Research Policy 55 (2026) 105568

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Hope, signals, and silicon: A game-theoretic model of the pre-doctoral academic labor market in the age of AI

Shaohui Wang

J. Mack Robinson College of Business, Georgia State University, United States of America

ARTICLE INFO

Keywords:
Generative AI
Pre-doctoral labor market
Research evaluation
Signal extraction
Task-based production
Academic tournaments
AI augmentation

ABSTRACT

Generative AI can make early research work easier to produce and harder to interpret. This paper develops a compact game-theoretic model of this production–evaluation tension in the pre-doctoral academic labor market. In the model, PIs organize RA labor, allocate AI between routine and novel tasks, and choose mentoring intensity. RAs choose effort, while admissions committees infer research potential from noisy task-level signals under fixed admissions capacity. A mechanism-preserving simulation examines whether the model's qualitative mechanisms continue to hold when RAs and PIs are heterogeneous, research outcomes partly depend on luck, admissions evaluation is noisy, and elite Ph.D. capacity is fixed. The analysis yields three implications. First, routine-task AI can increase observable routine output while reducing the diagnostic precision of routine evidence. Second, heterogeneous PI objectives and task complementarity can lead laboratories to adopt different AI strategies, with some emphasizing scalable routine production and others emphasizing mentoring and novel-task augmentation. Third, when elite Ph.D. capacity is fixed, broad improvements in visible records can raise admissions cutoffs rather than expand access proportionally. The simulation reinforces these mechanisms by showing that AI can raise routine output while weakening the link between evaluated scores and latent ability, increasing the risk that high-ability or high-realized-merit candidates are missed. The paper suggests that as routine evidence loses diagnostic content, evaluation should place greater weight on less easily automated forms of contribution, including judgment, interpretation, research design, and process-based evidence.

1. Introduction

Generative AI creates a production–screening tension in early-career research markets. It can make research output easier to produce and harder to interpret. A cleaned dataset, polished table, well-documented script, or clear literature summary may still help a project, but it may reveal less about the researcher's own effort, judgment, and potential. This matters because early-career research output is also a screening device: it helps supervisors, admissions committees, and hiring institutions decide who has research promise. Evidence that generative AI can raise productivity while compressing performance differences or reducing output diversity makes this concern especially salient (Noy and Zhang, 2023; Doshi and Hauser, 2024; Dell'Acqua et al., 2026). This paper studies this production–screening tension in the pre-doctoral academic labor market.

The setting is important because pre-doctoral work has become an increasingly important gateway into elite doctoral training in economics and related quantitative social sciences (Stansbury and Schultz, 2023; Schultz and Stansbury, 2022). Admissions committees allocate scarce training slots under uncertainty about future research ability, while Principal Investigators (PIs) need skilled research labor for data work, coding, literature review, and project development (Conley and Önder, 2014; Card and DellaVigna, 2013; Heckman and Moktan, 2020). Following Mangematin (2000)'s view of doctoral training as both career investment for students and research labor for supervisors, we view pre-doctoral appointments as serving three roles at once: labor input, apprenticeship, and extended signal-generation stage.[1]

Before generative AI, this market already relied on a fragile informational arrangement. Research Assistants (RAs) accepted temporary positions for wages, experience, training, and recommendation capital, while PIs obtained motivated research labor and admissions committees used work records and recommendations to infer latent research potential (Stansbury and Schultz, 2023). Because effort, judgment, mentoring quality, and recommendation credibility are difficult to contract on fully, the PI–RA relationship is not a simple spot-market

E-mail address: swang83@gsu.edu.

[1] Mangematin studies the PhD stage rather than the U.S.-style pre-doctoral stage. We use the analogy only to motivate the broader research-training logic: early research positions can combine current research labor, career investment, and later evaluation.

https://doi.org/10.1016/j.respol.2026.105568
Received 29 October 2025; Received in revised form 16 May 2026; Accepted 18 June 2026
Available online 25 June 2026

exchange (Baker et al., 2002; Levin, 2003). These signals matter at the market level because elite opportunities are scarce and candidates compete not only against an absolute standard but also against one another (Lazear and Rosen, 1981; Hopkins, 2012, 2023).

This paper develops an analytical model of how generative AI changes this arrangement. PIs take an AI capability frontier as given and choose normalized team intensity, AI allocation between routine and novel tasks, and mentoring intensity. RAs differ in ability and choose effort. Routine and novel task outputs generate noisy task-level signals, and admissions committees use a linear evaluation rule to summarize research potential from those signals. Elite PhD admissions are modeled as a fixed-capacity tournament. The model links three objects that are usually considered separately: laboratory production, evaluation from pre-doctoral output, and capacity-constrained selection.

The model yields three main findings. First, routine-task AI can raise observable routine output while reducing the diagnostic precision of routine evidence. Polished routine artifacts become more abundant, but they may reveal less about the RA's own effort and judgment because the artifact contains a larger AI-generated component that evaluators cannot perfectly separate from human contribution. Second, AI can generate segmented laboratory strategies. Quantity-oriented PIs tilt toward scalable routine production, while quality-oriented PIs place greater value on mentoring and novel-task augmentation. Third, fixed-capacity admissions can convert broad improvements in visible records into a higher cutoff rather than broader access. Together, these findings show that AI can improve research production while changing the informational meaning of pre-doctoral output and intensifying competition for scarce academic opportunities.

The analysis distinguishes productivity, learning, and signaling effects. In a bottlenecked research-training market, these effects need not move together. Some AI-enabled gains may increase real research output or human-capital formation; others may mainly raise visible records in a fixed-capacity tournament. The baseline model therefore isolates the production-and-evaluation mechanism, while the mechanism-preserving simulation in Section 5 examines whether the same mechanisms appear when RA talent is distributed, PI orientation is continuous, true merit contains project luck, and admissions scores are noisy. The appendix treats learning, welfare, and project-risk extensions as supporting technical material.

The contemporary U.S.-style economics pre-doctoral market provides the institutional template because it is visible, formalized, and well documented (Stansbury and Schultz, 2023; Abramitzky et al., 2024). The mechanism is not meant to describe all academic labor markets. It applies most directly to settings in which prospective doctoral students work as temporary RAs, supervisors mediate evaluation through recommendations, and access to elite doctoral training is bottlenecked (Câmara et al., 2023; Hopkins, 2012; Lazear and Rosen, 1981).

2. Related literature

This section positions the paper at the intersection of four literatures that map onto the model's main margins: the pre-doctoral labor-market setting, task-based AI in scientific work, the PI–RA training relationship, and evaluation under noisy evidence and fixed capacity. Existing work has studied these margins largely as separate problems. Our model connects them by showing how generative AI changes both the production of early-career research work and the evidentiary value of that work in a bottlenecked selection market.

2.1. Research labor markets, academic pipelines, and capacity constraints

Pre-doctoral work sits between research production and doctoral selection: it provides labor to PIs, training to prospective doctoral students, and evidence for admissions committees. This setting is shaped by uncertainty about future research productivity and by bottlenecks in elite academic placement (Conley and Önder, 2014; Card and DellaVigna, 2013; Ellison, 2002; Heckman and Moktan, 2020). The growth of the economics pre-doctoral stage can be understood as institutional responses from Universities to these frictions, with consequences for access and sorting into the profession (Stansbury and Schultz, 2023; Schultz and Stansbury, 2022). These studies motivate our treatment of the pre-doctoral market as a competitive pipeline in which scarce elite opportunities are allocated under uncertainty.

The setting also connects to the literature on doctoral labor markets. Mangematin (2000) treats PhD training as both students' investment in future careers and supervisors' source of research labor. Our pre-doctoral setting plays an analogous role one step earlier: current research work is tied to later selection and professional trajectories. This motivates two core objects in the model: the RA's continuation value of elite PhD admission and the PI's demand for research labor.

The same literature motivates our treatment of PI heterogeneity. Concentrated rewards for elite placement and high-impact output motivate the quality-oriented side of the model (Card and DellaVigna, 2013; Heckman and Moktan, 2020; Manso, 2011). Publication pressure and measurable-output incentives motivate the quantity-oriented side (Hamermesh, 2013; Edwards and Roy, 2017). Work on scientific competition further shows that evaluation pressure can generate quality distortions or inefficient strategic behavior (Kapeller and Steinerberger, 2016; Hill and Stein, 2025). We therefore model quality-oriented and quantity-oriented PIs as facing different research incentives, rather than as morally different types.

2.2. AI in science, discovery, and task-based research production

The second literature motivates the production side of the model. We treat generative AI as task-specific rather than as a uniform productivity shock. This follows the routine/nonroutine task tradition in Autor et al. (2003) and the task-based automation framework in Acemoglu and Restrepo (2018, 2019, 2020, 2022, 2024). The core insight is that technologies affect activities differently: they may substitute for human labor in some tasks, complement it in others, and create new tasks. This task-based view motivates our distinction between routine and novel research work.

Research on AI in science sharpens this task-based framing. Emerging technologies are characterized not only by novelty and rapid growth, but also by uncertainty and the capacity to reorganize existing practices (Rotolo et al., 2015). AI has similarly been described as an emerging general method of invention that reshapes discovery processes and the organization of scientific work (Bianchini et al., 2022). In a formal model of scientific discovery, Agrawal et al. (2024) show that AI can improve prioritized search over hypothesis spaces, but that the value of better search depends on complementary testing capacity. These studies support treating AI in science as a reorganization of research practice, not as a uniform productivity shifter.

Recent generative-AI evidence motivates the model's distinction between automation, augmentation, and diagnostic compression. AI can increase speed and quality in structured writing and knowledge-work tasks, but its effects vary across tasks and workers (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Dell'Acqua et al., 2026). Evidence that AI can reduce output diversity or compress observable performance differences motivates the possibility that AI-assisted output may become less informative about individual contribution (Doshi and Hauser, 2024; Noy and Zhang, 2023). This supports the model's separation between the productive value of AI-assisted output and its evidentiary value for evaluation.

The model also treats novel research work as uncertain and partly nonroutine. Novel tasks may include problem formulation, research design, hypothesis generation, interpretation, and judgment. This treatment follows problem-solving models of knowledge production (Garicano, 2000), research on exploratory innovation (Manso, 2011), and

recent work that characterizes AI in science as reshaping discovery, hypothesis search, and methods of invention (Bianchini et al., 2022; Agrawal et al., 2024). It is also consistent with evidence that scientific value can emerge outside initially targeted categories: Aslan et al. (2024) document substantial unexpectedness in NIH-funded medical research.[2] This supports modeling breakthrough-oriented output as upper-tailed and not fully captured by routine, pre-specified tasks.

Finally, the CES research technology gives the routine–novel distinction a production interpretation. The CES form represents different degrees of substitutability across productive inputs (Arrow et al., 1961). When routine and novel outputs are relatively substitutable, routine execution can partly compensate for weaker novel input. When they are relatively complementary, routine execution is valuable only when paired with research design, interpretation, and judgment. The complementarity case is related to O-ring production logic, where complex output depends on multiple tasks being performed well together (Kremer, 1993). It is also consistent with innovation and R&D studies that use flexible production structures to examine whether research inputs are substitutes or complements (Ceccagnoli et al., 2014; Growiec et al., 2023). In our setting, the elasticity parameter governs whether AI-enabled routine scaling can support research production on its own, or whether its value depends on mentored novel contribution.

2.3. Mentoring, apprenticeship, and research organizations

The third literature motivates the PI–RA training relationship. Pre-doctoral research work is not a simple spot-market labor exchange. Many relevant margins are difficult to contract on directly: RA initiative, care in execution, PI guidance, feedback quality, and the interpretation of RA performance. Relational-contract models provide useful background for treating effort, supervision, and communication as meaningful margins under incomplete contracting (Baker et al., 2002; Levin, 2003; Board and Meyer-Ter-Vehn, 2015; Watson, 2021). In the baseline model of this paper, this literature supports mentoring and task assignment; it does not require a full dynamic reputation model.

The apprenticeship literature further motivates mentoring as more than a symbolic input. Chassang (2010) shows how routines and cooperation can be learned under incomplete contracting. Fudenberg and Rayo (2019) and Kostadinov and Kuvalekar (2022) model how training, effort, and future capability evolve together. This supports the paper's learning extension and the interpretation of mentoring as an input into both current performance and future human capital. It also clarifies why AI may change not only which tasks RAs perform, but also the value of supervision and guided learning.

Two related perspectives qualify the baseline model without becoming additional mechanisms. First, absorptive capacity suggests that a common technological frontier need not imply equal effective AI capability across laboratories (Cohen and Levinthal, 1990).[3] Second, research productivity is embedded in collaboration structures and cross-community ties (Rotolo and Messeni Petruzzelli, 2013). These perspectives clarify why mentoring, research context, and organizational capacity may matter when routine artifacts become easier to produce with AI.

2.4. Evaluation, signaling, cumulative advantage, and tournaments

The fourth literature motivates the information and allocation side of the model. Classic signaling and adverse-selection models show why evaluators infer latent quality from imperfect observables (Akerlof, 1970; Spence, 1973). Because pre-doctoral work generates graded task evidence rather than a binary credential, models of graded and noisy signals are especially relevant (Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021). This literature motivates our use of linear prediction weights based on task-level informativeness, rather than fixed exogenous weights.

Recommendation environments add credibility concerns. Câmara et al. (2023) show how reputation and competition shape labor-market recommendations. We use this insight in reduced form through a PI-environment credibility adjustment.[4]

This evaluation problem also connects to research assessment. Responsible-assessment frameworks warn against treating easily measured outputs as direct proxies for quality (Hicks et al., 2015; Cagan, 2013). Performance-based research funding can shape behavior through prestige and ranking incentives (Hicks, 2012), and structurally diverse teams can be disadvantaged in ex ante evaluation despite stronger ex post performance (Banal-Estañol et al., 2019). These studies support the model's concern that shifting away from routine evidence toward more complex evidence of judgment may make evaluation more difficult, not automatically better.

Fixed-capacity competition links the evaluation problem to ranking. Tournament models motivate our treatment of admissions as relative allocation under scarce positions (Lazear and Rosen, 1981; Hopkins, 2012, 2023). This links the model to broader evidence on publication bottlenecks and escalating academic standards (Card and DellaVigna, 2013; Ellison, 2002; Heckman and Moktan, 2020). Science studies also emphasizes cumulative advantage: recognition and resources can compound over time (Merton, 1968; Bol et al., 2018), and institutional position can affect grant allocation (Viner et al., 2004). We do not model this dynamic directly. The literature helps interpret one implication of the model: when routine evidence becomes less diagnostic, evaluators may rely more heavily on contextual cues such as PI credibility, institutional affiliation, or prior placement history.

Together, these literatures motivate the paper's mechanism: generative AI changes task production inside laboratories; task outputs become noisy evidence about early-career researchers; and scarce admissions capacity converts that evidence into competitive outcomes. What remains underdeveloped is the link among these margins. Existing work does not explain how the same AI-assisted output can be productive for research, less informative for evaluation, and strategically consequential under fixed capacity. The model provides that link.

3. The model

This section defines the primitives of the model. We focus on a stylized version of the contemporary U.S.-style pre-doctoral academic labor market, where temporary RA appointments, recommendation-based evaluation, and competition for a fixed number of elite PhD positions are especially salient (Stansbury and Schultz, 2023). The model captures a setting in which PIs organize research work, RAs generate observable performance, and admissions committees infer latent research potential from imperfect evidence.

In the baseline model, each PI takes the AI capability frontier $K \geq 0$ as given.[5] The parameter K summarizes the external AI environment available for research work. A larger K means that AI systems are more

[2] Aslan et al. study medical research funding rather than pre-doctoral labor markets. We use their finding to motivate the broader point that scientific value may emerge outside pre-specified categories, which supports the model's treatment of novel research output as uncertain and upper-tailed.

[3] A simple extension would write effective AI capability as $K_i = A_i K$, where A_i captures laboratory-level absorptive capacity. The baseline holds K common to isolate the allocation and evaluation mechanisms.

[4] The parameter r_λ is a reduced-form credibility adjustment, not an endogenous reputation stock. A dynamic model in which evaluators update PI credibility from later candidate performance is a natural extension but is outside the baseline.

[5] The common-frontier assumption should be read as a baseline device. It does not imply equal effective AI capability across laboratories. Differences in infrastructure, data access, training, and verification routines could be represented by laboratory-specific effective capability $K_i = A_i K$.

capable for tasks such as search, coding, drafting, and data analysis. The baseline does not model PI-level AI adoption or investment. Instead, it studies how laboratories allocate a given AI frontier across research tasks and mentoring choices. This convention follows the task-based view that new technologies affect the productivity and feasibility of particular tasks, rather than acting only as uniform total-factor-productivity shocks (Autor et al., 2003; Acemoglu and Restrepo, 2018, 2019; Agrawal et al., 2019; Korinek, 2023).

A PI has type $\lambda \in \{\lambda_Q, \lambda_N\}$, where λ_Q denotes a quality-oriented PI and λ_N denotes a quantity-oriented PI. Each PI chooses $(n_\lambda, \alpha_\lambda, g_\lambda)$. The first choice, $n_\lambda \geq 0$, is normalized team intensity: a continuous measure of filled RA labor operated by the PI in equilibrium, rather than a literal headcount or a stock of posted vacancies. Participation and sorting determine which RA types fill those positions; unfilled postings are outside the baseline model. The second choice, $\alpha_\lambda \in [0,1]$, is the share of the available AI frontier allocated to routine tasks. The third choice, $g_\lambda \geq 0$, is mentoring intensity, which captures the PI's time and effort devoted to supervision, feedback, and guidance.

Given K and α_λ, routine-task AI is

$$K^T_{R,\lambda} \equiv \alpha_\lambda K,$$

and novel-task AI is

$$K^T_{N,\lambda} \equiv (1-\alpha_\lambda)K.$$

Thus,

$$K^T_{R,\lambda} \in [0,K], \qquad K^T_{N,\lambda} \in [0,K], \qquad K^T_{R,\lambda} + K^T_{N,\lambda} = K.$$

The superscript T denotes task allocation; the subscripts R and N refer to routine and novel tasks, respectively.[6]

The model follows a pre-doctoral RA through this environment. A PI first chooses filled team intensity, AI allocation, and mentoring. The RA then enters the lab with an underlying research ability and chooses effort. Routine and novel tasks generate outputs, which enter the PI's research production and also create evidence about the RA. Admissions committees observe noisy task-level signals, evaluate research potential, and allocate a fixed mass of elite PhD slots through a capacity-constrained admissions stage.

The production side and the evaluation side are deliberately separated. On the production side, routine and novel task outputs are combined through a CES research technology, allowing the degree of substitutability between structured execution and novel judgment to differ across research environments (Acemoglu and Restrepo, 2018; Kremer, 1993). On the evaluation side, admissions committees place weight on observed task performance according to its informativeness, rather than according to fixed exogenous signal weights. This information structure is motivated by models of noisy signals, market inference, and tournament-like allocation under scarcity (Daley and Green, 2014; Hopkins, 2012). Learning and welfare are introduced later as extensions in the appendix. Fig. 1 provides an overview of the mechanism, and Table 1 summarizes the notation.

3.1. Players

The model consists of three types of players.

- **Principal Investigators (PIs).** The population of PIs is normalized to have total mass one. PIs organize research projects, supervise RAs, provide mentoring, and write recommendations. These roles involve non-contractible margins such as feedback quality, task assignment, and the interpretation of RA performance (Baker et al., 2002; Levin, 2003). In the baseline model, these margins motivate the PI's choices over team organization, AI allocation, and mentoring.
  PIs differ in their research objectives. A fraction $\mu \in (0,1)$ are quality-oriented, denoted by type λ_Q, and place greater weight on high-value or breakthrough research outcomes. The remaining fraction $1-\mu$ are quantity-oriented, denoted by type λ_N, and place greater weight on scalable research throughput. This distinction is a reduced-form representation of different research incentives, not a moral classification. For compact notation, let $\mu_{\lambda_Q} \equiv \mu$ and $\mu_{\lambda_N} \equiv 1-\mu$, so that μ_λ denotes the mass of PIs of type λ.
- **Research Assistants (RAs).** RAs enter the model as prospective doctoral students working in temporary pre-doctoral research positions. The pre-doctoral stage is modeled as a transitional period before PhD admissions: it provides labor input to the PI's research projects, but it also creates training opportunities and evidence about the RA's research potential. This treatment follows a broader view of early research training as a transitional labor-market stage in which current research work is tied to later professional trajectories (Mangematin, 2000). It is also consistent with human-capital theory and apprenticeship-style models in which effort, supervision, and learning evolve together during early-career training (Becker, 1962; Fudenberg and Rayo, 2019; Kostadinov and Kuvalekar, 2022).
  Each RA has underlying research ability
  $$\theta \sim \mathrm{TN}(\mu_\theta, \sigma^2_\theta; \theta_L, \theta_H), \qquad 0 < \theta_L < \theta_H < \infty,$$
  which is private information. Here $\mathrm{TN}(\cdot)$ denotes a normal distribution truncated to the compact ability support $[\theta_L, \theta_H]$.[7] During the pre-doctoral period, the RA chooses continuous effort $e \in [0,\infty)$. Ability is not directly observed by the admissions market, so it must be inferred from subsequent task-level performance and other evidence generated during the appointment (Conley and Önder, 2014; Spence, 1973; Daley and Green, 2014).
- **The Market.** The "Market" represents the collective of top-tier PhD admissions committees. It does not observe θ directly. Instead, it observes noisy task-level signals generated in the PI's lab and evaluates the candidate's research potential from those signals. This information structure follows signaling models and models of graded noisy evidence (Spence, 1973; Daley and Green, 2014; Heinsalu, 2018). Because admission is to a fixed number of elite PhD positions, the market also has the structure of a rank-order tournament: relative standing matters in addition to absolute performance (Lazear and Rosen, 1981; Hopkins, 2012). The admissions stage therefore combines signal extraction with relative ranking under scarcity.

3.2. Research production, observed signals, and AI

We model research as a set of tasks rather than as a single undifferentiated output. This choice reflects the task-based view of technological change: new technologies affect different activities in different ways, rather than simply raising productivity everywhere by the same amount (Autor et al., 2003; Acemoglu and Restrepo, 2018, 2019, 2020, 2024). We distinguish between two kinds of research work:

- **Routine tasks** (T_R): structured tasks such as literature search, data cleaning, standardized regressions, code debugging, formatting, and other forms of execution.
- **Novel tasks** (T_N): less structured tasks such as hypothesis generation, research design, causal reasoning, problem formulation, and interpretation of ambiguous findings.

[6] The subscript N in $K^T_{N,\lambda}$ denotes novel tasks, while λ_N denotes quantity-oriented PIs.

[7] The compact support is a production-domain regularity condition. It keeps task outputs and effort-capacity terms well defined on the relevant equilibrium domain.

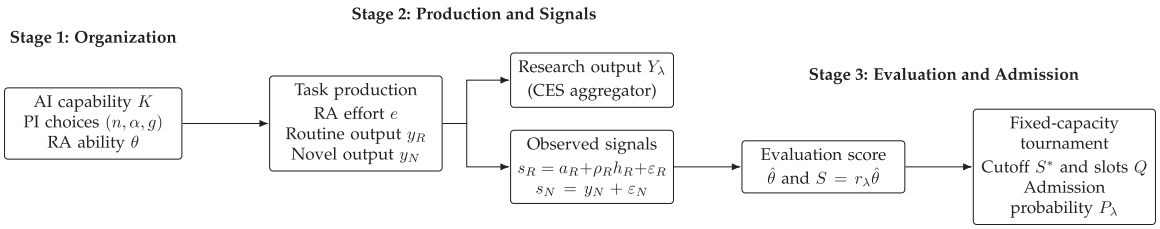

Fig. 1. Mechanism overview. Stage 1 combines the exogenous AI capability frontier K, PI organizational choices, and RA heterogeneity. Stage 2 separates into two branches. On the production side, RA effort and task outputs generate research output Y_λ through the CES aggregator. On the evaluation side, the market observes only noisy signals (s_R, s_N). Stage 3 converts the resulting evaluation score into admissions outcomes through a fixed-capacity cutoff S^*. The distinction between the production branch and the evaluation branch is a central feature of the model. This is a schematic overview of model objects, not a calibrated quantitative result.

Table 1
Core notation.

Symbol	Definition	Domain
Agents and types		
λ	PI type	$\{\lambda_Q, \lambda_N\}$
λ_Q	Quality-oriented PI type	–
λ_N	Quantity-oriented PI type	–
μ_λ	Mass of PIs of type λ	$\mu_{\lambda_Q} = \mu,\ \mu_{\lambda_N} = 1 - \mu$
Θ_λ	Ability set assigned to type-λ PIs	Subset of $[\theta_L, \theta_H]$
RA characteristics and choices		
θ	RA ability	$\mathrm{TN}(\mu_\theta, \sigma_\theta^2; \theta_L, \theta_H)$
e	RA effort	$\mathbb{R}_+$
$\bar{U}$	RA outside option	$\mathbb{R}$
V	Continuation value of elite PhD admission	$\mathbb{R}_+$
PI choices and technology		
K	Exogenous AI capability frontier	$\mathbb{R}_+$
n_λ	Filled normalized team intensity/mass of filled RA positions	$\mathbb{R}_+$
α_λ	Share of AI allocated to routine tasks	$[0, 1]$
g_λ	Mentoring intensity	$\mathbb{R}_+$
$K_{R,\lambda}^T$	Routine-task AI allocation	$\alpha_\lambda K$
$K_{N,\lambda}^T$	Novel-task AI allocation	$(1 - \alpha_\lambda)K$
Human-input terms and task technology		
h_R	Human input into routine tasks	$\theta + \eta_R e$
h_N	Mentored human input into novel tasks	$\theta + \eta_N e + \psi g_\lambda$
$A_R(\cdot)$	Inseparable AI-generated component of routine artifact	Mean zero
$\sigma_A^2(\cdot)$	Variance of the inseparable AI-generated component	Increasing in K_R^T
σ_h^2, σ_v^2	Residual human-input variance and artifact-noise variance	$\mathbb{R}_{++}$
$\rho_R(\cdot)$	Derived diagnostic loading of routine evidence	$(0, 1]$
$\kappa_\lambda(\cdot)$	Novel-task AI augmentation function	$(0, \infty)$
Outputs, signals, and admissions		
y_R, y_N	Deterministic routine and novel task outputs	$\mathbb{R}_+^2$
s_R, s_N	Market-observed noisy task signals	$\mathbb{R}^2$
$\hat{\theta}$	Linear evaluation of research potential	$\mathbb{R}$
S	Admissions-relevant score	$\mathbb{R}$
r_λ	Reduced-form credibility adjustment for PI environment	$(0, \infty)$
S^*	Fixed-capacity admissions cutoff	$\mathbb{R}$
$P_\lambda(\theta, e; S^*)$	Admission probability	$[0, 1]$
Q	Mass of elite PhD slots	$\mathbb{R}_+$
Research production		
Y_λ	CES research output	$\mathbb{R}_+$
$\bar{Y}_\lambda$	Expected per-RA CES research output in segment λ	$\mathbb{R}_+$
ς_λ	Elasticity of substitution between routine and novel outputs	$\mathbb{R}_+$
$\bar{p}_{B,\lambda}$	Average probability of breakthrough success	$[0, 1]$
Score-load coefficients		
A_λ^B	Deterministic component of the admissions score	$\mathbb{R}$
B_λ^B	Effort loading in the admissions score	$\mathbb{R}_+$
Σ_λ^B	Residual standard deviation of the admissions score	$\mathbb{R}_+$

Define human input for the production function as

$$h_R \equiv \theta + \eta_R e, \qquad h_N \equiv \theta + \eta_N e + \psi g,$$

where $\eta_R, \eta_N, \psi > 0$. The term h_R is the human input into routine tasks, while h_N is the mentored human input into novel tasks. The term ψg captures the incremental contribution of PI mentoring to novel-task human input, where g is mentoring intensity and $\psi > 0$ measures how effectively mentoring translates into research design, interpretation, and other judgment-intensive contributions.[8] Both h_R

[8] This reduced-form specification follows human-capital and apprenticeship models in which ability, effort, and training jointly shape current performance and future capability (Becker, 1962; Fudenberg and Rayo, 2019; Kostadinov

and h_N are reduced-form human-input terms rather than separately observed objects.

Following the task-based framework pioneered by Autor et al. (2003), Acemoglu and Autor (2011) and Acemoglu and Restrepo (2022), we model the output of routine tasks as a linear combination of AI's direct automation value and human contribution:

$$y_R = a_R(K_R^T) + m_R(K_R^T)h_R. \tag{1}$$

The first term, $a_R(K_R^T) \geq 0$, is the direct automation value of routine-task AI, with $a_R'(\cdot) > 0$. The second term, $m_R(K_R^T)h_R$, is the human contribution to routine output. We assume $m_R(K_R^T) > 0$ and $m_R'(\cdot) \geq 0$, so routine-task AI weakly increases the productive contribution of human routine input. The key modeling distinction in the paper is that productive routine output and evidentiary value need not move together.

The admissions market does not observe the RA's human input h_R directly. We summarize routine-task evidence by the reduced-form signal

$$s_R = a_R(K_R^T) + \rho_R(K_R^T)h_R + \varepsilon_R, \qquad \varepsilon_R \sim N(0, \sigma_R^2). \tag{2}$$

Here $a_R(K_R^T)$ is the observable automation component of routine-task output. It can raise the surface quality of routine artifacts, such as cleaned data, code, tables, or written summaries, but it is not itself diagnostic of the RA's own ability or effort. The term $\rho_R(K_R^T) > 0$ is the diagnostic loading of routine evidence on the RA's human input.[9] This loading differs from $m_R(K_R^T)$ in the production function: $m_R(\cdot)$ governs the productive contribution of human input to routine output, whereas $\rho_R(\cdot)$ governs the evidentiary value of routine output for evaluation. The noise term ε_R captures residual measurement and attribution uncertainty in the reduced-form signal. In the baseline, σ_R^2 is held fixed to isolate the loading channel; if residual signal noise also varied with K_R^T, the corresponding precision term would be $\rho_R(K_R^T)^2/\sigma_R^2(K_R^T)$.

This signal structure follows models in which evaluators infer latent quality from imperfect or graded evidence rather than observing ability and effort directly (Spence, 1973; Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021). It also makes precise why a polished table, clean code, or well-written summary may remain useful for the project while becoming less informative for evaluation: the routine artifact contains more AI-generated variation that cannot be perfectly attributed to the RA.[10]

Similarly, novel task output is

$$y_N = \kappa_\lambda(K_N^T)h_N, \tag{3}$$

where $\kappa_\lambda(0) = 1$ and $\kappa_\lambda'(K_N^T) > 0$. The function κ_λ captures AI augmentation in novel tasks. AI may help with ideation, coding support, feedback, interpretation, and problem formulation, but its value still depends on human judgment and mentoring. This view is consistent with problem-solving models of knowledge production, exploratory innovation, and recent work on generative AI in research and human–AI collaboration (Garicano, 2000; Manso, 2011; Korinek, 2023; Fügener et al., 2025).[11]

The market observes the routine signal in (2) and the novel signal

$$s_N = y_N + \varepsilon_N, \qquad \varepsilon_N \sim N(0, \sigma_N^2). \tag{4}$$

The noise terms capture luck, measurement error, and evaluative uncertainty. This follows stochastic production and noisy-signal models in which evaluators infer quality from imperfect observables (Just and Pope, 1978; Daley and Green, 2014; Heinsalu, 2018).

To capture how routine execution and novel judgment combine in research production, we use a CES aggregator:

$$Y_\lambda = \left[\tau y_R^{\frac{\varsigma_\lambda - 1}{\varsigma_\lambda}} + (1-\tau) y_N^{\frac{\varsigma_\lambda - 1}{\varsigma_\lambda}}\right]^{\frac{\varsigma_\lambda}{\varsigma_\lambda - 1}}, \qquad \tau \in (0,1),\ \varsigma_\lambda > 0. \tag{5}$$

The CES form represents different degrees of substitutability across productive inputs (Arrow et al., 1961). When $\varsigma_\lambda = 1$, (5) is interpreted by its Cobb–Douglas limit. When $\varsigma_\lambda > 1$, routine and novel outputs are relatively substitutable: additional routine execution can partly compensate for weaker novel input. When $\varsigma_\lambda < 1$, routine and novel outputs are relatively complementary: routine execution is valuable, but it cannot easily replace research design, interpretation, or original judgment. The complementarity case is related to O-ring production logic and to models of AI-assisted scientific discovery in which better hypothesis generation requires complementary testing and interpretation capacity (Kremer, 1993; Agrawal et al., 2024). Innovation and R&D studies also use CES or flexible production structures to study whether research inputs are substitutes or complements (Ceccagnoli et al., 2014; Growiec et al., 2023).[12]

The section therefore separates two objects that are often conflated. Task outputs (y_R, y_N) enter the PI's research production technology. Observed signals (s_R, s_N) enter the admissions market's evaluation problem. This distinction allows the model to capture both the productive value of AI-assisted work and the changing evidentiary value of that work for candidate evaluation (Acemoglu and Restrepo, 2018; Daley and Green, 2014; Bao and Duffy, 2021).

---

and Kuvalekar, 2022). The distinction between routine and novel human inputs is also consistent with the task-based view that different activities rely on different combinations of human input and technology (Autor et al., 2003; Acemoglu and Restrepo, 2018, 2019).

[9] A simple artifact-level microfoundation delivers a decreasing diagnostic loading. Suppose that, after conditioning on task allocation and deterministic routine-production scale, the evaluator observes a normalized routine artifact $z_R = h_R + A_R(K_R^T) + v_R$, $A_R(K_R^T) \sim N\left(0, \sigma_A^2(K_R^T)\right)$, $v_R \sim N(0, \sigma_v^2)$, where $A_R(K_R^T)$ is an inseparable AI-generated component and v_R is ordinary artifact noise. In the local residualized benchmark, $A_R(K_R^T)$, v_R, and the residual innovation in h_R are mutually independent. This representation follows the standard signal-extraction logic in which evaluators infer latent quality from noisy evidence rather than directly observing ability, effort, or contribution (Spence, 1973; Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021). Let $\sigma_h^2 > 0$ denote the local residual variance of h_R. If routine-task AI increases the variance of the inseparable AI-generated component, $\frac{d\sigma_A^2(K_R^T)}{dK_R^T} > 0$, then the best linear predictor of h_R from z_R places weight $\rho_R(K_R^T) = \frac{\sigma_h^2}{\sigma_h^2 + \sigma_A^2(K_R^T) + \sigma_v^2}$. Differentiating gives $\frac{d\rho_R(K_R^T)}{dK_R^T} = -\frac{\sigma_h^2 \, d\sigma_A^2(K_R^T)/dK_R^T}{\left[\sigma_h^2 + \sigma_A^2(K_R^T) + \sigma_v^2\right]^2} < 0$. Thus, $\rho_R'(K_R^T) < 0$ follows from the attribution-risk primitive that routine-task AI raises the variance of the inseparable non-human component of the observed artifact.

[10] The microfoundation is motivated by evidence that generative AI can raise output quality while compressing observable performance differences, reducing output diversity, or making individual contribution harder to separate from tool assistance in structured tasks (Noy and Zhang, 2023; Doshi and Hauser, 2024; Dell'Acqua et al., 2026; Brynjolfsson et al., 2025). It is not a claim that AI always reduces informativeness. In some settings, AI may improve documentation, traceability, or auditability; such process evidence would reduce the effective $\sigma_A^2(K_R^T)$ or add an independent signal.

[11] This treatment is also consistent with evidence that scientific projects can generate outputs outside initially targeted categories (Aslan et al., 2024), which motivates the uncertain and upper-tailed interpretation of novel research output. The mechanism-preserving simulation uses the linear approximation $\kappa_\lambda(K_N^T) = 1 + \phi_\lambda K_N^T$. The theoretical results do not depend on this linear form. We maintain $y_R > 0$ and $y_N > 0$ on the equilibrium support so that the CES aggregator is well defined.

[12] The elasticity ς_λ should be read as a reduced-form feature of the research environment rather than as a pure PI preference parameter. PI objectives may be correlated with research environments, but the model does not require preferences and physical task technology to be identical.

3.3. Evaluation from noisy signals

The admissions market does not observe an RA's research ability directly. Ability θ is private information, effort e is not directly contractible, and admissions committees observe only noisy evidence generated by the RA's work. This places the model in a standard setting of evaluation under imperfect information: evaluators infer latent quality from observed but noisy signals (Akerlof, 1970; Spence, 1973; Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021).

To keep the inference problem tractable, we use a linear-Gaussian benchmark. With a Gaussian prior and Gaussian noise, the posterior mean is a linear, precision-weighted average of the prior mean and observed signals (DeGroot, 1970; Gelman et al., 2013). Because ability has bounded support in the main model, the linear formula is exact only in the corresponding unbounded Gaussian benchmark. Under bounded support, we interpret it as a best linear prediction rule. The online appendix derives the exact truncated-normal posterior and shows that the main informativeness result is unchanged.[13]

The market is assumed to condition on the PI environment, the prevailing AI frontier, and the task-level AI allocation when forming its evaluation.[14] Define the centered routine signal as

$$\begin{aligned}\tilde{s}_R &\equiv s_R - a_R(K_R^T) - \rho_R(K_R^T)\mu_\theta \\ &= \rho_R(K_R^T)\big(\theta - \mu_\theta + \eta_R e\big) + \varepsilon_R. \end{aligned} \tag{6}$$

Similarly, define the centered novel signal as

$$\begin{aligned}\tilde{s}_N &\equiv s_N - \kappa_\lambda(K_N^T)\mu_\theta \\ &= \kappa_\lambda(K_N^T)\big(\theta - \mu_\theta + \eta_N e + \psi g\big) + \varepsilon_N. \end{aligned} \tag{7}$$

Neither centered signal is a pure measure of ability. Both combine ability, effort, mentoring, and noise. The weights below should therefore be read as benchmark linear prediction weights after deterministic effort and mentoring components are residualized, or equivalently as the market's best linear evaluation rule in the maintained linearized environment.[15]

The signal noises are assumed to be independent:

$$\varepsilon_R \sim N(0, \sigma_R^2), \qquad \varepsilon_N \sim N(0, \sigma_N^2).$$

Under the linear-Gaussian benchmark, the market's summary evaluation is

$$\hat{\theta} = \mu_\theta + \omega_R^B(\alpha, K, \lambda)\tilde{s}_R + \omega_N^B(\alpha, K, \lambda)\tilde{s}_N. \tag{8}$$

Here $\hat{\theta}$ is an admissions evaluation of research potential, written in ability units. It is the benchmark linear prediction rule used in the main text.

The coefficient on the routine signal is

$$\omega_R^B(\alpha, K, \lambda) = \frac{\rho_R(K_R^T)/\sigma_R^2}{\frac{1}{\sigma_\theta^2} + \frac{\rho_R(K_R^T)^2}{\sigma_R^2} + \frac{\kappa_\lambda(K_N^T)^2}{\sigma_N^2}}, \tag{9}$$

and the coefficient on the novel signal is

$$\omega_N^B(\alpha, K, \lambda) = \frac{\kappa_\lambda(K_N^T)/\sigma_N^2}{\frac{1}{\sigma_\theta^2} + \frac{\rho_R(K_R^T)^2}{\sigma_R^2} + \frac{\kappa_\lambda(K_N^T)^2}{\sigma_N^2}}. \tag{10}$$

These coefficients follow from the informativeness of the two signal channels. Routine evidence contributes precision $\rho_R(K_R^T)^2/\sigma_R^2$, while novel evidence contributes precision $\kappa_\lambda(K_N^T)^2/\sigma_N^2$. A signal is more informative when its loading on human contribution is larger and its noise variance is smaller. The coefficients in (9)–(10) are the corresponding linear prediction coefficients, adjusted for signal scale and prior uncertainty (Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021).

The key implication is simple. If routine-task AI makes routine evidence less diagnostic of the RA's own contribution, then routine evidence becomes less useful for evaluation. In the limiting case,

$$\lim_{\rho_R(K_R^T)\to 0} \omega_R^B(\alpha, K, \lambda) = 0.$$

Routine output may still be valuable for producing research, but it becomes less useful for evaluating the RA.

The admissions-relevant score scales the market's evaluation by the PI-environment credibility adjustment:

$$S = r_\lambda \hat{\theta}. \tag{11}$$

This term captures the idea that evaluators may interpret otherwise similar task-level evidence differently depending on the context in which it was produced (Câmara et al., 2023).[16]

The model focuses on pre-doctoral research evidence because that is the channel most directly affected by AI-assisted research work. Other admissions evidence, such as coursework, grades, standardized preparation, or additional letters, could be represented as additional noisy signals in the evaluation rule. Such signals would add independent precision and may dampen the effect of diagnostic compression in routine research evidence. They do not eliminate the mechanism whenever pre-doctoral research output remains an important signal of research potential (Spence, 1973; Daley and Green, 2014).

3.4. Payoffs

The RA payoff captures the trade-off between wages, effort costs, and the value of admission to an elite PhD program. The PI payoff captures the value of research output, the cost of organizing and mentoring RAs, and the difference between quantity-oriented and quality-oriented research environments.

RA payoffs. Pre-doctoral wages are treated as fixed in the short run.[17] Let

$$w = \bar{w}$$

denote the wage paid to an RA. We also let $\bar{U}$ denote the RA's outside option, which enters the participation constraint below. Changes in compensation and outside opportunities can therefore be represented by shifts in $\bar{w}$ and $\bar{U}$, respectively. A higher $\bar{w}$ raises the RA's payoff but also increases the PI's labor cost, while a higher $\bar{U}$ tightens the participation constraint. These shifts can affect participation and organizational scale, but they are not the source of the production–evaluation mechanism studied here.

Following the quadratic effort-cost specification used in multitask and relational-contract models in Holmstrom and Milgrom (1991), Baker et al. (2002), the RA incurs a convex effort cost:

$$c_\lambda(e; \theta, K_{R,\lambda}^T) = \frac{e^2}{2\chi(\theta, K_{R,\lambda}^T)}, \tag{12}$$

[13] Truncation adds a nonlinear boundary correction to the posterior mean. This correction matters near the lower and upper support boundaries, but it does not change the limiting result that routine evidence ceases to affect the posterior as $\rho_R(K_R^T) \to 0$.

[14] This is a benchmark information assumption. If evaluators are uncertain about the laboratory environment or AI allocation, that uncertainty can be represented as additional measurement noise in the task-level signals.

[15] The online appendix defines the residualized benchmark environment used for the technical derivation, without assuming that evaluators directly observe effort.

[16] The main signal-weight mechanism does not require variation in r_λ. Setting $r_\lambda = 1$ leaves the production-evaluation mechanism unchanged. The parameter is a reduced-form credibility adjustment, not an endogenous reputation stock.

[17] This is a short-run institutional simplification. Many pre-doctoral positions are posted with standardized salary bands, and the paper focuses on effort, task allocation, and evaluation rather than wage bargaining.

where $K^T_{R,\lambda} = \alpha_\lambda K$ is the amount of AI capability allocated to routine tasks in a type-λ laboratory. The function $\chi(\theta, K^T_{R,\lambda}) > 0$ is an effective effort-capacity term. We assume

$$\frac{\partial \chi(\theta, K^T_{R,\lambda})}{\partial \theta} > 0, \qquad \frac{\partial \chi(\theta, K^T_{R,\lambda})}{\partial K^T_{R,\lambda}} > 0.$$

Thus, higher-ability RAs face lower effective effort costs, and routine-task AI can make effort more productive in structured research work. This cost channel is separate from the signal-informativeness channel: routine-task AI may make effort easier while also making routine artifacts less diagnostic.

Using the evaluation rule in Section 3.3, the admissions-relevant score can be written in reduced form as

$$S = A^B_\lambda(\theta; K, \alpha_\lambda, g_\lambda) + B^B_\lambda(K, \alpha_\lambda, \lambda)e + \Sigma^B_\lambda(K, \alpha_\lambda, \lambda)\xi, \qquad \xi \sim N(0,1). \quad (13)$$

Here A^B_λ is the deterministic part of the score, B^B_λ is the marginal loading of RA effort into the admissions score, and $\Sigma^B_\lambda > 0$ is residual score noise. Closed-form expressions are reported in the online appendix.

Given an admissions cutoff S^*, the admission probability of an RA with ability θ and effort e in a type-λ laboratory is

$$P_\lambda(\theta, e; S^*) = 1 - \Phi\left(\frac{S^* - A^B_\lambda(\theta; K, \alpha_\lambda, g_\lambda) - B^B_\lambda(K, \alpha_\lambda, \lambda)e}{\Sigma^B_\lambda(K, \alpha_\lambda, \lambda)}\right). \quad (14)$$

The RA's utility is

$$U_{RA,\lambda}(\theta, e; S^*) = \bar{w} - \frac{e^2}{2\chi(\theta, K^T_{R,\lambda})} + \beta_{RA} V P_\lambda(\theta, e; S^*), \quad (15)$$

where $\beta_{RA} \in (0,1]$ is the RA's discount factor and $V > 0$ is the continuation value of elite PhD admission. The continuation value V captures the career value of successful placement, consistent with research on early research training and professional trajectories (Mangematin, 2000). The RA chooses effort according to

$$e^*_\lambda(\theta) \in \arg\max_{e \geq 0} U_{RA,\lambda}(\theta, e; S^*). \quad (16)$$

The online appendix gives sufficient regularity conditions under which this problem has a unique optimal solution in the maintained region. The solution may be interior or at the boundary; outside the maintained region, the effort choice should be interpreted as an optimal-effort correspondence.

Let

$$\widetilde{U}_\lambda(\theta) \equiv \max_{e \geq 0} U_{RA,\lambda}(\theta, e; S^*) \quad (17)$$

denote the RA's indirect utility in a type-λ environment. The RA participates only if

$$\widetilde{U}_\lambda(\theta) \geq \bar{U}, \quad (18)$$

where $\bar{U}$ is the outside option.

When RAs can choose across PI segments, equilibrium assignment is defined by

$$\Theta_\lambda = \left\{\theta \in [\theta_L, \theta_H] : \widetilde{U}_\lambda(\theta) \geq \widetilde{U}_{\lambda'}(\theta) \text{ for all } \lambda' \neq \lambda, \quad \widetilde{U}_\lambda(\theta) \geq \bar{U}\right\}. \quad (19)$$

Thus, Θ_λ is the set of abilities that choose the type-λ segment and participate.[18] Under this convention, Θ_λ determines the type composition of filled positions, while n_λ records their normalized mass. A model of posted positions would require a separate fill rate; the baseline abstracts from that layer. Under the monotonicity conditions stated in the online appendix, this set can be summarized by a lower cutoff $\underline{\theta}_\lambda$ within the relevant segment.

[18] The model abstracts from a separate RA-position matching market. The set Θ_λ should be read as the equilibrium ability composition of RAs hired into type-λ positions. Without monotonicity or single crossing, Θ_λ need not be an interval. The cutoff representation $\underline{\theta}_\lambda$ is used only in the maintained region described in the appendix.

PI payoffs. We model the PI's organizational costs as a convex function of team size and mentoring intensity. The quadratic cost of supervision and coordination captures the standard limits to the span of control driven by coordination frictions and communication costs within organizations, as formalized by Garicano (2000), Jones (2021), Fudenberg and Rayo (2019). A PI of type λ chooses normalized team intensity $n_\lambda \geq 0$, AI allocation $\alpha_\lambda \in [0,1]$, and mentoring intensity $g_\lambda \geq 0$. In both PI types, organizational costs are

$$C_\lambda(n_\lambda, g_\lambda) = \bar{w} n_\lambda + \frac{c_g}{2} g_\lambda^2 n_\lambda + \frac{c_n}{2} n_\lambda^2, \qquad c_g, c_n > 0. \quad (20)$$

The first term is the wage bill. The second term captures the cost of providing intensive mentoring to a larger RA team. The third term captures supervision and coordination congestion.

Given equilibrium effort $e^*_\lambda(\theta)$, task outputs for an RA of ability θ are

$$y_{R,\lambda}(\theta) = a_R(K^T_{R,\lambda}) + m_R(K^T_{R,\lambda})\left(\theta + \eta_R e^*_\lambda(\theta)\right), \quad (21)$$

$$y_{N,\lambda}(\theta) = \kappa_\lambda(K^T_{N,\lambda})\left(\theta + \eta_N e^*_\lambda(\theta) + \psi g_\lambda\right). \quad (22)$$

Let F_λ denote the conditional equilibrium distribution of abilities among RAs filling positions in the type-λ segment, normalized on Θ_λ. The per-RA expected CES research output is

$$\bar{Y}_\lambda = \int_{\Theta_\lambda} \left[\tau y_{R,\lambda}(\theta)^{\frac{\varsigma_\lambda - 1}{\varsigma_\lambda}} + (1-\tau) y_{N,\lambda}(\theta)^{\frac{\varsigma_\lambda - 1}{\varsigma_\lambda}}\right]^{\frac{\varsigma_\lambda}{\varsigma_\lambda - 1}} dF_\lambda(\theta). \quad (23)$$

This definition takes the expectation of the CES output itself, rather than applying the CES aggregator to average task outputs.

Quality-oriented research also has an upper-tail component. Let realized novel-project merit be

$$q_N = y_{N,\lambda}(\theta) + \zeta, \qquad \zeta \sim N(0, \sigma^2_\zeta),$$

where ζ captures project-level uncertainty that is distinct from the signal noises. A breakthrough occurs when $q_N > \bar{q}$. The average breakthrough probability in the type-λ segment is

$$\bar{p}_{B,\lambda} = \int_{\Theta_\lambda} \left[1 - \Phi\left(\frac{\bar{q} - y_{N,\lambda}(\theta)}{\sigma_\zeta}\right)\right] dF_\lambda(\theta). \quad (24)$$

With n_λ independent project attempts and average success probability $\bar{p}_{B,\lambda}$, the probability of at least one breakthrough is

$$1 - (1 - \bar{p}_{B,\lambda})^{n_\lambda}.$$

Because n_λ is normalized team intensity rather than literal headcount, this expression should be read as a continuous approximation to a portfolio success probability (Aghion et al., 2008). The threshold formulation captures the idea that exploratory research has an uncertain upper tail (Manso, 2011).

Our distinction between quantity-oriented and quality-oriented PIs builds on the fundamental trade-off between exploitation and exploration in the economics of science (Manso, 2011).

A quantity-oriented PI values scalable research output and solves

$$\Pi_N = \gamma_N n_N \bar{Y}_N - C_N(n_N, g_N), \qquad \gamma_N > 0. \quad (25)$$

The term $n_N \bar{Y}_N$ captures scalable research throughput. The quantity-oriented objective in Eq. (25) reflects the realities of the modern academic reward system, where hypercompetition and quantitative performance metrics often incentivize scholars to prioritize scalable, safe research output (Edwards and Roy, 2017; Azoulay et al., 2011).

A quality-oriented PI places primary value on breakthrough probability (Aghion et al., 2008):

$$\Pi_Q = \Omega\left[1 - (1 - \bar{p}_{B,Q})^{n_Q}\right] + \gamma_Q n_Q \bar{Y}_Q - C_Q(n_Q, g_Q), \qquad \Omega > 0, \quad \gamma_Q \geq 0. \quad (26)$$

The first term captures the value of at least one high-impact research outcome. The second term allows quality-oriented PIs to value ordinary CES research output as well.[19]

The two PI types differ in objectives and may operate in different research environments. Quantity-oriented PIs place greater weight on scalable throughput. Quality-oriented PIs place greater weight on upper-tail research success. In addition, the same AI frontier can have different effects depending on whether routine and novel tasks are more substitutable or more complementary in the laboratory's CES technology (Garicano, 2000; Kremer, 1993; Growiec et al., 2023).

Auxiliary coefficients such as A_λ^B, B_λ^B, and Σ_λ^B, as well as the fuller parameterization used in the technical derivations, are reported in the online appendix.

3.5. Timing and sequence of the game

Having defined the agents, production technology, signals, and payoffs, we now state the timing of the game. The sequence links the objects above into a single mechanism: laboratory organization affects task output, task output generates noisy evidence, evidence determines admissions scores, and fixed capacity determines placement.

At the beginning of the period, each PI observes the common AI capability frontier K and chooses laboratory organization,

$$x_\lambda \equiv (n_\lambda, \alpha_\lambda, g_\lambda).$$

The PI's AI-related decision is not how much AI exists, but how the available frontier is allocated across routine and novel research tasks.

After the PI has chosen the laboratory environment, a prospective RA in a type-λ segment observes that environment, decides whether to participate, and, if participating, chooses effort $e \geq 0$ to maximize $U_{RA,\lambda}(\theta, e; S^*)$. Effort affects both research production and evaluation. It enters task output through the routine and novel production functions, and it also affects the noisy task-level signals observed by the admissions market. In equilibrium, the chosen team-intensity policy is interpreted as filled intensity after participation and sorting, not as the number of initially posted slots.

At the end of the period, candidates enter the admissions market with score $S = r_\lambda \hat{\theta}$. The admissions market sets a cutoff S^* to allocate the fixed mass Q of elite PhD slots. Because capacity is fixed in the short run, admissions are determined by relative standing as well as absolute performance.

We characterize equilibrium by backward induction together with the market-clearing condition for S^*. In the final stage, the admissions market clears through the fixed-capacity cutoff. In the second stage, RAs choose effort and decide whether to participate, anticipating how effort affects their admissions score and how the cutoff affects admission chances. In the first stage, PIs choose x_λ, anticipating the induced RA effort response, participation decisions, and admissions environment. Individual PIs are atomistic in the continuum economy, so each PI takes S^* as given even though S^* is determined in aggregate equilibrium.[20]

4. Equilibrium analysis and propositions

For each PI type $\lambda \in \{\lambda_Q, \lambda_N\}$, an equilibrium consists of an RA effort schedule $e_\lambda^*(\theta)$, a sorting set Θ_λ, a PI organizational choice

$$x_\lambda^* \equiv (n_\lambda^*, \alpha_\lambda^*, g_\lambda^*),$$

and an admissions cutoff S^*. Here n_λ^* denotes filled normalized team intensity after participation and sorting, not posted capacity. The set Θ_λ is the primitive sorting object: it contains the RA ability types assigned to the type-λ segment. Under the maintained monotonicity and single-crossing conditions, Θ_λ can be represented by a lower participation cutoff $\underline{\theta}_\lambda$.[21]

The equilibrium has three requirements. First, RAs choose effort and participation optimally, taking the laboratory environment and admissions cutoff as given, as in signaling settings where agents respond to expected evaluation outcomes (Spence, 1973). Second, each PI chooses x_λ^* to maximize the type-specific payoff defined in Section 3.4, taking the induced RA behavior and admissions cutoff as given. Third, the admissions cutoff clears the fixed-capacity PhD market, following the tournament logic that scarce positions are allocated by relative standing (Lazear and Rosen, 1981; Hopkins, 2012). Together, these requirements combine individual optimality, organizational optimality, sorting, and market clearing under noisy evaluation (Daley and Green, 2014).

The final equilibrium object is the admissions cutoff, which aggregates the outcomes generated by PI choices and RA effort into a capacity-constrained admissions market. Let Q denote the exogenously fixed measure of available PhD slots. For each PI type λ, let F_λ denote the conditional equilibrium distribution of θ among RAs who fill positions in the type-λ segment, normalized on Θ_λ. Since n_λ^* is filled normalized team intensity, $\mu_\lambda n_\lambda^*$ is the aggregate mass of participating candidates from that segment reaching the admissions market. Given the induced PI choices, RA effort schedules, and sorting sets, the cutoff S^* clears the admissions market:

$$Q = \sum_{\lambda \in \{\lambda_Q, \lambda_N\}} \mu_\lambda n_\lambda^* \int P_\lambda\left(\theta, e_\lambda^*(\theta); S^*\right) dF_\lambda(\theta), \tag{27}$$

where $P_\lambda(\theta, e; S^*)$ is the admission probability defined in Section 3.4. No additional participation-rate multiplier appears because participation and sorting are already embedded in the filled team-intensity convention and in the conditional distribution F_λ. If n_λ were instead modeled as posted capacity, the leading mass would need to be replaced by $M_\lambda^* = \mu_\lambda n_\lambda^* \pi_\lambda^*$, with $\pi_\lambda^* = \Pr(\theta \in \Theta_\lambda)$. The right-hand side is therefore the total expected mass of admitted RAs across all PI types. For fixed induced PI choices, effort schedules, and sorting sets, this mass is decreasing in S^*. With equilibrium policy responses, the appendix states sufficient local conditions under which the induced aggregate admissions function is continuous, crosses Q, and yields a locally well-defined cutoff.

The equilibrium logic has three main implications. First, routine-task AI can increase routine output while reducing the diagnostic precision of routine evidence about human contribution; in the maintained parameter region, this also lowers the linear weight placed on routine evidence in the evaluation rule. Second, heterogeneous PI objectives and task complementarities can generate different laboratory organizations. Third, fixed admissions capacity can turn broad improvements in candidate records into a cutoff effect: when scores rise widely and the candidate-mass response does not offset the shift, the threshold for admission rises. The remainder of this section formalizes these implications as propositions.

4.1. Proposition 1: Routine-output expansion and reduced informativeness

The first proposition formalizes the paper's central informational tension. AI can make routine work easier to produce, but the same routine output may become less informative about the RA's own ability and effort. This distinction is important because pre-doctoral work is both productive labor and evidence for later evaluation (Autor et al., 2003; Noy and Zhang, 2023; Brynjolfsson et al., 2025; Doshi and Hauser, 2024; Dell'Acqua et al., 2026).

[19] Because $\bar{Y}_Q$ is defined as expected per-RA CES output, the ordinary-output component is written as $n_Q \bar{Y}_Q$. If $\bar{Y}_Q$ were defined as total portfolio output, the explicit n_Q factor would not be needed.

[20] This is a standard price-taking logic in a continuum economy. A single PI has zero mass and does not internalize the effect of their own laboratory choices on the aggregate admissions cutoff (Cole et al., 1998; Hopkins, 2012).

[21] Without monotonicity or single crossing, Θ_λ need not be an interval and the cutoff representation may fail. The equilibrium object is therefore the sorting set Θ_λ; $\underline{\theta}_\lambda$ is a convenient representation in the maintained region.

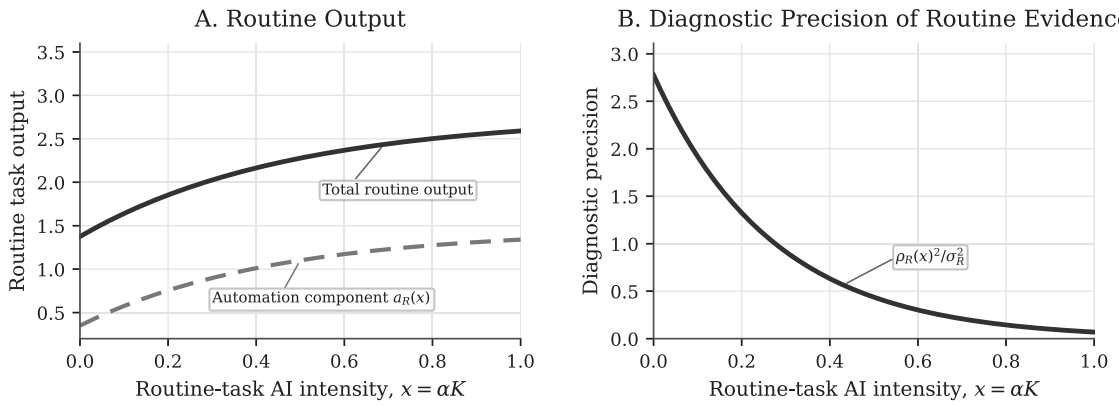

Fig. 2. Routine-output expansion and diagnostic compression. Panel A shows that, holding human input fixed, higher routine-task AI intensity raises the automation component and total routine output. Panel B shows that the diagnostic precision of routine evidence about human contribution, $\rho_R(x)^2/\sigma_R^2$, declines as routine-task AI adds inseparable AI-generated variation to routine artifacts and thereby lowers the derived loading on human contribution. This figure is illustrative and uses the parameterization reported in the online appendix. Panel B uses a smooth numerical path consistent with the derived projection loading from the artifact-level signal-extraction microfoundation in the online appendix's Assumption A4, not a separate imposed sign restriction on $\rho'_R(x)$.

Proposition 1 (*Routine-Output Expansion and Reduced Informativeness*). *Fix total AI capability K, and let $x \equiv K_R^T = \alpha K$ denote routine-task AI intensity. Under the routine production function in (1) and the routine signal in (2), suppose $a'_R(x) > 0$, $m'_R(x) \geq 0$, and $h_R > 0$. If the diagnostic loading $\rho_R(x)$ is generated by the artifact-level attribution problem described above, then $\rho'_R(x) < 0$. Holding the local human-input term h_R fixed:*

(i) routine-task AI weakly increases routine output:

$$\frac{dy_R(x)}{dx} = a'_R(x) + m'_R(x)h_R \geq 0;$$

(ii) routine-task AI reduces the diagnostic precision of routine evidence about the RA's human contribution:

$$\frac{d}{dx}\left(\frac{\rho_R(x)^2}{\sigma_R^2}\right) = \frac{2\rho_R(x)\rho'_R(x)}{\sigma_R^2} < 0;$$

(iii) as routine evidence ceases to load on human contribution, $\rho_R(x) \to 0$, the linear prediction coefficient on routine evidence also converges to zero:

$$\omega_R^B(\alpha, K, \lambda) = \frac{\rho_R(\alpha K)/\sigma_R^2}{\frac{1}{\sigma_\theta^2} + \frac{\rho_R(\alpha K)^2}{\sigma_R^2} + \frac{\kappa_\lambda((1-\alpha)K)^2}{\sigma_N^2}} \to 0.$$

Thus, routine-task AI can raise routine output while reducing the usefulness of routine evidence for evaluating the RA's own human contribution and, in the residualized benchmark, the RA's underlying ability.

Proposition 1 is a partial comparative static: it isolates the routine-AI allocation margin before equilibrium effort, sorting, and the admissions cutoff adjust. The result shows that routine output and routine evidence can move in opposite directions. Routine artifacts may become more abundant and more polished, while carrying less diagnostic information about the RA's own contribution. The loss of diagnostic content is not assumed directly; it follows from the projection problem created by an inseparable AI-generated artifact component.

Fig. 2 illustrates this production–evaluation wedge: routine-task AI raises observable routine output while compressing the diagnostic precision of routine evidence.

This does not imply that evaluation simply shifts to a better signal. Novel contribution, such as judgment, originality, and interpretation, may be more meaningful but also harder to assess ex ante than standardized execution. Evidence from research funding evaluation shows that structurally diverse teams can be penalized in ex ante review even when they perform better ex post (Banal-Estañol et al., 2019).[22]

[22] Grant review and pre-doctoral admissions are different institutions. The analogy is limited to the evaluation problem: complex or nonstandard contributions may be valuable ex post but difficult to assess ex ante.

This suggests that, if AI weakens routine evidence, evaluators need better ways to assess complex research contribution. The mechanism is also consistent with evidence that generative AI can raise productivity while compressing observed performance differences or reducing output diversity in structured knowledge tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Doshi and Hauser, 2024; Dell'Acqua et al., 2026; Tambe, 2025).

Implications. Three immediate implications follow from Proposition 1. First, visible output can improve while the overall precision of evaluation falls if the loss of routine precision is not offset by more informative novel evidence. In the residualized linear-Gaussian benchmark, total precision is

$$\frac{1}{\sigma_\theta^2} + \frac{\rho_R(K_{R,\lambda}^T)^2}{\sigma_R^2} + \frac{\kappa_\lambda(K_{N,\lambda}^T)^2}{\sigma_N^2}.$$

When routine-task AI lowers $\rho_R(K_{R,\lambda}^T)$, the routine component of precision falls. Second, independent process evidence becomes more valuable when baseline precision is lower. Third, the admissions return to effort shifts toward the signal channels that remain diagnostic. These implications reflect the same wedge: the AI allocation that increases research throughput need not be the allocation that best reveals candidate ability. The online appendix reports the corresponding precision and effort-loading algebra.

Fig. 3 provides an illustrative visualization of the signal-reweighting implication of Proposition 1. As routine-task AI reduces the diagnostic content of routine evidence, the linear evaluation rule places less weight on routine signals and relatively more weight on novel-task evidence. The figure is included only to illustrate how evaluation can move from routine-dominant to novel-dominant weighting as routine evidence loses diagnostic content; it is not used in the proof and should not be read as an estimated weighting surface.

4.2. Proposition 2: Local characterization of organizational segmentation

The second result explains why laboratories need not respond to AI in the same way. PI objectives differ, and routine and novel tasks may be more substitutable in some research environments than in others. The result should be read as a local characterization, not as a global monotone comparative static. Its role is to show how primitive differences in objectives and task technology can generate segmented laboratory strategies (Manso, 2011; Garicano, 2000; Kremer, 1993).

Proposition 2 (*Local Segmentation Under CES Task Production*). *Let $\Pi_N(n, \alpha, g)$ and $\Pi_Q(n, \alpha, g)$ denote the type-specific PI payoff functions*

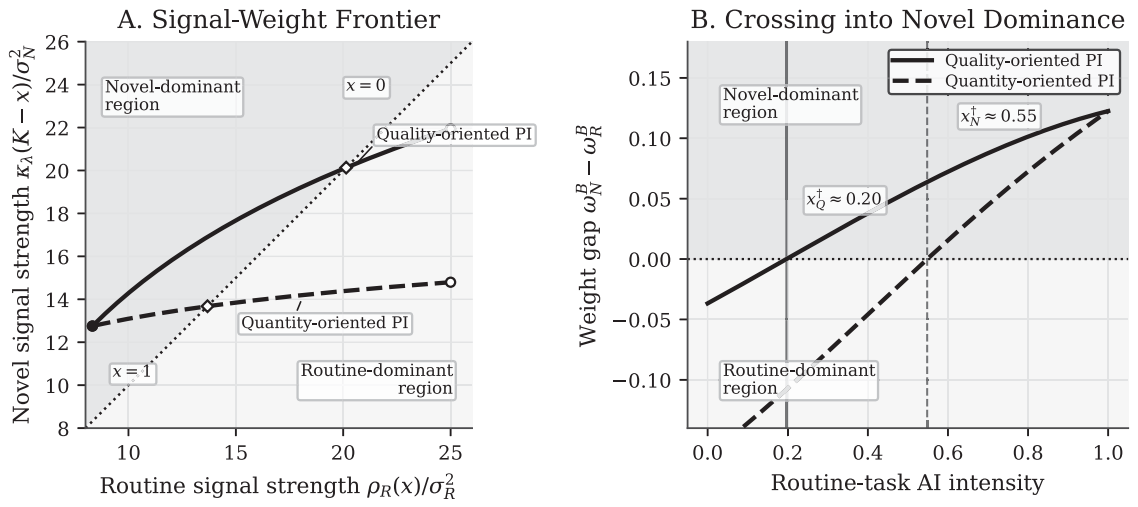

Fig. 3. Illustrative signal reweighting as routine evidence loses diagnostic content. Panel A plots the signal-weight frontier in the space of routine and novel signal strength. The horizontal axis is the routine signal strength, $\rho_R(x)^2/\sigma_R^2$, and the vertical axis is the novel signal strength, $\kappa_\lambda(K-x)^2/\sigma_N^2$. The dotted diagonal marks the equal-weight boundary, $\omega_R^B = \omega_N^B$. Points below the boundary correspond to routine-dominant evaluation, where routine evidence receives more weight; points above the boundary correspond to novel-dominant evaluation, where novel-task evidence receives more weight. The solid and dashed curves illustrate how the relative signal strengths move as routine-task AI intensity x changes for quality-oriented and quantity-oriented PI environments. Panel B plots the corresponding weight gap, $\omega_N^B - \omega_R^B$, against routine-task AI intensity x. The horizontal dotted line at zero marks the equal-weight benchmark, $\omega_N^B = \omega_R^B$. Values below this line indicate routine-dominant evaluation, where routine evidence receives more weight than novel-task evidence. Values above this line indicate novel-dominant evaluation, where novel-task evidence receives more weight than routine evidence. The solid curve traces the weight-gap path for the quality-oriented PI environment, and the dashed curve traces the path for the quantity-oriented PI environment. As x increases, routine evidence loses diagnostic content, so both curves move upward toward greater relative weight on novel-task evidence. The vertical solid line marks the illustrative crossing point $x_Q^* = 0.20$ for the quality-oriented environment, where its evaluation rule switches from routine-dominant to novel-dominant weighting. The vertical dashed line marks the corresponding crossing point $x_N^* = 0.55$ for the quantity-oriented environment. The earlier crossing of the solid curve indicates that the quality-oriented environment shifts toward novel-task evidence at a lower level of routine-task AI intensity. The shaded regions label the two regimes: the lower region is routine-dominant, and the upper region is novel-dominant. The figure uses a stylized parameterization and is intended only to visualize the signal-reweighting implication of Proposition 1; it should not be read as an estimated weighting surface.

defined in Section 3.4, where λ_N is the quantity-oriented type and λ_Q is the quality-oriented type. Fix total AI capability K, and consider a common interior comparison point

$$\bar{x} = (\bar{n}, \bar{\alpha}, \bar{g}), \qquad \bar{n} > 0, \quad \bar{\alpha} \in (0,1), \quad \bar{g} > 0.$$

All derivatives below are evaluated at $\bar{x}$, under the local-envelope convention that holds fixed the induced participant distribution and the admissions cutoff response.

Suppose that, in a neighborhood of $\bar{x}$, each type-specific payoff function is twice continuously differentiable, locally concave in (n, α, g), and has an isolated interior optimum. Suppose also that local cross-choice interactions are weak enough that the sign of each own marginal payoff at $\bar{x}$ determines the direction of the corresponding local optimum.[23] *Suppose further that the two PI types satisfy the following local payoff-gradient conditions:*

$$\frac{\partial \Pi_N}{\partial \alpha}(\bar{x}) > 0 > \frac{\partial \Pi_Q}{\partial \alpha}(\bar{x}), \tag{A}$$

$$\frac{\partial \Pi_Q}{\partial g}(\bar{x}) > 0 > \frac{\partial \Pi_N}{\partial g}(\bar{x}), \tag{G}$$

and

$$\frac{\partial \Pi_N}{\partial n}(\bar{x}) \geq 0 \geq \frac{\partial \Pi_Q}{\partial n}(\bar{x}). \tag{N}$$

Then there exists a local equilibrium neighborhood around $\bar{x}$ in which

$$\alpha_N^* > \alpha_Q^*, \qquad g_Q^* > g_N^*, \qquad n_N^* \geq n_Q^*.$$

If the scale condition in (N) holds strictly, then $n_N^* > n_Q^*$.

The economic interpretation is straightforward. Quantity-oriented PIs place greater value on scalable CES research output. When routine and novel tasks are sufficiently substitutable, routine-task automation has a stronger local payoff value for these PIs because it supports scalable research throughput. Quality-oriented PIs place greater value on upper-tail research success. When routine and novel tasks are sufficiently complementary, shifting AI capacity toward routine tasks is less attractive because routine execution cannot easily substitute for novel-task augmentation, research design, interpretation, and mentoring.

The team-size comparison requires an additional local scale condition. The value of upper-tail success must be high enough to make mentoring and novel-task preservation attractive for quality-oriented PIs, but not so high, relative to coordination costs and portfolio diminishing returns, that quality-oriented PIs also expand normalized team intensity at the comparison point. Under these primitive conditions, the local payoff-gradient signs in Proposition 2 hold. The online appendix derives the corresponding threshold conditions in terms of the CES derivatives, breakthrough-probability derivatives, and organizational cost primitives. These are sufficient local conditions, not universal predictions for arbitrary AI shocks.[24]

This maps naturally onto academic incentive environments in which some rewards concentrate around high-impact research outcomes while others reward sustained measurable output (Heckman and Moktan,

[23] A sufficient set of regularity conditions is that the Hessian of each type-specific payoff function with respect to (n, α, g) is negative definite and locally diagonally dominant around $\bar{x}$, so that own-curvature terms dominate cross-choice interactions. The appendix provides the corresponding local implicit-function argument and the primitive threshold conditions under which the displayed marginal payoff signs hold.

[24] A stronger global monotone-comparative-statics result could be obtained under additional increasing-differences and supermodularity assumptions. The paper uses the weaker local characterization because it is sufficient for the baseline mechanism.

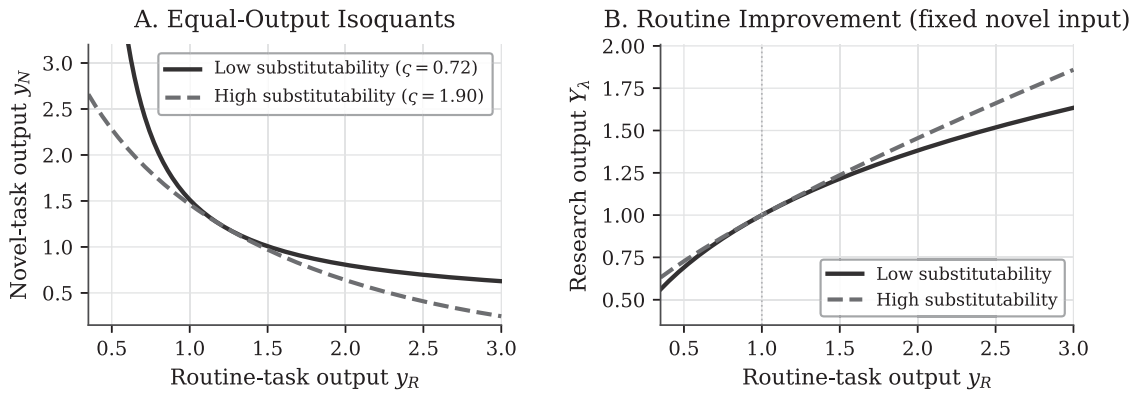

Fig. 4. CES task environments. Panel A plots equal-output isoquants for low- and high-substitutability environments. Lower substitutability implies stronger complementarity between routine and novel tasks. Panel B holds novel-task input fixed and shows that routine-task improvements translate more strongly into research output when routine and novel tasks are more substitutable. This figure is illustrative and uses the parameterization reported in the online appendix; it visualizes CES task environments rather than estimated production elasticities.

Table 2
Characterization of equilibrium strategies in the maintained local region.

Strategy Component	Quality-Oriented PI (λ_Q)	Quantity-Oriented PI (λ_N)
Primary Objective	Maximize the probability of at least one high-value novel outcome, with stronger task complementarity	Maximize scalable CES research output, with higher effective task substitutability
AI Allocation (α^*)	Lower routine-task share; stronger tilt toward novel-task augmentation	Higher routine-task share; stronger tilt toward automation/scaling
Mentoring (g^*)	Higher mentoring intensity; guidance complements novel-task performance	Lower mentoring intensity on average; guidance is relatively more costly at scale
Normalized Team Intensity (n^*)	Lower normalized team intensity; value concentrated in intensively developed RAs	Weakly higher normalized team intensity; pipeline production becomes more attractive
Implied RA Skill Focus	Idea generation, judgment, and mentored problem-solving	Workflow management, structured execution, and scalable production

Notes: The entries summarize the maintained local parameter region of Proposition 2. They are not universal predictions for all laboratories, estimated frequencies, or structural parameter estimates.

2020; Hamermesh, 2013; Edwards and Roy, 2017; Hicks, 2012). A continuous distribution of PI objectives would replace the two-type partition with a continuum of organizational choices; the result should then be read as identifying local sorting directions along that continuum rather than as requiring only two discrete laboratory types.

The equilibrium implication is organizational segmentation. One segment uses AI to scale structured research production; the other relies more on mentoring and novel-task support. The model does not claim that this segmentation is absolute in practice. It shows that such segmentation can arise from heterogeneous PI objectives interacting with task complementarity and AI allocation. Fig. 4 visualizes the production-side mechanism behind this result. Table 2 summarizes these local strategy differences across the two PI types.

4.3. Proposition 3: Fixed-capacity congestion

The third proposition states the fixed-capacity result. The key idea is simple: if many laboratories generate stronger candidate records at the same time, and elite admissions capacity remains fixed, some of the gain can be absorbed by a higher admissions cutoff rather than by a proportional expansion of opportunity (Lazear and Rosen, 1981; Hopkins, 2012; Card and DellaVigna, 2013).

Proposition 3 (*Location-Shift Congestion Under Fixed Capacity*). *Let $M(\bar{K}) \equiv \sum_\lambda \mu_\lambda n^*_\lambda(\bar{K})$ be the equilibrium candidate mass reaching the admissions market, and suppose $M(\bar{K}) > Q$. Suppose aggregate AI capability shifts admissions scores by*

$$S_i(\bar{K}) = \Delta(\bar{K}) + u_i, \qquad \Delta'(\bar{K}) > 0,$$

where $u_i \sim H_0$ and H_0 has density $h_0 > 0$ at the relevant upper quantile. The fixed-capacity cutoff is

$$S^*(\bar{K}) = \Delta(\bar{K}) + H_0^{-1}\left(1 - \frac{Q}{M(\bar{K})}\right).$$

For fixed Q, defining

$$q(\bar{K}) = H_0^{-1}\left(1 - \frac{Q}{M(\bar{K})}\right),$$

we have

$$\frac{dS^*(\bar{K})}{d\bar{K}} = \Delta'(\bar{K}) + \frac{QM'(\bar{K})}{M(\bar{K})^2 h_0(q(\bar{K}))}.$$

Thus, the cutoff rises whenever $M'(\bar{K}) \geq 0$, and entry amplifies the cutoff effect. If capacity also changes with $\bar{K}$, then

$$\frac{dS^*(\bar{K})}{d\bar{K}} = \Delta'(\bar{K}) + \frac{Q(\bar{K})M'(\bar{K})/M(\bar{K})^2 - Q'(\bar{K})/M(\bar{K})}{h_0(q(\bar{K}))},$$

so capacity expansion dampens the cutoff effect. For any candidate whose own score components do not rise enough to offset the higher standardized cutoff, admission probability falls. Hence, under fixed or slowly adjusting capacity, part of the private placement gain from AI is competed away through a congestion channel.

Proposition 3 is still a conditional statement, but the condition is primitive: AI shifts the market score distribution by a common component $\Delta(\bar{K})$. It does not say that every AI shock must shift the entire score distribution to the right, nor that AI necessarily lowers admission probabilities. It says that when AI produces a broad location shift in visible records, a fixed-capacity market converts that shift one-for-one into a higher cutoff; if more candidates enter the market, the relevant upper quantile moves further right. If capacity expands, or if candidate

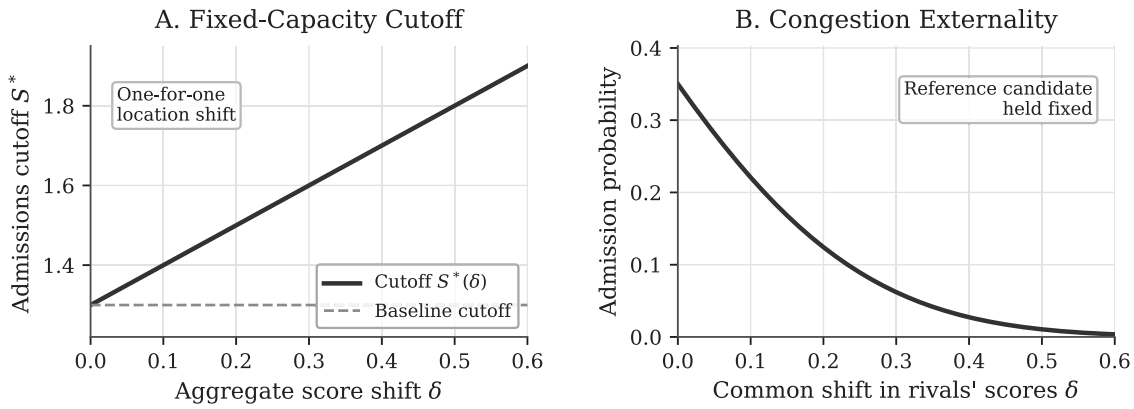

Fig. 5. Congestion under fixed admissions capacity. Panel A shows a simple location-shift benchmark: when the aggregate score distribution shifts upward and admissions capacity is fixed, the cutoff rises. Panel B holds a reference candidate's own score distribution fixed while rivals' scores shift upward; the rising cutoff lowers the reference candidate's admission probability. This figure is illustrative and uses a stylized location-shift benchmark; it is not a structural estimate of admissions cutoffs.

mass falls, the cutoff effect can be weakened. This is the same logic that makes rank-based competition different from evaluation against a fixed absolute standard (Lazear and Rosen, 1981; Hopkins, 2023). It is also related to evidence and models of escalating standards in academic and publication markets (Card and DellaVigna, 2013; Ellison, 2002).

The welfare implication is also conditional. AI may raise research productivity, candidate learning, and the quality of visible records. Proposition 3 isolates one opposing force: in a scarce admissions market, part of the private signaling gain can be competed away. The result is therefore a congestion mechanism, not a claim that AI is socially harmful.

Section 5 below reports a mechanism-preserving simulation that places this cutoff logic in a denser stochastic environment. Following the transparent simulation logic of Kapeller and Steinerberger (2016), the exercise is illustrative rather than structural: it uses the model's own task-output equations, signal weights, PI payoff, and fixed-capacity admissions rule while drawing RA talent, PI orientation, project luck, and assessment noise at high Monte Carlo density.

Fig. 5 visualizes the central comparative static of Proposition 3. In a fixed-capacity tournament, when many candidates receive a common upward shift in admissions scores, the cutoff rises with the shifted score distribution. This is the standard rank-order logic of tournament allocation under scarce positions (Lazear and Rosen, 1981; Hopkins, 2012, 2023). The figure also shows the associated congestion effect: a reference candidate whose own score distribution is held fixed can face a lower admission probability when rivals' scores rise and the admissions cutoff moves upward. Taken together, Propositions 1–3 yield the dominant comparative-static directions summarized in Table 3.

4.4. Extension 1: Apprenticeship learning and human capital accumulation

The baseline focuses on evaluation and congestion, but pre-doctoral work may also build human capital. The online appendix captures this apprenticeship channel with a recursive extension in which future research capability rises with effort, novel-task AI exposure, and mentoring (Becker, 1962; Fudenberg and Rayo, 2019; Kostadinov and Kuvalekar, 2022). The extension distinguishes a learning-dominant regime, where mentoring and difficult task exposure build durable capability, from a signaling-dominant regime, where effort mainly improves current admissions scores and may therefore reflect tournament congestion or signal escalation (Hopkins, 2023).

4.5. Extension 2: Social value, knowledge production, and welfare

The welfare extension separates private ranking incentives from the broader social value of research. The online appendix decomposes welfare into knowledge output, human-capital gains, effort costs, and mentoring costs. AI can raise social value by increasing research output and building future capability when paired with mentoring (Becker, 1962; Manso, 2011). But in a fixed-capacity tournament, some private gains may be absorbed by higher cutoffs and signal escalation (Lazear and Rosen, 1981). The CES structure matters because routine-task gains have greater social value when routine and novel tasks are complements rather than substitutes for judgment-intensive work (Kremer, 1993). Therefore, the welfare effect is ambiguous: the model identifies when AI-generated productivity gains become social value and when they are partly converted into competitive signaling.

5. Mechanism-preserving simulation with ability, luck, and bottlenecked selection

5.1. Simulation design and purpose

The analytical model isolates the production–evaluation mechanism in closed form. This section adds an agent-level simulation. The simulation is designed as a robustness exercise for the model's three mechanisms. It examines whether the model's qualitative mechanisms continue to hold in a richer setting with heterogeneous RAs and PIs, luck in research outcomes, noisy admissions evaluation, and a fixed number of elite PhD slots. This simulation follows a tradition that uses simple agent-level models to examine how evaluation rules and institutional bottlenecks can generate aggregate selection patterns that are not obvious from individual-level assumptions (Ahrweiler et al., 2015; Kapeller and Steinerberger, 2016).

The simulation therefore examines three mechanism-level claims. First, routine-task AI can raise routine output while reducing the diagnostic content of routine evidence. Second, PI organizational choices can segment over a continuous research-orientation space rather than only across two discrete types. Third, fixed capacity can convert noisy evaluation and project luck into false negatives among high-ability or high-realized-merit candidates. The last margin is motivated by work showing that observed success can combine talent with randomness, and that scarce evaluation systems can turn noisy assessments into misallocation (Pluchino et al., 2018; Kapeller and Steinerberger, 2016).

Parameter values are chosen to produce an informative baseline environment. Admission is neither nearly universal nor nearly impossible; PI choices vary across the relevant range; and admissions capacity is scarce enough for ranking errors to affect selection outcomes.

Table 3
Dominant comparative-static directions in the maintained parameter region.

AI-environment change	$\boldsymbol{\alpha}^{boldsymbol*}$	g^*	n^*	Routine evidence	S^*
Stronger routine-task automation	Higher in scalable/quantity-oriented segments	Lower on average where routine scaling substitutes for intensive guidance	Higher in scalable segments	Lower diagnostic precision when ρ_R falls; lower ω_R^B in the maintained region	Higher under a market-wide capability shift; otherwise no direct market-wide effect
Stronger novel-task augmentation	Lower, especially in quality-oriented segments	Higher	Often smaller, more intensive quality-oriented teams	Lower relative weight as novel signals become more informative	Higher under a market-wide capability shift; otherwise no direct market-wide effect
Stronger compression in observable routine performance	Ambiguous for production, but less evidentiary value from routine tilt	Higher in evaluation-intensive segments	Ambiguous	Diagnostic precision and routine weight lower	Higher if many candidates bunch in the upper tail
Higher task substitutability (ς_λ)	Higher, especially in pipeline-style environments	Lower	Higher	No direct effect	No direct effect except through induced equilibrium composition

Notes: The entries report dominant directions from the maintained parameter region rather than universal global comparative statics. The table is a qualitative map of analytical comparative statics, not a calibrated simulation table. Cells marked "ambiguous" indicate that the sign depends on PI type, on whether the shock is idiosyncratic or aggregate, or on whether production incentives or evaluation incentives are the relevant margin. The robust diagnostic-compression result in Proposition 1 is the decline in the human-contribution diagnostic precision ρ_R^2/σ_R^2; the decline in ω_R^B is reported for the maintained parameter region. The cutoff effect follows Proposition 3 and applies to broad market-wide score shifts under fixed capacity.

The online appendix reports the full parameter table and sensitivity checks over alternative capacity ratios, uncertainty scales, diagnostic-compression strengths, effort rules, assignment rules, and a finer λ-grid. We make two simplifying choices to keep the simulation focused. First, as K changes, we keep the number of candidates and the admissions capacity ratio fixed, so the simulation captures changes in scores and cutoffs rather than changes in market entry. Second, we do not simulate a full RA–PI matching market. Instead, RAs are assigned to lab environments in proportion to the filled RA positions in those environments, while the analytical model provides the formal sorting condition through Θ_λ. The appendix therefore reports a stronger rank-sorted assignment check that keeps the same filled-position masses but sorts RAs by comparative novel-versus-routine talent across the PI-orientation continuum.

5.2. Agents, production, and evaluation

Each replication contains J PIs and I RAs. PI j draws a continuous research-orientation index

$$\lambda_j \sim U[0,1],$$

where $\lambda_j = 0$ corresponds to the quantity-oriented endpoint and $\lambda_j = 1$ corresponds to the quality-oriented endpoint. Each type-specific primitive $v \in \{\phi, \varsigma, r, \Omega, \gamma\}$ is interpolated between the analytical endpoints:

$$v(\lambda_j) = (1-\lambda_j)v_N + \lambda_j v_Q.$$

Thus, the two-type analytical model should be read as an endpoint representation of a continuum. To make quality-oriented work differ in more than subjective payoff weights, the simulation also lets project uncertainty and reception uncertainty rise with the novelty-oriented endpoint:

$$\sigma_\zeta(\lambda_j) = \sigma_{\zeta,0}(1 + a_\zeta \lambda_j), \qquad \sigma_\xi(\lambda_j) = \sigma_{\xi,0}(1 + a_\xi \lambda_j).$$

The interpretation is that more exploratory research can have higher upper-tail potential, but also higher project risk and noisier ex ante assessment. This maps the simulation to work on innovation incentives, unexpected research outcomes, and AI-assisted scientific search, where novelty is valuable partly because its payoffs are uncertain and upper-tailed (Manso, 2011; Aslan et al., 2024; Agrawal et al., 2024). This is an exploratory-research specification, not a universal claim that novelty always increases evaluation noise. Settings with better documentation, audit trails, or process evidence could have flatter $\sigma_\xi(\lambda)$ schedules.

RA talent is distributed and multidimensional. Each RA draws routine-execution talent and novel-research talent,

$$\begin{pmatrix} \theta_i^R \\ \theta_i^N \end{pmatrix} \sim TN_2(\mu_\theta, \Sigma_\theta; [\theta_L, \theta_H]^2),$$

with positive but imperfect correlation. The simulation summarizes latent ability as

$$\theta_i^L = \frac{\theta_i^R + \theta_i^N}{2}.$$

Human input is

$$h_{ij}^R = \theta_i^R + \eta_R e_{ij}, \qquad h_{ij}^N = \theta_i^N + \eta_N e_{ij} + \psi g_j.$$

For computational transparency, effort is represented by a monotone effort-loading approximation based on the analytical score loading:

$$e_{ij}^* = \max\{0,\ \beta_{RA}\chi(\theta_i^L, \alpha_j K)\, B_j^B\},$$

where

$$B_j^B = r(\lambda_j)\left[\omega_R^B \rho_R (K_{R,j}^T)\eta_R + \omega_N^B \kappa_{\lambda_j}(K_{N,j}^T)\eta_N\right].$$

This keeps the simulation tied to the model's effort channel while allowing realized effort to vary with ability and the PI environment. This effort rule is a computational approximation rather than a full solution to the RA's cutoff-dependent effort problem in (16). The analytical model and the appendix provide the formal effort-choice problem; the simulation uses this simpler rule to preserve the main effort margin without adding another equilibrium fixed-point calculation. As a check, the online appendix replaces this rule with a cutoff-sensitive best-response approximation: given a simulated cutoff, each RA maximizes expected admission value net of quadratic effort cost over a one-dimensional effort grid, and the cutoff is updated from the induced score distribution. The qualitative selection-error results are unchanged.

Given the PI's choices $x_j = (n_j, \alpha_j, g_j)$, task allocations are $K_{R,j}^T = \alpha_j K$ and $K_{N,j}^T = (1-\alpha_j)K$. Task outputs follow the main model,

$$y_{ij}^R = a_R(K_{R,j}^T) + m_R(K_{R,j}^T)h_{ij}^R, \qquad y_{ij}^N = \kappa_{\lambda_j}(K_{N,j}^T)h_{ij}^N,$$

Table 4
Ability, luck, and evaluation-noise decomposition at $K = 2$.

Scenario	Compression	Luck	Noise	$\mathrm{corr}(S^{obs}, \theta^{L})$	$FN^{ability}$	FN^{merit}	Main channel
Idealized	No	No	No	0.725	0.501	0.757	Fixed-capacity benchmark
Luck only	No	Yes	No	0.724	0.503	0.781	Ability–merit realization risk
Evaluation noise only	No	No	Yes	0.626	0.558	0.764	Score–ability reception risk
Full noisy AI	Yes	Yes	Yes	0.510	0.617	0.838	All channels with diagnostic compression

Notes: All rows use $K = 2$ and fixed capacity $Q/M = 0.17$. The idealized row removes the simulation's added diagnostic-compression, project-luck, and reception-noise layers while retaining the model's task-evidence structure, heterogeneous PI environments, and fixed-capacity tournament. It is therefore a model-only benchmark, not a perfect-information ranking benchmark. $FN^{ability}$ is the rejection probability among top-quintile latent-ability RAs. FN^{merit} is the rejection probability among top-quintile realized-merit RAs.

and are aggregated into Y_{ij} using the CES research technology. PIs choose x_j by grid search over the analytical PI payoff on a scalar latent-ability grid. This policy-precomputation step keeps the organizational choice close to the closed-form model, but it is not a full optimization over multidimensional RA matching. RAs are then assigned to PI environments with probability proportional to filled team intensity n_j, the simulation analogue of the filled-position convention in the equilibrium definition. This assignment rule deliberately avoids imposing ability-based sorting mechanically in the baseline simulation. The appendix reports a rank-sorted assignment robustness check that assigns higher comparative novel talent, $\theta_i^N - \theta_i^R$, to higher-λ PI environments while preserving the same filled team-intensity masses.

The key measurement extension separates three objects that coincide only in the most stylized benchmark. Latent ability θ_i^L captures underlying research potential. Realized true merit is

$$M_{ij}^{\mathrm{true}} = Y_{ij} + \zeta_{ij}, \qquad \zeta_{ij} \sim N(0, \sigma_\zeta(\lambda_j)^2),$$

where ζ_{ij} captures project-level luck and discovery uncertainty. Admissions committees do not observe M_{ij}^{true}. They observe noisy routine and novel signals, form the same linear evaluation $\hat{\theta}_{ij}$ as in the model, and receive a noisy perceived score

$$S_{ij}^{\mathrm{obs}} = r(\lambda_j)\hat{\theta}_{ij} + \xi_{ij}, \qquad \xi_{ij} \sim N(0, \sigma_\xi(\lambda_j)^2).$$

This separation follows stochastic production and noisy-signal models in which latent quality, realized output, and evaluated evidence need not coincide (Just and Pope, 1978; Daley and Green, 2014; Heinsalu, 2018; Bao and Duffy, 2021). The extra term ξ_{ij} is a simulation-only reception shock layered on top of task-level signal noise; it is included to study assessment uncertainty, not to change the analytical propositions. Admissions are allocated to the top Q candidates by S_{ij}^{obs}, while ex post selection quality is evaluated using M_{ij}^{true}. The top-Q rule follows rank-order tournament logic under scarce positions (Lazear and Rosen, 1981; Hopkins, 2012, 2023). This distinction brings the bottleneck logic of Kapeller and Steinerberger (2016) into the admissions setting: when scarce slots are allocated using noisy evaluations, some high-ability candidates and high-realized-merit work may be overlooked.

5.3. Decomposition benchmarks

The simulation reports both ability-based and merit-based selection errors:

$$FN^{ability} = \Pr(A_i = 0 \mid \theta_i^L \geq q_{0.8}^{\theta}), \qquad FN^{merit} = \Pr(A_i = 0 \mid M_i^{\mathrm{true}} \geq q_{0.8}^{M}),$$

where $A_i = 1$ denotes admission, $q_{0.8}^{\theta}$ is the 80th percentile of latent ability, and $q_{0.8}^{M}$ is the 80th percentile of realized true merit. Because the baseline capacity ratio is $Q/M = 0.17$, which is below the top-quintile threshold, false negatives among top-quintile candidates are mechanically positive even under a highly informative ranking rule. The informative quantity is therefore the change relative to the idealized benchmark, not the level alone.

The simulation uses $K = 0.00, 0.10, \ldots, 2.00$, 80 Monte Carlo replications at each scenario–K point, 4000 RAs per replication, 200 PIs per replication, and 31 grid points for continuous-λ policy precomputation. The no-compression benchmark uses the same data-generating distributions for agent populations, PI orientations, project shocks, and the capacity rule while setting $\rho_R(K_R^T) = 1$.

Table 4 separates the channels. Project luck mainly affects the ability-to-merit link: realized merit can diverge from latent ability even when score-ability alignment remains strong. Assessment noise mainly affects the score-to-ability and score-to-merit links. Diagnostic compression weakens the informativeness of routine evidence inside the score. Fixed capacity turns these forms of misalignment into false negatives. Because the idealized row still contains heterogeneous PI environments and a capacity ratio below the top-quintile threshold, it should be read as the model-only benchmark rather than as the minimum feasible false-negative rate. The full noisy-AI environment combines the channels and produces the largest selection-error rates.

5.4. Results

Fig. 6 reports the integrated mechanism. Routine output rises as the AI frontier advances. The near overlap between the two lines in upper-left panel is intentional: diagnostic compression affects the evidentiary loading of routine output, not the production technology itself. To make the evidentiary scale interpretable, upper-right panel reports routine signal informativeness relative to its $K = 0$ value rather than the raw precision $\rho_R(\alpha_\lambda K)^2/\sigma_R^2$. Under diagnostic compression, relative routine informativeness falls from 1.00 at $K = 0$ to about 0.15 at $K = 2$, so routine evidence retains only about 15 percent of its baseline informativeness. Lower-left panel reports score-ability alignment rather than score-merit alignment, because realized merit also contains project luck and novelty-related uncertainty. The alignment remains positive but declines from about 0.63 to about 0.51. Thus, the simulation does not imply that admissions become random; it shows that evaluated scores become less tightly linked to latent ability. Lower-right panel reports cutoff pressure as the required cutoff increase in $K = 0$ score standard deviations. In the diagnostic-compression environment, this standardized cutoff pressure rises to about 0.38 baseline score standard deviations by $K = 2$. The no-compression benchmark produces slightly stronger cutoff pressure because routine-output gains transmit more fully into evaluated scores when routine evidence remains diagnostic. Lower cutoff pressure under compression should not be interpreted as a welfare improvement. It reflects the fact that AI-enabled routine gains are discounted by the evaluation rule; the main cost of compression appears in weaker alignment and higher false negatives, not necessarily in a larger raw cutoff.

The continuous PI-orientation margin is reported in the online appendix. The same payoff function that generates the two-type analytical segmentation also generates monotone policy regions when λ is distributed: lower-λ laboratories choose higher routine-task AI shares and larger team intensity, while higher-λ laboratories preserve more

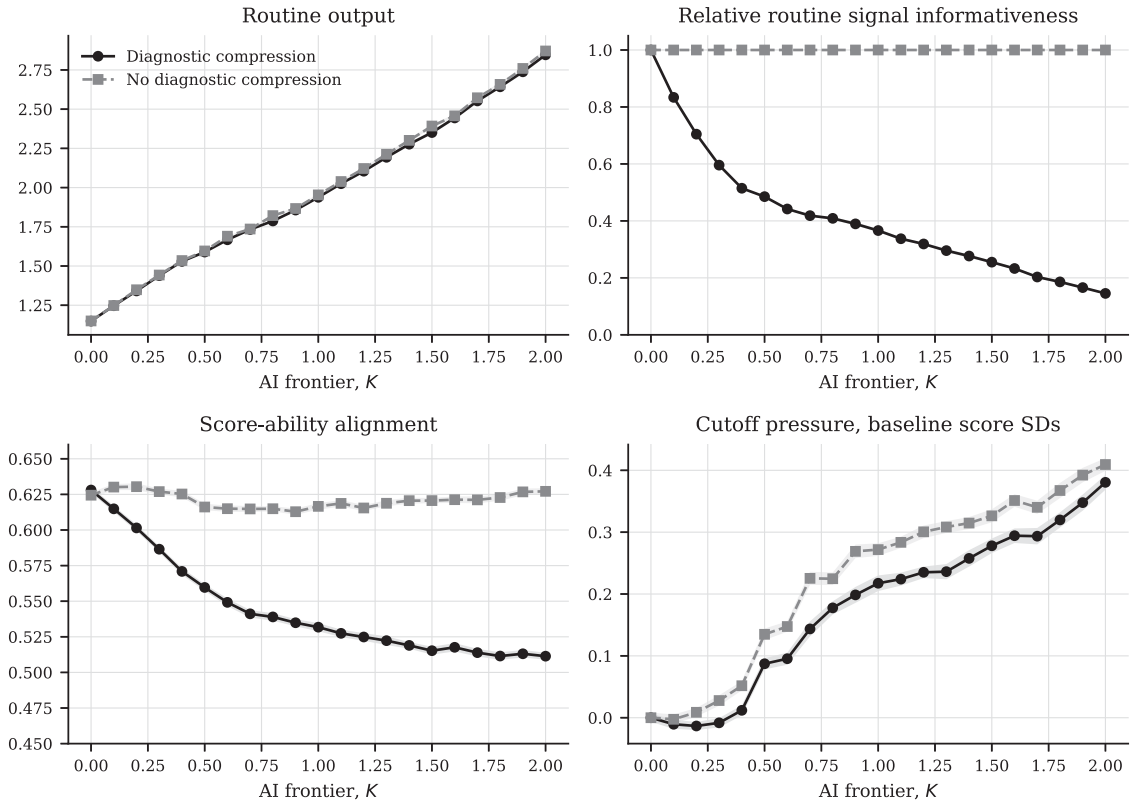

Fig. 6. Unified simulation mechanisms. The horizontal axis in all panels is the AI frontier K. Solid lines use the diagnostic-compression environment, in which routine-task AI lowers the informativeness of routine evidence; dashed lines use the no-compression counterfactual, which uses the same data-generating distributions and capacity rule while setting $\rho_R(K_R^T) = 1$. Upper-left panel plots mean routine output, showing the production gain from AI. The two lines nearly overlap because diagnostic compression changes how routine evidence is evaluated, not how routine output is produced. Upper-right panel plots routine signal informativeness, $\rho_R(\alpha_\lambda K)^2/\sigma_R^2$, normalized to one at $K = 0$; a value of 0.15 means that routine evidence retains about 15 percent of its baseline informativeness. Lower-left panel plots score-ability alignment, $\mathrm{corr}(S^{\mathrm{obs}}, \theta^L)$, where S^{obs} is the perceived admissions score and θ^L is latent RA ability. Lower-right panel plots cutoff pressure, $(S^*(K) - S^*(0))/\mathrm{sd}(S^{\mathrm{obs}}_{K=0})$, so the admissions cutoff increase is measured in baseline score standard deviations. Shaded bands are 95 percent Monte Carlo intervals. The figure is illustrative and is not a structural estimate.

mentoring and less routine-task automation. The plot is placed in the appendix because it is a robustness check for Proposition 2, whereas the main simulation result concerns how distributed ability, luck, and noisy evaluation affect selection under fixed capacity. The appendix also reports a 61-point λ-grid check showing that the segmentation pattern is not an artifact of the baseline 31-point grid, and effort/assignment checks showing that the selection-error results do not depend on the baseline effort-loading rule or on filled-random RA assignment.

Fig. 7 focuses on ability, realized merit, and bottlenecked selection. In the diagnostic-compression environment, the false negative rate among top latent-ability RAs rises from about 0.55 at $K = 0$ to about 0.62 at $K = 2$. The corresponding false negative rate among top realized-merit RAs rises from about 0.67 to about 0.84. These levels should not be interpreted as calibrated rejection rates for a real admissions market; the relevant result is the increase relative to the common-capacity benchmark. Score-ability alignment falls from about 0.63 to about 0.51, while merit-ability alignment falls from about 0.57 to about 0.52. Thus, AI raises production, but diagnostic compression, project luck, and noisy assessment weaken both links in the selection chain: ability to realized merit, and realized merit to evaluated score.

Additional sensitivity checks in the online appendix separate bottleneck and uncertainty effects. At $K = 2$, increasing the capacity ratio from $Q/M = 0.10$ to $Q/M = 0.35$ lowers the ability-based false negative rate from about 0.74 to about 0.38, while also reducing the cutoff. Holding $K = 0$ to isolate assessment uncertainty from AI diagnostic compression, higher reception noise raises ability-based false negatives and lowers score-merit alignment. Higher project luck has little direct effect on ability-based false negatives because admissions do not observe ζ, but it sharply reduces score-merit alignment by widening the gap between latent ability and realized project outcomes.

Overall, the simulation strengthens rather than replaces the propositions. Proposition 1 explains why routine evidence can lose diagnostic content. Proposition 2 explains why organizational choices differ across PI orientations. Proposition 3 explains why fixed capacity converts broad score improvements into cutoff pressure. The simulation shows that these mechanisms continue to operate when PI orientation and RA talent are distributed, true merit contains project luck, and admissions committees observe noisy perceived scores rather than true quality. The simulation uses simple agent-level rules to show how noisy evaluation under scarce capacity can generate aggregate misallocation, even when evaluators do not intentionally misclassify candidates.

6. Discussion and policy implications

The central implication of the model is that generative AI can separate three objects that were previously more closely aligned: what research work produces, what that work reveals about an early-career researcher, and how scarce opportunities are allocated. Pre-doctoral research work is therefore not only project labor. It is also evidence used to evaluate future research potential.

Proposition 1 shows that routine-task AI can raise visible routine output while lowering its diagnostic value. This gives an AI-specific reason for a broader concern in responsible research assessment: easily measured outputs should not be treated as direct substitutes for expert judgment (Cagan, 2013; Hicks et al., 2015). The implication

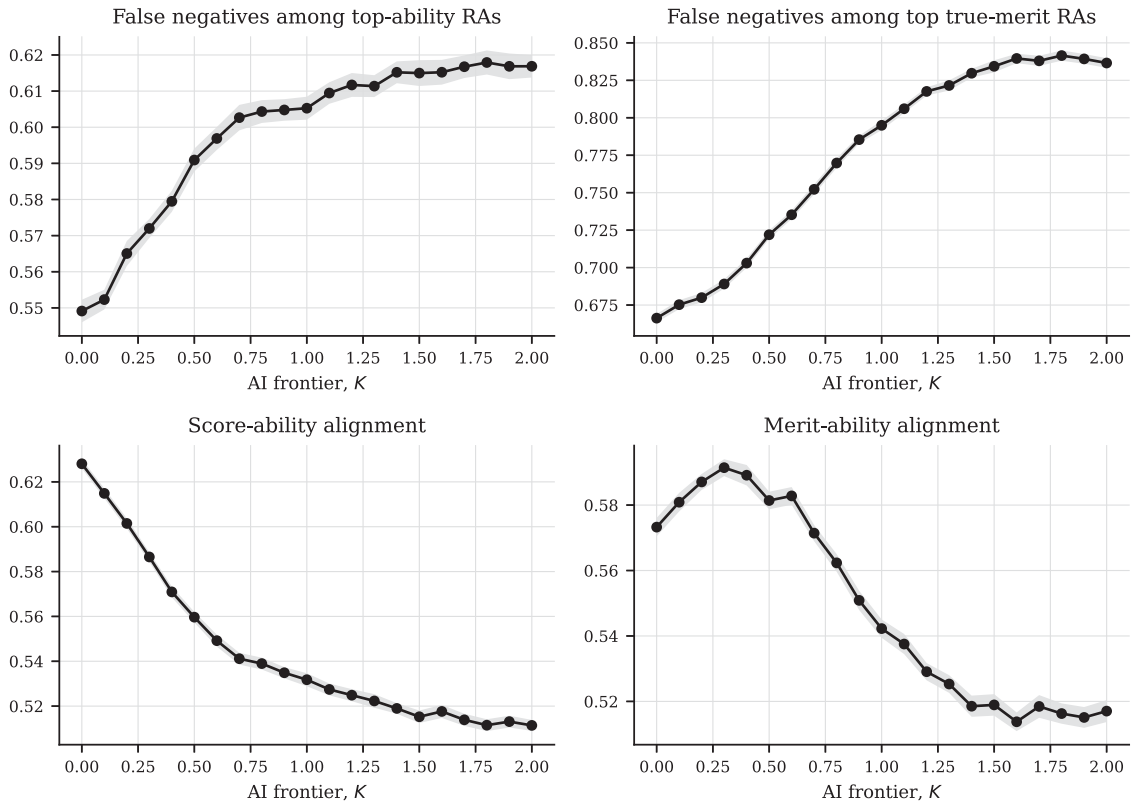

Fig. 7. Ability, realized merit, and selection errors. All panels use the diagnostic-compression environment and vary the AI frontier K. Admissions are assigned to the top Q candidates by perceived score S^{obs}, while ex post merit is measured by $M^{\text{true}} = Y + \zeta$. Upper-left panel plots $FN^{ability}$, the probability that a candidate is rejected conditional on latent ability θ_i^L being in the top quintile. Upper-right panel plots FN^{merit}, the probability that a candidate is rejected conditional on realized merit M_i^{true} being in the top quintile. Because the baseline capacity ratio is $Q/M = 0.17 < 0.20$, these false-negative rates are mechanically positive even under very accurate ranking. They should be read as simulation-indexed selection-error measures, not calibrated market rejection rates; the informative object is their change as K rises. Lower-left panel plots score-ability alignment, $\text{corr}(S^{\text{obs}}, \theta^L)$. Lower-right panel plots merit-ability alignment, $\text{corr}(M^{\text{true}}, \theta^L)$, which captures how project luck and PI environments make realized merit diverge from latent ability. Shaded bands are 95 percent Monte Carlo intervals.

is not to ignore routine evidence. Routine execution still matters. But when routine artifacts become cheaper and more polished, evaluation needs additional evidence about process and individual contribution. This raises the value of independent process signals: oral research defenses, code and analysis walk-throughs, short replication tasks, research-design interviews, AI-use statements, and offline tests that separate routine execution from independent reasoning.[25] These signals are valuable not only because they are immune to AI assistance, but also because they provide evidence about reasoning, interpretation, and ownership. The model therefore suggests a shift from evaluating polished artifacts alone toward asking candidates to explain, reproduce, defend, or extend their work.

This shift should be designed carefully. Oral or process-based assessment is not automatically fairer. If implemented informally, it may introduce new biases related to language, confidence, advising access, or institutional background. Its value depends on structured tasks, clear rubrics, and an effort to evaluate reasoning rather than polish or status. This point is especially important because judgment, originality, interpretation, and problem formulation may be harder to assess ex ante than standardized execution (Banal-Estañol et al., 2019; Aslan et al., 2024).

Proposition 2 shows that AI need not make laboratories converge on one organizational form. Some laboratories may use AI to scale routine execution. Others may use AI to support mentoring, research design, and less standardized work. The model does not rank these forms universally. It shows that PI objectives and task technologies can make different AI-enabled organizations privately attractive. Evaluation systems can then reinforce one direction or the other. If they reward visible throughput, they strengthen incentives for scalable routine pipelines. If they reward research design, interpretation, mentoring, and individual contribution, they support apprenticeship-intensive work. This is consistent with evidence that performance-based evaluation systems shape organizational priorities through rankings, prestige, and measurable targets (Hicks, 2012).

The same AI frontier may also create unequal effective capability across laboratories. The baseline treats K as common to isolate the allocation mechanism, but real laboratories differ in data access, software infrastructure, verification routines, AI expertise, and mentoring capacity. This maps onto absorptive capacity: organizations differ in

[25] A simple way to formalize this point is to add an independent process signal to the evaluation rule. Suppose the market observes

$$s_I = \theta + \varepsilon_I, \qquad \varepsilon_I \sim N(0, \sigma_I^2),$$

where s_I represents an oral research defense, code walk-through, replication task, or research-design interview. In the linear-Gaussian benchmark, the weight on this independent signal is

$$\omega_I = \frac{1/\sigma_I^2}{1/\sigma_\theta^2 + \rho_R (K_R^T)^2/\sigma_R^2 + \kappa_\lambda (K_N^T)^2/\sigma_N^2 + 1/\sigma_I^2}.$$

Holding the precision of s_I fixed, a decline in the diagnostic precision of routine evidence, $\rho_R (K_R^T)^2/\sigma_R^2$, raises the relative weight placed on s_I. Thus, independent process evidence becomes more valuable when routine artifacts lose diagnostic content.

their ability to recognize, assimilate, and use external knowledge (Cohen and Levinthal, 1990). The policy implication is that AI access is not only a software problem. Training, documentation norms, shared infrastructure, and verification support matter if AI is to improve research training rather than widen organizational differences.

Proposition 3 shows that stronger records are not the same as broader access. When elite admissions capacity is fixed, a broad upward shift in visible records raises the cutoff rather than proportionally expanding opportunity. If AI also draws more candidates into the market, cutoff pressure can rise further. If capacity expands, that pressure is weakened. This follows the rank-order logic of scarce academic tournaments (Lazear and Rosen, 1981; Hopkins, 2012, 2023) and is consistent with work on publication bottlenecks and escalating standards in academic careers (Card and DellaVigna, 2013; Ellison, 2002). The implication is not simply that PhD slots should expand. Advising resources, funding, and field-level demand constrain capacity. The more general point is that institutions should monitor whether AI raises the cost of becoming competitive even when it improves absolute preparation.

A related risk is greater reliance on contextual credibility.[26] If routine artifacts become less informative, evaluators may lean more heavily on PI prestige, institutional affiliation, or prior placement history. Science studies has long emphasized that recognition and resources can cumulate over time (Merton, 1968; Bol et al., 2018), and institutional position can matter in scientific resource allocation (Viner et al., 2004). A better response is to document individual contribution while evaluating research context without reducing it to prestige.

The apprenticeship and welfare extensions qualify the baseline mechanism. Pre-doctoral work may build human capital, not only signals. When AI-supported work is paired with mentoring and exposure to difficult novel tasks, it can raise durable research capability.[27] In a learning-dominant regime, AI helps RAs build skills beyond the current admissions cycle. In a signaling-dominant regime, effort mainly improves current admissions scores and may reflect tournament congestion. The welfare effect is therefore not mechanically positive or negative. AI can raise social value through knowledge output and human-capital formation, but fixed-capacity selection can convert part of the private gain into higher cutoffs and signal escalation as Lazear and Rosen (1981) shows. The model identifies when AI-supported work becomes durable capability and when it mainly intensifies competitive signaling.

The model also suggests an empirical agenda. The diagnostic-compression mechanism predicts changes in both evaluation language and evaluation procedures. Recommendation letters and RA evaluations should place less weight on routine execution and more weight on judgment, interpretation, research design, ownership, and initiative after the diffusion of generative AI. Admissions and hiring processes may also add independent process signals, such as oral defenses, code walk-throughs, replication tasks, AI-use statements, and research-design interviews. The segmentation mechanism predicts uneven change across laboratories: AI-assisted pipeline environments should emphasize documentation and verification, while mentoring-intensive environments should emphasize research design and independent reasoning. These predictions can be studied with pre-doctoral job descriptions, recommendation-letter text, admissions procedures, placement outcomes, and measures of laboratory AI infrastructure.

[26] The model captures this only in reduced form through r_λ, not as a dynamic reputation process.

[27] The online appendix formalizes this channel with a recursive extension in which future research capability increases with effort, mentoring, and exposure to novel-task AI.

7. Conclusion

This paper develops a model of how generative AI changes the pre-doctoral academic labor market. The central mechanism is a wedge between productive value and evidentiary value. AI can make routine research output easier to produce, while making that same output less informative about the RA's own effort, judgment, and research potential. Because pre-doctoral work serves both as research labor and as evidence for PhD admissions, this wedge matters for both laboratory organization and candidate evaluation.

The analytical model yields three main implications. First, routine-task AI can raise observable routine output while reducing the diagnostic precision of routine evidence. Second, heterogeneous PI objectives and task complementarity can generate segmented laboratory strategies: some environments tilt toward scalable routine production, while others rely more on mentoring and novel-task augmentation. Third, when elite admissions capacity is fixed, broad improvements in visible records can raise the admissions cutoff rather than expand access proportionally.

The simulation shows that these mechanisms continue to operate in a richer environment with heterogeneous RAs and PIs, luck in realized research outcomes, noisy admissions evaluation, and fixed-capacity selection. It shows that when ability, realized merit, and evaluated scores can diverge, diagnostic compression and scarce capacity can produce selection errors: high-ability or high-realized-merit candidates may be missed even as AI raises routine output.

The broader lesson is that productivity gains, learning gains, and signaling gains need not coincide. AI may improve research output and help RAs build durable human capital, but it may also intensify competition when stronger visible records are evaluated in a fixed-capacity admissions market. The paper provides a benchmark for studying how AI changes the link between research work, evaluation, and access to scientific careers, and suggests that research-training institutions may need more process-based evidence of individual contribution as routine artifacts become less diagnostic.

CRediT authorship contribution statement

Shaohui Wang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

Ethical approval

No ethical approval was required for this theoretical research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and acknowledgment statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.respol.2026.105568.

Data availability

No data was used for the research described in the article.

References

Abramitzky, R., Greska, L., Pérez, S., Price, J., Schwarz, C., Waldinger, F., 2024. Climbing the ivory tower: How socio-economic background shapes academia. http://dx.doi.org/10.3386/w33289, Working Paper No. 33289. National Bureau of Economic Research.

Acemoglu, D., Autor, D., 2011. Skills, tasks and technologies: Implications for employment and earnings. In: Handbook of Labor Economics. vol. 4, pp. 1043–1171. http://dx.doi.org/10.1016/s0169-7218(11)02410-5.

Acemoglu, D., Restrepo, P., 2018. Modeling automation. AEA Pap. Proc. 108, 48–53. http://dx.doi.org/10.1257/pandp.20181020.

Acemoglu, D., Restrepo, P., 2019. Automation and new tasks: How technology displaces and reinstates labor. J. Econ. Perspect. 33 (2), 3–30. http://dx.doi.org/10.1257/jep.33.2.3.

Acemoglu, D., Restrepo, P., 2020. Unpacking skill bias: Automation and new tasks. AEA Pap. Proc. 110, 356–361. http://dx.doi.org/10.1257/pandp.20201063.

Acemoglu, D., Restrepo, P., 2022. Tasks, automation, and the rise in US wage inequality. Econometrica 90 (5), 1973–2016. http://dx.doi.org/10.3982/ecta19815.

Acemoglu, D., Restrepo, P., 2024. A task-based approach to inequality. Oxf. Open Econ. 3 (Supplement_1), i906–i929. http://dx.doi.org/10.1093/ooec/odad082.

Aghion, P., Dewatripont, M., Stein, J.C., 2008. Academic freedom, private-sector focus, and the process of innovation. Rand J. Econ. 39 (3), 617–635. http://dx.doi.org/10.1111/j.1756-2171.2008.00031.x.

Agrawal, A.K., Gans, J., Goldfarb, A. (Eds.), 2019. The economics of Artificial Intelligence: An Agenda. University of Chicago Press, http://dx.doi.org/10.7208/chicago/9780226613475.001.0001.

Agrawal, A.K., McHale, J., Oettl, A., 2024. Artificial intelligence and scientific discovery: A model of prioritized search. Res. Policy 53 (5), 104989. http://dx.doi.org/10.1016/j.respol.2024.104989.

Ahrweiler, P., Schilperoord, M., Pyka, A., Gilbert, N., 2015. Modelling research policy: Ex-ante evaluation of complex policy instruments. J. Artif. Soc. Soc. Simul. 18 (4), 5. http://dx.doi.org/10.18564/jasss.2927.

Akerlof, G.A., 1970. The market for "lemons": Quality uncertainty and the market mechanism. Q. J. Econ. 84 (3), 488–500. http://dx.doi.org/10.2307/1879431.

Arrow, K.J., Chenery, H.B., Minhas, B.S., Solow, R.M., 1961. Capital-labor substitution and economic efficiency. Rev. Econ. Stat. 43 (3), 225–250. http://dx.doi.org/10.2307/1927286.

Aslan, Y., Yaqub, O., Sampat, B.N., Rotolo, D., 2024. Unexpectedness in medical research. Res. Policy 53 (8), 105075. http://dx.doi.org/10.1016/j.respol.2024.105075.

Autor, D.H., Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: An empirical exploration. Q. J. Econ. 118 (4), 1279–1333. http://dx.doi.org/10.1162/003355303322552801.

Azoulay, P., Graff Zivin, J.S., Manso, G., 2011. Incentives and creativity: Evidence from the academic life sciences. Rand J. Econ. 42 (3), 527–554. http://dx.doi.org/10.1111/j.1756-2171.2011.00140.x.

Baker, G., Gibbons, R., Murphy, K.J., 2002. Relational contracts and the theory of the firm. Q. J. Econ. 117 (1), 39–84. http://dx.doi.org/10.1162/003355302753399445.

Banal-Estañol, A., Macho-Stadler, I., Pérez-Castrillo, D., 2019. Evaluation in research funding agencies: Are structurally diverse teams biased against? Res. Policy 48 (7), 1823–1840. http://dx.doi.org/10.1016/j.respol.2019.04.008.

Bao, T., Duffy, J., 2021. Signal extraction: Experimental evidence. Theory and Decision 90 (2), 219–232. http://dx.doi.org/10.1007/s11238-020-09785-x.

Becker, G.S., 1962. Investment in human capital: A theoretical analysis. J. Political Econ. 70 (5, Part 2), 9–49. http://dx.doi.org/10.1086/258724.

Bianchini, S., Müller, M., Pelletier, P., 2022. Artificial intelligence in science: An emerging general method of invention. Res. Policy 51 (10), 104604. http://dx.doi.org/10.1016/j.respol.2022.104604.

Board, S., Meyer-Ter-Vehn, M., 2015. Relational contracts in competitive labour markets. Rev. Econ. Stud. 82 (2), 490–534. http://dx.doi.org/10.1093/restud/rdu036.

Bol, T., de Vaan, M., van de Rijt, A., 2018. The Matthew effect in science funding. Proc. Natl. Acad. Sci. 115 (19), 4887–4890. http://dx.doi.org/10.1073/pnas.1719557115.

Brynjolfsson, E., Li, D., Raymond, L.R., 2025. Generative AI at work. Q. J. Econ. 140 (2), 889–942. http://dx.doi.org/10.1093/qje/qjae044.

Cagan, R., 2013. The San Francisco declaration on research assessment. Dis. Model. Mech. 6 (4), 869–870. http://dx.doi.org/10.1242/dmm.012955.

Câmara, O., Jia, N., Raffiee, J., 2023. Reputation, competition, and lies in labor market recommendations. Manag. Sci. 69 (11), 7022–7043. http://dx.doi.org/10.1287/mnsc.2022.4654.

Card, D., DellaVigna, S., 2013. Nine facts about top journals in economics. J. Econ. Lit. 51 (1), 144–161. http://dx.doi.org/10.1257/jel.51.1.144.

Ceccagnoli, M., Higgins, M.J., Palermo, V., 2014. Behind the scenes: Sources of complementarity in R&D. J. Econ. Manag. Strat. 23 (1), 125–148. http://dx.doi.org/10.1111/jems.12048.

Chassang, S., 2010. Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts. Am. Econ. Rev. 100 (1), 448–465. http://dx.doi.org/10.1257/aer.100.1.448.

Cohen, W.M., Levinthal, D.A., 1990. Absorptive capacity: A new perspective on learning and innovation. Adm. Sci. Q. 35 (1), 128–152. http://dx.doi.org/10.2307/2393553.

Cole, H.L., Mailath, G.J., Postlewaite, A., 1998. Class systems and the enforcement of social norms. J. Public Econ. 70 (1), 5–35. http://dx.doi.org/10.1016/S0047-2727(98)00058-9.

Conley, J.P., Önder, A.S., 2014. The research productivity of new PhDs in economics: The surprisingly high non-success of the successful. J. Econ. Perspect. 28 (3), 205–216. http://dx.doi.org/10.1257/jep.28.3.205.

Daley, B., Green, B., 2014. Market signaling with grades. J. Econom. Theory 151, 114–145. http://dx.doi.org/10.1016/j.jet.2013.10.009.

DeGroot, M.H., 1970. Optimal statistical decisions. McGraw-Hill.

Dell'Acqua, F., McFowland, III, E., Mollick, E., Lifshitz, H., Kellogg, K.C., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K.R., 2026. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organ. Sci. 37 (2), 403–423. http://dx.doi.org/10.1287/orsc.2025.21838.

Doshi, A.R., Hauser, O.P., 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci. Adv. 10 (28), eadn5290. http://dx.doi.org/10.1126/sciadv.adn5290.

Edwards, M.A., Roy, S., 2017. Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. Environ. Eng. Sci. 34 (1), 51–61. http://dx.doi.org/10.1089/ees.2016.0223.

Ellison, G., 2002. The slowdown of the economics publishing process. J. Political Econ. 110 (5), 947–993. http://dx.doi.org/10.1086/341868.

Fudenberg, D., Rayo, L., 2019. Training and effort dynamics in apprenticeship. Am. Econ. Rev. 109 (11), 3780–3812. http://dx.doi.org/10.1257/aer.20171939.

Fügener, A., Walzner, D.D., Gupta, A., 2025. Roles of artificial intelligence in collaboration with humans: Automation, augmentation, and the future of work. Manag. Sci. 72 (1), 538–557. http://dx.doi.org/10.1287/mnsc.2024.05684.

Garicano, L., 2000. Hierarchies and the organization of knowledge in production. J. Political Econ. 108 (5), 874–904. http://dx.doi.org/10.1086/317671.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, 3rd ed. CRC Press, http://dx.doi.org/10.1201/b16018.

Growiec, J., McAdam, P., Mućk, J., 2023. R&D capital and the idea production function. http://dx.doi.org/10.18651/rwp2023-05, Research Working Paper No. 23-05. Federal Reserve Bank of Kansas City.

Hamermesh, D.S., 2013. Six decades of top economics publishing: Who and how? J. Econ. Lit. 51 (1), 162–172. http://dx.doi.org/10.1257/jel.51.1.162.

Heckman, J.J., Moktan, S., 2020. Publishing and promotion in economics: The tyranny of the top five. J. Econ. Lit. 58 (2), 419–470. http://dx.doi.org/10.1257/jel.20191574.

Heinsalu, S., 2018. Dynamic noisy signaling. Am. Econ. J.: Microeconomics 10 (2), 225–249. http://dx.doi.org/10.1257/mic.20160336.

Hicks, D., 2012. Performance-based university research funding systems. Res. Policy 41 (2), 251–261. http://dx.doi.org/10.1016/j.respol.2011.09.007.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I., 2015. Bibliometrics: The Leiden manifesto for research metrics. Nature 520 (7548), 429–431. http://dx.doi.org/10.1038/520429a.

Hill, R., Stein, C., 2025. Race to the bottom: Competition and quality in science. Q. J. Econ. 140 (2), 1111–1185. http://dx.doi.org/10.1093/qje/qjaf010.

Holmstrom, B., Milgrom, P., 1991. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. J. Law, Econ. Organ. 7 (special_issue), 24–52. http://dx.doi.org/10.1093/jleo/7.special_issue.24.

Hopkins, E., 2012. Job market signaling of relative position, or Becker married to Spence. J. Eur. Econ. Assoc. 10 (2), 290–322. http://dx.doi.org/10.1111/j.1542-4774.2010.01047.x.

Hopkins, E., 2023. Is everything relative? A survey of the theory of matching tournaments. J. Econ. Surv. 37 (3), 688–714. http://dx.doi.org/10.1111/joes.12508.

Jones, B.F., 2021. The rise of research teams: Benefits and costs in economics. J. Econ. Perspect. 35 (2), 191–216. http://dx.doi.org/10.1257/jep.35.2.191.

Just, R.E., Pope, R.D., 1978. Stochastic specification of production functions and economic implications. J. Econometrics 7 (1), 67–86. http://dx.doi.org/10.1016/0304-4076(78)90006-4.

Kapeller, J., Steinerberger, S., 2016. Emergent phenomena in scientific publishing: A simulation exercise. Res. Policy 45 (10), 1945–1952. http://dx.doi.org/10.1016/j.respol.2016.08.004.

Korinek, A., 2023. Generative AI for economic research: Use cases and implications for economists. J. Econ. Lit. 61 (4), 1281–1317. http://dx.doi.org/10.1257/jel.20231736.

Kostadinov, R., Kuvalekar, A., 2022. Learning in relational contracts. Am. Econ. J.: Microeconomics 14 (1), 284–329. http://dx.doi.org/10.1257/mic.20190203.

Kremer, M., 1993. The O-ring theory of economic development. Q. J. Econ. 108 (3), 551–575. http://dx.doi.org/10.2307/2118400.

Lazear, E.P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. J. Political Econ. 89 (5), 841–864. http://dx.doi.org/10.1086/261010.

Levin, J., 2003. Relational incentive contracts. Am. Econ. Rev. 93 (3), 835–857. http://dx.doi.org/10.1257/000282803322157115.

Mangematin, V., 2000. PhD job market: Professional trajectories and incentives during the PhD. Res. Policy 29 (6), 741–756. http://dx.doi.org/10.1016/s0048-7333(99)00047-5.

Manso, G., 2011. Motivating innovation. J. Financ. 66 (5), 1823–1860. http://dx.doi.org/10.1111/j.1540-6261.2011.01688.x.

Merton, R.K., 1968. The Matthew effect in science: The reward and communication systems of science are considered. Science 159 (3810), 56–63. http://dx.doi.org/10.1126/science.159.3810.56.

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

Pluchino, A., Biondo, A.E., Rapisarda, A., 2018. Talent vs luck: The role of randomness in success and failure. Adv. Complex Syst. 21 (03n04), 1850014. http://dx.doi.org/10.1142/s0219525918500145.

Rotolo, D., Hicks, D., Martin, B.R., 2015. What is an emerging technology? Res. Policy 44 (10), 1827–1843. http://dx.doi.org/10.1016/j.respol.2015.06.006.

Rotolo, D., Messeni Petruzzelli, A., 2013. When does centrality matter? Scientific productivity and the moderating role of research specialization and cross-community ties. J. Organ. Behav. 34 (5), 648–670. http://dx.doi.org/10.1002/job.1822.

Schultz, R., Stansbury, A., 2022. Socioeconomic diversity of economics PhDs. Working Paper No. 22-4 Peterson Institute for International Economics, URL https://www.piie.com/publications/working-papers/2022/socioeconomic-diversity-economics-phds.

Spence, M., 1973. Job market signaling. Q. J. Econ. 87 (3), 355–374. http://dx.doi.org/10.2307/1882010.

Stansbury, A., Schultz, R., 2023. The economics profession's socioeconomic diversity problem. J. Econ. Perspect. 37 (4), 207–230. http://dx.doi.org/10.1257/jep.37.4.207.

Tambe, P.B., 2025. Reskilling the workforce for AI: Domain expertise and algorithmic literacy. Manag. Sci. 72 (1), 515–537. http://dx.doi.org/10.1287/mnsc.2022.03968.

Viner, N., Powell, P., Green, R., 2004. Institutionalized biases in the award of research grants: A preliminary analysis revisiting the principle of accumulative advantage. Res. Policy 33 (3), 443–454. http://dx.doi.org/10.1016/j.respol.2003.09.005.

Watson, J., 2021. Theoretical foundations of relational incentive contracts. Annu. Rev. Econ. 13 (1), 631–659. http://dx.doi.org/10.1146/annurev-economics-090820-110736.