NBER WORKING PAPER SERIES

A MODEL OF THE BABBAGE FIRM

Joshua S. Gans

Working Paper 34145
http://www.nber.org/papers/w34145

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2025

Thanks to Refine.ink, Claude Opus 4 and ChatGPT-o3-pro for excellent research assistance and seminar participants at the University of Toronto and Ajay Agrawal and Philippe Aghion for useful discussions. Responsibility for all errors remains my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

A Model of the Babbage Firm
Joshua S. Gans
NBER Working Paper No. 34145
August 2025
JEL No. D83, J24, L23, L25

ABSTRACT

This paper develops a unified model of the cognitive division of labour in a knowledge economy. Building on recent frameworks for knowledge creation and decision making under uncertainty, it distinguishes between specialists, who engage in costly "on the spot" reasoning to generate tacit knowledge around a focal point, and generalists, who search for and interpolate existing knowledge but deliver answers subject to error. The model characterises how these two types of workers should be allocated across a continuum of questions, given the location of codified knowledge points and the distribution of problems. It shows that optimal task assignment depends on the cognitive process through which information is processed rather than on skill endowments or task complexity. When specialists operate around both knowledge points, their allocation is shaped by their absolute advantage over generalists, leading to non contiguous specialist domains interspersed with generalist regions. When specialists cluster around a single point, a natural boundary emerges between specialist and generalist domains that shifts but persists despite changes in question distribution. Extending the analysis to a two period setting, the paper identifies when specialists should sacrifice static efficiency to codify their tacit discoveries, creating bridges that allow generalists to operate more effectively in the future. These results provide a formal microfoundation for Babbage's insights into the division of cognitive labour and offer predictions about how knowledge work responds to changes in the knowledge environment, the distribution of questions, and the patience of capital.

Joshua S. Gans
University of Toronto
Rotman School of Management
and NBER
joshua.gans@rotman.utoronto.ca

1 Introduction

The division of labour has long been understood as a way to increase productivity by allowing workers to specialise. In his landmark work Babbage (1835) observed that, beyond the mechanical gains highlighted by Adam Smith, division of labour allows organisations to match worker skills to task requirements. By decomposing production into finer tasks, firms can purchase precisely the skills needed for each component. This insight is particularly salient in the modern knowledge economy, in which problems are heterogeneous and different types of expertise are differentially valuable across the domain of questions that an organisation faces.

Understanding how cognitive labour should be allocated is a central challenge for knowledge-intensive organisations. When should a problem be assigned to a specialist whose tacit knowledge is deep but narrow, and when to a generalist who draws on a wider set of codified insights? How do these assignments adjust as the distribution of problems shifts or as the knowledge base expands? Much of the existing literature on task assignment and hierarchies takes skills, attention, or knowledge endowments as the primary drivers of division of labour. The theory of knowledge-based hierarchies emphasises that organisations must manage the acquisition, communication and use of knowledge because knowledge is an indispensable input embedded in individuals with limited time. Hierarchical structures allow less knowledgeable workers to handle routine tasks while experts direct attention to more complex questions, thereby economising on the time of experts. Mainstream economic models often abstract from this organisational problem, jumping directly to production functions that depend on a fixed set of inputs. Recent research has sought to incorporate organisations into equilibrium frameworks by modelling agents' knowledge, communication and span of control (e.g., Calvo and Wellisz, 1978; Lucas, 1978; Rosen, 1982; Garicano, 2000; Garicano and Rossi-Hansberg, 2004; Garicano and Hubbard, 2006), and empirical work links organisational changes to wage inequality through the scale-of-operations effect.

This paper takes a different perspective. The central insight is that the division of labour depends not on skill endowments or time allocation constraints per se but on the cognitive processes through which workers deploy knowledge. Specialists engage in "on the spot" reasoning: they incur costs to construct novel, tacit knowledge tailored to the question at hand, and once that cost is sunk, they deliver an answer with perfect precision. Generalists, by contrast, search for and leverage existing knowledge. Their decision rule depends on the variance of their interpolation across codified reference points, so they deliver answers that are subject to error. This distinction between creating and using knowledge implies that specialists will not benefit from additional codified points, whereas generalists

are highly sensitive to where knowledge is codified and how uncertainty evolves. In other words, the division of cognitive labour arises from how information is processed rather than from differences in innate ability or the complexity of tasks, and knowledge itself is not ordered by complexity or frequency as in "management by exception" frameworks. Capturing this requires modelling how information about a continuum of problems is organised and used in decision-making, which is the contribution of the present framework.

Building on the general theory of knowledge and uncertainty developed by Carnehl and Schneider (2025), I model knowledge as a set of codified question–answer pairs along a one-dimensional domain. *Generalists* can interpolate between existing knowledge points but deliver answers with variance that increases with distance from known points. *Specialists* pay a fixed cost to develop tacit knowledge around a focal point and answer questions there with no error, but their advantage decays with distance. The firm must decide which questions to assign to each worker type and whether specialists should invest in codifying their tacit discoveries so that generalists can use them in the future. The dynamic choice of codification introduces a trade-off between static efficiency and knowledge creation: specialists can "stretch" towards interior regions of the problem domain, sacrificing immediate advantage to codify new points, thereby building bridges that generalists will exploit later. This trade-off resembles the exploration–exploitation problem studied in learning models (e.g., Callander, 2011), yet it arises here because of the difference in cognitive processes rather than because agents face stochastic payoffs or uncertain technologies.

The baseline model implies sharp boundaries between specialist and generalist domains as a function of the location of codified knowledge points and the distribution of questions. When specialists anchor around both points, non-contiguous specialist regions emerge, separated by interior zones served efficiently by generalists; when specialists cluster around a single point, a persistent boundary separates domains even as the question distribution shifts. These predictions map naturally to settings in which workers differ by *cognitive technology* rather than by innate skill: specialists create task-specific knowledge on the spot; generalists search, interpolate, and apply codified knowledge subject to variance. In medicine, primary-care physicians (generalists) resolve high-frequency presentations using guidelines and risk scores, while cardiologists and other subspecialists (specialists) take frontier cases and some routine procedures for which experience confers an absolute advantage; procedure-volume and specialisation effects on performance and efficiency are well documented, and codified decision rules such as the HEART Pathway demonstrably reduce unnecessary testing and admissions (Starfield et al., 2005; Birkmeyer et al., 2003; Mahler et al., 2015). In cybersecurity operations, tier-1 SOC analysts (generalists) triage alerts with playbooks and signatures, escalating novel or ambiguous incidents to reverse engineers and threat

hunters (specialists); the standard incident-response playbooks codify known patterns, while zero-day analysis remains a frontier activity (National Institute of Standards and Technology, 2025; Tounsi and Rais, 2018; Bilge and Dumitras, 2012). In legal services, associates and generalist practitioners handle routine drafting and discovery, whereas senior partners and narrow specialists concentrate on cutting-edge issues and selectively retain repetitive tasks where their know-how yields exceptional efficiency; evidence on knowledge hierarchies in law quantifies large productivity returns from leveraging specialist expertise (Garicano and Rossi-Hansberg, 2006; Garicano and Hubbard, 2016).

A dynamic extension pushes this forward. A two-period extension formalises when specialists optimally divert effort from static advantage to codify answers that lower future generalist variance—especially when knowledge points are far apart and δ is high. Customer support offers a clear analogue: level-1 agents (generalists) rely on knowledge bases and answer catalogues; escalation engineers (specialists) craft one-off fixes and then codify them into articles or automated flows. Field evidence shows that playbooks and answer ranking systems reduce median resolution times and shrink the mass that requires escalation (Molino et al., 2018; Xu et al., 2024). Site-reliability engineering exhibits the same logic: on-call SREs (generalists) follow runbooks and SLO-driven procedures, while system owners and deep debuggers (specialists) resolve novel failure modes and convert them into troubleshooting guides (TSGs) and automation; large-scale evidence indicates extensive runbook adoption and institutionalisation of codified fixes (Beyer et al., 2016; Narayan et al., 2022). In AML/KYC compliance, frontline analysts (generalists) apply risk-based rules and red-flag indicators, escalating emerging typologies to financial-crime specialists, who design and tune the typologies before pushing them back as codified rules—textbook risk-based governance (Financial Action Task Force, 2014, 2024). Finally, in industrial yield engineering, line operators and shift leads (generalists) use standard operating procedures and statistical process control (SPC) to handle common deviations; yield engineers (specialists) investigate rare excursions, then institutionalise countermeasures into SOPs and control-chart rules, progressively shrinking variance and moving the specialist boundary outward (Montgomery, 2019; Adler et al., 1999).

My approach thus contributes conceptually by shifting attention from skills and task complexity to cognitive processes in the organisation of knowledge work. It also contributes technically by integrating learning and uncertainty into a model of task assignment and knowledge creation. The framework relates to research on specialisation and human capital (e.g., Becker, 1964; Rosen, 1983), on knowledge-based hierarchies and span of control (Garicano, 2000; Garicano and Rossi-Hansberg, 2004; Garicano and Hubbard, 2006; Calvo and Wellisz, 1978; Lucas, 1978; Rosen, 1982), and on optimal task assignment and matching (Roy,

1951; Acemoglu et al., 2007), but it differs by explicitly modelling how information is used in decision-making and how knowledge codification endogenously evolves over time. It also distinguishes itself from "management by exception" results by recognising that knowledge is not ordered by complexity or frequency; rather, specialists and generalists are distinguished by their cognitive technology—creating versus using knowledge. In sum, the paper offers a unified framework in which the location of knowledge points, the variance of generalist interpolation, and the dynamic creation of new knowledge jointly determine the division of cognitive labour.

The remainder of the paper is organised as follows. Section 2 presents the model, defining the knowledge domain, uncertainty structure and worker types. Section 3 analyses the optimal allocation of specialists and generalists when specialists can anchor around both existing knowledge points, and Section 4 studies the case in which specialists cluster around a single knowledge point. Section 5 extends the model to a competitive setting with organisational dynamics. Section 6 looks at whether the cognitive division of labour is limited by the extent of the market (scale). Section 7 introduces the two-period problem of dynamic codification and characterises when specialists sacrifice static efficiency to create knowledge bridges for the future. A final section concludes.

2 Model Setup

2.1 Knowledge and Uncertainty

Before turning to formal details, it is helpful to motivate the modelling choices. Knowledge-intensive organisations constantly face a continuum of problems: some lie near areas where codified answers exist, while others are sufficiently novel that solving them requires genuine creativity. To capture this continuum in the simplest possible way, I represent the domain of questions along a one-dimensional line and map each question to a unique answer. Proximity along this line is intended to reflect the intuition that questions that are "close" are likely to have related answers, whereas questions far from established knowledge represent more speculative problems. This abstraction strips away multi-dimensional complexity but preserves the key trade-off between leveraging existing knowledge and creating new knowledge.

Following Carnehl and Schneider (2025), I model knowledge as consisting of a set of question–answer pairs distributed along the real line. Each question $x \in \mathbb{R}$ has a unique answer $y(x) \in \mathbb{R}$. The relationship between questions and answers follows a standard Brownian motion, capturing the intuition that questions that are "close" to each other tend to have related answers, with this relationship becoming more uncertain as the distance between

them increases.

Formally, knowledge at a point in time consists of a finite collection of known question-answer pairs, denoted by $\mathcal{F} = \{(x_i, y(x_i))\}_{i=1}^{k}$ with $x_1 < x_2 < \ldots < x_k$. For simplicity, I focus on the case where $k = 2$, so there are two knowledge points at positions x_0 and x_1 with $x_0 < x_1$. The domain of questions is bounded by $x_L < x_0$ and $x_H > x_1$.

For questions that haven't yet been answered, one can form conjectures based on existing knowledge. For a question x located in the interval $[x_i, x_{i+1}]$ between two adjacent known questions, the conjecture follows a normal distribution with mean:

$$\mu_x(\mathcal{F}) = y(x_i) + \frac{x - x_i}{x_{i+1} - x_i}(y(x_{i+1}) - y(x_i)) \tag{1}$$

and variance:

$$\sigma_x^2(\mathcal{F}) = \frac{(x_{i+1} - x)(x - x_i)}{x_{i+1} - x_i}\,. \tag{2}$$

For questions outside the bounds of existing knowledge, the variance increases linearly with distance from the nearest known point:

$$\sigma_x^2(\mathcal{F}) = \begin{cases} x_0 - x & \text{for } x < x_0 \\ \\ x - x_1 & \text{for } x > x_1 \end{cases}\,. \tag{3}$$

The value of answering a question depends on how precisely one can address it. I assume there is a threshold $q > 0$ representing the organisation's tolerance for imprecision. When the uncertainty (variance) associated with a question exceeds this threshold, it becomes preferable to abstain from answering rather than risk providing a highly inaccurate answer. Equivalently, q can be interpreted as a cut-off between routine problems, for which extrapolations from known points are acceptable, and frontier problems that require original insight. In practice, for example, general practitioners refer patients to specialists when cases are sufficiently far from their expertise; the parameter q plays a similar role in the model.

This framework allows us to divide the knowledge domain into five distinct regions:

1. Region 1: $(x_L, x_0 - q)$ – questions where uncertainty strictly exceeds threshold q when measured from x_0;

2. Region 2: $[x_0 - q, x_0)$ – questions where uncertainty increases with distance from x_0 but remains at or below threshold q;

3. Region 3: $[x_0, x_1]$ – questions between knowledge points, with uncertainty highest at the midpoint;

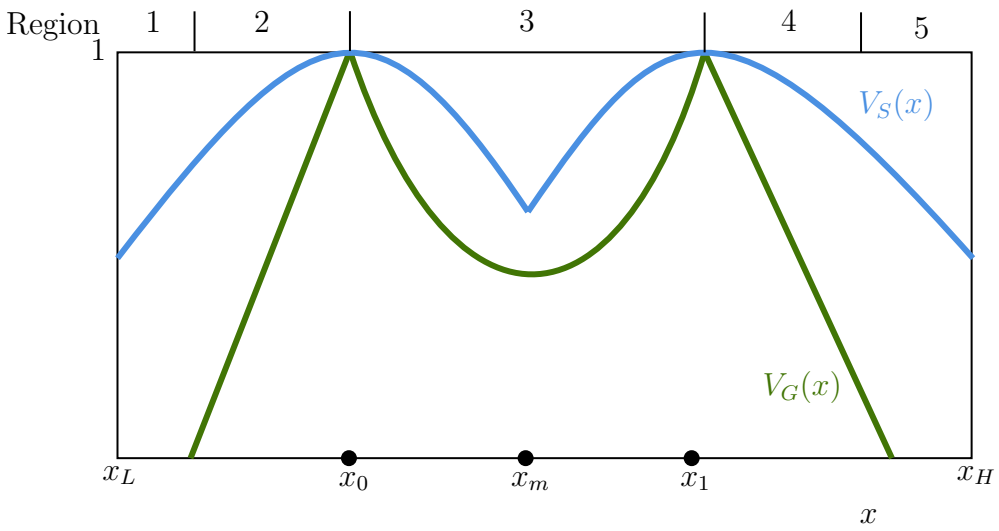

Figure 1: Value Created by Generalists and Specialists with Two Knowledge Points

4. Region 4: $(x_1, x_1 + q]$ – questions where uncertainty increases with distance from x_1 but remains at or below threshold q;

5. Region 5: $(x_1 + q, x_H)$ – questions where uncertainty strictly exceeds threshold q when measured from x_1.

These regions are shown in Figure 1. Throughout the paper, I assume that $x_1 - x_0 \leq 4q$, which ensures that even at the midpoint between knowledge points, the variance remains at or below the threshold q. This assumption allows questions in Region 3 to be potentially addressed based on existing knowledge.

I restrict attention to the case where the two knowledge points are not too far apart relative to the precision threshold. The inequality $x_1 - x_0 \leq 4q$ ensures that even the most uncertain question between the knowledge points (at the midpoint) has variance below q, so generalists can in principle address such interior questions. If the knowledge points were more widely separated, the midpoint variance would exceed q and generalists would never be allocated to any interior question; in that case, the problem would collapse to two independent specialist domains. Focusing on the case $x_1 - x_0 \leq 4q$ therefore isolates the interesting interaction between generalist and specialist capabilities without complicating the allocation rules.

2.2 Worker Types

I consider two types of knowledge workers: generalists and specialists. The terminology reflects the different mechanisms by which they produce answers. *Generalists* exploit codified,

commonly available knowledge within the organisation and therefore build their conjectures by interpolating between the known knowledge points x_0 and x_1. They are *knowledge users* rather than *knowledge creators*—they do not attempt to generate new insights beyond what is already codified. *Specialists*, by contrast, rely on tacit knowledge acquired through their own experience and judgment. Such knowledge cannot be codified or easily communicated, and thus it is not incorporated into the firm's shared knowledge base. In the spirit of Polanyi's paradox (Autor, 2014), specialists often "know more than they can tell," so their expertise remains localised and does not spill over to other workers.[1] Because tacit expertise must be built through costly experience, I assume that specialists are scarce relative to the number of questions, whereas generalists are abundant; the latter draw on codified knowledge that can be replicated and deployed widely.

2.2.1 Specialists

Specialists create new (tacit/private/ephemeral) knowledge centred around specific reference points. Specialists can create new insights, allowing them to operate across the entire domain. However, their effectiveness diminishes with distance from their focal points:

Definition 1 (Specialists)**.** *Specialists centered at knowledge point x_i incur cost $C_S(x) = \eta \cdot (x - x_i)^2$ where $\eta > 0$ is a cost parameter. Their value function is*

$$V_{Si}(x) = 1 - \eta \cdot (x - x_i)^2 ,$$

where the subscript $i \in \{0, 1\}$ denotes the knowledge point around which the specialist is centred.

This formulation captures the intuition that creating new knowledge becomes increasingly difficult as questions become more distant from reference points; see Figure 1.[2]

Let L_i denote the capacity of specialists associated with knowledge point $x_i \in \{x_0, x_1\}$. I assume specialists are scarce, with total capacity $L_S = L_0 + L_1 < x_H - x_L$, meaning they

[1] In the model, specialists have deep knowledge of a particular knowledge point and its conjectures, but only one such point; hence the term specialists. Therefore, it is assumed they do not know, take into account, or use any other available knowledge point to improve their answers.

[2] It is instructive to note that this "knowledge creation" cost function differs from that utilised by Carnehl and Schneider (2025). They consider new knowledge as arising out of a targeted search process and, consequently, the costs of finding knowledge increase with the variance associated with the answer process described earlier. That variance, as we already noted, is concave in the distance from existing knowledge, which is what gives the generalist value its linear shape for knowledge outside knowledge bounds. Here, however, I consider the generation of new answers to questions posed as a result of cognitive effort or capacity, and thus it is appropriate that the costs of doing so are convex. Nonetheless, the broad conclusions of the paper would be unchanged if specialist costs were linear.

cannot handle all questions in the domain. When endogenous hiring is considered later in the paper, specialist scarcity will be represented through a market-given hiring cost (c_S) per unit of specialist capacity.

2.2.2 Generalists

Generalists apply the existing codified knowledge from multiple available points to interpolate answers. They are thus *users* of knowledge rather than *creators*: unlike specialists, they do not attempt to generate new knowledge beyond what is already codified. Their value is determined by how precisely they can extrapolate from the known points to the question at hand:

Definition 2 (Generalists)**.** *Generalists can effectively handle questions where uncertainty is below the threshold q:*

$$V_G(x) = \begin{cases} 1 - \frac{\sigma_x^2(\mathcal{F})}{q} & \text{if } \sigma_x^2(\mathcal{F}) \leq q\,, \\ 0 & \text{otherwise.} \end{cases}$$

Generalists create maximum value ($V_G(x) = 1$) when answering questions directly at known points where uncertainty is zero; their value decreases linearly as uncertainty increases, reaching zero at the boundaries where $\sigma_x^2(\mathcal{F}) = q$.[3] Given our assumption that $x_1 - x_0 \leq 4q$, generalists can operate across Regions 2, 3, and 4, creating positive but varying value in each region (see Figure 1). Because codified knowledge can be replicated and diffused widely, I assume that generalists are abundant relative to the number of questions; there is sufficient generalist capacity to handle all questions where they can create positive value.

2.3 The Firm's Problem

Questions arise uniformly over the domain $[x_L, x_H]$. While the analysis can be extended to general distributions $f(x)$ at the cost of additional notational complexity, the main economic insights remain unchanged. The uniform distribution provides the clearest exposition of the key trade-offs.

[3] As Carnehl and Schneider (2025) argue the cutoff, q, can be microfounded in a straightforward manner. Let $\sigma^2(x)$ denote the posterior variance of a generalist at case x. (i) *Loss-based:* Suppose a correct answer yields benefit $\Delta > 0$ relative to an outside option, while the squared error e^2 entails expected loss $\alpha\,\mathbb{E}[e^2] = \alpha\,\sigma^2(x)$. The client accepts a generalist answer iff $\Delta - \alpha\sigma^2(x) \geq 0$, i.e. iff $\sigma^2(x) \leq q$ with $q := \Delta/\alpha$. Normalizing $\Delta = 1$ gives $V_G(x) = \max\{0, 1 - \sigma^2(x)/q\}$. (ii) *Chance-constraint:* Alternatively, suppose quality standards require $\Pr(|e| > \tau) \leq \beta$. If $e \sim \mathcal{N}(0, \sigma^2(x))$, this is equivalent to $\sigma^2(x) \leq q$ with $q := \tau^2/z_{1-\beta/2}^2$, where z_p is the p-quantile of the standard normal. In both cases q has the units of a variance and can be interpreted as a tolerance/standard implied by primitives (Δ, α) or (τ, β).

The firm must allocate generalists and specialists to questions to maximise total value creation, given the constraint on specialist capacity. When multiple specialist types exist (centered at x_0 and x_1), the firm's optimisation problem is

$$\max_{A_G, A_{S0}, A_{S1}} \frac{1}{x_H - x_L} \left[\int_{A_G} V_G(x)\, dx + \int_{A_{S0}} V_{S0}(x)\, dx + \int_{A_{S1}} V_{S1}(x)\, dx \right]$$

subject to

$$|A_{Si}| \leq L_i \quad \text{for } i \in \{0, 1\} \quad \text{(specialist capacity constraints)}, \tag{4}$$

$$A_G \cap A_{S0} = \emptyset, \quad A_G \cap A_{S1} = \emptyset, \quad A_{S0} \cap A_{S1} = \emptyset \quad \text{(non-overlapping allocation)}, \tag{5}$$

where A_G, A_{S0}, and A_{S1} are the sets of questions assigned to generalists and each type of specialist, respectively, and $|A|$ denotes the Lebesgue measure of set A.

To solve this problem, we must compare the absolute advantage of each specialist type over generalists:

$$A_i(x) \;=\; V_{Si}(x) - V_G(x) \quad \text{for } i \in \{0, 1\}.$$

The optimal allocation follows a threshold rule: specialist type i is assigned to questions where their absolute advantage exceeds a threshold value λ_i^*, which is determined by the capacity constraint:

$$\left| \{x \in [x_L, x_H] : A_i(x) \geq \lambda_i^*\} \right| \;=\; L_i.$$

When only one specialist type exists (say, centred at x_0), the problem simplifies to choosing between that specialist type and generalists, with the allocation determined by a single threshold λ^*.

3 Optimal Knowledge Worker Allocation

This section studies how the firm allocates its scarce specialist capacity and abundant generalist capacity when there are two potential focal points of expertise, x_0 and x_1. The case of two points already captures the fundamental trade-off between leveraging existing codified knowledge and creating new tacit knowledge without cluttering the exposition with additional parameters. Specialists may anchor their tacit knowledge around either x_0 or x_1. I assume both types have symmetric cost functions and equal capacity $L_S/2$, and that the knowledge points are sufficiently close—specifically $x_1 - x_0 < 4q$—so that generalists can span the gap between them. Generalists in practice handle a wide range of problems and coordinate care across settings; in contrast, tacit expertise is difficult to codify, reflecting

Polanyi's insight that "we can know more than we can tell."

3.1 Model and Absolute Advantage

Given a question $x \in [x_L, x_H]$, the firm must decide whether to allocate a generalist, a type-x_0 specialist, a type-x_1 specialist, or leave the question unanswered. Generalists simply interpolate from the known points, so their value function $V_G(x)$ is given in Section 2. A specialist of type $i \in \{0, 1\}$, anchored at x_i, generates value $V_{Si}(x) = 1 - \eta(x - x_i)^2$. The net gain from employing a specialist instead of a generalist is captured by the *absolute advantage* function

$$A_i(x) \;=\; V_{Si}(x) \;-\; V_G(x).$$

At the knowledge points themselves, both types are equally productive because generalists have zero variance there, so $A_i(x_i) = 0$. This function is depicted in Figure 2.

The shape of A_i is central to the allocation problem. Outside the region where generalists operate, $A_i(x) = 1 - \eta(x - x_i)^2$ is strictly concave and peaks at the boundary of that region. Within the generalist region, the trade-off between declining generalist precision and rising specialist cost can yield a single interior extremum. The following lemma summarises these properties.

Lemma 1 (Structure of absolute advantage)**.** *For each $i \in \{0, 1\}$, there are at most two local maxima of the absolute advantage function $A_i(x)$. If the specialist cost parameter satisfies $\eta > \frac{1}{2q^2}$, then A_i has an interior maximum at $x_i \mp \frac{1}{2\eta q}$ inside Region 2 (respectively Region 4), and a boundary maximum at the point where generalist variance reaches the precision threshold (i.e. at $x_i \mp q$). When $\eta \leq \frac{1}{2q^2}$, the interior extremum coincides with the boundary and the function is monotone between the boundary and x_i. In all cases $A_i(x_i) = 0$, so knowledge points are never strictly optimal for specialist deployment.*

The proofs of all results are in the appendix. Intuitively, a specialist's advantage over a generalist is small near their own knowledge point because generalists are equally precise there. As we move away, the specialist gains ground because they draw on tacit expertise, while the generalist's extrapolation becomes noisier. Beyond a point, however, further distance makes the specialist's effort prohibitively costly. This trade-off yields a "double-peaked" absolute advantage with one peak just outside the range of generalist competence and possibly another just inside it when knowledge creation costs are sufficiently convex.

3.2 Why Don't Specialists Use General Knowledge?

A reader might naturally ask why a specialist would ignore codified knowledge points that lie outside their immediate domain. In the framework developed here, specialists bear a fixed cost to build tacit expertise around a given reference point. Once that cost is sunk, the specialist delivers answers that are, by construction, perfectly accurate for the questions assigned. There is therefore no additional value to be gained by consulting public knowledge bases: incorporating a neighbouring reference point into the decision rule would not improve the precision of an answer that is already exact. Indeed, the modelling assumption captures the idea that tacit knowledge reflects experience and know-how that is difficult to articulate or augment with textbook insight. This assumption also clarifies the role of the dynamics considered later in the paper. As shown in Section 7, specialists invest in codifying some of the questions they answer not because they expect to use that codified knowledge themselves, but because doing so lowers future variance for generalists and hence raises the overall value of the organisation. In other words, the dynamic codification problem concerns the allocation of specialist time between immediate tasks and knowledge creation for others; it does not alter the fact that specialists do not consult generalist knowledge when forming their own answers.

This modelling choice can be relaxed. In practice, even the most expert clinicians, lawyers or engineers occasionally consult external databases or colleagues, especially when cases stray from familiar territory. If specialists produced answers with some residual error, either because their tacit knowledge is less than perfect or because the precision of their answer declines with distance from their core, then public knowledge points might become relevant. In such a setting, specialists would still outperform generalists in their home region, but would find it worthwhile to use codified information when venturing further afield. The boundaries between specialist and generalist domains could then become more diffuse, and the dynamic analysis in Section 7 would need to account for the possibility that codified knowledge improves specialists' future performance as well as that of generalists. As long as specialists retain a comparative advantage within some neighbourhood of their expertise, however, the main insights of the model — including the endogenous division of cognitive labour and the dynamic trade-offs associated with codification — would continue to hold.

3.3 Optimal Allocation

With equal specialist capacities and a uniform distribution of questions, the planner solves the assignment problem described in (4). Because specialists are scarce, it is optimal to allocate them to those questions where their absolute advantage over a generalist is greatest.

Let λ_i denote the shadow value of an additional unit of specialist capacity of type i. A threshold rule emerges: a question x is served by a type-i specialist if and only if $A_i(x) \geq \lambda_i$ and $A_i(x) - \lambda_i \geq A_{1-i}(x) - \lambda_{1-i}$. The thresholds (λ_0, λ_1) are endogenously chosen so that each specialist type is utilised exactly up to its capacity $L_S/2$.

Proposition 1 (Optimal assignment with two specialist types)**.** *Suppose $L_0 = L_1 = L_S/2$ and that questions are uniformly distributed over $[x_L, x_H]$. There exist unique thresholds $(\lambda_0^*, \lambda_1^*)$ such that in the optimal allocation:*

1. *A type-i specialist serves exactly those questions x for which $A_i(x) \geq \lambda_i^*$ and $A_i(x) - \lambda_i^* \geq A_{1-i}(x) - \lambda_{1-i}^*$. By construction, each type serves a set of measure $L_S/2$.*

2. *Generalists serve all remaining questions with $\sigma_x^2 \leq q$ that are not assigned to specialists. On questions where their variance exceeds q, generalists create zero value and are therefore never used.*

3. *Any question outside the generalist region that is not assigned to a specialist remains unanswered. Unserved questions arise only if the domain extends beyond $x_0 \pm q$ or $x_1 \pm q$, and scarce specialist capacity cannot cover those tails.*

The allocation rule in Proposition 1 has a stark interpretation. Generalists cover the "middle ground" where existing knowledge provides some traction, while specialists are deployed to the fringes of that domain or to pockets within it where their tacit expertise yields the greatest incremental value. Because the absolute advantage functions are non-monotonic, the specialist assignment sets need not be contiguous: scarcity may force the firm to leave a gap around each knowledge point even though specialists operate both just inside and just outside the range of generalist competence.[4]

[4]It is instructive to note that the analysis still goes through with imperfect specialist precision if specialists have small residual error. To see this, replace $V_{S,i}(x) = \max\{0,\, 1 - \eta_i \|x - x_i\|^2\}$ with $V_{S,i}^{\varepsilon}(x) = \max\{0,\, 1 - \varepsilon_i(x) - \eta_i \|x - x_i\|^2\}$, where $\varepsilon_i(x) \geq 0$ is unitless. Two convenient cases:

(i) *Constant misspecification.* If $\varepsilon_i(x) \equiv \bar{\varepsilon}_i$, then $A_i^{\varepsilon}(x) \equiv V_{S,i}^{\varepsilon}(x) - V_G(x) = A_i(x) - \bar{\varepsilon}_i$ is a pure level shift. The planner's pointwise rule, the threshold characterisation, and capacity clearing is unchanged; only the multiplier moves: $\lambda_i^* \mapsto \lambda_i^* + \bar{\varepsilon}_i$. In the hiring formulation introduced below, this is equivalent to replacing c_S by $c_S + \bar{\varepsilon}_i$. Assignment sets and their geometry (one or two intervals around x_i) are unchanged.

(ii) *Distance-dependent residual.* If $\varepsilon_i(x) = \tilde{\varepsilon}_i(\|x - x_i\|)$ is bounded and locally flat—say $0 \leq \varepsilon_i(x) \leq \bar{\varepsilon}_i$ and $\tilde{\varepsilon}_i'(r) \leq \kappa_i$ for all r with κ_i small relative to the curvature parameter η_i—then $A_i^{\varepsilon}(x) = A_i(x) - \varepsilon_i(x)$ preserves the single-peaked shape around x_i, so the super-level sets $\{x : A_i^{\varepsilon}(x) \geq \lambda\}$ remain unions of at most two intervals. Hence, the threshold rule in Proposition 1, the capacity/hiring equivalence $\lambda_i^* = c_S$, and all comparative statics are unchanged up to level shifts in the multipliers; only numerical cutoffs move.

Throughout, we continue to assume specialists do not use codified points directly, allowing $\varepsilon_i(x)$ to weakly *decrease* with additional codification near x_i strengthens specialists uniformly and leaves the qualitative assignment geometry and comparative statics unchanged under the same bounded-slope condition.

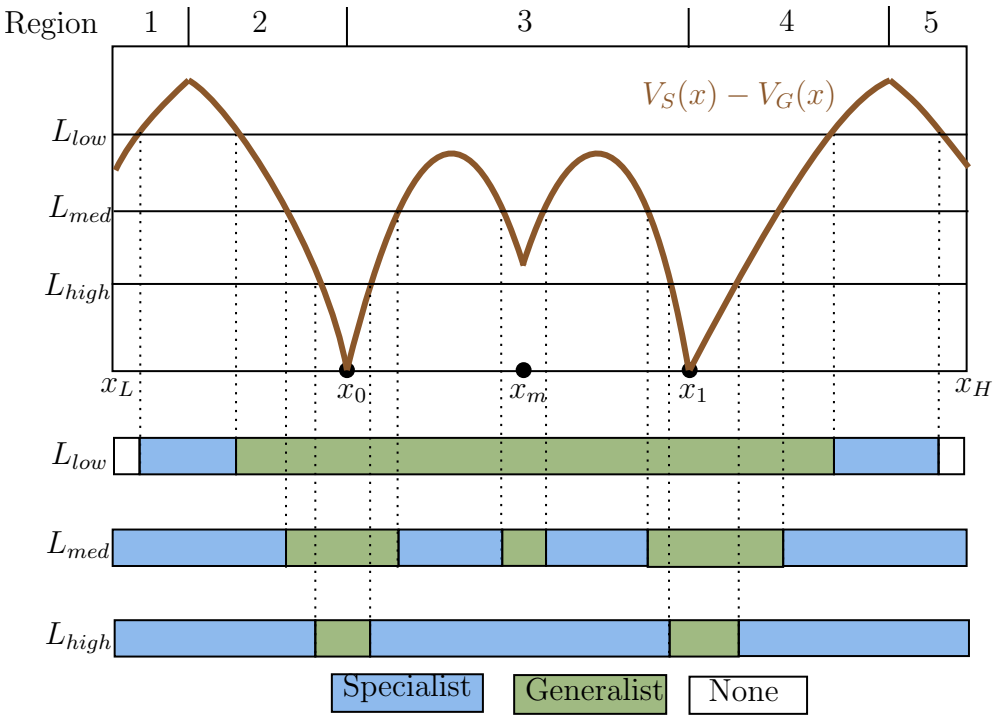

Figure 2: Absolute Advantage and Optimal Allocation

3.4 Comparative Statics

The threshold characterisation also allows sharp comparative statics. Two comparative statics of particular interest are changes in specialist capacity and changes in the width of the knowledge domain.

Specialist capacity. Starting from zero capacity, an increase in L_S lowers the threshold λ_i and causes the specialist assignment set to expand outward from the peak of A_i. Because A_i is double-peaked, this expansion occurs in phases. At low capacity, specialists concentrate in a single interval around the interior peak (if it exists) or the boundary. Once that interval is fully utilised, further increases in capacity open a second interval in the tail. Eventually, when capacity is sufficiently high, the two intervals merge, and the only unserved questions are in a small neighbourhood of each knowledge point where the absolute advantage is negative.

Proposition 2 (Expansion of specialist assignment)**.** *Fix one specialist type and let* $L \in [0, L_S/2]$ *denote its capacity. There exist thresholds* $0 \leq L^{\text{int}} \leq L^{\text{tail}} \leq L_S/2$ *such that:*

- *If* $L \leq L^{\text{int}}$*, the assignment set is a single interval contained in Region 2 (or Region 4), starting at the interior maximum of* A_i *and expanding outward as* L *rises.*
- *If* $L^{\text{int}} < L \leq L^{\text{tail}}$*, the assignment set consists of two disjoint intervals—one in Re-*

gion 2 and one in Region 1 (respectively one in Region 4 and one in Region 5). Additional capacity is devoted to the tail interval because its marginal advantage now exceeds the interior.

- *If $L^{\text{tail}} < L \leq L_S/2$, the two intervals grow and eventually merge into a connected set that excludes a neighbourhood of x_i where $A_i < 0$.*

The thresholds L^{int} and L^{tail} depend on η, q and the domain boundaries but not on the distribution of questions.

Domain width. Expanding the distance between x_0 and x_1 weakens generalist performance in the original interval because extrapolation must span a greater gap, but it improves generalist performance on the newly created questions just beyond x_1. Specialists respond by reoptimising their thresholds.

Proposition 3 (Effect of a wider domain)**.** *Consider an increase in the distance between the knowledge points from $X = x_1 - x_0$ to $X' = x_1' - x_0$ with $x_1' > x_1$. Then:*

1. *For $x \in [x_0, x_1)$, the variance of generalist conjectures increases from $\frac{(x_1-x)(x-x_0)}{x_1-x_0}$ to $\frac{(x_1'-x)(x-x_0)}{x_1'-x_0}$. Generalist precision deteriorates for all questions strictly between the original knowledge points.*

2. *For $x \in (x_1, x_1']$, the variance falls relative to single-point extrapolation because the question is now bounded by two knowledge points rather than one. Hence, generalist precision improves on these new questions.*

3. *The thresholds $(\lambda_0^*, \lambda_1^*)$ adjust, causing the specialist assignment sets to shift. In particular, more capacity may be devoted to the deteriorating interior as the domain widens, while some of the tail may be relinquished if generalists can now cover it more effectively.*

These comparative statics underscore that the firm's allocation of human capital depends not only on the relative costs and benefits of generalist and specialist expertise but also on the spatial structure of the knowledge landscape. When new knowledge arrives or the domain widens, the firm must reconsider where tacit expertise adds most value relative to codified knowledge.

3.5 Relationship to Knowledge-Based Hierarchies

The preceding comparative statics results can be reinterpreted through the lens of the knowledge-based hierarchies of Garicano (2000). Garicano's model treats workers as holders

of tacit solutions to problems and views organisational design as a matching technology: common problems are solved on the "production floor," while increasingly rare or difficult problems are escalated up a hierarchy of specialist problem solvers. The fundamental trade-off in that model is between the cost of acquiring knowledge and the cost of communicating with someone who has it; reductions in the cost of transmitting knowledge tilt the organisation toward relying more heavily on specialists, whereas reductions in the cost of acquiring knowledge increase the discretion of production workers and reduce their reliance on experts.

Our framework starts from a different technology: a generalist can interpolate between codified "knowledge points," while a specialist can generate tacit knowledge at a particular reference point. Nevertheless, the allocation rules derived above can be nested within a hierarchical theory by making the referral of a problem from a generalist to a specialist costly. Formally, suppose that a problem initially observed by a generalist can be "escalated" to the appropriate type–i specialist by paying a referral cost $m > 0$ (interpreted as time or cognitive effort). For a question at location x the organisation may either

(i) let the generalist answer directly, obtaining value $V_G(x)$; or

(ii) refer the question to a specialist, paying m, and then receive $V_{S_i}(x)$.

The net benefit from escalation is $\tilde{A}_i(x;m) \equiv V_{S_i}(x) - m - V_G(x)$; a question is escalated precisely when $\tilde{A}_i(x;m)$ exceeds the equilibrium threshold λ_i that clears the capacity constraint. Because $\partial \tilde{A}_i(x;m)/\partial m = -1$, the set of questions assigned to specialists shrinks monotonically as the cost of communication rises. Moreover, the shape of this set mirrors the double-peaked structure studied in the comparative statics above (see Proposition 2).[5]

Two cutoffs in the referral cost summarise how the organisation moves from a generalist–dominant regime to a Garicano–style hierarchy. When the cost m is small, the net advantage $\tilde{A}_i$ remains positive throughout the interior peaks of A_i and the specialist domains coincide with those in the baseline model. As m rises beyond a threshold m^{int}, the interior peak of $\tilde{A}_i$ falls below the cutoff, and specialists cease to be used in the middle of the knowledge domain: their allocation becomes "double–peaked," with specialists deployed only in the tails (Regions 1 and 5) and in narrow regions around the knowledge points. Finally, for m above a larger threshold m^{tail}, even the tail peaks of $\tilde{A}_i$ lie below the threshold, and specialists are used only in the extreme tails, exactly as in Garicano's hierarchy. Since the thresholds m^{int} and m^{tail} are decreasing functions of specialist capacity, higher capacity can partially offset referral costs by lowering λ_i.

[5]Here we suppress the specialist index i to lighten notation. The capacity constraint still pins down a threshold λ_i, and the relevant assignment set is $\{x : \tilde{A}_i(x;m) \geq \lambda_i\}$. When $m = 0$ we recover the baseline allocation of Proposition 1.

This comparative static connects the two theories in a precise way: increasing the cost of communicating with a specialist shifts the optimal assignment from the broad use of tacit expertise to a structure in which generalists handle all but the most novel questions. In the limit as $m \to \infty$, the organisation devolves to the canonical knowledge-based hierarchy described by Garicano (2000), with production workers solving common problems and passing rare ones up the chain. Conversely, when communication is cheap and knowledge points are available, the firm can flatten the hierarchy and rely on generalists.

Finally, note that the width of the knowledge domain interacts with communication costs. When the distance between knowledge points $X = x_1 - x_0$ grows sufficiently large, the variance of generalist conjectures in the interior exceeds the imprecision threshold even before adding any referral cost (Proposition 3). In this regime $\partial \tilde{A}_i(x; m)/\partial X < 0$: expanding the domain reduces the net advantage of specialists more sharply when $m > 0$. Intuitively, in a novel domain, the bridge created by interpolation becomes long enough that a generalist's conjecture is highly imprecise, yet referring the question to a specialist also requires paying m, so the incremental value of tacit expertise diminishes rapidly. This formalises the intuition that when specialists are costly, the firm may prefer to invest in codified knowledge only over relatively narrow domains, leaving truly novel areas to be handled by the few experts at the top of the hierarchy.

3.6 Endogenous Hiring Interpretation

The preceding analysis has taken the number of specialists of each type as fixed. This reduced-form assumption is convenient, but it also has a natural micro-foundation: in practice, firms hire specialists at a market rate and expand headcount until the last specialist hired adds no more value than the cost of hiring that specialist. Generalists, by contrast, are assumed to be in unlimited supply. Let $c_S > 0$ denote the cost per unit of specialist capacity. In an endogenous hiring model, the firm chooses capacities (L_0, L_1) and then assigns workers according to the absolute-advantage rule studied above. The objective is

$$\max_{L_0, L_1} \left\{ \frac{1}{x_H - x_L} \int_{x_L}^{x_H} \max\{V_G(x), V_{S0}(x), V_{S1}(x)\}\, dx \;-\; c_S\big(L_0 + L_1\big) \right\},$$

where the integrand, given (L_0, L_1), implements the optimal assignment from Proposition 1. Hiring more specialists expands the set of questions on which tacit expertise is deployed, but each additional unit must cover questions of strictly lower absolute advantage, so the marginal benefit of capacity is declining.

Proposition 4 (Equivalence of fixed capacity and endogenous hiring)**.** *For each fixed–capacity allocation with thresholds* $(\lambda_0^*, \lambda_1^*)$ *and specialist capacities* (L_0, L_1)*, there exists a hiring cost* $c_S = \lambda_i^*$ *such that* (L_0, L_1) *solves the endogenous hiring problem above. Conversely, for any hiring cost* $c_S > 0$*, the equilibrium capacities* (L_0^*, L_1^*) *satisfy*

$$A_i(x) \;=\; c_S \quad \textit{for every boundary point } x \in \partial S_i,$$

so that c_S *coincides with the threshold* $\lambda_i^*(L_i^*)$ *in the corresponding fixed–capacity model.*

The proposition shows that the fixed–capacity model is not ad hoc; it can be viewed as the reduced form of a standard optimal headcount problem. Hiring one more unit of specialist capacity is profitable only if there are questions where the absolute advantage of a specialist over a generalist exceeds the hiring cost. At the optimum, the firm expands S_i until the marginal question at the boundary has $A_i(x) = c_S$. Lower hiring costs lead to larger specialist capacities and therefore to lower thresholds λ_i^*, mirroring the inverse relationship between capacity and threshold in the fixed–capacity model.

4 Knowledge Creation

Building on the endogenous hiring framework developed above, I now examine how an organisation responds when new knowledge points can be created. Two contrasting strategies are particularly salient. One, which Carnehl and Schneider (2025) call *deepening*, enriches the existing domain by inserting a new knowledge point between two established ones. The other, *expansion*, pushes the frontier outward by adding a knowledge point beyond the current boundary. These choices mirror real-world dilemmas: a hospital might decide whether to invest in further training to improve performance in a specialised clinic or instead open a new clinic in a neighbouring city; a technology firm might choose between refining a proven product line and launching an entirely new product area. The analysis below formalises the trade-offs and shows how the relative cost of specialist expertise shapes the optimal strategy.

Consider two initial knowledge points x_0 and x_1 separated by distance $X = x_1 - x_0$. For concreteness, focus on the boundary case $X = 4q$ in which the midpoint variance exactly equals the precision threshold, $\sigma^2_{(x_0+x_1)/2}(\mathcal{F}) = q$. At this distance, generalists are just indifferent between operating or withdrawing at the centre of the interval, so it is exactly at this boundary that the choice between deepening and expansion becomes most acute.

4.1 Deepening the existing interval

The deepening strategy inserts a new knowledge point at the midpoint $x_m = (x_0 + x_1)/2$. Intuitively, this halves the length of the uncertain region and makes interpolated conjectures more precise throughout. Prior to deepening, the variance of a conjecture at $x \in [x_0, x_1]$ is

$$\sigma_x^2(\mathcal{F}) = \frac{(x - x_0)(x_1 - x)}{x_1 - x_0},$$

which is maximised at the midpoint and takes the value $X/4$. After inserting x_m, the interval splits into two sub-intervals of length $X/2$, and the maximum variance on each is halved to $X/8$. Because $X \leq 4q$, generalists can operate across the entire interval both before and after deepening, and their minimum value increases from $1 - X/(4q)$ to $1 - X/(8q)$. The next result summarises the welfare consequences.

Proposition 5 (Effect of knowledge deepening)**.** *Let $x_1 - x_0 = X \leq 4q$, and suppose a new knowledge point is added at the midpoint $x_m = (x_0 + x_1)/2$. Then:*

(i) The maximum conjecture variance falls from $X/4$ to $X/8$, so generalists' minimum value rises from $1 - X/(4q)$ to $1 - X/(8q)$.

(ii) The total value created by generalists on $[x_0, x_1]$ increases by $\Delta V_G = \frac{X^2}{12q}$.

(iii) The improvement ΔV_G is strictly increasing in X and achieves its maximum $4q/3$ when $X = 4q$.

The proof, given in the appendix, shows that the increase in generalist value arises entirely from reduced variance: splitting the interval shortens the "span" over which conjectures must be extrapolated. Deepening does not extend the domain on which generalists can work—the outer boundaries $x_0 \pm q$ and $x_1 \pm q$ remain unchanged. Instead, the key economic effect is to improve generalist productivity within the existing domain. When generalists become more productive, specialists are reallocated toward the extremities of the domain. In the high-cost regime where hiring costs exceed $1/(4\eta q)$, no specialists are ever hired at x_m, so the only benefit of deepening is improved generalist value. When costs are lower, deepening may also support a new type of specialist centred at x_m, though the returns to such specialists shrink quickly as X decreases.

In real terms, deepening corresponds to actions like investing in training or codifying tacit knowledge so that a wider range of workers can handle complex tasks. Hospitals employ such strategies to widen the scope of general practice, allowing more cases to be handled without specialist referral. Proposition 5 shows that the benefits are largest when existing knowledge

points are far apart (so variance reduction is greatest) and when specialist expertise is scarce and expensive.

4.2 Expanding the frontier

The expansion strategy introduces a new knowledge point beyond the current boundary. Following Carnehl and Schneider (2025), I take this new point to lie three precision units beyond x_1, at $x_2 = x_1 + 3q$. The choice of $3q$ maximises the direct value of interpolation on the new interval and will be justified formally below.

Proposition 6 (Effect of knowledge expansion)**.** *Assume $X = x_1 - x_0 \in [3q, 4q]$, and add a new knowledge point at $x_2 = x_1 + 3q$. Then:*

(i) The variance structure on the original interval $[x_0, x_1]$ is unchanged. Its maximum value remains $X/4$, which lies in $[3q/4, q]$ when $X \in [3q, 4q]$.

(ii) The new interval $[x_1, x_2]$ has length $3q$, and the maximum variance there is $3q/4 < q$. Thus, generalists can operate throughout $[x_1, x_2]$, and the total value they create on that segment is $3q/2$.

(iii) The point x_2 defines a new frontier beyond $x_2 + q$ on which generalists cannot operate. Specialists centred at x_2 can therefore serve this region and generate additional value, depending on the radius $r_S = \sqrt{(1 - c_S)/\eta}$ over which they can profitably operate.

Unlike deepening, expansion does not ameliorate any existing variance problem; the original interval $[x_0, x_1]$ is untouched and remains difficult if X is close to $4q$. Instead, expansion creates a new domain of questions that generalists can handle with relatively low variance. It also opens up a new region beyond x_2 where specialists face no competition from generalists, since single-point interpolation is too noisy to meet the precision threshold. In economic terms, expansion is akin to opening a new branch office or launching a new product category: it extends the organisation's reach but does not improve performance on existing tasks.

4.3 Deepening versus expansion

Which strategy is preferable? The answer depends on three primitives: the initial width $X = x_1 - x_0$, the size of the frontier beyond x_1,

$$\mathcal{F} \;:=\; \max\{0,\, x_H - x_1 - q\},$$

and the cost of specialist capacity c_S. Deepening inserts a codified point at $x_m = (x_0+x_1)/2$, which halves the span generalists must bridge and raises their interior value by $X^2/(12q)$ without changing the outer boundaries. Expansion instead adds a codified point at $x_2 = x_1+3q$, which creates a fresh $3q$ segment of low-variance questions that generalists can handle (worth $3q/2$) and also opens a specialist-only frontier beyond $x_2 + q$. Specialists centred at x_2 benefit from expansion because they can work in this frontier without competing with generalists, but they cannot substitute for the specialists who cover $[x_0, x_1]$.

To keep expressions compact, we let $a := \frac{2}{qX} + \eta$ and $r_S := \sqrt{\frac{1-c_S}{\eta}}$ and define the potential frontier allocation after expansion by $S_2 := [x_2 - r_S,\ x_2 + r_S] \cap (x_1 + q,\ x_H]$, which automatically covers both the symmetric and left-truncated cases. After deepening, the midpoint specialist's absolute advantage at distance $t = |x - x_m|$ is $A_m(t) = \frac{t}{q} - a\,t^2$; a concave quadratic whose superlevel sets are either empty or two symmetric bands around x_m. Using this, we can prove the following:

Proposition 7 (Comparing deepening and expansion)**.** *Suppose $X \in [3q, 4q]$ and questions are uniformly distributed over $[x_L, x_H]$. Let*

$$t_\pm := \frac{\frac{1}{q} \pm \sqrt{\frac{1}{q^2} - 4a\,c_S}}{2a} \quad \textit{(real iff } c_S \le \frac{1}{4aq^2}\textit{)}.$$

Then:

(i) High specialist cost $c_S > \frac{1}{4aq^2}$. *No midpoint specialist is hired after deepening. Expansion is strictly more profitable than deepening iff*

$$\frac{1}{x_H - x_L} \int_{S_2} \big[\, 1 - \eta(x - x_2)^2 - c_S \,\big]\, dx \;>\; \frac{X^2}{12q} \;-\; \frac{3q}{2}.$$

(ii) Otherwise $0 \le c_S \le \frac{1}{4aq^2}$. *After deepening, the midpoint specialist (if hired) is allocated on*

$$[x_m - t_+,\ x_m - t_-] \;\cup\; [x_m + t_-,\ x_m + t_+],$$

with net contribution

$$\Phi_m(c_S) \;=\; \frac{2}{x_H - x_L} \int_{t_-}^{t_+} \big[A_m(t) - c_S\big]\, dt = \frac{2}{x_H - x_L} \Big[\frac{t^2}{2q} - \frac{a t^3}{3} - c_S t\Big]_{t_-}^{t_+}.$$

Let the expansion frontier contribution be

$$\Psi_2(c_S) \;=\; \frac{1}{x_H - x_L} \int_{S_2} \big[\, 1 - \eta(x - x_2)^2 - c_S \,\big]\, dx.$$

Then expansion is strictly more profitable than deepening iff

$$\Psi_2(c_S) \;>\; \frac{X^2}{12q} \;-\; \frac{3q}{2} \;+\; \Phi_m(c_S).$$

If this fails, deepening is optimal.

Remark. *There is no $c_S \geq 0$ for which a midpoint specialist can profitably serve the entire interval $[x_0, x_1]$ after deepening: indeed $A_m(X/2) = -\eta X^2/4 < 0$, so the specialist loses to the generalist near the ends regardless of c_S.*

Intuitively, when c_S is high, the only effect of deepening is to upgrade generalists in $[x_0, x_1]$, so expansion wins only if the frontier is big enough and specialists reach far enough into it. As c_S falls, deepening begins to support a midpoint specialist on two interior bands; whether expansion beats deepening then hinges on the (discounted) length and quality of the frontier relative to the interior gain created by halving the span.

4.4 Choosing where to expand

A natural next question is where to place a *new* knowledge point when starting from a single point x_0. The choice matters for two reasons. First, a larger span between anchors raises the value created by generalists on the interior because interpolation becomes more informative; second, pushing the right anchor outward also enlarges the one-sided frontier beyond $x_1 + q$ on which generalists cannot operate and where specialists anchored at x_1 may be deployed. In short, greater spacing helps generalists inside but creates more (or preserves) specialist work outside.

Formally, pick a distance $d \in [0, 4q]$ for the new point $x_1 = x_0 + d$. The upper bound $4q$ ensures that generalists can operate throughout $[x_0, x_1]$.[6] Let $C := x_H - x_0 - q$ denote the available room to the right beyond the knee x_0+q. Specialists hired at cost $c_S \in [0, 1)$ per unit deliver value $V_{S1}(x) = 1-\eta(x-x_1)^2$ near their anchor; they are deployed only where this net value is non-negative. Let the one-sided specialist reach be written as $r_S := \sqrt{(1-c_S)/\eta}$ and the frontier length to the right of $x_1 + q$ as $F(d) := \big(C - d\big)_+$. Since the frontier is one-sided, the x_1-specialist can cover at most length $\ell(d) := \min\{r_S, F(d)\}$ on $(x_1 + q, x_H]$. Up to the uniform normalisation constant $1/(x_H - x_L)$, the d-dependent objective is

$$W(d) \;=\; \underbrace{\Big[d - \frac{d^2}{6q}\Big]}_{\text{interior generalist value on } [x_0,x_1]} \;+\; \underbrace{\int_0^{\ell(d)} \big[1 - c_S - \eta s^2\big]\, ds}_{\text{frontier specialist value to the right of } x_1}.$$

[6] With two anchors at distance L, the interior value generated by generalists is $V(L) = L - \frac{L^2}{6q}$; this is strictly concave in L and peaks at $L = 3q$.

The integral equals $\ell(d)\,(1-c_S) - \frac{\eta}{3}\,\ell(d)^3$.

Proposition 8 (Optimal placement of a new knowledge point)**.** *Starting from x_0, choose $d \in [0, 4q]$ to place $x_1 = x_0 + d$. Let $C := x_H - x_0 - q$ and $r_S := \sqrt{(1-c_S)/\eta}$. The maximiser d^* is unique and given by:*

(i) (Frontier slack) If $C - 4q \geq r_S$, then $d^* = 3q$.

(ii) (Frontier binds somewhere) If $C - 4q < r_S$, define

$$d_L := \max\{0,\, C - r_S\}, \qquad d_U := \min\{4q,\, C\},$$

and

$$d^\circ \;:=\; C + \frac{1}{6\eta q} \;-\; \frac{1}{2\eta}\sqrt{\left[\tfrac{1}{9q^2} + \tfrac{4\eta C}{3q} - 4\eta c_S\right]_+}.$$

Then

$$d^* = \begin{cases} d^\circ, & \text{if } d^\circ \in [d_L, d_U], \\ d_L, & \text{if } d^\circ < d_L \text{ or the square root is not real}, \\ d_U, & \text{if } d^\circ > d_U. \end{cases}$$

In all cases $d^* \leq 4q$. In regime (ii), d^* is (weakly) increasing in c_S, in η, and in C.

The proposition makes transparent why "$3q$ spacing" emerges as a robust benchmark: when there is ample room to the right, the frontier specialist can always work up to reach r_S, so moving x_1 only changes the interior—whose concave payoff is maximised at $d = 3q$. When the right frontier is *not* ample, however, pushing x_1 outward trades off interior gains against lost frontier: increasing d shrinks $F(d)$ one-for-one. The FOC equates the marginal interior benefit, $1 - d/(3q)$, to the marginal frontier loss, $(1 - c_S) - \eta(C - d)^2$, the latter falling in the square of the remaining room $(C - d)$. Two comparative statics follow. First, a higher c_S weakly raises d^*—when specialists are expensive, it is optimal to make more of the interior and lean less on the frontier. Second, a larger η (steeper decay of specialist value with distance) has the same effect. Both are consistent with the qualitative message of Section 4.3: interior strengthening is most attractive when specialist reach is short or dear.

5 Competition and Organisational Dynamics

The analysis above derives an optimal internal allocation of knowledge workers within a single firm. In practice, however, organisations compete in markets where some firms are explicitly specialist and others are broadly generalist. Boutique law practices specialise in

narrow fields such as antitrust or family law, while full-service firms offer a wide range of legal expertise. Similar patterns emerge in consulting, engineering and medicine, where some organisations build deep expertise in a single domain and others assemble broad portfolios of capabilities. This section develops a model of competition between such specialist and generalist firms, showing how integration costs and evolving demand can lead to entry, co-existence and exit of different organisational forms. The empirical work of Henderson (1993) on the photolithographic alignment equipment industry provides an illustrative benchmark. She documents how incumbent firms, initially dominant in contact and scanning projection aligners, failed to adapt to new technologies not because the innovations were economically radical, but because the organisational knowledge required to exploit them differed fundamentally from the incumbent firms' routines and information filters. Her case studies of Kasper, Perkin–Elmer and GCA reveal how organisations misinterpreted early signals of change and mis-evaluated entrants' designs, and her econometric evidence shows that incumbents invested less in projects that were radical in the organisational sense than in incremental projects and captured only a small fraction of the market for such innovations (Henderson, 1993). These findings underscore the importance of organisational frictions in determining who succeeds when technological domains shift.

5.1 Integration costs and organisational specialisation

Coordinating workers with fundamentally different cognitive approaches is costly. Specialists build deep, tacit knowledge around a single reference point and are evaluated on their precision in a narrow domain. Generalists cultivate broad pattern-recognition skills and are valued for their ability to interpolate between multiple reference points. Training, performance evaluation and decision-making systems must accommodate both modes of cognition, and communication between specialists and generalists is difficult because they literally speak different professional languages. The experiments recounted by Henderson provide vivid examples: Kasper dismissed complaints about the accuracy of its proximity aligner because it interpreted them through the lens of its experience with contact aligners and thus misdiagnosed the source of error; Perkin–Elmer and GCA failed to appreciate the architectural advantages of stepper technology because their engineers were organised around the components of the incumbent design. Such cases suggest that integrating divergent knowledge frameworks can be more difficult than acquiring new technical components.

To capture these frictions, we introduce a fixed integration cost $\gamma_S > 0$ that any firm incurs if it employs both specialists and generalists. This cost reflects the organisational infrastructure required to support parallel cognitive systems. Suppose a monopolist chooses

among three organisational forms: a mixed firm employing specialists and generalists, a pure generalist firm and a pure specialist firm. Let $V_G(x)$ be the quality delivered by a generalist at location x, and let $V_{Si}(x) = 1 - \eta(x - x_i)^2$ be the quality delivered by a specialist centred at $x_i \in \{x_0, x_1\}$. A mixed firm earns

$$\Pi^M_{\text{mixed}} = \int_{x_L}^{x_H} \max\{V_G(x), V_{S0}(x), V_{S1}(x)\} f(x)\, dx - c_S\big(L_0^* + L_1^*\big) - \gamma_S, \tag{6}$$

where L_i^* denotes the measure of locations at which specialist i is strictly better than the generalist. A pure generalist firm earns

$$\Pi^M_{\text{gen}} = \int_{\mathcal{G}} V_G(x) f(x)\, dx, \tag{7}$$

and a pure specialist firm earns

$$\Pi^M_{\text{spec}} = \max_{L_0, L_1} \left[\int_{\mathcal{S}} \max\{V_{S0}(x), V_{S1}(x)\} f(x)\, dx - c_S(L_0 + L_1) \right], \tag{8}$$

where the maximisation is over the number of specialists. Let

$$\Delta\Pi = \int_{x_L}^{x_H} \max\{V_G(x), V_{S0}(x), V_{S1}(x)\} f(x)\, dx - \max\{\Pi^M_{\text{gen}}, \Pi^M_{\text{spec}}\}, \tag{9}$$

be the additional value created by mixing relative to the best pure form before paying the integration cost. The monopolist chooses specialisation if and only if $\gamma_S > \Delta\Pi$. This threshold is increasing in the width of the domain and in the distance between the specialists' knowledge points, because a wider or more uneven domain creates more opportunities for complementarity between specialists and generalists. Conversely, high specialist costs c_S reduce $\Delta\Pi$ by making specialist deployment less attractive even when specialists have a quality advantage. When integration costs exceed the endogenous threshold γ_S^*, even a monopolist that internalises all value created chooses to operate either as a pure specialist or as a pure generalist. Henderson (1993) provides empirical support for this modelling assumption: she shows that incumbents in photolithography invested significantly less in projects that were radical in the organisational sense and captured virtually none of the market for such innovations. One interpretation is that the integration cost of combining old and new cognitive frameworks discouraged them from pursuing frontier technologies at all.

5.2 The competitive framework

We now embed the technology in a competitive market for knowledge services. Consumers are uniformly distributed over the domain $[x_L, x_H]$ and require answers to questions at their location. A consumer with quality valuation V who pays price p obtains utility $u = V - p$. Firms compete through personalised pricing; each consumer receives a bespoke quote reflecting the quality differential across firms. Under Bertrand competition with personalised pricing, a consumer at x is served by the firm offering the highest quality and pays

$$p_i^*(x) = V_i(x) - \max_{j \neq i} V_j(x), \tag{10}$$

which extracts exactly the consumer's switching cost. Three potential firms may operate: a generalist firm G with quality $V_G(x)$, an incumbent specialist S_0 located at x_0, and a frontier specialist S_1 located at x_1. Generalists have zero marginal cost and can serve an unlimited number of consumers. Specialist firms, by contrast, must hire one worker per customer at a cost of c_S and therefore serve location x only if the quality advantage over the next best alternative exceeds c_S. Entry into the specialist market involves choosing a capacity L_{Si} before observing rivals' decisions; capacity is sunk because hiring and training specialists takes time. Personalised pricing eliminates cross-subsidisation across customers, so competitive allocation coincides with the monopolist's optimal allocation: each question is answered by the worker type who maximises value net of hiring costs. The difference lies in rents. In a monopoly, the firm captures the full value of each answer, whereas in competition, the specialist's profit at x is only the quality difference relative to the next best firm. As a result, the integration cost that is just large enough to induce specialisation under monopoly is even more than enough to induce specialisation under competition: competition redistributes the monopolist's rents to consumers and leaves a narrower margin for mixed firms.

5.3 Domain evolution and market structure

Knowledge markets rarely remain static. New technologies, regulations and customer needs continually redraw the space of relevant questions. We model this evolution by letting a parameter $\theta \geqslant 0$ index time or environmental change. The lower and upper boundaries of the domain evolve according to

$$x_L(\theta) = x_L^0 + \theta,$$
$$x_H(\theta) = x_H^0 + \beta\theta,$$

where $\beta > 1$ captures the asymmetric nature of knowledge evolution: new questions emerge at the frontier faster than old questions disappear. The width of the domain is $W(\theta) = (x_H^0 - x_L^0) + (\beta - 1)\theta$, and the density of consumers falls to $f(x;\theta) = 1/W(\theta)$. In fields such as semiconductor lithography, alignment equipment and integrated circuits, new generations of technology arrive faster than old generations retire, so practitioners face an ever-widening frontier. Generalist organisations can deploy the same infrastructure across a broad range of locations, whereas specialists must adjust the location of their knowledge base or expand their repertoire.

5.4 Competitive dynamics

The evolution of the domain gives rise to predictable phases of competition. At small values of θ, the incumbent specialist S_0 enjoys a substantial exclusive territory around x_0 where generalists cannot operate. It hires a number of specialists equal to the mass of consumers in this region and prices at its quality advantage. As θ increases, the left boundary $x_L(\theta)$ moves right, shrinking the exclusive territory of S_0, while consumer density declines. At the same time, the right boundary $x_H(\theta)$ expands, enlarging the exclusive territory of the frontier specialist S_1. Because the right boundary expands faster than the left boundary contracts, there is a moment when S_1 first has a profitable region. Once this region is large enough to cover the hiring cost of specialists, S_1 enters. Eventually, the incumbent specialist's exclusive region vanishes, at which point S_0 can no longer recover its hiring cost and exits. Formally, the following proposition characterises these dynamics.

Proposition 9 (Market evolution under domain shift)**.** *Let $x_1 - x_0 \leqslant 4q$ and $\beta > 1$. There exist thresholds $0 < \theta^*_{\text{entry}} < \theta^*_{\text{exit}}$ such that*

1. *for $0 \leqslant \theta < \theta^*_{\text{entry}}$ the equilibrium is a duopoly between the generalist firm G and the incumbent specialist S_0;*
2. *for $\theta^*_{\text{entry}} \leqslant \theta < \theta^*_{\text{exit}}$ the equilibrium involves all three firms G, S_0 and S_1;*
3. *for $\theta \geqslant \theta^*_{\text{exit}}$ the equilibrium is a duopoly between G and the frontier specialist S_1.*

The proof shows that the length of the region where S_0 has an absolute quality advantage shrinks continuously with θ, while the length of the region where S_1 has such an advantage expands. Since the right boundary of the domain moves faster than the left boundary, the frontier specialist obtains a non-empty exclusive region strictly before the incumbent specialist loses its exclusive region. Equilibrium entry and exit decisions follow from the

requirement that a specialist hires exactly one worker per customer and breaks even at the margin.

To understand how profits evolve across these phases, let $\pi_i(\theta)$ denote firm i's equilibrium profit and let $m_i(\theta) = |\mathcal{M}_i(\theta)| / W(\theta)$ be its market share. The following corollary summarises the comparative statics.

Corollary 1 (Firm performance evolution)**.** *Under the conditions of Proposition 9, profits satisfy the following properties.*

1. *At $\theta = \theta^*_{\text{entry}}$ the incumbent specialist earns positive profit and the frontier specialist makes zero profit upon entry.*

2. *On $[0, \theta^*_{\text{exit}})$ the incumbent specialist's profit declines monotonically in θ, while on $(\theta^*_{\text{entry}}, \infty)$ the frontier specialist's profit increases monotonically. The generalist's profit increases initially as the incumbent retreats and then declines as the frontier entrant expands, attaining a unique maximum at some $\theta^*_G \in (\theta^*_{\text{entry}}, \theta^*_{\text{exit}})$.*

3. *There exist unique $\theta_1, \theta_2 \in (\theta^*_{\text{entry}}, \theta^*_{\text{exit}})$ with $\theta_1 < \theta_2$ such that $\pi_{S_1}(\theta_1) = \pi_G(\theta_1)$ and $\pi_{S_1}(\theta_2) = \pi_{S_0}(\theta_2)$. For $\theta > \theta_1$ the frontier specialist earns more than the generalist, and for $\theta > \theta_2$ the frontier specialist earns more than the incumbent specialist.*

4. *At $\theta = \theta^*_{\text{exit}}$ the frontier specialist earns strictly positive profit and the incumbent specialist earns zero profit upon exit.*

Figure 3 illustrates these patterns. The incumbent specialist experiences a steady decline in profit as its exclusive territory disappears and as consumer density falls. The frontier specialist initially earns nothing, but eventually grows to dominate the market. The generalist benefits from the incumbent's retreat but loses ground to the entrant; its profit profile is non-monotonic. These dynamics capture three distinct modes of organisational resilience. Incumbent specialists enjoy high margins in the early phase but are vulnerable to domain shift; their deep expertise becomes obsolete as knowledge moves away from their centre. Frontier specialists are patient: they incur entry costs only when a sufficiently large market appears at the frontier, but once they enter, their position strengthens over time. Generalist organisations never command the highest margins but survive every phase by flexibly reallocating effort to the interior of the domain. The sequence from incumbent dominance to entrant dominance, with generalists as persistent survivors, mirrors the evolution observed in photolithographic alignment equipment and in other knowledge-intensive industries (Henderson, 1993).

Although incumbents often decline and new entrants take over, Henderson's narrative also highlights the possibility of sustained survival through adaptation. Canon, for instance,

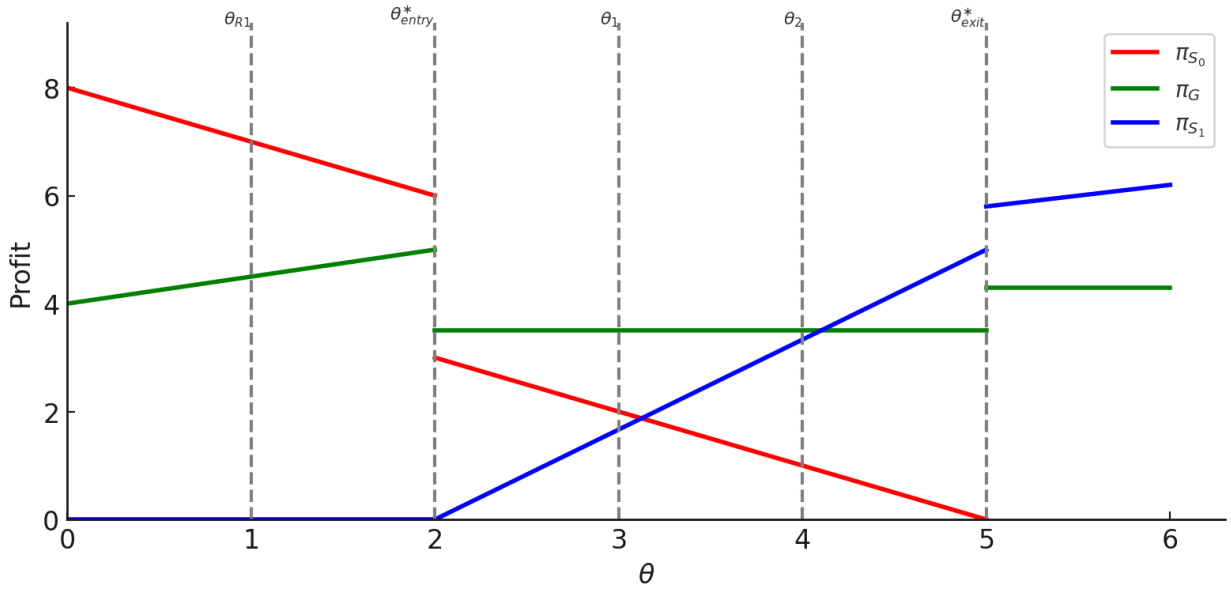

Figure 3: Evolution of firm profits under domain shift. The incumbent specialist's profit (red) declines as the domain shifts, the frontier specialist's profit (blue) rises following entry, and the generalist's profit (green) is non-monotonic but never zero. Canon's long-run persistence in photolithography, discussed in the text, resembles the non-zero dotted line.

introduced a proximity aligner in the late 1970s that redefined the market and became widely adopted. Unlike Kasper and GCA, which failed to adjust their knowledge base, Canon continued to compete across subsequent generations of photolithography equipment. In the context of our model, such persistence resembles the trajectory of the generalist firm in Figure 3: its profits never fall to zero even as specialist profits rise and fall, reflecting an ability to reposition at the frontier and maintain a non-trivial share of the market. Canon's experience thus suggests that firms willing to invest in new knowledge points and to reinterpret their organisational capabilities can survive radical change, even if they do not dominate the market (Henderson, 1993).

5.5 Strategic implications

The model highlights the trade-offs inherent in choosing an organisational form. Specialisation yields operational simplicity and high returns when questions cluster near a fixed knowledge point, but it exposes firms to positional risk when the domain shifts. Generalist organisations sacrifice depth for breadth; they rarely out-earn well-positioned specialists but retain a presence in all phases and have stronger incentives to invest in new knowledge points. In stable markets with concentrated demand, a specialist firm may be optimal; in rapidly evolving markets, a generalist organisation offers adaptive resilience. The history

of photolithography is again instructive: incumbents invested heavily in incremental improvements but were reluctant to reconfigure their knowledge base, while entrants invested in technologies that eventually redefined the industry (Henderson, 1993). Organisations anticipating similar patterns should therefore align their structure with expected market evolution.

5.6 Knowledge investment under competition

We now turn to examine the impact of competition on knowledge creation. As already studied, a firm can *deepen* the existing domain by inserting a codified point between x_0 and x_1 (which reduces interpolation variance for generalists everywhere in $[x_0, x_1]$) or *expand* the frontier by adding a point beyond x_1 (which creates a new interior segment $[x_1, x_2]$ that generalists can serve and opens an exclusive tail beyond $x_2 + q$). In the baseline (monopoly) environment, deepening raises generalist value on $[x_0, x_1]$ by $\Delta V_G = X^2/(12q)$ with $X \equiv x_1 - x_0$, while expansion adds $3q/2$ on the new interval and, crucially, creates a frontier region where specialists face no generalist competition.

Under personalised pricing in competition, the private return to an investment at location x is the induced increase in the firm's local quality advantage, $V_i(x) - \max_{j \neq i} V_j(x)$. This changes incentives asymmetrically. For a *generalist* firm, deepening raises V_G throughout the contested interior without strengthening rivals, and hence increases the extractable margin wherever the generalist already wins or is close to the threshold. By contrast, generalist-led expansion cannibalises that margin by inviting frontier specialist entry, so its private return can fall short of its social return even when $3q/2$ is large. For a *specialist* firm, deepening tends to compress its absolute-advantage peaks inside $[x_0, x_1]$—tightening price competition against the generalist—while forcing the firm to bear integration cost γ_S to maintain multiple cognitive systems. Expansion, instead, carves out an exclusive tail in which the specialist's price equals its quality, net of hiring cost c_S, and so is more appropriable. These forces rationalise why mixed specialist organisations underinvest in interior codification relative to generalists and why they tilt toward frontier-oriented investments, echoing Henderson (1993)'s evidence that incumbents captured a small share of radical (frontier) innovations while entrants—organisations naturally aligned with expansion—prevailed.[7]

Two comparative-static predictions follow directly from the model's primitives. First, competition tilts investment toward deepening for generalists and toward expansion for specialists. When knowledge points are sparse (large X), deepening's variance reduction is

[7]Empirically, Henderson (1993) finds that entrants captured more than half of the market for radical innovations while incumbents obtained less than seven percent, suggesting that specialists did not invest sufficiently to remain at the frontier.

first-order for generalists ($X^2/(12q)$ from Proposition 5), while specialists capture higher rents by expanding into tails where $V_G = 0$ (Proposition 6), avoiding both γ_S and interior price compression; the cross-over conditions are characterised in Proposition 7. Second, appropriability interacts with organisational form. The mixed form's incremental value over pure forms, $\Delta\Pi$ in equation (9), is eroded under competition because personalised pricing transfers gains to consumers in contested regions; as a result, the investment that maximises *private* value for a specialist is more frontier-seeking than under monopoly, whereas a generalist's private argmax coincides with interior deepening. As explored in the next section, a larger μ supports deeper cognitive division; these predictions imply that thicker markets produce more generalist deepening (denser interior knowledge points) and more specialist expansion (faster outward movement of the frontier). Empirically, this aligns with settings in which guideline proliferation and playbooks (deepening) shift routine volume to generalists, while new subspecialties and greenfield units (expansion) emerge at the edge.

5.7 Conclusion

The cognitive division of labour that drives internal efficiency also shapes market structure. Integration costs create an ecology of specialist and generalist firms, and the evolution of the domain of knowledge determines which organisational forms thrive. Real-world experiences, from stepper technology in photolithography to the rise of new specialties in medicine and consulting, reflect the patterns derived here: incumbents excel when questions cluster near their expertise, new specialists enter when the frontier expands, and generalists persist by flexibly serving the space between. Recognising these dynamics helps organisations choose appropriate structures and informs policy choices affecting the cost of coordinating diverse knowledge. The central message is that cognitive diversity is valuable but costly; understanding how that cost interacts with market evolution is essential for designing resilient knowledge-intensive organisations.

6 Market Size and Cognitive Specialisation

Classical economics emphasises that the scope for division of labour increases with market size. Adam Smith famously argued that "the division of labour is limited by the extent of the market," and Becker and Murphy formalised this idea by showing that the optimal physical division of labour is increasing in the scale of output (Becker and Murphy, 1992). In the context of cognitive work, a similar mechanism emerges once the organisation endogenises headcounts. When only a small flow of questions arrives each period, the fixed (or quasi-fixed)

cost of employing specialists cannot be justified, and all tasks are handled by generalists. As the volume of questions grows, the firm can amortise the cost of bringing specialists "online" over more tasks, supporting a deeper division of cognitive labour. This section formalises that intuition and derives comparative statics analogous to those in Becker and Murphy (1992).

6.1 Model with endogenous head-counts

Suppose the firm hires specialists at unit cost c_S and generalists at cost $c_G < c_S$ per unit of capacity. Let the flow of questions be μ per period; we interpret μ as a measure of market size or, at the firm level, as the size of a single organisation's problem set. Denote by $L_S(\mu)$ and $L_G(\mu)$ the capacities of specialists and generalists, respectively, that maximise expected surplus:

$$\max_{L_S, L_G} \left\{ \mu \int_{x_L}^{x_H} \max\{V_G(x),\, V_S(x)\}\, f(x)\, dx \;-\; c_S\, L_S \;-\; c_G\, L_G \right\}.$$

where $V_S(x) = \max\{V_{S0}(x), V_{S1}(x)\}$. Because specialists can answer at most one question at a time, L_S represents the measure of locations assigned to specialists in a given period, and L_G analogously denotes the measure served by generalists. The objective scales linearly in μ: a thicker market produces more questions, but hiring costs are per-unit and do not depend on μ. Let $A(x) = V_S(x) - V_G(x)$ denote the absolute advantage of specialists; by Proposition 1, the optimal static allocation assigns specialists exactly to those locations x where $A(x)$ exceeds a threshold determined by the shadow price of specialist capacity.

6.2 Comparative statics

The next result shows that market size governs the equilibrium mix of specialists and generalists in exactly the way that Becker and Murphy's model predicts.

Proposition 10 (Market size and cognitive specialisation)**.** *There exists a critical market size*

$$\mu^\star \;=\; \frac{c_S}{\displaystyle\int A(x)\, f(x)\, dx},$$

such that

$$L_S(\mu) \;=\; \begin{cases} 0, & \mu \le \mu^\star, \\ \text{strictly increasing in } \mu, & \mu > \mu^\star, \end{cases} \qquad \text{and} \quad L_G(\mu) \text{ is weakly decreasing in } \mu.$$

Moreover, if $c_S > c_G$ then the ratio $L_S(\mu)/L_G(\mu)$ is strictly increasing in μ whenever $\mu > \mu^\star$. In other words, a thicker market supports a deeper division of cognitive labour: beyond the critical scale, hiring more specialists and fewer generalists maximises value.

The proof uses the endogenous-hiring equivalence established in Proposition 4 to characterise the optimal capacities. When μ is small, the marginal revenue from assigning a question to a specialist is less than the marginal cost c_S, so $L_S(\mu) = 0$. As μ increases, the shadow price of specialist capacity rises, and the firm gradually expands specialist coverage until, in the limit, generalists are relegated only to those locations where they have a smaller efficiency loss relative to specialists. Because generalists cost less to employ, the firm always retains some generalist capacity, but the ratio of specialists to generalists rises monotonically in market size.

Interpreting μ as the size of a single firm's knowledge problem set, Proposition 10 implies that larger firms will exhibit deeper cognitive hierarchies: the ratio of specialists to generalists increases with the number of questions answered. Empirical patterns in law, medicine and consulting are consistent with this prediction. Multi-office law firms often field numerous narrow specialists (e.g. antitrust, tax, environmental) alongside generalist partners, while small firms rely primarily on generalists. Multi-specialty hospitals employ a wide range of experts because high patient volumes justify investing in costly specialist talent, whereas rural clinics must rely on general practitioners. At the industry level, μ can be interpreted as the extent of the external market. In trade-exposed sectors, global demand shocks and trade liberalisation raise μ, allowing firms to amortise the fixed cost of specialists over a larger customer base and thereby reshaping organisational form. In our model, a larger market leads to a higher specialist-to-generalist ratio, offering a microfoundation for the idea that the division of labour deepens as the market expands.

The critical threshold $\mu^\star$ depends inversely on the integral of the absolute advantage $A(x)$ over the distribution of questions. When specialists have a large quality advantage on average (i.e., $\int A(x)f(x)\,dx$ is large), even modest question flows justify hiring specialists. Conversely, when the specialist advantage is small or the cost c_S is large, the market must be substantial before specialists are employed. Thus, both the cognitive landscape (captured by A) and the cost structure determine how the division of labour responds to market size.

It is instructive to contrast two interpretations of μ. At the firm level, μ is the flow of tasks faced by a single organisation. When μ increases, for instance, because a firm acquires another or because its clients require more services, the firm responds by hiring more specialists relative to generalists. At the market level, μ measures the total number of questions across all consumers. An increase in μ then represents a broader market, perhaps due to population growth or trade liberalisation. In a competitive equilibrium, the interplay

of entry and exit (analysed in Section 5) determines how many specialist and generalist firms operate. The propositions in this section apply to each firm's internal allocation; when aggregated across firms, they predict that the overall supply of specialist services rises disproportionately with market size. This offers a new perspective on Smith's theorem: the extent of the market not only increases output but also deepens cognitive hierarchies within organisations.

This result aligns with Becker and Murphy's assertion that "the extent of the market limits the division of labour" by providing an explicit mechanism based on cognitive processes and the costs of knowledge work. Whereas Becker and Murphy emphasised coordination costs and the increasing returns to specialisation, Proposition 10 highlights the role of absolute advantage and hiring costs in mediating the effect of market size. The underlying intuition is similar: larger markets support more specialisation because fixed or quasi-fixed costs can be spread over more transactions. However, the cognitive context introduces new comparative statics. For example, if generalists become cheaper relative to specialists (i.e. c_S falls only slightly but c_G falls substantially), the ratio L_S/L_G may rise more slowly with μ. Conversely, if technological change increases the absolute advantage of specialists, the critical threshold $\mu^\star$ falls, and specialisation intensifies even in moderate markets. These predictions differentiate cognitive specialisation from physical division of labour and suggest empirical tests using data on organisational hiring patterns.

7 Dynamic Knowledge Codification

The core of the paper has thus far taken the knowledge base as given and asked how to allocate specialists and generalists across that domain. Yet a key attribute of knowledge-intensive organisations is their ability to convert *tacit* specialist insights into codified knowledge that generalists can later use. Such investments are dynamic: they impose an immediate opportunity cost, as specialists spend less time answering current questions, but they create a richer knowledge base in the future. In this section, I develop a two-period model to capture this trade-off.

The approach draws inspiration from the literature on trial–and–error learning in continuous choice spaces. In particular, Callander (2011) models a sequence of entrepreneurs who observe noisy outcomes from past choices and update beliefs about an underlying function drawn from a Brownian motion. Early experiments in his model are bold and traverse wide segments of the state space, while later experiments become incremental as learning accumulates. Similarly, the model below shows that organisations may initially "stretch" to codify new knowledge between existing expertise, but only if the expected future value

justifies the immediate loss from reassigning specialists.

7.1 Model Setup

Time is discrete with two periods $t \in \{1, 2\}$, and the firm discounts period-2 payoffs at rate $\delta \in (0, 1)$. At the beginning of period 1 the organisation's shared knowledge consists of two codified question–answer pairs $\mathcal{F}_1 = \{(x_0, y(x_0)), (x_1, y(x_1))\}$ with $x_0 < x_1$. Specialists of type $i \in \{0, 1\}$ possess deep tacit knowledge centred at x_i, exactly as in the static model. Generalists use only codified knowledge. For clarity, I normalise the measure of type-i specialists to $L_i > 0$ and the measure of generalists to be arbitrarily large.

Each specialist must divide her time between two activities in period 1: answering questions (which yields immediate value) and codifying the answers she produces (which has no period-1 payoff but expands the knowledge base for period 2). Let $\tau_i \in [0, 1]$ denote the *codification intensity* of type-i specialists: a fraction τ_i of their questions are codified for future use and the remaining $1 - \tau_i$ are simply answered. Formally, type-i specialists choose an *allocation set* $I_i \subset \mathbb{R}$ on which they operate in period 1; the set has Lebesgue measure $|I_i| = (1 - \tau_i)L_i$, reflecting the fact that codification reduces the measure of questions answered in period 1. For any question $x \in I_i$ assigned to a specialist, that specialist provides the answer in period 1 and codifies it with probability τ_i. The period-2 knowledge base is then

$$\mathcal{F}_2 \;=\; \mathcal{F}_1 \;\cup\; \big\{\,(x, y(x)) : x \in I_0 \cup I_1 \text{ and the answer was codified}\big\}.$$

Thus period-2 generalists can extrapolate between x_0, x_1 and any codified points in I_0 and I_1.

The distribution of period-1 questions is given by a continuous density $f\colon \mathbb{R} \to \mathbb{R}_+$ with compact support $[x_L, x_H]$; the firm's objective is the discounted sum of the value of all answered questions. I assume the same knowledge/variance structure from Section 2: generalists can only answer questions x if the variance of their interpolated answer does not exceed a threshold $q > 0$, while specialists can answer any question but suffer decreasing advantage with distance from their centre. Denote by $V_G(x; \mathcal{F})$ the value generated by a generalist at x given knowledge base $\mathcal{F}$ and by $V_{Si}(x)$ the value produced by a type-i specialist at x. As in the static model, $V_{Si}(x) = V_G(x; \mathcal{F}_1) + A_i(x)$, where A_i is the absolute advantage of specialists.

The firm chooses $(I_0, I_1, \tau_0, \tau_1)$ in period 1 to maximise

$$\Pi(I_0, I_1, \tau_0, \tau_1) \;=\; V_1(I_0, I_1, \tau_0, \tau_1) \;+\; \delta\, V_2(\mathcal{F}_2),$$

where V_1 is the measure-weighted value of answering period-1 questions (given codification intensities), and V_2 is the expected value generated in period 2 using the expanded knowledge base. Because generalists are abundant, period-2 questions are handled either by generalists or by specialists exactly as in the static allocation problem given $\mathcal{F}_2$.

7.2 Marginal Analysis

For any allocation set I_i and intensity τ_i, define

$$W_i(x) \;=\; A_i(x)\, f(x)$$

as the period-1 *weighted advantage* of allocating a type-i specialist to question x. Let x_i^* denote the point that maximises W_i under the static allocation. The opportunity cost of reallocating a type-i specialist to $x \in I_i$—instead of placing her at the best static point—is

$$\mathcal{L}_i(x) \;=\; W_i(x_i^*) \;-\; W_i(x).$$

Lemma 1 established that W_i is single-peaked, so $\mathcal{L}_i(x)$ is smallest near the peaks of A_i (either in the tails or in the interior) and largest near the knowledge points. Because $(1-\tau_i)L_i$ questions must be answered in period 1, reallocating a specialist to any x comes at this opportunity cost.

The marginal benefit of codifying at x arises in period 2. Fix an allocation (I_0, I_1) and let I_{-i} denote the allocation of the other specialist type. When a question $x \in I_i$ is codified, the variance of generalists' extrapolation at each point z weakly decreases. Let $\sigma_z^2(\mathcal{F})$ denote the generalist variance at z given knowledge $\mathcal{F}$. Define

$$\Delta V_G(z \mid x, I_{-i}) \;=\; \max\Big\{0,\, \frac{\sigma_z^2\big(\mathcal{F}_1 \cup I_{-i}\big) - \sigma_z^2\big(\mathcal{F}_1 \cup I_{-i} \cup \{x\}\big)}{q}\Big\}$$

as the increase in generalist value at z from adding $(x, y(x))$ to the knowledge base, normalised by the precision threshold q. Integrating over period-2 question densities yields the total marginal gain from codifying at x:

$$\mathcal{G}_i(x \mid I_{-i}) \;=\; \int_{x_L}^{x_H} \Delta V_G(z \mid x, I_{-i})\, f(z)\, dz.$$

The interdependence across specialists enters through the dependence of ΔV_G on I_{-i}. For instance, if type-1 specialists heavily codify in a region, then adding a nearby point by type-0 yields little additional variance reduction; but if type-1 codification is sparse, then type-0's

codification creates significant bridges that generalists can exploit.

7.3 First-Order Conditions

An interior optimum must equate marginal costs and marginal benefits. If $x \in I_i$ is an interior point of the allocation set, codifying at x yields expected period-2 gain $\delta\,\mathcal{G}_i(x \mid I_{-i})$ but displaces a specialist from her best static task at cost $\mathcal{L}_i(x)$. Therefore a necessary condition for an interior optimum is

$$\mathcal{L}_0(x_0) = \delta\,\mathcal{G}_0(x_0 \mid I_1), \tag{11}$$

$$\mathcal{L}_1(x_1) = \delta\,\mathcal{G}_1(x_1 \mid I_0), \tag{12}$$

for any interior points $x_0 \in I_0$ and $x_1 \in I_1$. These two equations are coupled because $\mathcal{G}_i$ depends on the other specialist's allocation. Boundary solutions obtain when I_i is empty or when the capacity constraint binds at a corner.

7.4 Structure of Variance Reduction

To analyse the marginal gains $\mathcal{G}_i$, it is useful to characterise how new knowledge points alter the variance landscape faced by generalists. If the only codified knowledge is $\mathcal{F}_1$, the variance at z is $\sigma_z^2(\mathcal{F}_1) = \min\{d(z,\mathcal{F}_1), \sigma^2_{\text{bridge}}(z)\}$, where $d(z,\mathcal{F}_1)$ is the distance to the closest knowledge point and σ^2_{bridge} is the variance when z lies between two knowledge points. Adding codified points in I_0 and I_1 can reduce this variance in two ways: it shrinks distances to the nearest knowledge point and it shortens the span over which generalists must interpolate. The next lemma formalises this structure.

Lemma 2 (Variance reduction structure)**.** *Fix allocation sets I_0 and I_1 and let $\mathcal{F}_2 = \mathcal{F}_1 \cup I_0 \cup I_1$. For any point z, the variance of a generalist's interpolated answer in period 2 is*

$$\sigma_z^2(\mathcal{F}_2) = \min\big\{\, d(z,\mathcal{F}_2),\ \sigma^2_{\text{bridge}}(z \mid \mathcal{F}_2)\big\},$$

where $d(z,\mathcal{F}_2)$ is the distance from z to the nearest point in $\mathcal{F}_2$, and $\sigma^2_{\text{bridge}}(z \mid \mathcal{F}_2)$ is the variance of linear interpolation between the two knowledge points that bracket z (if they exist).[8]

Lemma 2 shows that codification yields three sorts of benefits: *direct coverage*, as points in I_i have zero variance; *unilateral spillovers*, as points near I_i have reduced distance to their

[8] If z lies to the left of all points in $\mathcal{F}_2$, then the bracketing term is omitted and the first term applies; similarly if z lies to the right.

nearest knowledge point; and *bridges*, where interpolation spans are shortened when I_0 and I_1 both extend toward one another. The bridging effect is non-linear in the distance between codified sets and is the source of the strategic complementarity analysed next.

7.5 Strategic Complementarity

Because $\mathcal{G}_i$ depends on the other specialist's allocation, the optimal allocations are interdependent. When one specialist stretches toward the other, it reduces the span over which generalists must interpolate and thereby increases the benefit of the other specialist also moving toward that span. The following result formalises this complementarity. Let $x_0 \in I_0$ and $x_1 \in I_1$ with $x_0 < x_1$. For simplicity, I focus on uniform question densities and on symmetric specialists (i.e., $L_0 = L_1$ and $A_0(x) = A_1(x_1 + x_0 - x)$). Define

$$\Pi(x_0, x_1) \;:=\; V_1\big(\{x_0\}, \{x_1\}, \tau_0 = 1, \tau_1 = 1\big) + \delta\, V_2\big(\mathcal{F}_1 \cup \{x_0, x_1\}\big)$$

as the value of codifying single points x_0 and x_1 (relative to best static allocations elsewhere). Treating Π as a function of (x_0, x_1), we can examine its cross-partial derivative.

Proposition 11 (Strategic complementarity)**.** *Under a uniform question distribution and symmetric specialists, the cross-partial derivative of* $\Pi(x_0, x_1)$ *satisfies*

$$\frac{\partial^2 \Pi}{\partial x_0\, \partial x_1} \;<\; 0 \quad \textit{for} \quad 2q < x_1 - x_0 < 4q.$$

Thus when codified points are neither too close nor too far apart, specialists' allocations are strategic complements: moving one codification point closer to the other increases the marginal value of the other also moving inward.[9]

Note that while the mixed partial derivative is negative, this implies complementarity in terms of moving knowledge points closer, i.e., $\frac{\partial^2 \Pi}{\partial x_0\, \partial(-x_1)} > 0$. The proof demonstrates that the bridging term in $V_2(\mathcal{F}_1 \cup \{x_0, x_1\})$ is strictly supermodular in (x_0, x_1) for $x_1 - x_0$ in the stated interval. Intuitively, reducing the span from $x_1 - x_0$ to $x_1 - x_0 - \epsilon$ yields a second-order improvement in generalist variance near the midpoint, and this improvement is larger when the span is moderate than when it is very small or very large. Because V_1 is separable in (x_0, x_1), the sign of the cross-partial is governed by the bridging effect in V_2.

[9] If $x_1 - x_0 < 2q$, the interval between codified points is already sufficiently narrow that generalists have low variance throughout, so further bridging yields little gain; if $x_1 - x_0 > 4q$, the span is so wide that the marginal benefits of bridging are linear in the positions and the cross-partial is zero.

7.6 Optimal Codification Patterns

Equations (11) characterise interior solutions. To solve for the structure of optimal allocations, it is useful to consider symmetric specialists with identical capacities $L_0 = L_1 =: L$ and a uniform distribution of period-1 questions. Denote by $x_m = (x_0 + x_1)/2$ the midpoint between existing knowledge points, and normalise so that $x_1 - x_0 = 4q$; this ensures that generalists' variance at the midpoint equals the threshold q when no codification occurs. The next proposition summarises the patterns of optimal allocations as a function of the discount factor.

Proposition 12 (Optimal codification patterns)**.** *Suppose $x_1 - x_0 = 4q$, specialists are symmetric, and questions are uniformly distributed on $[x_L, x_H]$. Then there exist thresholds $0 < \underline{\delta} < \bar{\delta} < 1$ such that the optimal allocation has one of three forms:*

1. *If $\delta \leq \underline{\delta}$ (impatient firms), no codification occurs: each specialist concentrates on their static advantage region, which lies in the tails beyond $[x_0 + q, x_1 - q]$, and generalists operate in the interior as before.*

2. *If $\underline{\delta} < \delta < \bar{\delta}$ (intermediate patience), each specialist "stretches" toward the interior. There exist $\epsilon > 0$ and lengths $\ell_0, \ell_1 > 0$ such that*

$$I_0 = [x_0 + q - \epsilon,\ x_0 + q + \ell_0], \qquad I_1 = [x_1 - q - \ell_1,\ x_1 - q + \epsilon],$$

   and codification occurs on these intervals (i.e., $\tau_0, \tau_1 > 0$). Specialists thus operate in regions of moderate absolute advantage where bridging creates the greatest variance reduction for generalists.

3. *If $\delta \geq \bar{\delta}$ (very patient firms), specialists maximally stretch subject to non-negativity of payoffs. The allocation intervals expand until the marginal cost $\mathcal{L}_i$ equals the maximal marginal benefit $\delta\,\mathcal{G}_i$, so that specialists operate as close as possible to the midpoint without crowding out generalists entirely.*

With $x_1 - x_0 = 4q$, generalists can cover the whole interior $[x_0, x_1]$ but are weakest at the midpoint; their interpolation error there sits exactly at the threshold q. In period 1, specialists have static "peaks" in the tails (where their absolute advantage over generalists is largest), so pulling them inward sacrifices immediate value. Codifying, however, changes period 2: between any two codified points a generalist's variance scales with the *span* L between them (for the Brownian bridge, $\sigma^2(z) = t(1-t)L$ with $t \in [0, 1]$). Shrinking the residual gap that generalists must bridge, therefore, lowers variance everywhere inside that gap and raises generalist value across a whole interval, not just at a point.

Why codify near the "knees" x_0+q and x_1-q? Codifying right at x_0 or x_1 buys nothing (variance is already zero at the anchors), while codifying too deep into the middle is costly because specialists are far from their home bases. The knees mark locations that are still close enough for specialists to reach at modest static cost, yet far enough from the anchors that each unit of inward movement shortens the generalists' span by a full unit. That is where a small stretch creates the biggest *per-unit* reduction in the interior variance landscape.

The two sides reinforce one another. If the left specialist moves inward by Δ while the right does not, the gap contracts by Δ; if the right also moves by Δ, the gap contracts by 2Δ, and the induced variance drop is larger near the midpoint where generalists are tightest. This "zipper" effect is the strategic complementarity formalised in Proposition 11: an inward move on one side raises the marginal return to an inward move on the other.[10] As patience δ rises, the discounted *future* gain from shortening the span eventually outweighs the *current* static loss from relocating specialists. Hence the three regimes in Proposition 12: for low δ no codification (specialists stay at their static peaks); for intermediate δ short codification intervals I_0 and I_1 open around the knees and begin to "stretch" toward the midpoint; and for high δ these intervals expand further until the boundary condition $L_i(x) = \delta\, G_i(x \mid I_{-i})$ binds—i.e., until the specialist's marginal period-1 opportunity cost equals the marginal discounted variance reduction created for period-2 generalists. Throughout, generalists remain active in the interior; codification makes them *better* there by shortening the remaining gap they must bridge.

7.7 Codification Intensity

Given the optimal allocation intervals described in Proposition 12, one can determine the optimal codification intensities τ_i. When specialists operate only in the tails, codification yields no benefit because those points are already close to existing knowledge points; thus $\tau_i = 0$ in case (1). When specialists stretch toward the interior, codifying all answered questions may still be suboptimal if the marginal benefits vary within the allocation interval. Let

$$\bar{\mathcal{L}}_i := \frac{1}{|I_i|}\int_{I_i} \mathcal{L}_i(x)\,dx, \qquad \bar{\mathcal{G}}_i := \frac{1}{|I_i|}\int_{I_i} \mathcal{G}_i(x \mid I_{-i})\,dx$$

denote the average marginal cost and benefit over I_i. The next result provides the optimal intensities.

Proposition 13 (Codification intensity)**.** *In any allocation profile (I_0, I_1) that includes in-*

[10] Formally, the period-2 bridging term has increasing differences in $(x_0, -x_1)$, so inward moves are complements.

terior points where generalist variance exceeds q, the optimal codification intensities satisfy

$$\tau_i^* \;=\; \begin{cases} 0, & \text{if } I_i \subseteq \text{tail regions}, \\ \dfrac{\delta\,\bar{\mathcal{G}}_i - \bar{\mathcal{L}}_i}{\delta\,\bar{\mathcal{G}}_i}, & \text{if } I_i \cap (x_0+q,\, x_1-q) \neq \emptyset, \end{cases}$$

provided $\delta\,\bar{\mathcal{G}}_i > \bar{\mathcal{L}}_i$ so that codifying some points yields a net gain. In particular, $\tau_i^* > 0$ only when specialists operate in regions where the marginal future benefit of creating bridges outweighs the opportunity cost of shifting away from static advantage.

The expression for τ_i^* is reminiscent of a standard rule in investment problems: codify up to the point where the discounted marginal benefit equals the marginal cost. Observe that τ_i^* increases with δ and with the dispersion of I_{-i}, since both raise the expected variance reduction.

7.8 The Value of Patience

How much does patience increase firm value? Let $V^*(\delta)$ denote the maximised value as a function of the discount factor. Because the optimal allocation depends on δ, differentiating under the optimum requires taking account of how I_0^* and I_1^* change with δ. The following result decomposes the derivative into a direct effect of discounting and an indirect effect via allocation changes.

Proposition 14 (Value of patience)**.** *Let $V^*(\delta) = \Pi(I_0^*(\delta), I_1^*(\delta), \tau_0^*(\delta), \tau_1^*(\delta))$ be the optimal value. Then*

$$\frac{dV^*}{d\delta} \;=\; \frac{\partial V^*}{\partial \delta} \;+\; \sum_{i\in\{0,1\}} \frac{\partial V^*}{\partial I_i}\,\frac{dI_i^*}{d\delta} \;+\; \sum_{i\in\{0,1\}} \frac{\partial V^*}{\partial \tau_i}\,\frac{d\tau_i^*}{d\delta}.$$

Moreover, when $x_1 - x_0 \approx 4q$ the indirect term $\sum_i \frac{\partial V^*}{\partial I_i}\frac{dI_i^*}{d\delta}$ is strictly positive and dominates the direct term for sufficiently small q, indicating that increased patience raises firm value mainly by encouraging specialists to stretch and build bridges.[11]

The proof applies the envelope theorem and the fact that the interior first-order conditions pin down the optimal positions of specialists. The result underscores that patient investors not only weigh period-2 payoffs more heavily but also induce organisational choices that expand the future knowledge base. This mechanism links the cognitive division of labour to dynamic investment incentives, in line with the learning-by-experimenting literature.

[11] When q is small, generalists have low tolerance for variance, so the bridging effect of codification is particularly valuable.

7.9 Discussion

Two themes echo throughout the results above. First, optimal codification involves sacrificing static efficiency by reallocating specialists to points of moderate advantage. This is a canonical "exploration vs. exploitation" trade-off: specialists forsake some immediate productivity in order to transform tacit knowledge into public knowledge that generalists can leverage in the future. Second, the value of such exploration is highly sensitive to the distance between existing knowledge points. When $x_1 - x_0$ is of order $4q$, small stretches by both specialists create bridges that substantially reduce generalist variance in the interior. When the gap is smaller, the interior is already well served; when the gap is larger, bridging is too costly.

Callander (2011) analyses a dynamic search problem in which entrepreneurs sequentially choose points on a line to maximise profit generated by an unknown function drawn from a Brownian motion. Early entrepreneurs experiment with bold leaps that provide maximal information gain, whereas later entrepreneurs take smaller steps as knowledge accumulates. Although the formal environments differ—Callander's agents observe outcomes rather than interpolate between known points—the qualitative insights are parallel. In both models, incremental investments in knowledge creation connect previously isolated regions of the state space and produce outsized returns precisely when the residual uncertainty is neither trivial nor overwhelming. The two-period structure here captures one step of such learning dynamics: specialists choose whether to invest in bridges that generalists will exploit in the future. Extending the model to many periods—with codified points in one period becoming anchors for further codification in subsequent periods—would generate cascades of knowledge accretion similar to the product life cycle dynamics documented by Callander (2011). Doing so is beyond the scope of the present analysis but offers an intriguing direction for future research.

8 Conclusion

This paper develops a microfoundation for the cognitive division of labour in knowledge-intensive organisations. Rather than recapitulate the results, it is worth looking ahead to the questions that this framework opens up. A particularly compelling avenue concerns the role of artificial intelligence in shaping cognitive hierarchies.

Recent evidence from academia and industry suggests that generative artificial intelligence is lowering the cost of expertise (Agrawal et al., 2024a). Such tools are also impacting knowledge creation (Agrawal et al., 2024b). These developments raise the possibility that

the shadow cost of specialist capacity in our model, c_S, may fall dramatically over time (Agrawal et al., 2025). At the same time, AI can rapidly ingest and synthesise vast amounts of data, helping non-experts bridge disparate knowledge sources (Gans, 2025). From the perspective of this paper, AI technologies thus act on both sides of the cognitive hierarchy: they improve the interpolation technology available to generalists and reduce the cost of answering frontier questions for specialists.

Future work could study how the organisational equilibria derived here respond when AI lowers both c_S and the variance of generalist extrapolation. If AI reduces the absolute advantage of specialists by making generalist errors negligible, the critical market size at which specialist hierarchies emerge may increase, and organisations may flatten as generalists armed with AI handle a broader swath of problems. On the other hand, if AI primarily reduces the cost of codifying and searching for specialised answers, specialisation could deepen: more questions would be routed to highly leveraged experts, and tacit knowledge may be codified and reused more efficiently (Ide and Talamas, 2025). There is also a dynamic aspect: AI may allow knowledge to be codified in real time, altering the trade-off between exploration and exploitation analysed in our two-period extension.

A further research agenda involves endogenous adoption of AI tools. Firms could choose not only how many generalists and specialists to hire but also whether to invest in AI systems that augment each group. AI may substitute for some tasks while complementing others, leading to new boundary problems between human and machine cognition. In sectors where tacit knowledge is rapidly codified via AI, organisational learning could accelerate, compressing the distance between knowledge points and reshaping the geography of questions. Conversely, if AI commoditises routine expertise, human effort may shift toward truly novel or creative tasks where tacit reasoning still commands a premium (Autor and Thompson, 2025).

By embedding these technological forces into the model, one can explore how the division of labour adapts to improvements in the tools of thought. Such extensions would speak directly to current debates over the future of work and the organisation of knowledge production. Far from rendering human expertise obsolete, artificial intelligence may alter the margins along which specialists and generalists are valuable. Understanding these margins is a promising challenge for future research.

References

Acemoglu, D., P. Aghion, C. Lelarge, J. Van Reenen, and F. Zilibotti (2007): "Technology, information, and the decentralization of the firm," *Quarterly Journal of Economics*, 122, 1759–1799.

Adler, P. S., B. Goldoftas, and D. I. Levine (1999): "Flexibility vs. Efficiency? A Case Study of Model Changeovers in the Toyota Production System," *Organization Science*, 10, 43–68.

Agrawal, A., J. S. Gans, and A. Goldfarb (2024a): "The Turing Transformation: Artificial Intelligence, Intelligence Augmentation, and Skill Premiums," *Harvard Data Science Review*.

——— (2025): "The Economics of Bicycles for the Mind," Tech. rep., National Bureau of Economic Research.

Agrawal, A., J. McHale, and A. Oettl (2024b): "Artificial intelligence and scientific discovery: A model of prioritized search," *Research Policy*, 53, 104989.

Autor, D. (2014): "Polanyi's paradox and the shape of employment growth," Tech. rep., National Bureau of Economic Research.

Autor, D. and N. Thompson (2025): "Expertise," *Journal of the European Economic Association*, jvaf023.

Babbage, C. (1835): *On the Economy of Machinery and Manufactures*, Charles Knight.

Becker, G. S. (1964): *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, University of Chicago Press.

Becker, G. S. and K. M. Murphy (1992): "The division of labor, coordination costs, and knowledge," *Quarterly Journal of Economics*, 107, 1137–1160.

Beyer, B., C. Jones, J. Petoff, and N. R. Murphy, eds. (2016): *Site Reliability Engineering: How Google Runs Production Systems*, Sebastopol, CA: O'Reilly Media.

Bilge, L. and T. Dumitras (2012): "Before We Knew It: An Empirical Study of Zero-Day Attacks in the Real World," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS)*, 833–844.

Birkmeyer, J. D., T. A. Stukel, A. E. Siewers, P. P. Goodney, D. E. Wennberg, and F. L. Lucas (2003): "Surgeon Volume and Operative Mortality in the United States," *The New England Journal of Medicine*, 349, 2117–2127.

Callander, S. (2011): "Searching and learning by trial and error," *American Economic Review*, 101, 2277–2308.

Calvo, G. A. and S. Wellisz (1978): "Hierarchy, Ability, and Income Distribution," *Journal of Political Economy*, 86, 991–1010.

Carnehl, C. and J. Schneider (2025): "A Quest for Knowledge," *Econometrica*, 93, 623–659.

Financial Action Task Force (2014): "Risk-Based Approach for the Banking Sector," Tech. rep., FATF.

——— (2024): "Money Laundering National Risk Assessment Guidance," Tech. rep., FATF.

Gans, J. S. (2025): "A Quest for AI Knowledge," Tech. rep., National Bureau of Economic Research.

Garicano, L. (2000): "Hierarchies, markets, and the limits of knowledge-based organizations," *Journal of Economic Literature*, 38, 874–904.

Garicano, L. and T. N. Hubbard (2006): "The organization of knowledge in small firms," *Quarterly Journal of Economics*, 123, 1349–1384.

——— (2016): "Returns to Knowledge Hierarchies," *Journal of Law, Economics, and Organization*, 32, 653–684.

Garicano, L. and E. Rossi-Hansberg (2004): "Inequality and the Organization of Knowledge," *American Economic Review*, 94, 197–202.

Garicano, L. and E. Rossi-Hansberg (2006): "Organization and Inequality in a Knowledge Economy," *Quarterly Journal of Economics*, 121, 1383–1435.

Henderson, R. (1993): "Underinvestment and incompetence as responses to radical innovation: Evidence from the photolithographic alignment equipment industry," *The Rand journal of economics*, 248–270.

Ide, E. and E. Talamas (2025): "Artificial intelligence in the knowledge economy," *Journal of Political Economy*, forthcoming.

Lucas, R. E. (1978): "On the Size Distribution of Business Firms," *Bell Journal of Economics*, 9, 508–523.

Mahler, S. A., R. F. Riley, B. C. Hiestand, et al. (2015): "The HEART Pathway Randomized Trial: Identifying Emergency Department Patients With Acute Chest Pain for Early Discharge," *Circulation: Cardiovascular Quality and Outcomes*.

Molino, P., H. Zheng, and Y. Wang (2018): "COTA: Improving the Speed and Accuracy of Customer Support through Ranking and Deep Networks," arXiv:1807.01337.

Montgomery, D. C. (2019): *Introduction to Statistical Quality Control*, Hoboken, NJ: Wiley, 8th ed.

Narayan, A. et al. (2022): "AutoTSG: Learning and Synthesis for Incident Troubleshooting," arXiv:2205.13457.

National Institute of Standards and Technology (2025): "Computer Security Incident Response Guidance (SP 800-61, Revision 3)," Tech. rep., NIST.

Rosen, S. (1982): "Authority, Control, and the Distribution of Earnings," *Bell Journal of Economics*, 13, 311–323.

——— (1983): "Specialization and Human Capital," *Journal of Labor Economics*, 1, 43–49.

Roy, A. D. (1951): "The distribution of earnings and of individual output," *The Economic Journal*, 60, 489–505.

Starfield, B., L. Shi, and J. Macinko (2005): "Contribution of Primary Care to Health Systems and Health," *The Milbank Quarterly*, 83, 457–502.

Tounsi, W. and H. Rais (2018): "A Survey on Technical Threat Intelligence in the Age of Sophisticated Cyber Attacks," *Computers & Security*, 72, 212–233.

Xu, Z., M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li (2024): "Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering," arXiv:2404.17723.

A Proofs

Proof of Lemma 1 (Structure of absolute advantage)

Fix a specialist type i and, without loss of generality, focus on $i = 0$; the proof for $i = 1$ is symmetric. Outside the domain where generalists operate (Region 1), $V_G(x) = 0$ and $A_0(x) = 1 - \eta(x - x_0)^2$ is a downward-opening quadratic. When constrained to $x \leq x_0 - q$ its maximum occurs at the boundary $x = x_0 - q$, since A_0 is decreasing as one moves away from x_0.

Now consider the region where generalists operate and uncertainty is below the threshold (Region 2: $x \in [x_0 - q, x_0]$). There $V_G(x) = 1 - \frac{x_0 - x}{q}$ and hence

$$A_0(x) \;=\; 1 - \eta(x - x_0)^2 - \left[1 - \tfrac{x_0 - x}{q}\right] \;=\; \frac{x_0 - x}{q} - \eta(x - x_0)^2.$$

Let $t = x - x_0 \leq 0$. Then $A_0(t) = -\frac{t}{q} - \eta t^2$. Differentiating yields $A_0'(t) = -\frac{1}{q} - 2\eta t$, which equals zero at $t^* = -\frac{1}{2\eta q}$. This critical point lies in the interval $t \in [-q, 0]$ if and only if $\eta \geq \frac{1}{2q^2}$. When $\eta > \frac{1}{2q^2}$, t^* is a unique interior maximiser (the second derivative $A_0''(t) = -2\eta < 0$), giving an interior peak at $x_0 - \frac{1}{2\eta q}$. When $\eta \leq \frac{1}{2q^2}$, the critical point lies outside the region and the maximum occurs at the boundary $x_0 - q$. At $x = x_0$, both specialist and generalist deliver value 1, so $A_0(x_0) = 0$. The analysis for Region 4 and Region 5 is identical, with the roles of x_0 and x_1 reversed. Hence, each absolute advantage function has at most two local maxima: one at the boundary of the generalist region and, when $\eta > \frac{1}{2q^2}$, another strictly inside it. In all case,s $A_i(x_i) = 0$, so specialists are never optimally allocated exactly at the knowledge points.

Proof of Proposition 1 (Optimal Allocation)

Because questions are uniformly distributed and the objective is linear in the allocation sets, the planner solves the assignment problem point by point. Introduce shadow prices $\lambda_0, \lambda_1 \geq 0$ for specialist capacities. At any location x the planner chooses an assignment $\tau(x) \in \{G, S_0, S_1, \emptyset\}$ to maximise

$$v(x, \tau) \;=\; V_\tau(x) \;-\; \lambda_i \mathbf{1}\{\tau = S_i\},$$

where $V_\emptyset(x) = 0$ and there is no shadow price on generalists because they are abundant. If $\sigma_x^2 \leq q$ and $V_G(x) \geq \max\{V_{S0}(x) - \lambda_0, V_{S1}(x) - \lambda_1, 0\}$, a generalist serves x; otherwise type-i specialists serve those x at which $V_{Si}(x) - \lambda_i$ is maximal and non-negative. Since

$A_i(x) = V_{Si}(x) - V_G(x)$, a type-$i$ specialist is assigned at x if and only if $A_i(x) \geq \lambda_i$ and $A_i(x) - \lambda_i \geq A_{1-i}(x) - \lambda_{1-i}$; if no specialist has non-negative net advantage and $\sigma_x^2 > q$, the question remains unanswered.

The capacity constraints require that the measure of the set on which a type-i specialist is assigned equals $L_S/2$. Because A_i is continuous with a finite number of local maxima, the function $\lambda \mapsto |\{x : A_i(x) \geq \lambda\}|$ is continuous and strictly decreasing for λ between the peak value and zero. Therefore there exist unique thresholds λ_i^* such that $|\{x : A_i(x) \geq \lambda_i^*\}| = L_S/2$. These thresholds are the equilibrium shadow prices. The resulting allocation maximises value at each location subject to capacity, and any deviation would either violate the capacity constraints or assign a question to a less productive worker. The statements on generalist assignment and unserved questions follow directly from this pointwise comparison.

Proof of Proposition 2 (Allocation under Different Capacities)

Fix a specialist type (suppress the subscript for clarity) and let $\bar{A}^{\text{int}}$ and $\bar{A}^{\text{tail}}$ denote the values of A at its interior and tail maxima (if the interior maximum exists, $\bar{A}^{\text{int}} \geq \bar{A}^{\text{tail}}$). For each $\lambda \in [0, \bar{A}^{\text{int}}]$ define the upper contour set $E(\lambda) = \{x : A(x) \geq \lambda\}$. By Lemma 1, $E(\lambda)$ is the union of at most two intervals: one surrounding the interior maximum and one in the tail. As λ falls from $\bar{A}^{\text{int}}$ to $\bar{A}^{\text{tail}}$, only the interior interval is non-empty and its measure increases monotonically. Define L^{int} as the measure of $E(\bar{A}^{\text{tail}})$; if capacity L does not exceed L^{int}, the optimal assignment concentrates entirely on this interior interval. When λ drops below $\bar{A}^{\text{tail}}$, a second interval in the tail appears and expands as λ declines further. Let L^{tail} denote the capacity level at which these two intervals meet; for capacities between L^{int} and L^{tail} the assignment set is the union of two disjoint intervals. Once L exceeds L^{tail}, the two intervals merge and the assignment set becomes connected, expanding until its complement shrinks to a neighbourhood of the knowledge point where $A < 0$. Because capacity is measured by the Lebesgue measure of $E(\lambda)$ and λ is pinned down by the condition $|E(\lambda)| = L$, these thresholds are uniquely defined.

Proof of Proposition 3 (Domain Comparative Static)

Take any x between the original knowledge points x_0 and x_1. Before expansion, the generalist variance is

$$\sigma_x^2 \;=\; \frac{(x_1 - x)(x - x_0)}{x_1 - x_0}.$$

After moving the upper knowledge point from x_1 to $x_1' > x_1$, the variance becomes $\frac{(x_1' - x)(x - x_0)}{x_1' - x_0}$. Since both the numerator and the denominator increase and the ratio $\frac{x_1' - x}{x_1' - x_0}$ exceeds $\frac{x_1 - x}{x_1 - x_0}$

for all $x < x_1$, the variance strictly increases for $x \in (x_0, x_1)$ and generalist precision deteriorates on those questions. For $x \in (x_1, x_1']$, generalists previously extrapolated from a single point, so their variance was $x - x_1$. The new two-point variance $\frac{(x_1'-x)(x-x_0)}{x_1'-x_0}$ is strictly lower because the presence of two bounding points reduces extrapolation error, so precision improves on these questions. Because $A_i(x) = V_{Si}(x) - V_G(x)$, these changes in generalist precision shift the absolute advantage functions: A_0 and A_1 decrease on (x_0, x_1) and increase on $(x_1, x_1']$. Given fixed specialist capacities, the thresholds $(\lambda_0^*, \lambda_1^*)$ adjust so that the measure of each assignment set remains $L_S/2$. As the domain widens, specialists reallocate effort towards the now more difficult interior questions and may reduce coverage of the tail if generalists can handle it more effectively.

Proof of Proposition 4 (Endogenous Hiring)

I prove both directions of the correspondence.

From fixed capacity to endogenous hiring. Fix capacities (L_0, L_1) and let (S_0, S_1) denote the specialist assignment sets that solve the allocation problem in Proposition 1. By construction there exist thresholds λ_0^* and λ_1^* such that $S_i = \{x : A_i(x) \geq \lambda_i^*\}$ and $|S_i| = L_i$. Let $c_S = \lambda_i^*$; I claim that (L_0, L_1) maximises the endogenous hiring objective when the cost per unit of specialist capacity is c_S.

Given the allocation rule in Proposition 1, increasing L_i by a small amount ΔL_i allows a type-i specialist to cover additional questions with $A_i(x) \geq \lambda_i^*$ at the boundary of S_i. The value created by deploying a specialist on those marginal questions is $A_i(x)$, while the value of a generalist there is zero (either because $\sigma_x^2 > q$ or because the question was previously unserved). By the envelope theorem, the marginal benefit of expanding L_i is

$$\frac{d}{dL_i}\Big[\tfrac{1}{x_H - x_L}\int_{S_i} V_{Si}(x)\,dx + B(\text{generalist allocation})\Big] \;=\; A_i(x) \quad \text{for } x \in \partial S_i.$$

Since $A_i(x) = \lambda_i^*$ at all boundary points, the derivative of the objective with respect to L_i equals λ_i^*. Choosing $c_S = \lambda_i^*$ therefore satisfies the first-order condition for capacity choice. Because the absolute advantage function is continuous and single-peaked on each side of x_i, the marginal benefit of capacity is strictly decreasing in L_i, so (L_0, L_1) maximises the hiring objective at cost c_S.

From endogenous hiring to fixed capacity. Now fix a hiring cost $c_S > 0$ and let (L_0^*, L_1^*) denote the capacities that solve the endogenous hiring problem. Given (L_0^*, L_1^*), the firm allocates specialists according to the absolute-advantage rule. Suppose, toward a

contradiction, that there exists a boundary point $x \in \partial S_i$ with $A_i(x) \neq c_S$. If $A_i(x) > c_S$, then expanding L_i by a small amount and assigning a specialist to cover x would increase total value by $A_i(x) - c_S > 0$, contradicting optimality of (L_0^*, L_1^*). If $A_i(x) < c_S$, then decreasing L_i slightly and replacing the specialist at x with a generalist (or leaving the question unanswered) would save cost c_S and lose value $A_i(x)$, again improving the objective. Therefore $A_i(x) = c_S$ for every $x \in \partial S_i$. In the fixed–capacity model with capacities (L_0^*, L_1^*), the threshold for type i is $\lambda_i^*(L_i^*) = c_S$, and the resulting assignment coincides with the equilibrium of the endogenous hiring problem.

Proof of Proposition 5 (Knowledge Deepening)

Let $\mathcal{F} = \{(x_0, y(x_0)), (x_1, y(x_1))\}$ and write $X = x_1 - x_0 \leq 4q$. On $[x_0, x_1]$ the conjecture variance under two-point interpolation is

$$\sigma_x^2(\mathcal{F}) = \frac{(x - x_0)(x_1 - x)}{X},$$

which is maximised at $x = (x_0 + x_1)/2$ with value $X/4$. Adding a midpoint $x_m = (x_0 + x_1)/2$ divides $[x_0, x_1]$ into $[x_0, x_m]$ and $[x_m, x_1]$ of length $X/2$. On each subinterval, the variance becomes

$$\sigma_x^2(\mathcal{F} \cup \{x_m\}) = \frac{(x - x_0)(x_m - x)}{X/2},$$

attaining its maximum $X/8$ at the midpoint of the subinterval. Since $X \leq 4q$, the maximum variance before deepening is at most q, and generalists can operate across $[x_0, x_1]$ both before and after deepening. Their minimum value, equal to $1 - \sigma_{\max}/q$, therefore rises from $1 - X/(4q)$ to $1 - X/(8q)$, establishing part (i).

For part (ii), note that a generalist answering questions on an interval of length L creates value

$$V(L) = \int_0^L \left[1 - \frac{z(L - z)}{qL}\right] dz = L - \frac{1}{qL} \int_0^L z(L - z)\, dz = L - \frac{L^2}{6q},$$

where the substitution $z \mapsto L - z$ shows that $\int_0^L z(L - z)\, dz = L^3/6$. Before deepening the value created on $[x_0, x_1]$ is $V(X) = X - X^2/(6q)$. After deepening, there are two intervals of length $X/2$, yielding $2V(X/2) = X - X^2/(12q)$. The difference, $X^2/(12q)$, is the increase in generalist value, proving part (ii).

Part (iii) follows because the function $X^2/(12q)$ is strictly increasing in X for $X > 0$; subject to the constraint $X \leq 4q$, it attains its maximum at $X = 4q$, where $\Delta V_G = 4q/3$.

Proof of Proposition 6 (Knowledge Expansion)

Let $X = x_1 - x_0 \in [3q, 4q]$ and set $x_2 = x_1 + 3q$. On $[x_0, x_1]$ nothing changes when x_2 is added, so the conjecture variance remains $\sigma_x^2 = \frac{(x-x_0)(x_1-x)}{X}$ with maximum $X/4$. Since $X \in [3q, 4q]$, this maximum lies in $[3q/4, q]$, showing part (i).

On the new interval $[x_1, x_2]$ the variance under two-point interpolation between x_1 and x_2 is

$$\sigma_x^2 = \frac{(x - x_1)(x_2 - x)}{x_2 - x_1} = \frac{(x - x_1)(x_2 - x)}{3q},$$

which is maximised at the midpoint $x = (x_1 + x_2)/2$ with value $\frac{(3q/2)^2}{3q} = 3q/4 < q$. Hence, generalists can answer all questions in this interval. To compute the total value they generate, substitute $u = x - x_1$ to get

$$\int_{x_1}^{x_2} \left[1 - \frac{\sigma_x^2}{q}\right] dx = \int_0^{3q} \left[1 - \frac{u(3q-u)}{3q^2}\right] du = 3q - \frac{1}{3q^2} \int_0^{3q} \left(3qu - u^2\right) du.$$

Evaluating the integral $\int_0^{3q}(3qu - u^2)\, du$ gives $\frac{9q^3}{2}$, so the total value created by generalists on $[x_1, x_2]$ is $3q - \frac{1}{3q^2} \cdot \frac{9q^3}{2} = \frac{3q}{2}$. This proves part (ii).

For part (iii), note that for $x > x_2 + q$ the variance under two-point interpolation between x_2 and x exceeds q, so generalists abstain. A specialist anchored at x_2 delivers value $V_{S2}(x) = 1 - \eta(x - x_2)^2$, which remains positive for $|x - x_2| < \sqrt{(1 - c_S)/\eta}$ if the firm is willing to hire specialists at cost c_S. Thus x_2-specialists can operate on the frontier region $(x_2 + q, \min\{x_2 + r_S, x_H\}]$, where $r_S = \sqrt{(1 - c_S)/\eta}$.

Proof of Proposition 7 (Deepening vs. Expansion)

Write $a := 2/(qX) + \eta$ and $A_m(t) := t/q - at^2$ for the midpoint specialist's absolute-advantage profile after deepening. By Proposition 5, deepening raises generalist value on $[x_0, x_1]$ by $\Delta V_G^{\text{deep}} = X^2/(12q)$; by Proposition 6, expansion raises generalist value on $[x_1, x_2]$ by $\Delta V_G^{\text{exp}} = 3q/2$. We compare specialist terms and add these generalist components.

(i) High specialist cost $c_S > 1/(4aq^2)$. Since $\max_t A_m(t) = 1/(4aq^2) < c_S$, no midpoint specialist is hired under deepening. Under expansion, an x_2-specialist operates on $S_2 = [x_2 - r_S, x_2 + r_S] \cap (x_1 + q, x_H]$, $r_S = \sqrt{(1 - c_S)/\eta}$, with contribution

$$\Psi_2(c_S) \;=\; \frac{1}{x_H - x_L} \int_{S_2} \left[\, 1 - \eta(x - x_2)^2 - c_S \,\right] dx.$$

Expansion strictly dominates deepening iff $\Psi_2(c_S) > \Delta V_G^{\text{deep}} - \Delta V_G^{\text{exp}}$, i.e.

$$\Psi_2(c_S) \;>\; \frac{X^2}{12q} - \frac{3q}{2}.$$

(ii) Otherwise $0 \le c_S \le 1/(4aq^2)$. Solve $A_m(t) = c_S$ to obtain

$$t_\pm = \frac{\frac{1}{q} \pm \sqrt{\frac{1}{q^2} - 4a\,c_S}}{2a},$$

so the midpoint specialist (if hired) is assigned on $[x_m - t_+, x_m - t_-] \cup [x_m + t_-, x_m + t_+]$. Its net contribution is

$$\Phi_m(c_S) = \frac{2}{x_H - x_L} \int_{t_-}^{t_+} \big[A_m(t) - c_S\big]\, dt = \frac{2}{x_H - x_L} \Big[\frac{t^2}{2q} - \frac{at^3}{3} - c_S t\Big]_{t_-}^{t_+}.$$

The expansion–frontier specialist contributes $\Psi_2(c_S)$ as above (covering both the symmetric and truncated cases). Expansion strictly dominates deepening iff

$$\Psi_2(c_S) \;>\; \frac{X^2}{12q} - \frac{3q}{2} \;+\; \Phi_m(c_S).$$

The remark follows since $A_m(X/2) = -\eta X^2/4 < 0$, so the midpoint specialist cannot profitably cover the entire $[x_0, x_1]$ for any $c_S \ge 0$. □

Proof of Proposition 8 (Optimal Placement)

Place $x_1 = x_0 + d$ with $d \in [0, 4q]$ so generalists can operate on $[x_0, x_1]$. Two terms vary with d.

Generalists. On the interior $[x_0, x_1]$ of length d, generalists create

$$V(d) = \int_{x_0}^{x_1} \Big[1 - \tfrac{(x - x_0)(x_1 - x)}{q\,d}\Big] dx = d - \frac{d^2}{6q},$$

which is strictly concave with derivative $V'(d) = 1 - \frac{d}{3q}$.

Frontier specialists. Let $C := x_H - x_0 - q$ be the available room to the right of $x_0 + q$. The *one-sided* frontier beyond $x_1 + q$ has length $F(d) := (C - d)_+$. A specialist anchored at x_1 earns net density $g(s) = 1 - c_S - \eta s^2$ at distance $s \ge 0$ beyond $x_1 + q$, and has reach

$r_S := \sqrt{(1-c_S)/\eta}$. Hence

$$S(d) \;=\; \int_0^{\min\{r_S,\,F(d)\}} \big[\,1-c_S-\eta s^2\,\big]\,ds \;=\; \ell(d)\,(1-c_S) \;-\; \frac{\eta}{3}\,\ell(d)^3, \qquad \ell(d) := \min\{r_S, F(d)\}.$$

Thus S is concave in d. Its derivative is piecewise:

$$S'(d) = \begin{cases} 0, & F(d) \ge r_S \quad \text{(slack frontier)},\\ -\,g\big(F(d)\big) \;=\; -\big[1-c_S-\eta(C-d)^2\big], & 0 < F(d) < r_S \quad \text{(binding frontier)}. \end{cases}$$

In the binding region $S''(d) = -2\eta(C-d) \le 0$.

FOC and solution. The objective $W(d) := V(d) + S(d)$ is strictly concave on $[0, 4q]$, so the maximiser is unique and solves

$$0 = W'(d) = \Big(1 - \frac{d}{3q}\Big) + S'(d).$$

Slack frontier. If the frontier is slack at the maximiser, $S'(d) = 0$ and $W'(d) = 1 - \frac{d}{3q}$ yields $d^* = 3q$. This is guaranteed, for example, when $r_S \le F(d)$ for all feasible d (e.g. when $C - 4q \ge r_S$).

Binding frontier. When $0 < F(d) \le r_S$ near the maximiser, the FOC becomes

$$1 - \frac{d}{3q} \;=\; 1 - c_S - \eta(C-d)^2 \quad\Longleftrightarrow\quad \eta(C-d)^2 + c_S = \frac{d}{3q}. \tag{$*$}$$

Solving $(*)$ gives the interior candidate

$$d^\circ \;=\; C + \frac{1}{6\eta q} - \frac{1}{2\eta}\sqrt{\frac{1}{9q^2} + \frac{4\eta C}{3q} - 4\eta c_S}\,.$$

Feasibility of a binding frontier requires $0 < F(d) \le r_S$, i.e. $d \in [d_L, d_U]$ with

$$d_L := \max\{0,\, C - r_S\}, \qquad d_U := \min\{4q,\, C\}.$$

Therefore

$$d^* = \begin{cases} d^\circ, & \text{if } d^\circ \in [d_L, d_U],\\ d_L, & \text{if } d^\circ < d_L \text{ or the discriminant is negative},\\ d_U, & \text{if } d^\circ > d_U. \end{cases}$$

Comparative statics (binding case). Implicitly differentiating $(*)$ gives

$$\frac{\partial d^*}{\partial c_S} = \frac{1}{\frac{1}{3q} + 2\eta(C - d^*)} > 0, \quad \frac{\partial d^*}{\partial \eta} = \frac{(C - d^*)^2}{\frac{1}{3q} + 2\eta(C - d^*)} > 0, \quad \frac{\partial d^*}{\partial C} = \frac{2\eta(C - d^*)}{\frac{1}{3q} + 2\eta(C - d^*)} \in (0, 1),$$

so d^* is (weakly) increasing in c_S, in η, and in C. □

Proof of Proposition 9 (Market Evolution)

We begin by analysing the geometry of the problem. Let x_0 and x_1 denote the incumbent and frontier knowledge points with $x_1 - x_0 \leqslant 4q$, and consider the evolving domain $[x_L(\theta), x_H(\theta)]$ defined by $x_L(\theta) = x_L^0 + \theta$ and $x_H(\theta) = x_H^0 + \beta\theta$ with $\beta > 1$. For each $\theta \geqslant 0$ the domain can be partitioned into regions where only specialists can operate (exclusive regions) and regions where both specialists and the generalist can operate (contested regions). Because generalists can interpolate between knowledge points up to a distance q on either side, the left exclusive region for the incumbent specialist S_0 is $\mathcal{E}_0(\theta) = [x_L(\theta), x_0 - q]$ and the right exclusive region for the frontier specialist S_1 is $\mathcal{E}_1(\theta) = [x_1 + q, x_H(\theta)]$. The lengths of these intervals are

$$\ell_0(\theta) = x_0 - q - x_L^0 - \theta \quad \text{and} \quad \ell_1(\theta) = x_H^0 + \beta\theta - x_1 - q.$$

The left exclusive region shrinks linearly in θ and vanishes at $\theta_{\text{exit}}^{\text{geo}} := x_0 - q - x_L^0$, while the right exclusive region expands linearly and becomes positive at $\theta_{\text{entry}}^{\text{geo}} := \frac{x_1 + q - x_H^0}{\beta}$. Because $\beta > 1$ the right boundary expands faster than the left boundary contracts, so $\theta_{\text{entry}}^{\text{geo}} < \theta_{\text{exit}}^{\text{geo}}$.

In the exclusive regions, the price a specialist can charge is simply $V_{S_i}(x)$ because there is no competition. In contested regions around x_0 and x_1 the price is the quality advantage over the generalist, $V_{S_i}(x) - V_G(x)$, which falls with the distance from x_i. For any cost level $c_S > 0$ there exists a unique distance $d_c > 0$ such that $V_{Si}(x_i \pm d_c) - V_G(x_i \pm d_c) = c_S$; beyond this distance the specialist earns zero profit. These distances do not depend on θ, but the positions of the contested intervals shift as the domain evolves.

Let $m_{S0}(\theta)$ (respectively $m_{S1}(\theta)$) denote the measure of locations at which the incumbent specialist S_0 (respectively the frontier specialist S_1) has strictly positive profit. This measure is the sum of the mass of the exclusive region and the mass of contested locations within distance d_c of x_i where $p_{Si}(x) \geqslant c_S$. Because $x_L(\theta)$ increases linearly while $x_H(\theta)$ increases faster, $m_{S0}(\theta)$ is continuous and strictly decreasing in θ, and $m_{S1}(\theta)$ is continuous and strictly increasing. Define

$$\theta_{\text{entry}}^* = \inf\{\theta \geqslant 0 : m_{S1}(\theta) > 0\}, \qquad \theta_{\text{exit}}^* = \inf\{\theta \geqslant 0 : m_{S0}(\theta) = 0\}.$$

The assumptions $x_1 - x_0 \leqslant 4q$ and $\beta > 1$ ensure that $0 < \theta^*_{\text{entry}} < \theta^*_{\text{exit}}$: the frontier specialist obtains a positive exclusive region strictly before the incumbent specialist loses its exclusive region. In equilibrium, a specialist enters if and only if $m_i(\theta) > 0$ because entry is profitable exactly when there exist customers who generate revenue at least equal to the hiring cost c_S. Similarly, a specialist exits once $m_i(\theta) = 0$. Therefore, for $\theta < \theta^*_{\text{entry}}$ only the incumbent specialist can earn positive profits, leading to a duopoly between S_0 and the generalist. For $\theta^*_{\text{entry}} \leqslant \theta < \theta^*_{\text{exit}}$, both specialists have non-empty markets, so all three firms operate. Finally, for $\theta \geqslant \theta^*_{\text{exit}}$ the incumbent specialist has no profitable customers and exits, leaving a duopoly between the frontier specialist and the generalist. This establishes the three phases described in Proposition 9.

Proof of Corollary 1 (Competitive Performance)

Let $\pi_i(\theta)$ denote the equilibrium profit of firm $i \in \{S_0, S_1, G\}$ and let $m_i(\theta)$ be the equilibrium measure of consumers it serves. For specialist firms the per-customer price is $p^*_{Si}(x) = V_{Si}(x) - \max_{j \neq i} V_j(x)$, which declines with distance from x_i. Each specialist hires exactly one worker per customer at a cost c_S, so $\pi_i(\theta) = \int_{\mathcal{M}_i(\theta)} p^*_{Si}(x) f(x;\theta)\, dx - c_S m_i(\theta)$.

(i) *Profitability ordering at entry.* At $\theta = \theta^*_{\text{entry}}$ the frontier specialist S_1 has just enough customers to break even, so $\pi_{S_1}(\theta^*_{\text{entry}}) = 0$. The incumbent specialist still has a strictly positive profitable region and therefore earns strictly positive profits, whereas the generalist earns zero profits. Hence $\pi_{S_0}(\theta^*_{\text{entry}}) > \pi_G(\theta^*_{\text{entry}}) > \pi_{S_1}(\theta^*_{\text{entry}}) = 0$.

(ii) *Monotonicity.* On $[0, \theta^*_{\text{exit}})$, the incumbent specialist's exclusive region and contested region both shrink as θ increases, and the density of consumers falls because $W(\theta)$ grows. Both effects reduce $m_{S_0}(\theta)$ and the average price per customer, so $\pi_{S_0}(\theta)$ is strictly decreasing on this interval. For $\theta > \theta^*_{\text{entry}}$ the frontier specialist's exclusive region and contested region expand monotonically in θ, while the price it can charge in contested regions does not decrease. Consequently $\pi_{S_1}(\theta)$ is strictly increasing for $\theta > \theta^*_{\text{entry}}$. The generalist's market share is the residual of the specialists' shares. It initially increases as the incumbent specialist retreats from marginal territories, but once the frontier specialist expands sufficiently, it begins to decline. Continuity therefore implies the existence of a unique turning point θ^*_G at which $\pi_G(\theta)$ attains its maximum.

(iii) *Profit crossing points.* The functions π_{S_1}, π_{S_0} and π_G are continuous on $(\theta^*_{\text{entry}}, \theta^*_{\text{exit}})$, with π_{S_1} strictly increasing, π_{S_0} strictly decreasing and π_G unimodal. By the intermediate value theorem, there exist unique thresholds θ_1 and θ_2 in $(\theta^*_{\text{entry}}, \theta^*_{\text{exit}})$ such that $\pi_{S_1}(\theta_1) = \pi_G(\theta_1)$ and $\pi_{S_1}(\theta_2) = \pi_{S_0}(\theta_2)$. For $\theta > \theta_1$ one has $\pi_{S_1}(\theta) > \pi_G(\theta)$ because π_{S_1} increases while π_G eventually decreases. Similarly, for $\theta > \theta_2$ one has $\pi_{S_1}(\theta) > \pi_{S_0}(\theta)$ because π_{S_1} increases

and π_{S_0} decreases.

(iv) *Profitability at exit.* At $\theta = \theta^*_{\text{exit}}$ the incumbent specialist's profitable region has shrunk to zero, so $\pi_{S_0}(\theta^*_{\text{exit}}) = 0$. Because the frontier specialist continues to gain new customers beyond this point, $\pi_{S_1}(\theta^*_{\text{exit}}) > \pi_G(\theta^*_{\text{exit}})$. Thus $\pi_{S_1}(\theta^*_{\text{exit}}) > \pi_G(\theta^*_{\text{exit}}) > \pi_{S_0}(\theta^*_{\text{exit}}) = 0$.

The four statements in Corollary 1 follow directly from these observations, completing the proof.

Proof of Proposition 10 (Extent of the Market)

Recall that $A(x) = V_S(x) - V_G(x)$ is the absolute advantage of specialists over generalists and that by Proposition 4 the endogenous-hiring problem is equivalent to the fixed-capacity problem of Section 3. Let $\lambda(\mu)$ denote the shadow price of specialist capacity when the question flow is μ. Under a fixed shadow price λ, the measure of locations at which a specialist has a non-negative net advantage equals the specialist capacity L_S, and the measure of locations served by generalists is L_G. The first part of Proposition 10 shows that there exists a unique threshold $\mu^\star$ at which $\lambda = 0$.

To prove the existence and uniqueness of $\mu^\star$, note that the firm's total expected surplus in the endogenous-hiring problem can be written as

$$\mu \int A(x)\, f(x)\, \mathbf{1}\{A(x) \geq \lambda(\mu)\}\, dx \;\; \lambda(\mu)\, L_S(\mu),$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and $\lambda(\mu)$ is the Lagrange multiplier on specialist capacity. The optimality condition for $\lambda(\mu)$ equates the marginal revenue from expanding specialist coverage with the marginal cost c_S:

$$\mu \int A(x)\, f(x)\, dx \;=\; c_S \quad \text{when } \lambda(\mu) = 0.$$

Define $\bar{A} = \int A(x) f(x)\, dx$. When $\mu < c_S/\bar{A}$, the left-hand side is strictly less than c_S, so the optimum sets $\lambda(\mu) > 0$ and $L_S(\mu) = 0$; when $\mu > c_S/\bar{A}$, the left-hand side exceeds c_S, so the firm benefits from expanding specialist coverage and drives $\lambda(\mu)$ down to zero. This establishes the threshold $\mu^\star = c_S/\bar{A}$ and the properties of $L_S(\mu)$ stated in the proposition.

To establish monotonicity, observe that for $\mu > \mu^\star$ the capacity constraint is binding with $\lambda(\mu) = 0$. Differentiating the specialist hiring condition with respect to μ yields

$$\frac{dL_S}{d\mu} \;=\; \int \mathbf{1}\{A(x) \geq 0\}\, f(x)\, dx \;>\; 0,$$

so L_S is strictly increasing in μ. Specialist capacity expands because a larger market scales up the benefit of placing specialists at any location where they have a non-negative absolute advantage. Since total coverage $L_S + L_G$ scales linearly with μ, it follows that L_G must be weakly decreasing. Finally, when $c_S > c_G$ the firm substitutes towards specialists whenever the marginal value of a question exceeds c_G, so the ratio L_S/L_G increases in μ; differentiating the ratio and using the monotonicity of L_S and L_G yields the stated result.

Proof of Lemma 2 (Variance Reduction)

Let $\mathcal{F}_2 = \mathcal{F}_1 \cup I_0 \cup I_1$ and fix $z \in \mathbb{R}$. By definition, a period-2 generalist uses the two codified points in $\mathcal{F}_2$ that are closest on either side of z to interpolate an answer. If z lies to the left of all codified points (including x_0), then a generalist extrapolates from the nearest knowledge point, so the variance equals $d(z, \mathcal{F}_2)$. A symmetric argument holds if z lies to the right of all codified points. Now suppose z lies between two codified points $a < b$ with no codified points strictly between. Conditional on $\mathcal{F}_2$, the Brownian motion property implies that $y(z)$ has mean $y(a) + \frac{z-a}{b-a}\big(y(b) - y(a)\big)$ and variance $\frac{(b-z)(z-a)}{b-a}$; this is the variance $\sigma^2_{\text{bridge}}(z \mid \mathcal{F}_2)$ used in the statement of the lemma. A generalist chooses the smaller of this variance and the distance to the nearest knowledge point, because linear interpolation inside (a, b) yields lower variance than extrapolation outside that interval only when z is sufficiently close to the midpoint. Combining the three cases yields $\sigma^2_z(\mathcal{F}_2) = \min\{d(z, \mathcal{F}_2), \sigma^2_{\text{bridge}}(z \mid \mathcal{F}_2)\}$, as claimed.

Proof of Proposition 11 (Strategic Complementarities)

For simplicity set $x_0 = 0$ and write $x_1 = d > 0$. Let f be uniform on $[0, d]$ and assume symmetric specialists. Consider the value $\Pi(x_0, x_1)$ of codifying points s_0 and s_1 with $0 < s_0 < s_1 < d$. Codifying at (s_0, s_1) affects only period-2 generalist value on $[s_0, s_1]$, because outside that interval the nearest codified point remains at one of $\{0, d, s_0, s_1\}$ and the interpolation span does not depend on s_0 and s_1. Inside (s_0, s_1), the variance at z is $\frac{(s_1-z)(z-s_0)}{s_1-s_0}$; without these codified points it would have been $\frac{(d-z)z}{d}$ because the only knowledge points would be 0 and d. The gain in generalist value at z is therefore

$$\frac{\sigma^2_z(\{0, d\}) - \sigma^2_z(\{0, d, s_0, s_1\})}{q} \;=\; \frac{1}{q}\Big[\frac{(d-z)z}{d} - \frac{(s_1-z)(z-s_0)}{s_1-s_0}\Big].$$

Integrating over $z \in [s_0, s_1]$ and differentiating twice with respect to s_0 and s_1 yields the cross-partial

$$\frac{\partial^2}{\partial s_0 \, \partial s_1} \int_{s_0}^{s_1} \Big[\frac{(d-z)z}{d} - \frac{(s_1 - z)(z - s_0)}{s_1 - s_0} \Big] dz \;=\; - \frac{(s_1 - s_0)^2 - 4q(s_1 - s_0) + 4q^2}{3q\,(s_1 - s_0)^3}.$$

(Details of the algebra are routine; differentiating under the integral sign is justified by uniform convergence on compact intervals.) The numerator is negative exactly when $2q < s_1 - s_0 < 4q$, and otherwise the expression is non-positive. Because period-1 value $\int W_i(s_i) ds_i$ does not depend on s_{-i}, the cross-partial of Π inherits this sign. Restoring the original labels x_0, x_1 proves the proposition.

Proof of Proposition 12 (Optimal Codification Patterns)

Fix $d = x_1 - x_0 = 4q$ and a uniform question distribution. For any allocation (I_0, I_1), period-1 value is linear in the measure of I_i weighted by W_i, and period-2 value depends only on the codified points. When δ is very small, codification yields negligible future value; the planner therefore maximises period-1 value by assigning each specialist to her static advantage peak. Because A_i is single-peaked and questions are uniformly distributed, the optimal static assignment places type-i specialists in the tail region where $x \leq x_0 + q$ (for $i = 0$) and $x \geq x_1 - q$ (for $i = 1$). This yields case (1) with no codification.

As δ increases, codification becomes valuable. Consider moving the left specialist's assignment from $x_0 + q$ inward by a small $\epsilon > 0$ and simultaneously moving the right specialist's assignment inward by ϵ from $x_1 - q$. By Proposition 11, when $2q < x_1 - x_0 < 4q$ the cross-partial of Π is negative, so these moves are mutually reinforcing: the marginal benefit of the left specialist moving inward is greater when the right specialist also moves inward. However, the opportunity cost of moving inward rises with ϵ because W_i declines away from its peak. For intermediate values of δ, the planner therefore chooses a small positive ϵ and contiguous allocation intervals I_0 and I_1 centred around $x_0 + q$ and $x_1 - q$, respectively. The lengths ℓ_0, ℓ_1 are determined by the first-order conditions $\mathcal{L}_i(x) = \delta \, \mathcal{G}_i(x \mid I_{-i})$ at the boundaries of I_i.

When δ becomes large, the marginal benefit of codifying dominates the opportunity cost for any x sufficiently close to the midpoint. As a result, the allocation intervals expand toward each other until further expansion would yield a negative net value because $A_i(x)$ becomes too small. Formally, let $\bar{x}$ be the point where $W_i(\bar{x}) = 0$; beyond $\bar{x}$, assigning a specialist yields no period-1 value. The maximal stretching in case (3) sets the boundary of I_i at $\bar{x}$, and codification occurs on the entire interior of $[x_0 + q, \bar{x}]$ for type-0 and symmetrically on

$[\bar{x}, x_1 - q]$ for type-1. Thresholds $\underline{\delta}$ and $\bar{\delta}$ at which the planner switches regimes are determined by solving $\delta\,\mathcal{G}_i = \mathcal{L}_i$ at the relevant boundaries; uniqueness follows from monotonicity of the functions involved. This completes the proof.

Proof of Proposition 13 (Codification Intensity)

Given allocation intervals I_i, codification intensity τ_i determines the measure of questions whose answers are codified. Because codification is costly only in that it reduces the number of period-1 questions answered, we can treat τ_i as a continuous decision between 0 and 1. Denote by $g_i(\tau_i)$ the period-2 gain from codifying a fraction τ_i of points in I_i, holding I_{-i} fixed. By linearity of expectation, $g_i(\tau_i) = \tau_i \int_{I_i} \mathcal{G}_i(x \mid I_{-i})\,dx$. The opportunity cost from period 1 is $(1-\tau_i)\int_{I_i} \mathcal{L}_i(x)\,dx$, because each uncodified answer displaces a specialist from her static peak. Hence, the planner solves

$$\max_{0\le\tau_i\le 1}\ \tau_i\delta \int_{I_i} \mathcal{G}_i(x \mid I_{-i})\,dx\ -\ (1-\tau_i)\int_{I_i} \mathcal{L}_i(x)\,dx.$$

Differentiating with respect to τ_i yields the first-order condition

$$\delta \int_{I_i} \mathcal{G}_i(x \mid I_{-i})\,dx\ +\ \int_{I_i} \mathcal{L}_i(x)\,dx\ =\ 0,$$

which rearranges to

$$\tau_i^* = \frac{\delta\bar{\mathcal{G}}_i - \bar{\mathcal{L}}_i}{\delta\bar{\mathcal{G}}_i}.$$

When I_i lies in the tail region, $\mathcal{G}_i(x \mid I_{-i}) = 0$ because codifying points close to existing knowledge yields no variance reduction; thus $\bar{\mathcal{G}}_i = 0$ and $\tau_i^* = 0$. When I_i includes interior points, $\bar{\mathcal{G}}_i > 0$ and the formula applies provided $\delta\,\bar{\mathcal{G}}_i > \bar{\mathcal{L}}_i$. The second-order condition (that the objective is concave in τ_i) holds because the objective is linear in τ_i. This proves the proposition.

Proof of Proposition 14 (Value of Patience)

Define $V(\delta, I_0, I_1, \tau_0, \tau_1) = \Pi(I_0, I_1, \tau_0, \tau_1)$ for notational convenience. By definition,

$$V^*(\delta) = V\big(\delta, I_0^*(\delta), I_1^*(\delta), \tau_0^*(\delta), \tau_1^*(\delta)\big).$$

Applying the chain rule and the envelope theorem gives

$$\frac{dV^*}{d\delta} = \frac{\partial V}{\partial \delta} + \frac{\partial V}{\partial I_0}\frac{dI_0^*}{d\delta} + \frac{\partial V}{\partial I_1}\frac{dI_1^*}{d\delta} + \frac{\partial V}{\partial \tau_0}\frac{d\tau_0^*}{d\delta} + \frac{\partial V}{\partial \tau_1}\frac{d\tau_1^*}{d\delta}.$$

At an interior optimum $(I_0^*, I_1^*, \tau_0^*, \tau_1^*)$, the envelope theorem implies that $\frac{\partial V}{\partial I_i} = 0$ and $\frac{\partial V}{\partial \tau_i} = 0$ because small variations in I_i or τ_i around their optimum values do not affect value to first order. However, when δ changes, the optimal I_i and τ_i may move, and the envelope theorem alone does not capture this. The terms $\frac{\partial V}{\partial I_i}\frac{dI_i^*}{d\delta}$ and $\frac{\partial V}{\partial \tau_i}\frac{d\tau_i^*}{d\delta}$ account for the fact that optimal allocations shift with δ. Differentiating the first-order condition $\mathcal{L}_i(x_i) = \delta \mathcal{G}_i(x_i \mid I_{-i})$ with respect to δ and solving for $\frac{dx_i}{d\delta}$ show that $\frac{\partial V}{\partial I_i}\frac{dI_i^*}{d\delta} > 0$ when $x_1 - x_0 \approx 4q$: increasing δ induces specialists to stretch toward each other, raising period-2 value at a second-order rate. A similar argument holds for τ_i^*. When q is small, the variance threshold is tight and the improvement from bridging dominates, so the positive indirect term outweighs the direct effect $\frac{\partial V}{\partial \delta}$ of increased discounting. This proves the proposition.