take in a task: automation, in which AI replaces the human, and augmentation, in which AI works with the human. They show when each role pays off. This work asks how to divide tasks between people and AI to perform well now. We ask what delegating the practice task does to the skill a learner builds.

The studies closest to our question ask what delegation does to the person who delegates. People who defer to AI scrutinize less and bring less of their own knowledge to the task. Advice from AI makes people's answers converge (Fügener et al. 2021). Each person becomes more accurate, but people lose some of what they know that the AI does not, so groups that work with AI are far less effective than groups without it. Incorrect AI advice led to more wrong diagnoses than no advice at all, among novice physicians and experienced radiologists alike, often because they monitored their own reasoning or the system poorly (Jussupow et al. 2021). Physicians reflected more critically when the system's output conflicted with their own assessment, and more shallowly when it agreed (Abdel-Karim et al. 2023). Recruiters given a high-quality AI tool disengaged, deferred to the model, and stopped improving (Dell'Acqua 2022). In preregistered experiments,

people adopted AI suggestions with little scrutiny and reported higher confidence even after the AI erred, a pattern Shaw and Nave (2026) call cognitive surrender. How people see AI also shapes their effort. In a survey of 1,943 U.S. workers, those who reported that AI complements their own effort worked harder after its introduction, whereas those who reported that AI substitutes for their effort did not (Kim et al. 2026).

Other studies find that the effect of AI depends on how it is used. Our two modes draw the same distinction. In ad writing, an LLM that drafted the copy lowered the quality of experts' ads, whereas an LLM that gave feedback on the writer's own draft improved nonexperts' ads (Chen and Chan 2024). In design work, generative AI raised the creativity of all users when used for brainstorming, but when used to execute the design it helped only novices and cost experts time without adding creativity (Hou et al. 2025). Wang and Lee (2026) observe the delegation choice directly in 1.8 million prompts: when users can see what their use costs, they shift from letting the model search for a solution to guiding it with detailed prompts.

We add a further outcome to this stream: the skill a person acquires. We follow learners through years of practice where they could delegate to AI and measure how they later perform in evaluations without AI access.

2.3 Backdrop: generative AI and productivity

We take one premise from the productivity evidence: with the tool present, generative AI raises output. ChatGPT raised productivity on professional writing tasks and narrowed the gap between workers (Noy and Zhang 2023). In a call center, a generative AI assistant raised productivity most among novice and lower-skilled agents (Brynjolfsson et al. 2025). On a task outside the capabilities of GPT-4, however, consultants using it were 19% less likely to produce a correct solution (Dell'Acqua et al. 2026). In an experiment, developers with GitHub Copilot completed a programming task 55.8% faster than developers without it (Peng et al. 2023). Using Copilot also raised participation and code contributions in open-source projects (Song et al. 2026). In three field experiments at large firms, an AI coding assistant raised the number of completed tasks, most for less experienced developers (Cui et al. 2026). AI also changes how programmers work. With Copilot, undergraduate computer science students moved from reading, understanding, and implementing code to prompting, viewing the response, and implementing it (Shihab et al. 2025). Our AI-prompt signature looks for the trace of this workflow. Access to ChatGPT led Stack Overflow users to give more answers, but shorter ones (Shan and Qiu 2026). These studies run over sessions to months and mostly measure output with the tool present. The call-center study is an exception: during outages of the AI assistant,

workers who had used AI still handled chats faster than before they had it (Brynjolfsson et al. 2025). We turn to multi-year practice histories and to performance where AI is prohibited by rule and by proctoring.

2.4 Evaluation, signals, and screening under AI

Our study's implications concern where to use AI-prohibited evaluations, and two findings from research on evaluation under AI bear on that question. The first is that AI changes what signals of skill reveal. In signaling and screening theory, credentials are costly actions or tests that separate more able candidates from less able ones (Spence 1973, Stiglitz 1975). Records of practice can serve as such signals: in open-source projects, rank earned through a merit-based system is associated with wages up to 18% higher (Hann et al. 2013). Such records can lose their value even without AI, such as through institutional change (Yao 2026). Generative AI may exacerbate this loss, because it makes impressive signals cheaper to produce. Using experiments on hiring and start-up investing, Cowgill et al. (2026) show that screening accuracy dropped by 4 to 9% due to applicants' access to ChatGPT. At the extreme, cheap AI-written applications destroy the signal (Galdin and Silbert 2025). By contrast, in a field experiment, nongenerative writing assistance raised jobseekers' hires and wages with no sign that employers were less satisfied (Wiles et al. 2025). Evaluation design can also help: when applicants can choose between an AI and a human evaluator, their choice restores the signal, because stronger and weaker applicants choose differently (Jabarian and Reshidi 2025). In short, an AI-prohibited evaluation as a screening design choice may restore a signal's value.

The second finding is that evaluation rules change how people work. After Stack Overflow banned AI-generated content, answers became longer and signaled more expertise, but participation fell (Borwankar et al. 2026). Creators who expected their AI use to be disclosed to a lay audience withdrew from the creative process and gave the AI more control, because they feared the audience would discount their own creativity (Jussupow et al. 2026). Knowledge workers who know their work leaves trace data, and care how the data represent them, change their work to shape the data (Aaltonen and Stelmaszak 2024). The AI-prohibited evaluations we study are rules of this kind. They may therefore work in two ways, by screening learners or by changing how learners practice, and Section 6.2 discusses both.

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. 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, running contests several times a week since 2010 and issuing individual Elo-style ratings, like chess, that measure each user's *relative* skill based on contest performance. Anyone can create an account and practice on the platform. Practice submissions do not enter the rating system and serve as a training sandbox. The platform has no gate, our term for an AI-prohibited evaluation that a competitor must pass to advance, such as the ICPC and IOI qualifying rounds described below. Nor is practice proctored, and nothing prevents a user from running an AI tool while solving problems, so both modes of AI use mix in the practice pool. Rated contests are unproctored, and until September 14, 2024, no rule restricted AI use during contests. On that date, Codeforces introduced a rule prohibiting AI tools during contests, but it cannot verify that contestants comply and the rule is not enforced.[1] AI may also be a weaker substitute during contests: 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).

[1] "Rule Restricting the Use of AI [revision 2024-09-14]," Codeforces blog entry 133941, `codeforces.com/blog/entry/133941`; the rule applies to registered and out-of-competition participants of rated rounds and states that AI tools may be used freely outside such contests.

ICPC and IOI sit on the other end of the institutional spectrum. Both run multi-stage qualification rounds before the finals, and both finals are in-person, proctored, and AI-prohibited. The published rules enforce the prohibition through a closed environment rather than an AI-specific clause: they bar personal computing devices and printed materials other than a reference document, and they cut contestants off from outside help, through communication rules or a network ban.[2] ICPC builds teams of three from a single university, with qualification rounds generally held in person under rules that restrict outside assistance; IOI selects individual contestants through national olympiads whose formats and proctoring vary, and its international finals are held in person (online only in 2020 and 2021, due to COVID-19). We do not observe proctoring intensity at the earlier stages and treat screening strength as unobserved. However, users whose CF performance leans heavily on AI are unlikely to pass these qualification rounds, so the populations that reach the ICPC and IOI finals have been screened of the most AI-dependent practitioners. Medals in these high-stakes contests are recruiting signals cited on university and job applications, so the non-AI-aided scores measure frontier skill under real stakes.

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, where the data are also dense. Even though we do not observe a verified measure of AI use, the signature proxies for how far a user's practice has shifted toward an AI-style prompt-and-fix workflow (Shihab et al. 2025). 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 introduce a simple framework that defines the two modes of AI use, formalizes why an AI-prohibited gate separates them, and yields the three testable predictions in Section 3.4.

[2] ICPC Europe Championship 2026 Official Rules (updated March 4, 2026), `euc.icpc.global`; IOI Competition Rules (IOI 2022), `ioi2022.id`; both accessed August 26, 2026. Neither document names AI tools; the prohibition operates through the device, materials, network, and communication rules summarized here, under in-person proctoring.

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

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

$$L_e > 0, \qquad L_{ea} > 0, \qquad L(0, a) = 0. \tag{2}$$

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 ($L(0, a) = 0$): AI can deliver a task's output in place of the learner's own effort, but that output is not skill.

What separates learners is whether they invest effort in practice. We call practice that pairs AI with genuine effort a *complement mode* and reliance on AI without effort a *substitute mode*: under the same AI access, both learners get the same help with output, but the complement-mode learner accrues $z = L(e, a)$ with $e > 0$, while the substitute-mode learner delegates and ends with $z = L(0, a) = 0$. Throughout, skill *erosion* refers to this shortfall: unaided skill that fails to form relative to the practice path, appearing as slower skill growth, not as decay of skill already formed. This gap is what our empirical strategy exploits. More AI use raises contemporaneous output in either mode, as AI route suggestions raise low-skilled taxi drivers' productivity without raising their skill (Kanazawa et al. 2026), so the AI-prompt signature (proxying the magnitude of a) is a footprint of AI use that can reflect either mode, not a direct measure of skill. The gate, which reveals z, is what distinguishes the two modes.

The practice effort e is costly at level κ; in the open, AI-permitted environment, leaning on AI saves κ, which is why it is attractive. 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 complement-mode learners with high z and turns away substitute-mode learners with low z. How a learner came to be one type or the other is a separate question, and two explanations 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 practice, while one who does not substitutes AI for effort. Our data are consistent with either explanation, so they identify the sorting but not how the types arise. The two explanations differ in policy implications (Section 6.2).

This framework maps onto the two empirical margins of the signature that our design uses: the *sorting margin*, the signature's level across users, and the *adaptation margin*, its within-user change. The sorting margin proxies the magnitude of a: it shows how far practice has shifted toward AI reliance, but not which mode lies behind it. Among users who have passed the gate, an increase in a raises z by raising the skill yield of practice ($L_{ea} > 0$). In the mixed pool, the same increase can equally be deepening substitution, so the framework predicts the direction of the link between the adaptation margin and skill *only* among the screened. This reading assumes that complement-mode learners keep practicing with effort after they qualify, which is plausible because the finals, not the qualifying rounds, are the prize.

3.4 Testable predictions

The framework of Section 3.3 yields three predictions, which we test 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. Because AI raises output in either mode, the shift appears for substitute- and complement-users alike and does not by itself separate the user types.

Prediction 2 (asymmetry in the unscreened pool). In the open CF pool, where both user types coexist, a stronger AI-prompt signature, measured as its post-ChatGPT level (the sorting margin), predicts smaller rating gains among unscreened (non-rostered) users, a group that still contains substitute-users, but not among roster-linked users, whom AI-prohibited gates have screened. The signature-to-rating slope is therefore more

negative in the unscreened subpopulation than in the screened one. Because the mechanism runs through modes of AI use, the asymmetry should also hold within the AI-era entry cohort alone, and the unscreened rating-gain shortfall should deepen across entry cohorts with growing AI exposure.

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 within-user shift toward AI-intensive practice (the adaptation margin) predicts *higher* non-AI-aided contest scores. 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. One screening mechanism predicts the direction of all three cells: a negative level gradient among the unscreened, a null one among the screened, and a positive change slope inside the AI-prohibited contest. A uniform mechanism does not explain all three. Pure erosion would put a negative gradient in both arms of the open pool and a negative change slope inside ICPC, while pure enhancement would produce no negative gradient anywhere.

All three predictions assume, rather than measure, that the signature is a proxy for AI use in either mode (Section 5.1 and Appendix B bound the leading alternative, that it proxies skill). Predictions 2 and 3 make this assumption testable: if the signature were not a proxy for AI use, nothing would produce the opposite signs on the two margins, the asymmetry within the AI-era entry cohort alone, or a shortfall that deepens across entry cohorts with AI exposure.

4 Data

4.1 Data sources and panel construction

Table 1 lists our three data sources.[3] The finals rosters of the two AI-prohibited contests, ICPC and IOI, are public. The contestants on these rosters anchor the CF panel. We build the CF panel from the CF API. We first link CF accounts to the ICPC and IOI rosters and collect the full rating and submission histories of these rostered contestants. We then add a non-rostered expansion.

[3]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. Appendix A.5 describes data provenance and research ethics.

Table 1: Data sources and roles.

Source	Records	Contents	Role
Competitive Programming Hall of Fame (`cphof.org`)	1,055 team-year records; 3,163 contestant-year records (2018–2025)	Team-level problems solved, ranks, and medals; team rosters	Outcome in an AI-prohibited contest; rosters for linking to CF accounts
IOI results pages (`stats.ioinformatics.org`)	2,779 contestant-year records (2018–2025)	Individual-level points, ranks, and medals; country delegation rosters	Outcome in an AI-prohibited contest; rosters for linking to CF accounts
CF API	10,419 users; 793,920 user-month records; 2018-01 to 2026-04	Rating histories, submission verdicts and timestamps, and profiles	The CF open pool panel; accounts for the roster linkage

ICPC and IOI results are collected from the yearly pages `cphof.org/standings/icpc/<year>` *and* `stats.ioinformatics.org/results/<year>`.

The analysis uses two unbalanced panels: the panel from the AI-prohibited ICPC and IOI contests, and the CF open pool panel. Appendix Table 20 reports the sample sizes, cohort composition, and summary statistics of CF ratings for both panels. It also reports how many CF accounts are linked to the ICPC and IOI rosters.

The AI-prohibited panel comprises the ICPC and IOI contestants whose first rated CF contest falls between 2015 and 2024. ICPC contestants are linked to CF accounts through the CF usernames published on their Competitive Programming Hall of Fame profiles; IOI contestants are linked by name-and-country matching. Appendix A.6 documents the linkage construction, its confidence tiers, and the full accounting from the linked accounts to the analysis panel.

The CF open pool panel contains 10,419 users: 2,480 roster-linked users and a non-rostered expansion of 7,939 users scraped directly from the CF API among active accounts with a maximum rating of at least 1,400 (the rating floor applies only to the non-rostered expansion). Of these, 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.[4]

Each user is assigned to a cohort by the date of their first rated CF contest. The split captures training vintage relative to the GitHub Copilot launch on June 29, 2021: pre_AI users (first rated 2015 through 2018) trained before AI, transition users (2019 through June 28, 2021) were partway through training when AI arrived, and AI_era users (June 29, 2021, through 2024) trained with AI available throughout. The cohort

[4]The exclusions documented in Appendix A.6 are based on linkage confidence, the date of a user's first rated contest, and whether the user has any practice-mode submissions. None of them is an outcome variable. The only rating-based criterion is the 1,400 floor on the non-rostered expansion. Imposing the same floor on roster-linked users strengthens the asymmetry of Section 5.2 (Appendix Table 13).

labels mark entry vintage relative to the two AI rollouts; they are not measures of individual AI exposure or use (Section 5.1 tests, and does not support, a causal AI reading of the differences across cohorts). Appendix Table 19 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 Appendix Table 20.

4.2 Outcome measurement: scoring system properties

The CF rating is a modified, self-calibrating Elo-rating system inspired by chess (Mirzayanov 2015): after each contest a competitor's rating moves according to how they ranked, with opponents' ratings entering the calculation directly, so the rating reflects field strength without additional controls. An inflation correction procedure also keeps the scale roughly stable across years. The rating is relative, measuring a contestant's standing in the pool rather than absolute skill, so it does not move when the whole pool improves by the same amount; our open-pool analysis (Section 5.2) therefore concerns relative performance, whether a stronger AI-prompt signature predicts losing ground against one's peers.

In contrast, ICPC and IOI provide no self-calibrating relative rating comparable to CF's: their relative measures (rank, medal) depend on the strength of the then-current contestant pool. We therefore use the absolute scores as outcomes, problems solved at the ICPC finals and total points earned at the IOI, which reflect a competitor's own performance on a fixed problem set. We then use contest-year fixed effects to absorb score variation caused by year-to-year changes in problem-set difficulty.

4.3 Descriptive statistics

Table 2 reports descriptive statistics for the variables of the three analyses of Section 5, one panel each. In Panel A, the cohort differences in pre-to-post changes preview the practice-shift analysis of Section 5.1, which also defines the signature. In Panel B, the signature varies widely within each subpopulation, which is the variation the rating analysis of Section 5.2 hinges on, and the two subpopulations differ sharply in baseline rating, which is why the analyses control for it. Appendix Figure 1 complements Panel A by tracing the three behaviors by cohort over the full window; its wide confidence intervals obscure any clear change at the AI releases, so the formal analysis of the shift is left to Section 5.1.

Table 2: 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 difference-in-differences (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-prompt 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 $\bar{r}_h^{\text{pre}}$	1,377 (381)	1,766 (451)	1,459 (427)

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

	pre_AI	transition	AI_era
Signature shift (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 $\bar{r}_h^{\text{pre}}$	2,268 (372)	1,946 (371)	1,338 (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 Section 4.1. Panel A cells are unadjusted cohort × period means over user-quarters; the estimated DiD (Table 3) 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 3. Panel B reports the cross-subpopulation sample of Section 5.2; "AI-prompt 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; "signature shift" is the within-user first-difference principal component* `shift_pc1`*. The IOI sister sample (Section 5.3; 345 obs, 243 contestants) has cohort signature-shift 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 tests Predictions 1–3 of Section 3.4 in turn. Section 5.1 documents a cross-cohort shift in CF practice toward the AI-prompt signature. Section 5.2 tests whether the signature predicts weaker rating gains in the open CF pool but not among roster-linked users. Section 5.3 tests whether, inside the AI-prohibited ICPC, a shift toward the signature predicts higher scores for AI_era entrants.

All three analyses share the cohort structure of Section 4.1, with the pre_AI cohort as the reference.

The three components of the signature (Section 5.1) enter the later analyses through two composites, each a first principal component: `ai_pc1_post`, the post-ChatGPT *level* of the signature, compares users where

the modes mix (Section 5.2), and `shift_pc1`, the pre-to-post *change* in a user's signature, tracks users' own shifts among the screened (Section 5.3). The level and the change are the sorting and adaptation margins of the framework (Section 3.3).

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 (summary statistics in Table 2, Panel A): `frac_single_shot`, the fraction of problems solved on the first attempt; `median_attempts`, the median number of attempts to solve a problem; and `frac_quick_retry`, the fraction of quick retries after a failed submission, where a quick retry is 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 can instead submit a generated solution that often clears the problem on the first try, consistent with the workflow shift documented by Shihab et al. (2025). If users rely on AI during practice, all three patterns should move in the AI-prompt direction. The fall in quick retries is the least obvious of the three patterns. It has two causes: more first submissions succeed and leave nothing to retry, and when an AI-generated solution fails, the user often cannot quickly debug code they did not write.

Throughout, this signature is a behavioral proxy inferred from submission timing, 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, that the signature merely proxies programmer skill. Before AI was available, higher-rated (more skilled) users were 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. As a user's own rating rises, the signature moves an order of magnitude too little for skill gains alone to produce the cohort shift, and established users did not change their practice patterns at the ChatGPT release, so the cohort shift reflects who entered the platform rather than a platform-wide trend.

To assess shifts in the three behavioral patterns after AI releases, we consider a difference-in-differences (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 (ChatGPT's release quarter, 2022-Q4, is dropped from the panel), α_h and τ_t are user and quarter fixed effects, respectively. $n_{h,t}$ is the user's submission count in the quarter. We control for it because practice volume differs across cohorts and over time and all three patterns move with it (Table 3). 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.

The sample is the 10,419-user CF panel of Section 4.1, restricted to user-quarter observations on practice-mode submissions from 2018-Q1 through 2025-Q4. Table 3 reports the observation counts by outcome.

The two AI events anchor the design in different ways: the Copilot launch (2021-06-29) sets the cohort split (see Section 4.1), and the ChatGPT release (2022-11-30) sets the within-user pre/post split. Anchoring both at Copilot would leave AI_era users with essentially no pre-period: they enter rated competition after Copilot by construction, conflating their novice-to-competitor transition with AI exposure. Anchoring the time series at ChatGPT instead gives AI_era users who entered before ChatGPT a clean pre-period of up to five quarters (2021-Q3 through 2022-Q3), observed after they entered rated competition. Re-anchoring the split at the Copilot launch, which is identifiable only for the pre_AI and transition cohorts, finds the transition cohort already shifted on all three proxies (Appendix A.3), so the headline does not depend on the ChatGPT anchor.

The identification of β_1 and β_2 comes from each user's own change in a practice pattern from before to after the ChatGPT release, averaged within each cohort and then compared across cohorts. User fixed effects absorb time-invariant user characteristics (e.g., skill baseline, education, country), and quarter fixed effects absorb calendar-time shocks common to all users, such as platform policy changes and seasonal effects. Because cohort assignment reflects platform-entry timing rather than observed AI use, the specification does not identify a causal effect of AI; we read the coefficients as a descriptive cohort contrast.

Results. The AI_era cohort × post interaction (β_2) is significant on all three outcomes (Table 3). Each β_2 is a DiD term: the AI_era cohort's pre-to-post change in a specific outcome measured against the pre_AI cohort. All three changes 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 $p < 0.001$.

Table 3: 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. Without the* $\log(1 + n_{h,t})$ *control, the AI_era coefficient on* `frac_quick_retry` *is* +0.018*; the other two coefficients keep their signs. **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 versus total submissions). Each regression uses the natural sample for its outcome.*

The transition cohort lies between the other two, closer to the pre_AI cohort: on single-shot submissions and median attempts, β_1 is roughly a quarter of the corresponding AI_era effect (Table 3), so the gradient from pre_AI to transition to AI_era is monotonic, consistent with the signature strengthening with how much of a user's training occurred after AI arrived. For quick retries, the transition coefficient runs the opposite way (0.0090, SE 0.0029, $p = 0.002$), possibly because users who adopted AI mid-career kept debugging its output by hand. The shift appears in both subpopulations: re-estimating the `frac_single_shot` DiD separately, the AI_era coefficient is 0.0356 ($p = 0.004$) among roster-linked users and 0.0617 ($p < 0.001$) among non-rostered users.

The cohort contrast is not cleanly timed to AI. Three patterns support the descriptive rather than causal reading of the cohort gradient. First, parallel pre-trends fail: quarter-by-cohort interactions (Appendix Figure 2) show neither cohort moving discontinuously at the ChatGPT release, and for the transition cohort a joint test rejects parallel pre-trends on all three patterns ($\chi^2(18) = 96.7$, 76.8, and 84.6; all $p < 0.001$). Second, at matched career stage the transition and AI_era cohorts are statistically indistinguishable (single-shot difference 0.0002, $p = 0.95$), even though the transition cohort entered before either AI release. Third, the gap is largest on the problems AI handles worst, not on the easy, templated problems where AI is likeliest to supply

a correct first attempt: holding career stage fixed, the AI_era-versus-pre_AI single-shot difference is null on the easiest problems (+0.002, $p = 0.63$) and peaks at +0.039 ($p < 0.001$) on the 2,000–2,399 difficulty band. A gradual trend in practice norms unrelated to AI, together with novice-to-competitor maturation, could fit these facts as well as an AI shock, so we carry the practice shift forward as a descriptive pattern that AI alone does not explain (full specifications in the online replication package).

The shift is descriptive, but it is the premise for what follows: the AI-prompt signature is now a common feature of practice in this population. Practice patterns alone do not show whether the shift erodes or builds skill. Sections 5.2 and 5.3 ask that question in the CF open pool and inside the AI-prohibited ICPC.

5.2 A cross-subpopulation CF rating asymmetry

We assess CF rating changes under the AI-prompt shift discussed above, comparing two subpopulations: roster-linked users who participate in AI-prohibited ICPC and IOI contests, and non-rostered users who do not (we use subpopulations and arms interchangeably). Prediction 2 implies the asymmetry to test: the qualifiers screen substitute-users out of the roster-linked pool, so the AI-prompt signature should predict weaker rating gains only in the non-rostered pool, where substitute-users remain.

We implement this analysis in two 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 user h's mean CF rating after ChatGPT minus their mean rating before, computed from the monthly rating panel as $\bar{r}_h^{\text{post}} - \bar{r}_h^{\text{pre}}$. ChatGPT was released on 2022-11-30, so the post-ChatGPT window covers months $t \in$ [2022-12, 2025-12] and the pre window covers $t \in$ [2018-01, 2022-12). 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 the signature predicts the subpopulation's rating change. The covariate `ai_pc1_post` is the user's post-ChatGPT AI-prompt signature score: the first principal component of the three sign-aligned proxies of Section 5.1, oriented so that all three contribute positively to the AI-prompt direction (bootstrap-validated principal component

analysis (PCA) loadings 0.614, 0.535, and 0.580; the three proxies are only moderately correlated, with pairwise r between 0.45 and 0.68).

- 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).

Within a subpopulation, θ_s (Equation 5) compares users who face the same gate (or no gate). Across subpopulations, θ_1 (Equation 6) holds the signature fixed and varies the screening environment. The asymmetry test does not require roster status to be randomly assigned: it relies on the maintained assumption that the qualifying rounds have screened substitute-users out of the roster-linked pool, which the within-ICPC test of Section 5.3 probes. The baseline control $\bar{r}_h^{\mathrm{pre}}$ blocks omitted-variable bias, because high-signature users are on average lower-skill and $\bar{r}_h^{\mathrm{pre}}$ independently predicts Δr_h (coefficients near -0.33 in both arms, $p < 0.0001$). The specification still does not identify the *causal* effect of AI on rating change, since high-signature users may differ on unobservables.

Results. The within-subpopulation slopes on `ai_pc1_post` are starkly asymmetric (Table 4). The non-rostered slope is -12.24 rating points (SE 2.45, $p < 0.0001$). 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, $p = 0.048$): the AI-prompt signature predicts around ten points weaker rating gains per signature unit in the unscreened pool than in the screened roster-linked pool. One standard deviation of `ai_pc1_post` is 1.48 on the estimation sample. Accordingly, a one-SD-stronger AI-prompt signature is associated with an 18-point lower rating gain among non-rostered users (-12.24×1.48), roughly 6% of the 300-point mean post-ChatGPT

gain across all users (Panel B of Table 2), and with a 14.5-point lower rating gain in the unscreened pool relative to the screened pool (-9.82×1.48).

Table 4: 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 specification tests whether $\theta_{\text{non_rostered}}$ and θ_{rostered} differ. `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 specification (Equation 6)				
`ai_pc1_post×non_rostered` (θ_1)	**−9.82**	4.98	0.048	$[-19.58, -0.07]$

The asymmetry runs in the direction the screening mechanism predicts: weaker rating gains with the signature in the non-rostered pool, where substitute-users still appear, and no correlation in the roster-linked pool, consistent with the qualifiers having screened 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 test of Section 5.3 addresses this objection, since in-contest AI use is blocked there.

Where the asymmetry lives: within-cohort estimates. The screening mechanism carries a timing implication: pre-AI qualification rounds from 2018 through 2021 (the Copilot release year) could not screen on AI-era practice, so if the gate separates the modes, the asymmetry should concentrate among AI_era entrants and vanish among pre_AI entrants. Re-estimating Equations 5 and 6 within each entry cohort (Table 5) shows this pattern. The pre_AI cohort is a clean placebo: both subpopulation slopes are null and the interaction is −1.17 (p = 0.86). The non-rostered slope grows more negative with AI-era exposure, from −3.83 (pre_AI) to −10.71 (transition) to −23.14 (AI_era), while the roster-linked slope stays null in

Table 5: Within-cohort cross-subpopulation asymmetry. Each row re-estimates the Section 5.2 specifications inside one entry cohort: per-subpopulation slopes of Δr_h on `ai_pc1_post` (Equation 5) and the pooled interaction θ_1 (Equation 6), with the baseline-rating control and HC1 standard errors. The diagnostic pipeline reproduces the pooled headline at $\theta_1 = -9.88$ ($p = 0.049$), matching Table 4 up to a small cross-pipeline discrepancy documented in the replication materials.

Cohort	Roster-linked slope	Non-rostered slope	θ_1 (interaction)
pre_AI (n = 1,616)	−2.66 (SE 5.76)	−3.83 (SE 3.54)	−1.17 (SE 6.75, $p = 0.862$)
transition (n = 3,861)	−0.23 (SE 5.14)	−10.71 (SE 3.05)	−10.48 (SE 5.97, $p = 0.079$)
AI_era (n = 1,213)	+13.72 (SE 17.42)	−23.14 (SE 5.92)	−36.86 (SE 18.35, $p = 0.044$)

every cohort, and the AI_era-only interaction is −36.86 ($p = 0.044$). Because that comparison is among same-vintage AI_era entrants, it also rules out the explanation that the roster null merely reflects veterans whose already-formed skill is insensitive to post-2022 practice style, since that explanation predicts no asymmetry among AI_era entrants. The AI_era interaction rests on a 260-user roster cell and is only marginally significant, so we read the joint pattern, not any single cohort's p-value, as the evidence of screening. The AI_era contrast survives practice-volume controls, though less precisely (−27.2, $p = 0.14$), and sharpens when the quick-retry component is dropped from the composite (−52.0, $p = 0.024$), although the pooled all-cohorts θ_1 then attenuates to a null (Appendix A.1).

Selection robustness and bounds. The post-ChatGPT activity requirement (≥ 2 quarters with ≥ 5 practice submissions each), which `ai_pc1_post` needs to be measured reliably at the user level, cuts the sample from 10,419 users to 6,690. The practice-pattern shift of Section 5.1 survives on this sub-sample with 8 to 13% attenuation, the interaction θ_1 stays between −11.92 and −7.14 under looser filters, bracketing the headline −9.82, and imposing the non-rostered expansion's rating floor of 1,400 on both arms strengthens the asymmetry ($\theta_1 = -12.17$, $p = 0.013$). The asymmetry is also robust to selection that mimics the observables: 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 it to zero, and the asymmetry holds under coarsened exact matching ($\theta_1 = -9.73$, $p = 0.050$). It is not robust to the differential attrition. Roster-linked users drop out of the activity filter more often than non-rostered users (43.3% versus 33.5%), and under the most permissive filter the roster-linked slope itself turns significantly negative (−7.47, $p = 0.040$), so the roster-arm null is partly a property of which roster-linked users survive the filter. The Lee bound (Lee 2009) brackets θ_1 in $[-11.95, -0.35]$, and a confounder explaining 2.4% of the residual variance in treatment and outcome would zero θ_1, against the

15.7% that baseline rating explains (Cinelli and Hazlett 2020). We therefore read the asymmetry as suggestive and attrition-bounded (Appendix A.1 and Appendix Table 7). The mechanism evidence rests on it together with the within-ICPC test of Section 5.3, whose direction is stable across every robustness check.

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

The third set of analyses asks whether the shift toward the signature during CF practice predicts AI-prohibited ICPC performance differently across cohorts. We estimate a pooled cross-sectional regression with university fixed effects and team-clustered standard errors:

$$\begin{aligned} s_{h,y} = \alpha + \psi_0 \cdot \text{shift_pc1}_h + \sum_{c \in \{\text{trans,AI}\}} \psi_c \cdot \text{shift_pc1}_h \cdot \mathbb{1}[\text{cohort}_h = c] \\ + \gamma \cdot \bar{r}_h^{\text{pre}} + \delta_y + \nu_u + \varepsilon_{h,y}, \end{aligned} \tag{7}$$

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. The treatment `shift_pc1` measures each user's within-user shift toward AI-style CF practice.[5] For each user, we measure the three AI-prompt proxies of Section 5.1 over a pre window (2021-Q3 through 2022-Q3, the five quarters between the Copilot and ChatGPT releases) and over a contest-year-conditional post window (2023-Q1 through the quarter immediately preceding the contest). The user-level change on each proxy is its post-window mean minus its pre-window mean, and `shift_pc1` is the first principal component of the three sign-aligned changes. The loadings are fit on a broad CF reference group and then frozen, so that the axis is not defined by the sample on which it is tested (Appendix A.2 reports them and defines the group). We assign each user a single value, anchored at their first qualifying year, so that the treatment is measured before any contest outcome could feed back into practice and does not drift with the post-window length. 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 4.1, with pre_AI as the reference. Contest-year fixed effects δ_y absorb year-specific score-distribution variation (different problem sets, different

[5] The variable is called `ai_intensity_pc1` in the online replication files.

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 address two issues. The first is selection into university: top universities recruit more homogeneous high-skill students, so their teams may both score higher and include more AI_era members with large shifts toward the signature. With the fixed effects, the AI_era × `shift_pc1` coefficient is identified from within-university variation only. The second is the shared team score: residuals are mechanically correlated within team, and team-level clustering captures that correlation. University-level clustering ($n = 196$) leaves the headline marginal ($p = 0.068$) with the point estimate unchanged, and dropping the university fixed effects gives a larger coefficient (0.323, $p = 0.040$).

Appendix Table 14 reports the filter chain that produces the sample: the intersection of the three cohorts and the ICPC roster, restricted to contest years on or after 2023 where the shift toward the signature is meaningfully defined.

The coefficient of interest, ψ_{AI} on the interaction between `shift_pc1` and the AI_era cohort indicator, measures how the relationship between the shift and the score differs between the AI_era and pre_AI cohorts within a university and a contest year, holding baseline skill fixed. 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. The estimate is observational, since AI_era users who shifted more may differ on unobservables that baseline skill and the fixed effects cannot rule out.

Results. The AI_era cohort interaction is positive and significant under the baseline team clustering: $\psi_{\text{AI}} = 0.190$ (SE 0.087, $p = 0.028$; Table 6). The pre_AI placebo slope is small and null ($\psi_0 = -0.048$, $p = 0.303$), and so is the transition interaction ($\psi_{\text{trans}} = 0.060$, $p = 0.373$). The AI_era cohort's own slope is +0.142 (SE 0.074, $p = 0.055$, 95% CI $[-0.003, 0.287]$): of the predicted sign, though only marginally significant on its own. The design's test is the cohort contrast with the placebo, which also completes the two-margin comparison with Section 5.2: 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 scores in 2023–2025, controlling for baseline skill and university. Taking the interaction at face value, a one-standard-deviation larger shift (0.825 PC1 units; Table 2, Panel C) corresponds to about 0.16 more problems solved, roughly 3% of the AI_era mean score of 4.8. The effect is modest; its evidentiary weight lies in the sign and in the contrast with the placebo, not in the size.

Table 6: Within-ICPC signature shift 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 slope on `shift_pc1` (pre_AI reference slope plus the AI_era interaction) is +0.142 (SE 0.074, $p = 0.055$).

Term	ψ	SE	p	95% CI
`shift_pc1` (pre_AI reference)	−0.048	0.047	0.303	[−0.140, 0.043]
`shift_pc1`:transition	0.060	0.068	0.373	[−0.072, 0.193]
`shift_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 adaptation-margin counterpart to the sorting-margin gradient in the unscreened CF pool (Section 5.2): under the screening reading, the qualifier has already removed the substitute-users, so the within-user shift among the qualified reflects complement-style adaptation that raises score. The framework's remaining cell, the within-user change where the modes mix, carries no predicted sign, and empirically it is a null (cross-arm interaction +0.3, $p = 0.95$; Appendix A.1).

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. Section 5.2 relates rating growth to the signature's level, the sorting margin, which shows how far practice has shifted toward AI reliance but not which mode lies behind it. The complementarity $L_{ea} > 0$ concerns an increase in a for a given learner, which is the adaptation margin that this section measures. Prediction 2 called for a null there, since screening removes the erosion signal and rating is a relative measure that complement-style gains need not move among elite peers. And 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; Table 6). Appendix A.2 reconciles the two estimates directly on the same 444 users.

Construct and selection robustness. We read the within-ICPC evidence as establishing the *sign* of the effect rather than its precise size, and three checks bear on that sign. First, the AI_era change could partly capture novice-to-competitor maturation, since AI_era contestants are the novices in the 2023–2025 sample (mean tenure 3.9 years since their first rated contest, against 7.4 for the pre_AI cohort). Career-stage controls (tenure, tenure interacted with cohort, and total rated contests) leave the AI_era interaction at 0.195 to 0.196 ($p \leq 0.027$), and it survives even conditioning on the user's rating growth over the span (0.190, $p = 0.038$), an

over-control since rating growth is partly an outcome of AI use. Second, the filter chain (Appendix Table 14) drops 56% of ICPC participants for low activity and 647 users with pre-2023 ICPC history but no recent participation, an exit that could itself remove substitute-users. Retained and dropped participants do not differ in ICPC score ($p = 0.73$), but the direction of any bias cannot be determined from these data. Third, as the activity thresholds loosen, the interaction stays positive but loses significance (+0.148 to +0.084, with p rising to 0.307), consistent with noisier measurement. Further checks leave the sign unchanged. Dropping one team at a time, or one university at a time, keeps the interaction positive, and the largest p-value across these re-estimates is 0.067. Other variance estimators move p from 0.04 (clustering by user) to 0.18 (CRV3, a conservative cluster-robust estimator at the team level), against 0.028 under the baseline team clustering. Appendix A.2 reports these checks, along with controls for teammate composition and a composite without the quick-retry component.

IOI sister analysis. The IOI counterpart to the within-ICPC test runs the same specification on a smaller, younger pool (345 contestant-years from 243 CF-linked contestants). The AI_era interaction on IOI score is directionally consistent but noisy (18.1, SE 18.0, $p = 0.31$); we anchor the non-AI-aided performance headline on ICPC and treat IOI as a directional check (Appendix A.2).

Design-wide checks. Two checks concern the design as a whole. The COVID-19 disruption falls just before Copilot: quarter fixed effects absorb common pandemic shocks, the rating-change windows average over multiple years on each side of ChatGPT, and the within-ICPC contest years post-date the disruption (Appendix A.4). The within-subpopulation slopes are stable under all three handlings of the 1,374 lower-confidence name-only matches (Appendix A.6). Appendix A collects the full battery of checks.

6 Discussion

6.1 Contributions

First, we add platform-scale field evidence to research on technology-mediated learning: we relate the AI-prompt signature, computed from multi-year practice records, to performance both where AI is allowed and where it is prohibited. The closest studies in this stream are set inside organizations: in hospital emergency departments, adopting an algorithmic recommendation tool was associated with flatter learning curves

(Sundaresan and Güler 2025), and in an accounting firm, reliance on automated software eroded accountants' expertise in tasks for which they remained accountable (Rinta-Kahila et al. 2023). Our evidence is at the individual level, spans years, and measures performance where the tool is prohibited. At the skill frontier, it also gives a field counterpart to the macroeconomic human-capital depreciation mechanism of Acemoglu et al. (2026) and to the experimental evidence of Bastani et al. (2025), Stadler et al. (2024), and Shen and Tamkin (2026). Our two-margin result shares its structure with the model of Bondi and Johnson (2026), in which the same AI input reads differently over time depending on the design that measures it. We find the structure across people instead, because an AI-prohibited gate sorts complement-users from substitute-users.

Second, we add field evidence to research on delegation to AI: the substitute-versus-complement distinction is real, since over multiple years the two modes of AI use leave measurable traces in submission behavior and accumulate into divergent unaided performance. Experiments support the premise that the modes are indistinguishable in AI-aided output: in the patent-drafting experiment of Autor et al. (2026), treated junior lawyers matched seniors while assisted but showed no gain unaided, and the authors leave open which juniors gained durable skill. The AI-prompt signature, a practice-style marker measured from submission records, is one answer to that question. The closest findings on delegation to AI concern performance with AI in use: working with AI raises accuracy but reduces the diversity in people's knowledge (Fügener et al. 2021), and its benefits vary across users (Knight et al. 2026, Hou et al. 2025). We instead ask what capability remains after AI-assisted practice, measured in AI-prohibited environments.

For research on evaluation under AI, the implication is that screening decides where each pattern can be observed. AI-prohibited contests screen substitute-users out of the elite pool. On AI-allowed platforms, where substitute-users remain, the signature marks weaker rating gains. AI-prohibited contests show only the complement-users' pattern: among AI-era entrants, higher unaided scores alongside a larger shift toward the signature. Recognizing the screening reconciles two findings that would otherwise look contradictory. The screening also makes an AI-prohibited evaluation the place to check practice records against later performance.

The stakes are at the skill frontier, so it matters where in the skill distribution the erosion evidence sits. The open-pool pattern concentrates below the elite range: among users with a maximum rating of at least 1,800, the signature no longer predicts weaker gains in the non-rostered arm (Appendix Table 13). It is not a skill-band artifact, however: within the roster arm's own baseline-rating range, the asymmetry attenuates when pooled across cohorts but sharpens within the AI-era entry cohort (-50.6, $p = 0.005$; Appendix A.1). Those users share the entry era and the rating range, so the difference tracks the gate, not skill level. The open

pool thus shows the erosion risk in the population from which frontier talent is drawn, while the screened pool shows no sign of it: there the signature predicts no rating shortfall (Section 5.2), and, among AI-era entrants, a within-user shift toward it predicts higher unaided scores (Section 5.3).

Screening also gives organizations that evaluate talent a design principle. An institution sets the rules of an AI-prohibited gate (no AI, in-person proctoring, qualifying rounds), and those rules decide whether differences in AI use become visible or are screened out. The signature tracks how people use AI in practice rather than whether they have access to it, as field experiments that manipulate the mode of use also find (Chen and Chan 2024, Krakowski et al. 2026). The lever is therefore the design of the evaluation, which organizations can control directly. In organizational economics, too, the returns to a capable algorithm depend on how decision authority is organized around it, not on access alone (Kim et al. 2024). As Section 2.4 discussed, AI-assisted output can become either a weaker or a sharper signal of skill (Cowgill et al. 2026, Wiles et al. 2025). In our evidence, the AI-prohibited gate is one design under which practice records and later performance line up as screening would predict. Organizations that recruit from AI-era talent pools must then decide where in the evaluation pipeline to observe skill unaided.

These claims rest on different amounts of evidence. The evidence establishes a pair of associations at one gate of this type: within the AI-era entry cohort, the signature's level predicts weaker rating gains only where no gate has screened the pool, and its within-user change predicts higher unaided scores among those the gate admitted. We interpret that pair as evidence that the gate separates the two modes of AI use. It is the one mechanism consistent with both associations, but we infer it rather than measure it. For evaluators, an AI-era practice record says something about unaided skill only once they know whether the environment that produced it was screened. Finally, what a gate restores is the signal, not the skill. If types are fixed and the gate only selects among them (the intrinsic-traits explanation of Section 3.3), it reveals skill that already exists but does not build new skill.

6.2 Practical implications and policy levers

Our findings point to two policy levers for preserving unaided skill at the frontier. Both levers assume that substitute use can change. Three facts support that assumption: the signature is not skill (Section 5.1), the behavior moves within person (Section 5.3 measures each user's own shift toward the signature), and in the experiment that varied the design of the AI tool, the mode of use changed with the tool design (Bastani et al. 2025). Our gate evidence is about detecting and separating the modes, and it says nothing about

moving people from one mode to the other. Still, a distributional concern remains: if substitute-style practice concentrates among weaker users, the skill gap between them and stronger users will widen. The two levers respond to that concern in different ways. One shapes the mode within training, and the other screens it at the population level:

- **Lever 1: complement-style AI in training.** Among AI-era entrants who survived the qualification gate, a stronger shift toward AI-style practice is associated with *higher* non-AI-aided ICPC performance (Section 5.3), so the complement mode is visible among elite performers. While our findings do not establish whether the *mode* can be steered, the classroom experiment of Bastani et al. (2025) does supply that causal step: an AI tutor that gave problem-specific hints rather than full solutions avoided the loss in AI-free exams. Together they suggest a training principle for elite settings: let AI give hints, and leave the solving to the learner. The principle matches the view of the MIT committee that AI should augment learning rather than automate it (MIT Ad Hoc Committee 2026). A timing result from diagnostic work suggests one concrete form, in which the learner attempts first and consults the tool afterward (Yin et al. 2025).

- **Lever 2: AI-prohibited evaluation gates.** Substitute-mode programmers who delegate work to AI are unlikely to pass AI-prohibited qualification rounds, so the elite pipeline retains those complement-users who invested efforts in skill frontier. The lever is therefore the existence and design of AI-prohibited checkpoints, not the elimination of AI from training. The MIT committee recommends the same on pedagogical grounds, with in-person proctored exams, oral examinations, and semester portfolios, and it frames unaided work not as policing but as "productive struggle" that builds durable understanding (MIT Ad Hoc Committee 2026).

Intrinsic traits vs. anticipation. The mechanism our evidence supports for Lever 2 is *sorting*: AI-prohibited gates screen substitute-users out of the elite pipeline. Sorting happens under both explanations of Section 3.3 for how the types arise. Under intrinsic traits, the gate winnows a pool it did not shape, without changing anyone's behavior. Under *anticipation*, it would also shift the training distribution itself. Our data cannot separate the two, so the pessimistic interpretation, in which types are fixed, is as unproven as the optimistic one. The intrinsic-traits explanation has an experimental counterpart: Autor et al. (2026) suggest that foundational expertise may be a prerequisite for extracting durable skill from AI-assisted practice. If so, the gate screens on the expertise that makes AI use complementary, an interpretation our data are consistent

with but do not establish. The policy argument here therefore rests on sorting, which holds under either explanation.

Both levers rest on the screening we document, not on a proven erosion effect. Any long-run cost of substitute-style practice falls on the users the gate screens out, so the gate matters whether or not that cost materializes.

6.3 Alternative explanations

Factors 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 contests, and universities may have steered top students toward applied AI and machine learning coursework. Both could weaken average CF rating gains among non-rostered users. For non-rostered users with unusually strong pre-period ratings, part of the weaker post-2022 performance could be regression to the mean. These channels, however, would move participation and rating levels. Section 5.2 instead compares non-rostered users who start from the same baseline rating and finds that a stronger AI-prompt signature predicts weaker rating gains. The labor-market and curricular channels could produce that pattern only if they both pushed practice toward the AI-prompt signature and lowered rating gains among non-rostered users but not among roster-linked ones. Regression to the mean cannot produce it, because the comparison holds baseline rating fixed and the asymmetry survives coarsened exact matching (Section 5.2). None of the three channels predicts the positive within-ICPC result of Prediction 3 either. Screening at the AI-prohibited gate explains the shortfall among non-rostered users, the roster-linked null, and the within-ICPC result with one mechanism, whereas each non-AI alternative would have to explain them separately.

Three further alternatives target the roster-linked null, which is half of the asymmetry (Prediction 2). The first concerns the outcome measure. Roster-linked users step back from open CF practice after their qualifying year (Appendix A.1), so their CF rating may carry little signal. Neither the within-cohort AI-era contrast nor the positive within-ICPC result, however, depends on that rating alone. The second concerns training. Roster-linked users train inside university programs, with coaches, teammates, and curated problem sets. Such structured training could itself loosen the link between practice style and rating. Our data cannot separate this explanation from prior screening. Both predict the same null, so telling them apart does not change what an evaluator can read from a practice record. It matters instead for how the institutions that train and screen these contestants design their training. The third is disengagement. A rising signature could mark

users who stopped iterating for reasons unrelated to AI. Three observations limit this explanation. It cannot explain the positive within-ICPC result. It also predicts that, in the open pool, users who shift toward the signature gain less, but that association is null (Appendix A.1). Finally, adding controls for the level and trajectory of practice volume leaves the asymmetry unchanged (-9.8, $p = 0.03$).

6.4 Limitations

Identification and screening strength. Within-individual fixed effects handle only part of the selection on entry into the CF population. The design also leaves some selection on continued participation, on AI tool use, and on entry into the ICPC and IOI qualification rounds. The robustness checks of Appendix A address some of these concerns, but not all. The screening interpretation also assumes that qualification rounds screen out substitute-users. We do not observe the strength of proctoring. While imperfect proctoring at earlier qualification stages would weaken the screening, the positive within-ICPC result in our analysis is consistent with effective screening.

Generalizability. Our setting is competitive programming. We conjecture that other AI-prohibited evaluations, such as medical board and bar exams, would separate the two modes the same way. The gate we study is a tournament: entry is team-based with peer selection, the prize is a recruiting signal rather than a license, and there is no retake structure. Medical boards and bar exams are individual, pass/fail, retakeable, and license-granting. Their candidates have stronger incentives to prepare for the specific test, which plausibly makes anticipation matter more, and intrinsic traits less, than in our setting. The urgency of the risk also differs across fields. AI is unusually strong at code generation and debugging, the skills competitive programming trains, so the substitution risk is most immediate in our setting. Expanding model capabilities may erase that difference quickly.

Measurement. We do not observe which CF users adopt which AI tools, and no such record exists for this setting. The platform logs submissions but not the tools used to produce them, so none of our three designs measures AI use directly. The construct validation of Section 5.1 and Appendix B bounds the leading alternative, that the signature proxies skill. The AI interpretation of the signature remains the most likely one, but we cannot verify it.

6.5 Conclusion

We ask whether AI-prohibited evaluation gates can separate two coexisting modes of AI-assisted practice in elite competitive programming. The evidence assembles into three patterns: a cross-cohort shift in practice toward the AI-prompt signature; an asymmetry in which the signature predicts weaker rating gains only in the unscreened, non-rostered pool; and *higher* non-AI-aided ICPC performance with a larger shift toward AI-style practice among AI-era entrants inside the gate. The unified reading is that AI-prohibited qualification gates separate substitute-style from complement-style practice, removing substitute-users before they reach elite contests. Our evidence speaks to the gatekeeping organization that decides whether any long-run cost of AI-era training reaches the skill frontier; it does not establish the existence or the size of that cost. Erosion is not a property of the technology: within the screened pool, AI-style practice coexists with frontier skill, and the risk attaches to the substitute mode of use. The two levers of Section 6.2 follow, and our evidence supports the premise of the second, that the modes are real and institutionally separable.

Natural extensions include variation in gate introduction or removal, to separate intrinsic traits from anticipation; other elite skill markets with AI-prohibited gates (medical boards, bar exams, scientific peer review); and user-level measurement of AI use through platform telemetry or surveys. For research on technology-mediated learning and on delegation alike, the open question is how to steer learners toward the complement mode rather than merely sort on it.

References

Aaltonen, A. and M. Stelmaszak (2024). The Performative Production of Trace Data in Knowledge Work. *Information Systems Research 35*(3), 1448–1462.

Abdel-Karim, B., N. Pfeuffer, K. V. Carl, and O. Hinz (2023). How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work. *MIS Quarterly 47*(4), 1395–1424.

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

Alavi, M. and D. E. Leidner (2001). Research Commentary: Technology-Mediated Learning—A Call for Greater Depth and Breadth of Research. *Information Systems Research 12*(1), 1–10.

Autor, D., T. Rodchenko, J. Martin, Z. Iscenko, S. Strand, D. Pearl, and M. Ferere (2026, September). Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting. NBER Working Paper 35720, National Bureau of Economic Research. Pre-registered, AEARCTR-0015823.

Baird, A. and L. M. Maruping (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly 45*(1), 315–341.

Bao, L., D. Huang, and C. Lin (2026). Can Artificial Intelligence Improve Gender Equality? Evidence from a Natural Experiment. *Management Science 72*(1), 474–494.

Bapna, R., N. Langer, A. Mehra, R. Gopal, and A. Gupta (2013). Human Capital Investments and Employee Performance: An Analysis of IT Services Industry. *Management Science 59*(3), 641–658.

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.

Beane, M. (2019). Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. *Administrative Science Quarterly 64*(1), 87–123.

Belo, R., P. Ferreira, and R. Telang (2014). Broadband in School: Impact on Student Performance. *Management Science 60*(2), 265–282.

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.

Boh, W. F., S. A. Slaughter, and J. A. Espinosa (2007). Learning from Experience in Software Development: A Multilevel Analysis. *Management Science 53*(8), 1315–1331.

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

Borwankar, S., W. Khern-am nuai, A. Pocheptsova Ghosh, and K. Kannan (2026). Unraveling the Impact: An Empirical Investigation of ChatGPT's Exclusion from Stack Overflow. *Information Systems Research*. Articles in Advance.

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*.

Chen, Z. and J. Chan (2024). Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. *Management Science 70*(12), 9101–9117.

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.

Cowgill, B., P. Hernández-Lagos, and N. L. Wright (2026). Does AI Cheapen Talk? Theory and Evidence from Global Entrepreneurship and Hiring. *Management Science*. Articles in Advance.

Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2026). The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science*. Articles in Advance.

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.

Deng, Z., Z. A. Cheng, P. Ferreira, and P. A. Pavlou (2026). From Smartphones to Smart Students: Learning vs. Distraction Using Smartphones in the Classroom. *Information Systems Research 37*(1), 276–293.

Fügener, A., J. Grahl, A. Gupta, and W. Ketter (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly 45*(3), 1527–1556.

Fügener, A., J. Grahl, A. Gupta, and W. Ketter (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research 33*(2), 678–696.

Fügener, A., D. D. Walzner, and A. Gupta (2026). Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work. *Management Science 72*(1), 538–557.

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.

Hann, I.-H., J. A. Roberts, and S. A. Slaughter (2013). All Are Not Equal: An Examination of the Economic Returns to Different Forms of Participation in Open Source Software Communities. *Information Systems Research 24*(3), 520–538.

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

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

Hou, J. J., L. Wang, G. Wang, H. J. Wang, and S. Yang (2025). The Double-Edged Roles of Generative AI in the Creative Process: Experiments on Design Work. *Information Systems Research*. Articles in Advance.

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

Jussupow, E., K. Bauer, R. M. Heigl, B. Vogt, and O. Hinz (2026). The Indirect Disclosure Effect: How Disclosing Generative AI Use Impacts Human Creative Collaboration with AI. *Information Systems Research*. Articles in Advance.

Jussupow, E., K. Spohrer, A. Heinzl, and J. Gawlitza (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research 32*(3), 713–735.

Kanazawa, K., D. Kawaguchi, H. Shigeoka, and Y. Watanabe (2026). AI, Skill, and Productivity: The Case of Taxi Drivers. *Management Science 72*(2), 1376–1388.

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.

Knight, B., D. Mitrofanov, and S. Netessine (2026). Human-Algorithm Collaboration in Gig Work: The Role of Experience, Skill Level, and Task Complexity. *Information Systems Research*. Articles in Advance.

Krakowski, S., D. Haftor, J. Luger, N. Pashkevich, and S. Raisch (2026). Human-Centered Artificial Intelligence: A Field Experiment. *Management Science 72*(1), 57–72.

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.

MIT Ad Hoc Committee (2026, August). Report of MIT's Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training. Technical report, Massachusetts Institute of Technology. MIT Ad Hoc Committee on AI Use in Teaching, Learning, and Research Training; co-chairs Eric Klopfer and Sam Madden; dated August 13, 2026.

Narayanan, S., S. Balasubramanian, and J. M. Swaminathan (2009). A Matter of Balance: Specialization, Task Variety, and Individual Learning in a Software Maintenance Environment. *Management Science 55*(11), 1861–1876.

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

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*.

Piccoli, G., R. Ahmad, and B. Ives (2001). Web-Based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic IT Skills Training. *MIS Quarterly 25*(4), 401–426.

Rinta-Kahila, T., E. Penttinen, A. Salovaara, W. Soliman, and J. Ruissalo (2023). The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation. *Journal of the Association for Information Systems 24*(5), 1378–1412.

Santhanam, R., S. Sasidharan, and J. Webster (2008). Using Self-Regulatory Learning to Enhance E-Learning-Based Information Technology Training. *Information Systems Research 19*(1), 26–47.

Shan, G. and L. Qiu (2026). Examining the Impact of Generative AI on Users' Voluntary Knowledge Contribution: Evidence from a Natural Experiment on Stack Overflow. *Information Systems Research 37*(2), 1021–1041.

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)*.

Singh, P. V., Y. Tan, and N. Youn (2011). A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects. *Information Systems Research 22*(4), 790–807.

Song, F., A. Agarwal, and W. Wen (2026). The Impact of Generative AI on Collaborative Open-Source Software Development: Evidence from GitHub Copilot. *Information Systems Research*. Articles in Advance.

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.

Sundaresan, S. and I. Güler (2025). Algorithmic Recommendation Tools and Experiential Learning in Clinical Care. *Organization Science 36*(5), 1786–1802.

Tambe, P., X. Ye, and P. Cappelli (2020). Paying to Program? Engineering Brand and High-Tech Wages. *Management Science 66*(7), 3010–3028.

Wambsganß, T., A. Janson, M. Söllner, K. Koedinger, and J. M. Leimeister (2025). Improving Students' Argumentation Skills Using Dynamic Machine-Learning–Based Modeling. *Information Systems Research 36*(1), 474–507.

Wang, L. and H. C. B. Lee (2026). How Costs Influence Preferences for Control in Generative Artificial Intelligence (GenAI): Human-Guided vs. GenAI-Based Delegated Search. *Information Systems Research*. Articles in Advance.

Wiles, E., Z. Munyikwa, and J. J. Horton (2025). Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires. *Management Science 71*(12), 10144–10164.

Yao, S. (2026). Stranded Credentials: Keeping Online Reputation Systems Informative in the AI Era. *arXiv preprint*. v2, September 2026.

Yin, J., K. Y. Ngiam, S. S.-L. Tan, and H. H. Teo (2025). Designing AI-Based Work Processes: How the Timing of AI Advice Affects Diagnostic Decision Making. *Management Science 71*(11), 9361–9383.

Zhang, S., P. V. Singh, and A. Ghose (2019). A Structural Analysis of the Role of Superstars in Crowdsourcing Contests. *Information Systems Research 30*(1), 15–33.

Appendix

A Robustness

This appendix reports the robustness checks referenced in Section 5: 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 very 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 subsection documents nine checks. Three address 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. Two address the skill control and the roster-versus-skill confound. Four probe the asymmetry itself: rating-floor symmetry, the change construct in the open pool, volume conditioning and composite sensitivity, and the matched rating-range test.

Table 7 summarizes the selection bounds discussed in Section 5.2.

Filter sensitivity

We re-estimate Equation 5 (per-subpopulation) and Equation 6 (pooled, fully stratified) under three activity thresholds: the headline (≥ 2 quarters, ≥ 5 submissions each), an intermediate variant (≥ 2 quarters, ≥ 1 submission each), and a permissive variant (≥ 1 quarter, ≥ 1 submission). The first principal component is

Table 7: 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 in the replication package.

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.

Table 8: 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 4'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
(≥ 1 q, ≥ 1 sub)		(1.87)	(3.64)	($p = 0.081$)
INTERMEDIATE	7,753	-16.38^{***}	-4.46	$\mathbf{-11.92}^{*}$
(≥ 2 q, ≥ 1 sub)		(2.32)	(4.45)	($p = 0.017$)
HEADLINE	6,690	-12.42^{***}	-2.54	$\mathbf{-9.88}^{*}$
(≥ 2 q, ≥ 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The pooled interaction θ_1 is significant at HEADLINE and INTERMEDIATE (INTERMEDIATE larger in magnitude than HEADLINE), and attenuates to marginal significance at PERMISSIVE. The PERMISSIVE attenuation is consistent with classical errors-in-variables: at ≥ 1 quarter with ≥ 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} 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 9: 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 cutoff-reach indicator (whether the user has CF activity on both sides of the ChatGPT cutoff), the pre-window single-shot AC rate (a proxy for pre-window AI-prompt signature among the subset of dropped users with at least one pre-window practice submission), and cohort composition.

Table 10: 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
Cutoff-reach 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 11: 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 users who quit; 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. Most do not reach across the cutoff (cutoff-reach rate 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-prompt 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.

Table 12: 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%

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 of the naive expectation 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 4) could be partially produced by differential censoring of high-signature roster-linked users rather than by the screening mechanism alone. The filter-sensitivity result in Table 8 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 two-margin pattern completed by the within-ICPC test (Section 5.3) remains the more robust evidence of the screening mechanism.

Skill-control sensitivity

The headline specification includes the baseline-skill control $\bar{r}_h^{\mathrm{pre}}$ to block omitted-variable bias from the negative 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 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-prompt 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 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 $\bar{r}_h^{\mathrm{pre}}$, 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 $\bar{r}_h^{\mathrm{pre}}$ 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 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.

Rating-floor symmetry

The non-rostered expansion was scraped among accounts with career-maximum rating of at least 1,400, while the roster-linked arm carries no floor. Conditioning one arm on a realized career maximum could in principle create mean-reversion dynamics in Δr_h that differ between the two arms and that the pre-period mean control does not remove. Table 13 imposes the floor symmetrically (career maximum proxied by the maximum monthly rating in the observed panel) and then raises the common floor. Symmetry at 1,400 removes 170 users whose proxied career maximum falls below the floor (6,690 to 6,520): 19 roster-linked (1,407 to 1,388) and 151 non-rostered (5,283 to 5,132). The symmetric floor strengthens the asymmetry, from -9.88 to -12.17, because the 19 roster-linked users below it were pulling that arm's slope downward; removing them leaves a cleaner roster null (-0.02) and widens the gap between the two arms' slopes. Rising common floors attenuate θ_1 as the non-rostered arm converges to an elite-only composition: a pool restricted to users whose career peak already cleared a high bar is a quasi-screened pool, so the mechanism itself predicts the convergence, though the same pattern is equally consistent with the signature-rating association concentrating below the elite range. Floors below 1,400 are infeasible because the non-rostered arm is scrape-truncated there.

The change construct in the open pool

The within-ICPC treatment of Section 5.3 is a within-user change, while the open-pool tests of this appendix use a cross-user level. Estimating the open-pool specification with the change construct instead (the full-post-

Table 13: Rating-floor symmetry for the cross-subpopulation asymmetry. Each row re-estimates Equations 5 and 6 with a common career-maximum-rating floor imposed on both subpopulations. HC1 standard errors. The first row (floor on the non-rostered arm only) is the diagnostic pipeline's reproduction of the headline (Table 4: −9.82, $p = 0.048$).

Variant	N	Roster-linked slope	Non-rostered slope	θ_1	p
Floor on non-rostered arm only (headline)	6,690	−2.54 (SE 4.37)	−12.42 (SE 2.45)	−9.88 (SE 5.01)	0.049
Symmetric floor 1,400	6,520	−0.02 (SE 4.27)	−12.20 (SE 2.43)	−12.17 (SE 4.92)	0.013
Symmetric floor 1,600	5,242	+1.07 (SE 4.29)	−7.07 (SE 2.65)	−8.13 (SE 5.04)	0.107
Symmetric floor 1,800	3,738	+2.43 (SE 4.28)	−2.89 (SE 3.21)	−5.32 (SE 5.35)	0.320

window within-user first difference of the three sign-aligned proxies, headline thresholds, scored with the frozen broad-reference loadings; $N = 4{,}480$ users with sufficient activity in both windows) gives slopes of +2.6 among non-rostered users ($p = 0.38$) and +2.3 among roster-linked users ($p = 0.62$), with a cross-arm interaction of +0.3 ($p = 0.95$); within the AI_era cohort, both arms' slopes are positive and insignificant. The level and change constructs correlate at $r = 0.61$ on the overlapping sample. The null matches the framework's no-prediction cell for the change margin in a mixed pool (Section 3.4): where both modes are present, a within-user shift toward the signature carries no predicted direction.

Volume conditioning and composite sensitivity

Two further checks perturb the asymmetry specification. First, engagement conditioning: adding each user's log post-window practice volume and the pre-to-post change in log volume to the per-arm and (fully stratified) pooled specifications leaves the cross-arm interaction unchanged at −9.8 ($p = 0.033$). The conditioning shifts both arms' slope levels downward (non-rostered −18.1, roster-linked −8.3) while preserving their difference, the same pattern as in the skill-control check above: the arm levels move, their difference stays. Within the AI_era cohort the conditioned interaction is −27.2 ($p = 0.14$), with the non-rostered slope at −25.0 ($p < 0.001$) and the roster slope null. A volume-trajectory split is less informative because entry vintage and volume trajectory are confounded (late entrants mechanically have non-declining volume); the split point estimates are −14.7 (non-declining) and −5.5 (declining), neither precise. Second, composite sensitivity: refitting the level composite on the same fitting sample with the quick-retry component excluded attenuates the pooled all-cohorts interaction to −3.9 ($p = 0.45$) but sharpens the within-AI_era interaction to −52.0 ($p = 0.024$; non-rostered slope −18.7, $p = 0.016$). The within-cohort contrast therefore does not depend on the quick-retry component, whose sign Section 5.1 showed to be conditioning-dependent, while the pooled all-cohorts interaction partly does.

Matched rating-range test for near-qualifiers

An alternative holds that the asymmetry reflects where the two arms sit on the skill distribution rather than which side of the gate they are on: most non-rostered users sit below the roster arm's baseline-rating range, so the negative slope could belong to a skill band rather than to screening status. The test restricts both arms to the roster arm's baseline-rating range and re-estimates the Section 5.2 specifications (per-arm slopes and the pooled interaction, with the baseline-rating control and HC1 standard errors) inside the common band, at two widths: the roster arm's interquartile range ($\bar{r}_h^{\text{pre}} \in [1{,}478, 2{,}081]$) and its decile range ($[1{,}157, 2{,}295]$). Pooled across cohorts, the asymmetry attenuates (interquartile $\theta_1 = -4.7$, $p = 0.42$; decile -7.1, $p = 0.14$), consistent with the rating-floor finding that the all-cohorts pattern concentrates below the elite range. Within the AI-era entry cohort the pattern reverses: the decile band sharpens the asymmetry to $\theta_1 = -50.6$ ($p = 0.005$; 547 users), with the non-rostered slope at -19.9 ($p = 0.02$) and the roster-linked slope at $+30.7$ ($p = 0.054$), and the interquartile band shows the same signs on a sample too small for precision (-40.8, $p = 0.16$; 173 users). Matching the rating range removes the all-cohorts pattern but not the AI-era one: among entrants of the same era in the same baseline-rating range, the signature's level predicts weaker rating gains only on the unscreened side of the gate. The pipeline reproduces the pooled headline ($\theta_1 = -9.88$, $p = 0.049$) before the range restriction.

A.2 Robustness for the within-ICPC signature-shift test

The within-ICPC signature-shift test of Section 5.3 (Equation 7) is estimated on the 444-user, 579-observation sample described in Table 14. The headline $\psi_{\text{AI}} = 0.190$ ($p = 0.028$) depends on the activity filter that defines the qualifying practice window. This subsection documents six checks: filter sensitivity; a threshold check with the treatment pipeline held fixed, with influence and clustering diagnostics; a balance check on filter dropout by cohort; a cross-check with the Section 5.1 practice-pattern shift; teammate composition and the two-component composite; and the IOI sister analysis.

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 of the main specification, but aggregating at the user-year level rather than collapsing each user to a

Table 14: 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	`shift_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%

single value as the headline does. PCA loadings are fit on the headline-threshold sample and frozen across the INTERMEDIATE and PERMISSIVE thresholds so that the treatment is on a comparable scale across thresholds.

Table 15: Within-ICPC signature-shift test under three activity-filter thresholds, with the treatment built per (user, contest year) and aggregated as within-user changes. AI_era × `shift_pc1` 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
PERMISSIVE (≥ 1 quarter, ≥ 1 sub)	541	465	**0.434**	0.257	0.091
INTERMEDIATE (≥ 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 producing the positive direction. 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.

Threshold check with the treatment pipeline held fixed; influence and clustering diagnostics

The per-user-year table above varies the construction and the thresholds at once. Table 16 separates them: each variant re-runs the headline treatment pipeline itself, holding the ICPC participant pool, the first-qualifying-year anchor, and the frozen broad-reference loadings fixed, and changing only the three activity

thresholds (minimum active pre-window quarters / post-window quarters / submissions per active quarter). The headline row is the canonical per-user treatment. The interaction stays positive at every threshold; significance is specific to the headline filter, and the smooth attenuation as low-activity users enter is the pattern classical measurement error predicts when sparser windows measure the practice shift more noisily. The AI_era-observation column counts observations in the regression sample; the construction stage has four to six more rows per variant, removed because a few AI_era users lack the baseline-rating control. For the headline row, that is four rows from three users, leaving the 80 observations and 67 AI_era users reported in Table 2, Panel C, in the main text.

Table 16: Fixed-construction threshold sweep for the within-ICPC test, with influence and clustering diagnostics on the headline sample. Spec: Equation 7 (team-clustered standard errors unless stated). The AI_era × `shift_pc1` interaction is reported.

Thresholds (pre / post / subs)	N obs	N users	AI_era obs	ψ_{AI}	p
3 / 2 / 5 (headline)	579	444	80	+0.190 (SE 0.087)	0.028
3 / 2 / 1	718	554	90	+0.148 (SE 0.086)	0.084
3 / 1 / 1	748	582	90	+0.084 (SE 0.082)	0.307
2 / 2 / 5	675	517	101	+0.122 (SE 0.077)	0.111

On the headline sample, influence diagnostics show no single cluster drives the estimate: across 321 leave-one-team-out re-estimates, one per team cluster in the regression sample, ψ_{AI} ranges over $[+0.156, +0.227]$ with p in $[0.011, 0.067]$ (eight of the re-estimates push p above 0.05, none above 0.10); across 196 leave-one-university-out re-estimates, the range is $[+0.158, +0.233]$ with p in $[0.012, 0.066]$ (four push p above 0.05). Alternative clusterings on the same sample give team $p = 0.028$ (baseline), university $p = 0.068$, and user $p = 0.04$; the conservative CRV3 estimator, a jackknife-type small-sample correction at the team level, gives $p = 0.18$. Clustering simultaneously by team and by user returns a standard error smaller than either one-way estimate ($p = 0.0003$). Because 69.6% of the 444 user clusters are singletons on 579 observations, a setting in which two-way cluster-robust variance estimates are unreliable, we read that number as a small-sample artifact and do not rely on it. The CRV3 result supports the main text's reliance on the direction rather than the size of the within-ICPC estimate.

Balance check on filter dropout by cohort

Step 3 of Table 14 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 14 (`shift_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 versus 1,802, Welch $p < 0.0001$) and lower cutoff-reach rate (0.74 versus 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 versus 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 $\times$ `shift_pc1` slope upward or downward depends on the within-cohort relationship between the within-user change in AI-prompt 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 $\times$ 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).

The single-shot effect is *amplified* on the within-ICPC sub-sample to 126% of its full-panel magnitude (0.0736 versus 0.0586, $p < 0.001$). The median-attempts effect retains 83% of the full-panel magnitude (-0.2705 versus -0.3271, $p = 0.017$). The quick-retry effect has essentially the same magnitude as the full panel (-0.0193 versus -0.0190) but loses statistical significance on the smaller sample ($p = 0.15$). The

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, $p = 0.15$)

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 single-shot shift is larger in magnitude on this sub-sample than on the full panel.

Teammate composition and the two-component composite

ICPC score is a team outcome, so Equation 7's interaction could reflect teammate composition rather than the user's own practice change. Of the 579 estimation observations, 447 have at least one CF-linked teammate whose shift toward the signature is measurable (228 have two). On the covered sub-sample the interaction is +0.137 without and +0.137 with teammates' mean shift added as a control (the teammate term is +0.02, $p = 0.88$), so the attenuation from the headline +0.190 reflects the sub-sample, not the control; a full-sample variant that fills missing teammate means and adds a missing indicator gives +0.20 ($p = 0.016$). In the sub-sample where all three team members have a measurable shift, the teammate term is larger (+0.67, $p = 0.05$) and the interaction attenuates to +0.07 ($p = 0.06$), remaining positive; that cell is heavily selected on linkage and saturated with fixed effects, so we read it as bounding rather than headline evidence. Separately, rescoring the same builder rows with a two-component composite that excludes quick-retry (broad-reference loadings refit on the two remaining sign-aligned changes; the three-component rescoring reproduces the stored treatment at $r = 1.000$) gives +0.198 ($p = 0.076$): the interaction's sign and size do not depend on the quick-retry component.

Construction details for the change composite

We fit the `shift_pc1` 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 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. The group comprises the 4,959 practice-panel users with

≥ 3 active quarters in the pre-window 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 being out-of-window-cohort or name-only-matched users the panel excludes. Each user's single value is anchored at 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.

Reconciling the roster-linked rating null with the within-ICPC slope

Weighting the three within-ICPC cohort slopes (−0.048, +0.012, and +0.142) by their user counts (144, 233, and 67; Table 2, 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 for whom the within-ICPC AI_era interaction 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.

The IOI sister analysis

The IOI counterpart to the within-ICPC signature-shift test is weaker than ICPC's in two ways. First, precision: the IOI is a smaller pool, with only 345 contestant-year observations from 243 CF-linked contestants over 2023–2025, against 579 from 444 users for ICPC. 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 shift-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-aided-performance headline on ICPC and treat IOI as a directional check only.

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 that cohort. 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 positive coefficient (0.0090, $p = 0.002$), opposite to the AI-prompt direction, for the transition cohort in the ChatGPT-anchored headline. One reading is that Copilot's integration into the code editor, as autocomplete-style suggestions, reduced human retries directly, while ChatGPT's generate-and-debug workflow reintroduced debugging cycles among transition users. 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 3) 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.

Figure 1 shows the raw signature-component levels by cohort and quarter across both release dates, the descriptive backdrop for the two-shock timing discussed above.

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

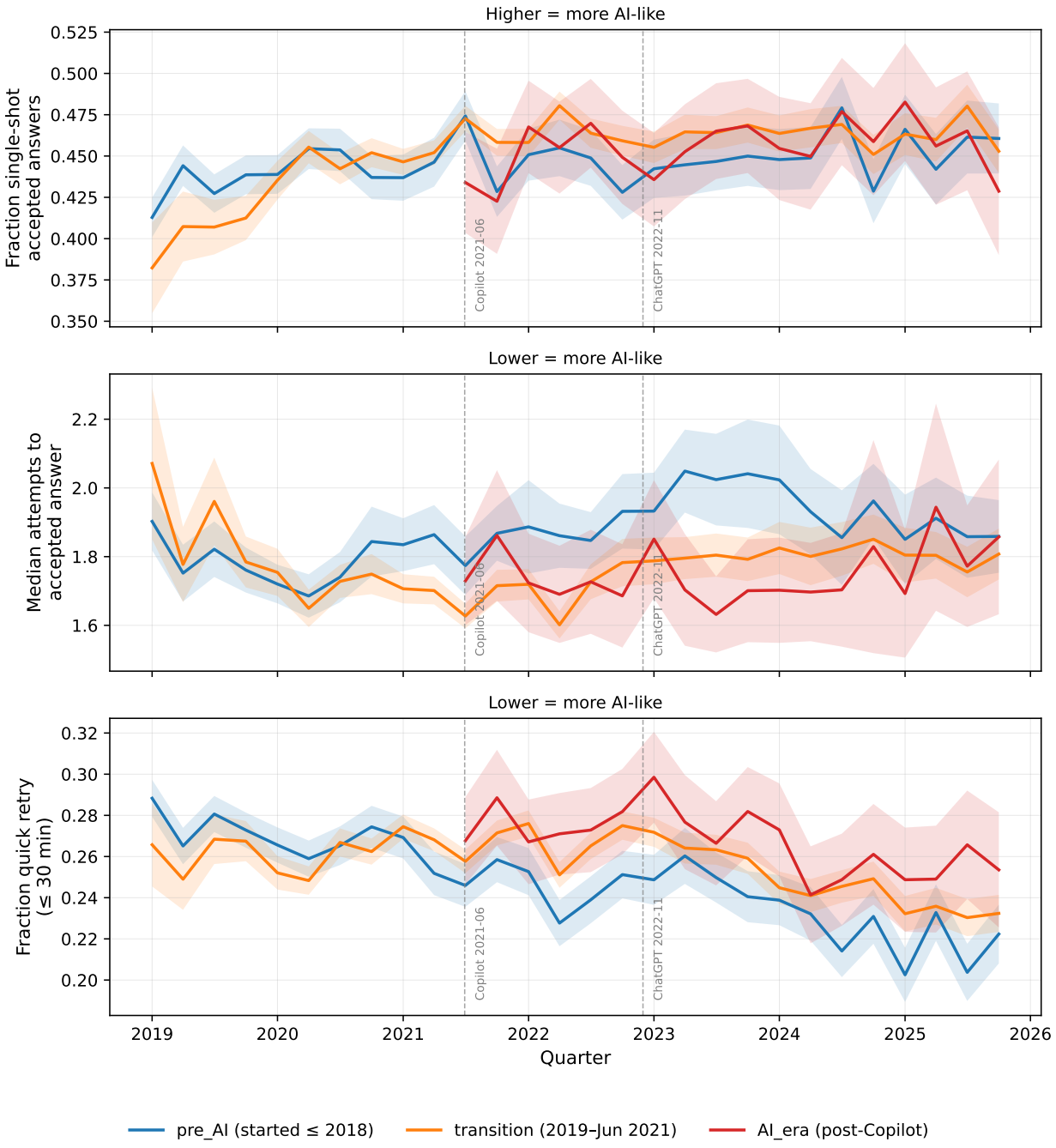


Figure 1: AI-prompt signature components by cohort and quarter. Each panel plots the cohort mean of one practice proxy by quarter with 95% confidence bands; the vertical axis names the metric and the panel title gives its AI-prompt direction. Vertical dashed lines mark the Copilot (2021-06) and ChatGPT (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 20) in the three focal cohorts.

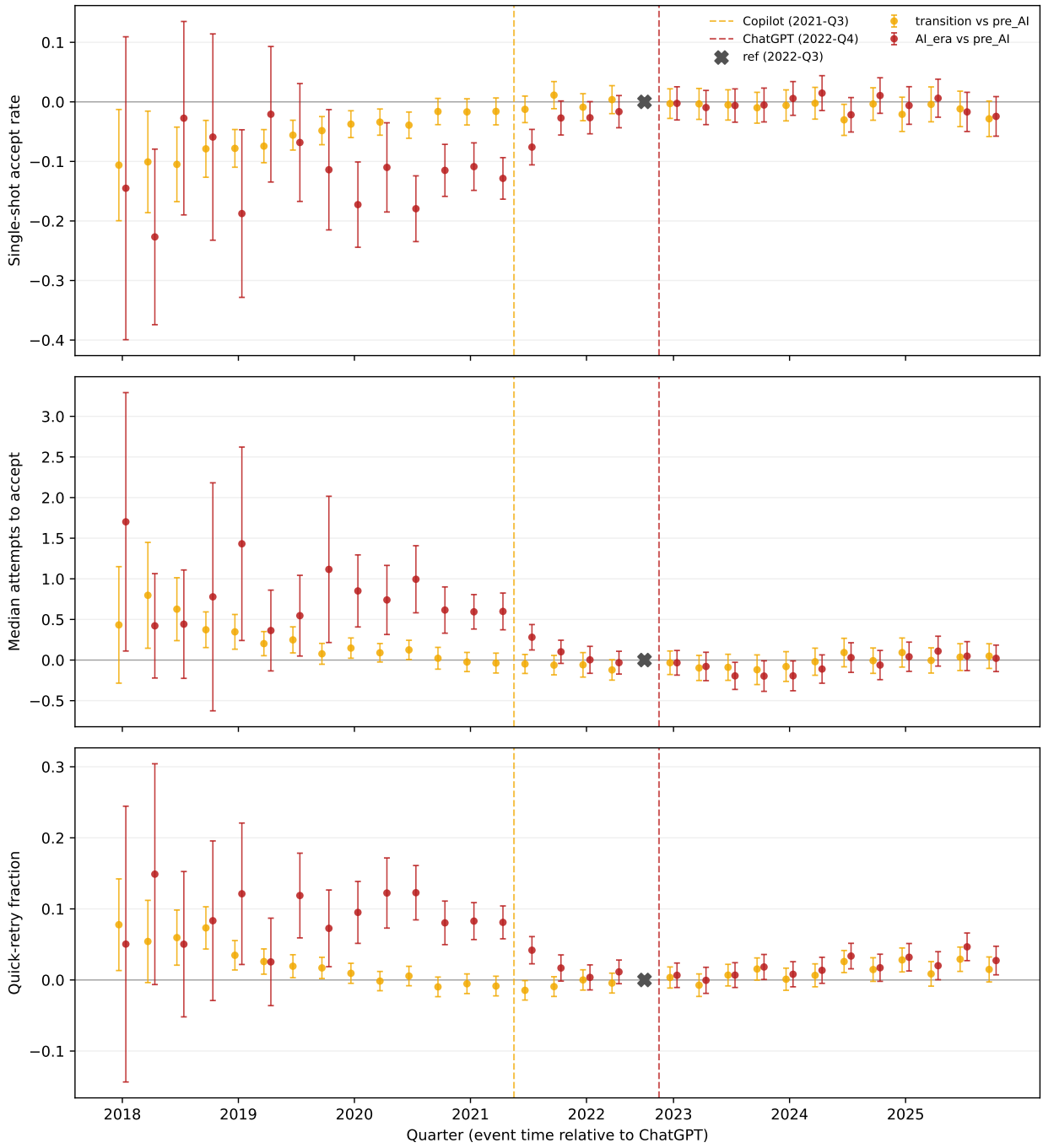


Figure 2: Cohort event study of the three AI-prompt signature proxies (Equation 4 with quarterly event-time interactions). Each panel plots event-time coefficients for the transition (orange) and AI_era (red) cohorts relative to the pre_AI cohort; the reference period is 2022-Q3 (gray X, normalized to zero), the vertical dashed lines mark the first post-Copilot quarter (2021-Q3) and the ChatGPT release quarter (2022-Q4), and bands are 95% confidence intervals (user-clustered standard errors). Neither cohort jumps at either release. Because the event study runs on the practice panel, whose submissions can predate a user's first rated contest, a few AI_era users appear before 2021; these few-user leads are uninformative, so the testable pre-trend contrast is the transition cohort, which rejects parallel trends on all three proxies (Section 5.1).

(participation in our rosters did not fall); multiple national qualification pipelines were disrupted across 2020 and 2021. The CF practice panel is the least exposed of the three samples to the pandemic disruption 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 removed with indicator variables. The cross-subpopulation CF rating asymmetry (Section 5.2) is similarly insulated: 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 signature-shift test (Section 5.3) restricts the sample to contest years ≥ 2023, which entirely post-date COVID, so the COVID years cannot directly contaminate the `shift_pc1` 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 Data provenance and research ethics

All data are collected from publicly accessible sources: submission metadata and rating histories through the documented CF public API and public contest pages, and ICPC results from `cphof.org` and IOI results from the contests' official public pages (`stats.ioinformatics.org`). Collection used rate-limited requests, and no login-gated content was accessed. CF usernames are public pseudonyms, and the roster linkage relies only on information the sources themselves publish (usernames and publicly listed contestant names). Working datasets are de-identified for analysis, results are reported only in aggregate, and no individual is named or identifiable in the paper. The study is secondary analysis of public records. We neither interact with nor intervene on the individuals observed, and we obtain no private identifiable information about them. The IRB of the author's institution determined that the study is not human subjects research (determination 202608190, August 25, 2026), and informed consent was not required.

A.6 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 processes 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 account-level arithmetic connects these person-level counts to the panel. The roster-candidate scrapes collected 4,491 CF accounts, targeting both link types: accounts named directly on cphof profiles and accounts proposed by the name-and-country matcher (a matched person can propose more than one candidate account, which is why the 2,215 match pairs exceed the 1,592 matched persons). Of the 4,491, the 3,117 confirmed roster links above are retained: 2,189 ICPC direct-username accounts and 1,164 IOI accounts whose name match is confirmed by an attached contest record, overlapping in 236. The remaining $4{,}491 - 3{,}117 = 1{,}374$ accounts, scraped as candidates whose best match could not be confirmed against a contest record, are the excluded name-only matches of the next paragraph. Together with the 8,748 accounts of the non-rostered expansion, the scrape universe is $4{,}491 + 8{,}748 = 13{,}239$ accounts, and the exclusions lead to the panel: $13{,}239 - 1{,}374$ (name-only) $-1{,}408$ (outside the cohort window) -38 (no practice-mode submissions) $= 10{,}419$.

Of the 3,117 roster-linked accounts, 2,494 are cohort-assignable (1,673 ICPC and 1,030 IOI contestants, 209 of whom competed in both): their first rated CF contest falls within the 2015–2024 cohort window of Section 4.1 in the main text; the remaining 623 fall outside the window and belong to no cohort. The linked contestants span 354 universities and 90 countries over 2018–2025. Three groups of scraped users are excluded from the analysis panel: the 1,374 lower-confidence name-only matches discussed below; 1,408 users whose first rated contest falls outside the cohort window (the 623 rostered contestants above and 785 non-rostered users); and 38 users with 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; the 14 reconcile the 2,494 cohort-assignable contestants with the panel's 2,480 roster-linked users). The panel is not otherwise filtered on any outcome variable.

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 4.1). 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 provided in the replication package.

For reference, Table 19 reports the cohort composition of the CF analysis panel underlying the Section 5.1 DiD.

Table 19: 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 counts 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 4.1), 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.*

A.7 Sample summary and practice-shift tables

Table 20 reports the headline summary statistics for the two contest panels, the CF panel, and the linkage; it is referenced from the main text.

Table 20: Summary statistics for the two contest panels, the CF panel, and the ICPC/IOI–CF linkage. Panels A and B are descriptive panels of the ICPC university-year and IOI country-year records; Panel C is the combined CF open pool 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 contest periods	112
Contest period: 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 contest periods	86
Contest period: 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 open pool 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 / transition / 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 4.1). Contest periods in Panels A and B group contest years: 2018–19, 2020–21, and 2022–25 for ICPC, and 2018–21, 2022, and 2023–25 for IOI. "Spanning" CF users have ≥30 practice submissions in both the pre-Copilot and post-ChatGPT windows; spanning universities and countries appear in both the pre-AI and the AI-era contest period. 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.6 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 sub-samples are documented in Section 4.1.

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 signature-shift test (Section 5.3) rest on the AI-prompt signature: the behavioral composite of higher single-shot AC 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 (a cross-sectional test and a within-user test of discriminant validity, and a within-user event study) 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 these 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-shock high-rated users would already exhibit the AI-prompt signature: higher single-shot AC 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 the ChatGPT release on 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 21.

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

Table 21: 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.

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.

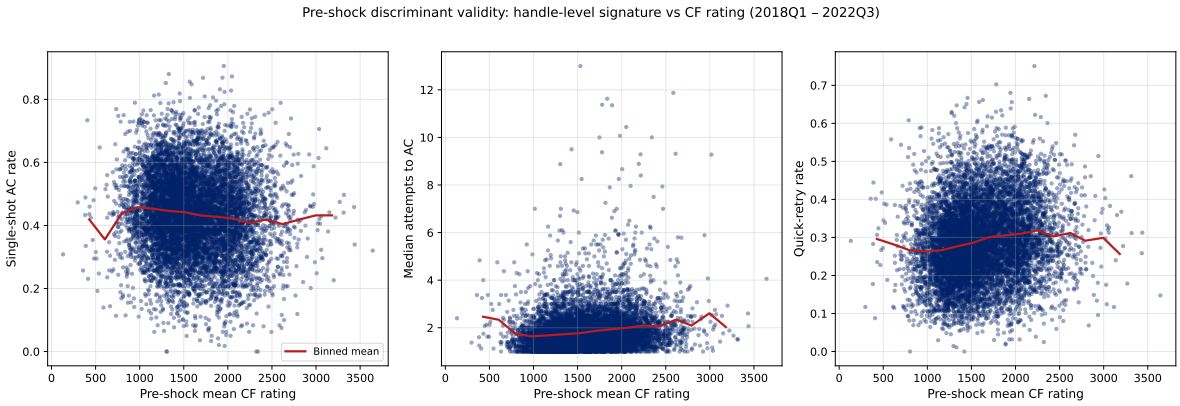


Figure 3: Pre-shock discriminant validity: user-level mean signature components vs. pre-shock mean CF rating (n = 8,535 users, 2018-Q1 to 2022-Q3). Red lines are binned means. The cross-sectional relationships have signs opposite to the AI-prompt prediction: higher pre-shock rating is associated with 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 relationship. 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 rating points. The sample is the pre-shock quarterly panel; standard errors are clustered at the user level. Table 22 reports the results.

Table 22: Pre-shock within-user elasticity: signature vs. rating

Signature component	n obs	δ per 100 rating points	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. Specification (8): user and quarter fixed effects; standard errors clustered at the user level. Pre-shock window: 2018-Q1 to 2022-Q3.

Within users, the signature does shift with rating, but the magnitudes are small. The single-shot AC elasticity is $\delta = 0.0043$ per 100 rating points, meaning that a user would need to gain roughly 1,350 rating points to produce a single-shot AC shift of magnitude comparable to the AI-era cohort × post coefficient of +0.0586 documented in Section 5.1. The median-attempts component has an elasticity of −0.033 per 100 rating points against the cohort coefficient of −0.327, roughly a 10:1 ratio of cohort contrast to skill elasticity. Skill movements at the magnitudes plausibly achievable within the 2022-Q4 to 2025-Q4 post-shock window (typically tens to low hundreds of rating points for established users, not 1,000+) cannot reproduce the cohort contrast.

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 contrast.

B.2 Within-user stability around AI release

The remaining alternative explanation for the cohort contrast is a smooth-trend story: practice patterns have 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 (full results are provided in the replication package).

The Copilot release (2021-06-29) points the same way but with pre-shock parallel-trends concerns (several pre-shock lead-month coefficients for the single-shot rate are significantly positive). We therefore treat the Copilot event study as inconclusive and relegate it to the replication package.

Cohort-stratified event studies on the ChatGPT shock show the same pattern: no within-user movement toward the signature among existing practitioners. Within the pre-AI cohort ($n = 733$ users, 8,791 user-month observations 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 user-month observations), the post-shock pattern matches the pre-AI cohort's direction, with no movement toward the signature at the release. The AI-era stratum (the remaining 204 of the 1,712 spanning-panel users, 2,144 user-month observations) is not shown. AI-era users enter rated competition after the Copilot release by construction, so their lead windows before the ChatGPT release are short and thin, and the event study was never run for that stratum.

Interpretation. Established practitioners did not shift their practice patterns at the 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 contrast must originate primarily in cohort *composition*: users who entered the platform after AI's broad availability arrived with practice patterns that differ from those of the pre-AI cohort. This is consistent with the paper's skill-formation framework: AI exposure during the formation period shapes practice habits; AI exposure among experienced practitioners does not change theirs.

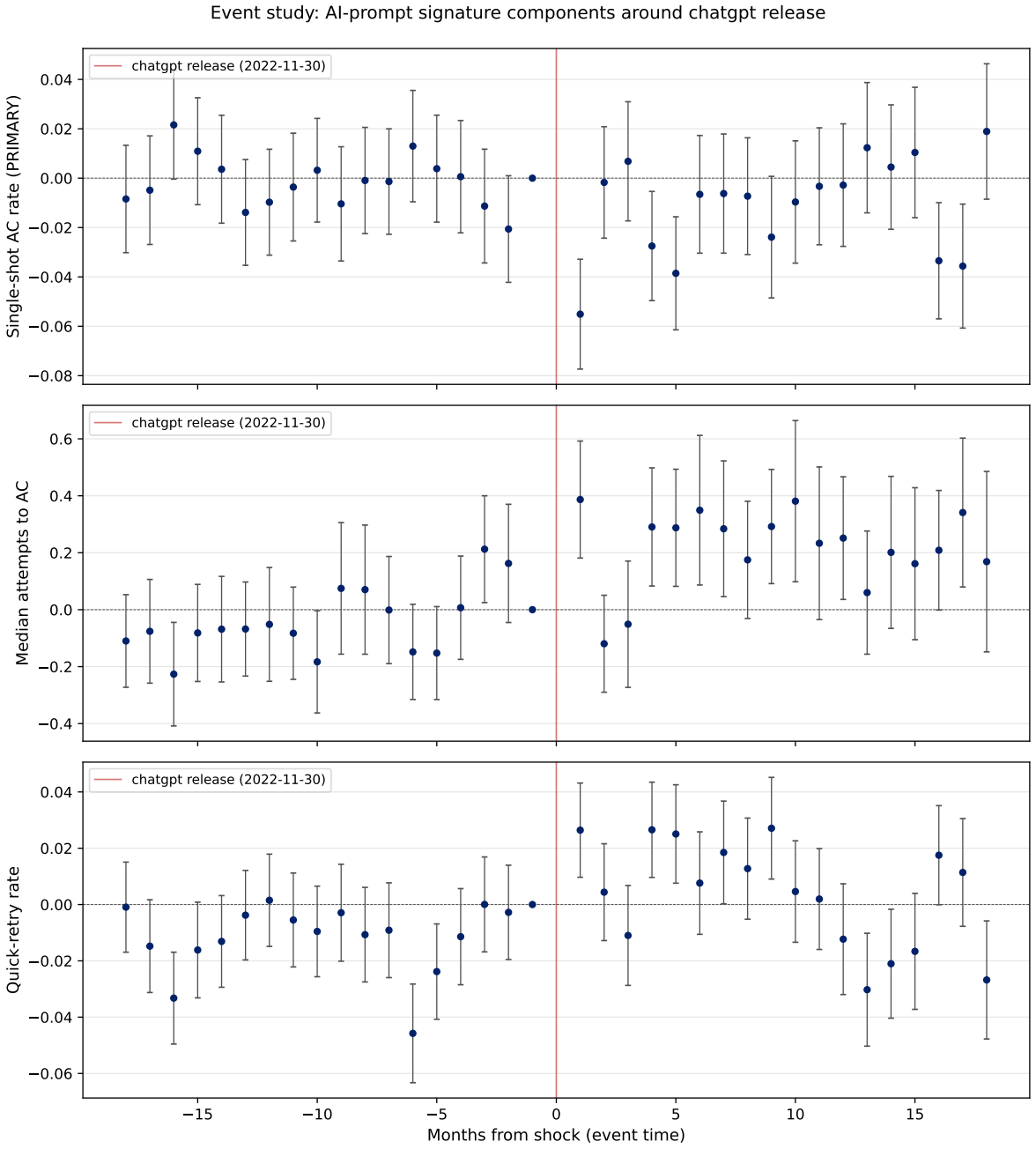


Figure 4: Within-user event study at the ChatGPT release. Spanning panel of 1,712 users active both pre- and post-shock. Reference period: $k = -1$ (one month before release, October 2022). Vertical red line marks release. The release month is dropped, so no $k = 0$ coefficient is estimated. The post-release 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.

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 contrast. The event-study test (Subsection B.2) rules out adoption by established practitioners: there is no within-user movement toward the signature at the ChatGPT release (the Copilot event study is inconclusive), which locates the cohort contrast 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 signature gap is largest on hard problems, where AI is weakest, rather than on the easy, templated problems AI handles best. 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.

The randomized experiment of Shen and Tamkin (2026) 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. In their exploratory clustering of usage patterns, participants in the low-engagement clusters (those delegating code generation or debugging to the AI) scored far below those in the cognitively engaged clusters, although the clusters are small and no formal test of where the effect concentrates is reported. That low-engagement, delegating way of working with AI is one of the usage patterns our signature tracks. Because AI use is randomly assigned in their design, the average effect on non-AI-aided performance is causally identified in that setting; the pattern-level contrast is descriptive. Our signature is the multi-year, field-observational analog of the usage patterns they observe.

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 the RCT of Shen and Tamkin (2026) shows that AI use during learning lowers later unaided performance, with the lowest scores among participants who delegated to the AI. What they do not establish is that the cross-cohort variation in the signature is AI-caused; the career-stage-matched 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.