AI Monoculture Transform(er)s Financial Markets

Wenjing Xu * Mingzhe Zheng †

August 2026

Abstract

Transformer-based AI is increasingly being integrated into various stages of investment decision-making. A natural question is whether AI is a harmless efficiency-enhancing tool. We show that it need not be. AI performance can be endogenously uneven across different dimensions of financial information. When many investors rely on the same pretrained technology, they inherit common strengths and blind spots in information processing, which in turn reshape the acquisition of human expertise. We refer to this phenomenon as AI monoculture. Without AI, human specialization follows a monotone cutoff across payoff dimensions, with expertise concentrated in high-exposure dimensions. As AI is introduced, it crowds out specialized research on the dimensions it covers. At the same time, dimensions beyond the AI coverage boundary retain lower information precision, raising the relative value of human expertise at the boundary and reallocating expertise toward risks not covered by AI. The equilibrium therefore features three regions: an AI-covered region with no additional human expertise, an intermediate uncovered region supported by human specialists, and a low-exposure "dark tail" with neither AI processing nor human expertise. AI therefore does not simply crowd out human expertise. It systematically reshapes the allocation of human expertise. As AI quality improves, the aggregate mass of human specialists varies non-monotonically. It thus makes the effects of AI on financial market quality generally non-monotone.

*Rotterdam School of Management, Erasmus University, e-mail: w.xu@rsm.nl

†Financial Technology Thrust, Society Hub, Hong Kong University of Science and Technology (Guangzhou), e-mail: mzheng842@connect.hkust-gz.edu.cn

1 Introduction

Transformer-based artificial intelligence is transforming how financial markets process information. Large language models and generative AI tools now assist sell-side analysts, portfolio managers, and systematic trading desks. Unlike earlier quantitative tools designed for narrow tasks, a pretrained transformer is a general-purpose inference engine. When market participants rely on the same pretrained model, or on models built from a common architectural family and broadly overlapping corpora, their information processing may become more homogeneous. At the same time, AI performance depends on the composition of the pretraining data. Improvements in average predictive ability can therefore coexist with persistent differences in performance across economically relevant topics.

This paper asks how such training-induced unevenness changes human information production and price formation. Instead of representing AI as an exogenous and uniformly more precise information technology, we derive its cross-dimensional coverage profile from a tractable transformer learning problem. We then embed that profile in a financial market in which investors can acquire targeted human expertise. The central tension is that selective AI breaks the monotone allocation of expertise that obtains in the no-AI benchmark. Under our cross-sectional ranking conditions, human research without AI follows a single cutoff: investors specialize in high-exposure dimensions and stop doing so once exposure becomes too small. With AI, expertise can instead follow a three-region pattern—no human research on AI-covered dimensions, active specialization immediately outside coverage, and no additional human research in a low-exposure dark tail. The analysis connects three decisions that are usually studied separately: which payoff dimensions a transformer learns sufficiently well, where investors continue to produce costly human information, and how a competitive market maker extracts that information from aggregate order flow.

Financial information is highly uneven. Routine earnings news, common macroeconomic developments, and widely discussed valuation signals appear repeatedly in a pretraining corpus, whereas less common topics appear less often. In our model, each payoff dimension corresponds to a sequence class in a next-token prediction problem. Although all classes follow the same decoder-learning dynamics, their frequencies determine the speeds of their local learning clocks. More frequent classes accumulate effective training exposure more quickly, while less frequent classes progress more slowly along the same learning path.

Uneven exposure alone does not imply that a less frequent class can never be learned. With sufficient training, the model continues to improve on every class. The key force is that training is costly. Once the model predicts the corpus well, further training produces only small reductions in aggregate loss. The remaining errors are disproportionately associated with infrequent classes, which receive fewer updates and carry less weight in the training objective. It can therefore be optimal

for the developer to stop before all classes reach a common performance threshold. The stopping decision generates an endogenous boundary between learned and undertrained payoff dimensions. This boundary arises from the interaction between the frequency distribution of the pretraining corpus and the cost of computation, rather than from an intrinsic inability to learn less frequent classes.

We translate this training outcome into an inference rule for the financial model. Dimensions that cross the learning threshold receive higher signal precision, whereas the remaining dimensions retain baseline precision. The transformer analysis determines which dimensions belong to each group; the financial model maps this endogenous coverage cutoff into a two-level Gaussian precision profile. In our benchmark, dimensions with larger direct payoff loadings are also more prevalent in the modeled corpus, so learned dimensions form the high-loading end of the payoff ranking. This alignment is an economic restriction rather than a relabeling.

The common coverage profile is our notion of *AI monoculture*. Investors share the dimensions that the pretrained model processes well and those that remain undertrained. They do not, however, observe identical signals: conditional on fundamentals, their signal errors remain independent. AI monoculture is therefore a common pattern of information-processing strengths and weaknesses, not mechanically correlated forecasts or identical orders. Moreover, an undertrained dimension is not devoid of information, because investors retain a baseline private signal about it.

We embed this inference rule in a competitive order-flow market. A risky asset loads on multiple independent fundamentals. Each investor observes private signals whose precision follows the pretrained coverage profile and, before signals are realized, may pay for an additional signal about the one dimension in which she can specialize. Human expertise is thus an overlay on AI information rather than a competing technology. Investors submit market orders before observing noise order flow, aggregate order flow, or the transaction price. A competitive market maker observes aggregate order flow and sets the price equal to expected payoff conditional on that flow. The resulting joint trading-and-specialization equilibrium links information production to price formation: specialization changes informed order flow and the pricing rule, which in turn changes the ex ante value of expertise.

We first solve the trading equilibrium for any specialization profile. The value of human expertise is the ex ante certainty-equivalent gain from adding the human signal after the associated trading equilibrium is solved. It depends on the residual payoff exposure of the specialized dimension, the information already supplied by AI, and the background risk that remains after conditioning on private information. A large direct payoff loading raises the value of research, but an informative order flow attenuates it. Consequently, payoff importance alone does not rank specialization incentives. Under a sufficient condition selecting a unique trading branch, additional expertise in any dimension makes order flow more informative and lowers pricing-error variance. Its effect on private research incentives nevertheless can be positive

or negative: the more informative price changes both total residual risk and the residual exposure of each particular dimension.

The learning cutoff creates the break from the no-AI allocation. Without AI, every dimension has the same baseline signal precision, so the cross-sectional ranking of residual information value generates a single specialization cutoff. Selective AI introduces a discrete precision change at the coverage boundary. It makes an additional human signal more redundant on covered dimensions, while the first uncovered dimension retains baseline precision. We derive the exact condition under which the value of expertise jumps upward at this boundary despite the decline in direct payoff exposure. When expertise is unprofitable on covered dimensions, profitable immediately outside coverage, and progressively less valuable farther down the uncovered ranking, equilibrium therefore has three regions: an AI-covered region without human research, an active uncovered region supported by specialists, and a low-exposure dark tail without the costly human overlay. Thus, AI does not merely move a monotone research cutoff inward or outward; it creates a nonmonotone allocation of human expertise across payoff dimensions. The dark tail is not uninformed: it retains baseline private signals but receives neither high-precision AI processing nor additional human expertise.

This nonmonotone expertise allocation separates the location of research from its total amount. Within strict cutoff equilibria, expanding coverage by one dimension removes the specialist at the old coverage boundary. Aggregate expertise rises only if induced entry farther down the uncovered tail more than offsets this displacement. With equal type masses, moving the tail boundary outward by one dimension merely relocates one specialist; aggregate expertise rises only if the boundary moves by at least two dimensions. This discrete accounting result shows why broader AI coverage can crowd out research, redirect it, or increase it, even though the direct precision improvement is concentrated on the newly covered dimension.

The effects on market quality reflect two distinct channels. Holding human specialization fixed and under our sufficient uniqueness condition, greater signal precision makes aggregate order flow more informative and reduces pricing-error variance. At the same time, it raises price impact and reduces market depth because the market maker responds more strongly to informative flow. Endogenous expertise adds a second channel. A coordinatewise expansion of specialization reinforces the direct improvement in price informativeness and further reduces depth, whereas a coordinatewise contraction offsets the information benefit and restores depth. When expertise exits some dimensions and enters others, neither effect has a general sign: the payoff exposures and prior precisions of the entering and exiting types matter, not only aggregate specialist mass.

Research cost isolates where this endogenous channel can overturn the direct technology effect. Under our sufficient uniqueness conditions, sufficiently low costs lead all types to specialize in both the AI and no-AI economies, while sufficiently high costs lead none to specialize. Because the research allocation is the same at

these corners, better technology makes prices more informative and reduces depth. At intermediate costs, types enter and exit at different thresholds. If the AI–no-AI market-quality comparison reverses, it must therefore do so in this intermediate region; heterogeneity can generate multiple crossings as different dimensions adjust.

Related Literature Our paper connects to three main strands of the literature: the effects of AI on financial markets, information acquisition in financial markets, and the theory of transformer architectures.

The effects of AI on financial markets. A rapidly growing literature studies how AI and big data reshape financial markets. Farboodi, Matray, Veldkamp, and Venkateswaran (2022) show that big data can reduce the incentive to acquire information, and Farboodi and Veldkamp (2020) study the long-run dynamics of financial data technology, finding that both fundamental and strategic data continue to be processed as technology advances. Dugast and Foucault (2018) find that more data can paradoxically reduce price informativeness when processing capacity is limited. Goldstein, Spatt, and Ye (2021) survey the broader implications of big data for finance. On the empirical side, Chinco, Clark-Joseph, and Ye (2019) show that stock returns load on sparse signals, and Lopez-Lira and Tang (2023) examine the predictive ability of large language models for stock returns.

Most closely related to our work is the recent series of papers by Dou, Goldstein, and Ji (2025a) and Dou, Goldstein, and Ji (2025b). Dou, Goldstein, and Ji (2025a) show that AI-powered traders using reinforcement learning can autonomously sustain collusive outcomes, undermining competition and price efficiency. Dou, Goldstein, and Ji (2025b) extend this analysis to a dynamic setting and study how AI planning—the ability of agentic AI to pursue long-horizon objectives—affects financial stability. Colliard, Foucault, and Lovo (2022) study algorithmic market makers using Q-learning and find that they adapt to adverse selection but fail to learn competitive pricing, earning supra-competitive rents. Kelly, Kuznetsov, Malamud, and Xu (2025) take a different approach, embedding a transformer directly in the stochastic discount factor to build AI asset pricing models, and derive an interpretable decomposition of the transformer's pricing mechanisms.

Our paper complements these contributions from a distinct angle. While Dou, Goldstein, and Ji (2025a) and Dou, Goldstein, and Ji (2025b) study AI traders who learn *strategies* through reinforcement learning, we study AI traders who learn *inference rules* through transformer learning. The frictions we identify—partial coverage and common noise—arise not from strategic interaction among AI agents but from the architecture of the pretrained model itself. Our focus is on the information structure of the transformer's attention mechanism and its equilibrium consequences for price informativeness and tail fragility. While Kelly, Kuznetsov, Malamud, and Xu (2025) embed a transformer in the SDF for empirical asset pricing, we embed the transformer in a theoretical trading model to study how its architectural properties affect market quality. In our framework, the key object is not the transformer's

predictive accuracy but its coverage boundary—which dimensions it processes and which it ignores.

Relative to the earlier big-data literature, our paper opens the black box of AI. Prior work treats AI as an information technology that is either better or worse than human processing. We show that the *form* of processing—transformer attention—matters for equilibrium outcomes, not just its quality.

Information acquisition in financial markets. A central theme of our paper is that shared AI creates correlated information and correlated ignorance across traders. This connects to the literature on information acquisition, crowded trading, and correlated strategies.

The foundational framework is Grossman and Stiglitz (1980), who show that prices cannot be fully efficient when information is costly. Verrecchia (1982) introduces endogenous precision choice. Van Nieuwerburgh and Veldkamp (2009) and Mondria (2010) study rational attention allocation under capacity constraints, and Veldkamp (2006) shows how the supply side of the information market generates correlated attention across investors. In our model, the coverage boundary is not a choice variable of the trader but a technological constraint determined by the transformer's pretraining. Traders choose whether to adopt AI, not which dimensions it covers. This makes the source of correlated attention fundamentally different: it arises from a shared technology, not from optimal individual choices.

On the crowding side, Stein (2009) argues that sophisticated investors can destabilize prices when they crowd into similar strategies. Khandani and Lo (2011) document the consequences of correlated unwinding by quant funds during August 2007. Kyle, Obizhaeva, and Wang (2026) study strategic trading when traders' private signals are correlated and show that misperception of market crowdedness distorts strategies and can create vulnerability to crashes. In these settings, crowding arises from correlated *positions* or correlated *signals*. In our model, monoculture arises from a shared *inference architecture*. The fragility we identify does not require correlated positions or correlated fundamental signals—it requires a common blind spot in how signals are processed. This distinction is economically important: crowding in the traditional sense can be mitigated by position diversification, while monoculture in our sense cannot, because the blind spot is embedded in the technology itself.

Our tail fragility results also connect to the systemic risk literature. Brunnermeier and Pedersen (2009) study the feedback between market liquidity and funding liquidity. Danielsson, Shin, and Zigrand (2012) analyze how endogenous risk arises from interacting strategies. Our dark-tail fragility result provides a new channel for systemic risk: AI monoculture creates blind spots that are invisible during normal times but become crisis triggers when a large shock realizes in an unmonitored dimension.

Transformer learning dynamics. Our analysis is related to theoretical literature that studies how transformer representations emerge during training. Vaswani,

Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017) introduce the transformer architecture based on self-attention. Edelman, Goel, Kakade, and Zhang (2022) characterize its ability to represent functions that depend on sparse parts of the input. Other work studies how the information selected by attention depends on the training data and the optimization process. Snell, Zhong, Klein, and Steinhardt (2021) develop a two stage account in which token co-occurrence is learned before attention concentrates on salient inputs. Li, Li, and Risteski (2023) show how transformer embeddings and attention recover topic structure from word co-occurrence. The closest paper to our analysis is Tian, Wang, Chen, and Du (2023), who characterize the training dynamics of a one layer transformer performing next token prediction. They represent the training data as a mixture of sequence classes and study how the decoder and self-attention layers learn from the contextual tokens associated with each class. The analysis in their paper shows how frequently occurring and class-specific contextual tokens become more prominent in the learned representation. Tian, Wang, Zhang, Chen, and Du (2024) extend this training dynamics approach to multilayer transformers by studying the joint evolution of attention and multilayer perceptrons. The framework shows how salient tokens are learned and combined into increasingly complex representations across layers. This work suggests that the learning dynamics identified in a one layer model capture forces that remain relevant in richer transformer architectures.

We build directly on the sequence class formulation and training dynamics of Tian, Wang, Chen, and Du (2023) and consider the learning process across classes and interpret each sequence class as the payoff dimension. Under a finite computation decision problem where continued training is costly, the model developer stops with a controlled loss. The tradeoff produces an endogenous boundary between learned and undertrained payoff dimensions.

Our contribution is to bring this endogenous learning boundary into a financial market. The learned dimensions receive higher inference precision, while the undertrained dimensions receive lower precision. When investors use the same pretrained model, they inherit the same pattern of precision and the same exposure to poorly learned tail risks. In this way, transformer training generates a common inference rule that becomes a source of AI monoculture in financial markets.

2 Model

2.1 Environment

There are four dates. At $t = -1$, a model developer pretrains a transformer on historical financial datasets and determines the resulting inference rule. At $t = 0$, heterogeneous investors observe the pretrained AI technology and choose whether to acquire human expertise. At $t = 1$, each investor observes private raw information, the AI model maps that evidence into private signals, and specialists can

receive an additional one-dimensional signal. Investors submit market orders, and a competitive market maker sets the price after observing aggregate order flow. At $t = 2$, the risky payoff is realized. The historical data contains no realization of the future payoff fundamentals defined below. All primitives and the AI technology are common knowledge when the financial-market stage begins.

There is a risk-free numeraire with gross return one and one risky asset in zero net supply. The risky payoff is

$$V = \sum_{n=1}^{N} \beta_n v_n, \qquad v_n \sim \mathcal{N}(0, \tau_n^{-1}), \qquad \beta_1 \geq \beta_2 \geq \cdots \geq \beta_N > 0, \qquad \tau_n > 0, \quad (1)$$

where the fundamentals are mutually independent. Let $\mathcal{N} \equiv \{1, \ldots, N\}$, with $N \geq 1$. Ordering the payoff exposures is without loss of generality. Ties are ordered arbitrarily. The index n therefore records payoff relevance rather than an otherwise empty label.

Investors form a unit continuum on an agent space that supports an exact law of large numbers. They have CARA utility $-\exp(-\gamma W)$, where $\gamma > 0$. Investors are partitioned into types $n \in \mathcal{N}$. Type n has mass $\mu_n > 0$, with $\sum_n \mu_n = 1$, and can acquire human information only about dimension n. Initial wealth is common and normalized to zero.

2.2 AI and human specialization

We study a ranked deployment rule in which AI first covers dimensions with larger payoff exposures. Define

$$\mathcal{C} = \{n \in \mathcal{N} : n \leq \bar{d}\}, \qquad \mathcal{U} = \mathcal{N} \setminus \mathcal{C}, \qquad \bar{d} \in \{0, \ldots, N\}, \quad (2)$$

where $\mathcal{C}$ is the AI-covered set. The selective-AI economy has $0 < \bar{d} < N$, while $\bar{d} = 0$ gives the no-AI benchmark. In Section 3, we show that the structure is endogenous due to the inherent limitations of transformer learning. The ordering of β_n is a relabeling. The ranked coverage rule is an economic assumption. It captures deployment toward risks with greater direct payoff exposure. Prior precisions and type masses remain unrestricted.

Every investor i observes an effective signal about each dimension,

$$s_{i,n}^A = v_n + \varepsilon_{i,n}^A, \qquad \varepsilon_{i,n}^A \sim \mathcal{N}(0, q_n^{-1}), \qquad q_n = \tau_L + \kappa_A \mathbf{1}_{\{n \in \mathcal{C}\}}, \quad (3)$$

where $\tau_L > 0$ is the precision available without AI and $\kappa_A > 0$ is AI's incremental precision on a covered dimension. Define $\tau_A \equiv \tau_L + \kappa_A$. Thus $q_n = \tau_A$ for $n \in \mathcal{C}$ and $q_n = \tau_L$ for $n \in \mathcal{U}$. One Gaussian microfoundation supplies a baseline observation with precision τ_L and, on covered dimensions, an independent observation with precision κ_A. Their precision-weighted sufficient statistic has precision τ_A. All investors face the same precision profile, while signal realizations remain investor specific.

Before signals are realized, a type-n investor either remains AI-only or pays $c > 0$ to acquire

$$s^H_{i,n} = v_n + \varepsilon^H_{i,n}, \qquad \varepsilon^H_{i,n} \sim \mathcal{N}(0, \tau_H^{-1}), \qquad \tau_H > 0. \tag{4}$$

The human signal is an overlay: a specialist retains the AI signal on every dimension. Let

$$\mathcal{F}^A_i = \sigma(s^A_{i,1}, \ldots, s^A_{i,N}), \qquad \mathcal{F}^{S,n}_i = \mathcal{F}^A_i \vee \sigma(s^H_{i,n}). \tag{5}$$

Let $x_n \in [0,1]$ be the fraction of type n that specializes and

$$m_n \equiv \mu_n x_n, \qquad \mu_A \equiv 1 - \sum_{n=1}^{N} m_n. \tag{6}$$

Acquisition occurs before signal realizations, so an interior share is ex ante mixing.

All signal errors are independent of fundamentals and, conditional on fundamentals, independent across investors, dimensions, and sources. The exact law of large numbers makes cross-sectional averages of idiosyncratic errors and private randomization vanish almost surely within every positive-mass group.

Gaussian updating gives

$$\widehat{v}^A_{i,n} = \mathbb{E}[v_n \mid \mathcal{F}^A_i] = \frac{q_n}{\tau_n + q_n} s^A_{i,n}, \qquad \delta^A_n = \mathrm{Var}(v_n \mid \mathcal{F}^A_i) = \frac{1}{\tau_n + q_n}, \tag{7}$$

$$\widehat{v}^{S,n}_{i,n} = \mathbb{E}[v_n \mid \mathcal{F}^{S,n}_i] = \frac{q_n s^A_{i,n} + \tau_H s^H_{i,n}}{\tau_n + q_n + \tau_H}, \quad \delta^S_n = \mathrm{Var}(v_n \mid \mathcal{F}^{S,n}_i) = \frac{1}{\tau_n + q_n + \tau_H}. \tag{8}$$

On every $k \neq n$, a type-n specialist has the AI-only posterior. Hence

$$D^A = \sum_{k=1}^{N} \frac{\beta_k^2}{\tau_k + q_k}, \qquad D^{S,n} = D^A - \Delta^H_n, \qquad \Delta^H_n = \frac{\beta_n^2 \tau_H}{(\tau_n + q_n)(\tau_n + q_n + \tau_H)}. \tag{9}$$

2.3 Financial market and equilibrium

At date 1, noise traders submit

$$u \sim \mathcal{N}(0, \sigma_u^2), \qquad \sigma_u^2 > 0, \tag{10}$$

independently of fundamentals and private information. Each informed investor submits an unrestricted market order $X_i \in \mathbb{R}$ before observing u, aggregate order flow, or the transaction price. Each investor is infinitesimal and takes the pricing rule and execution-price distribution as unaffected by her order. Aggregate order flow and competitive pricing are

$$Y = \int X_i \, di + u, \qquad p(Y) = \mathbb{E}[V \mid Y]. \tag{11}$$

Terminal wealth is $W^A_i = X^A_i(V - p)$ for an AI-only investor and $W^{S,n}_i = -c + X^{S,n}_i(V - p)$ for a specialist.

A linear rational-expectations equilibrium consists of specialization shares, linear order strategies, and an affine price such that orders maximize conditional expected utility, orders aggregate to Y, the price equals $\mathbb{E}[V \mid Y]$, and specialization is optimal. Investors with the same information use the same strategy. No equality across dimensions is imposed. The primitive vector is

$$\Theta = \left(N, \bar{d}, \{\beta_n, \tau_n, \mu_n\}_{n=1}^{N}, \tau_L, \kappa_A, \tau_H, \gamma, c, \sigma_u^2\right). \tag{12}$$

Every object introduced below is derived from Θ and, when the trading stage is considered separately, the given vector x.

3 Inference Rule of Generative AI

This section explains how generative AI learns financial information during pretraining at $t = -1$. We present the transformer learning mechanism and its financial interpretation in the main text. Appendix A provides the full mathematical derivation and proof.

3.1 Structured Sequence Classes

Transformer learning begins with a next token prediction task. In our setting, each training sample contains a sequence of observed tokens followed by a query token and a target token,

$$(x_1, \ldots, x_{T-1}, m, n), \qquad n \in [N],$$

where x_1 through x_{T-1} are contextual tokens drawn from a vocabulary $[M]$, m is the query token, and n is the correct next token that the transformer learns to predict.

A one layer transformer processes this sequence in three steps. Self-attention aggregates the contextual tokens according to the query token. A normalization layer converts the resulting output into a representation. The decoder then maps this representation into a probability distribution over all candidate next tokens. Training rewards the model for assigning a high probability to the correct token.

Following Tian, Wang, Chen, and Du (2023), two matrices summarize the key objects learned by the transformer. The matrix Z governs self-attention. Its entry z_{ml} measures how heavily the query token m weights contextual token l. The matrix Y governs the decoding step and determines which next token the model favors after the representation has been formed. Thus, attention determines which evidence receives weight, while the decoder determines what that evidence implies. The upper panel of Figure 1 shows the architecture.

We next explain these elements' financial interpretation in our research question. The query token m represents the common valuation task for the asset. The possible next token labels $n \in [N]$ index the payoff information channels associated

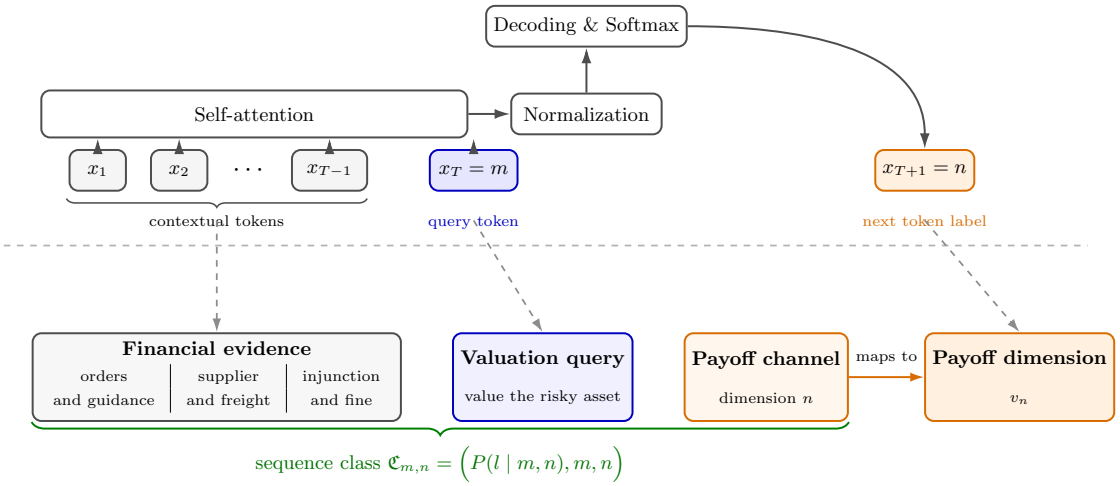


Figure 1: Transformer learning and its financial reading. The upper panel shows the model architecture. Contextual tokens enter self attention, the query token directs the weighting, and the decoder predicts the next token. The lower panel is the associated financial meaning. Contextual tokens carry the observable evidence from financial document, the query token is the common valuation question and the next token links the payoff dimension that the evidence points to. Together these objects form the sequence class $\mathfrak{C}_{m,n}$, which is the pretraining environment of payoff dimension v_n.

with $v_1, \ldots, v_N$. Contextual tokens preceding m represent the financial evidence available to the transformer. Based on this evidence, the transformer predicts which payoff information channel n is relevant for the valuation task. The lower panel of Figure 1 illustrates this interpretation.

A single valuation task can involve several payoff information channels. We formalize this structure through the mapping used by Tian, Wang, Chen, and Du (2023).

Assumption 1 (Common valuation query)**.** *There exists a mapping ψ from the next token labels to the query vocabulary such that*

$$\psi(n) = m \qquad \text{for every } n \in [N].$$

Consequently,

$$\psi^{-1}(m) = [N].$$

The risky asset therefore generates one overall valuation question, while its value depends on several payoff information channels. Conditional on the query token m and the next token label n, the contextual token distribution is

$$P(l \mid m, n), \qquad l \in [M].$$

A high value of $P(l \mid m, n)$ means that contextual token l appears frequently in sequences whose evidence points to payoff channel n. The conditional distribution therefore describes the composition of financial evidence associated with that channel. We denote this sequence class by $\mathfrak{C}_{m,n} = \big(P(l \mid m, n), m, n\big)$.

$\mathfrak{C}_{m,n}$ is the complete pretraining environment associated with payoff component v_n. The transformer learns about v_n through the sequences belonging to this class. The pretraining corpus is a mixture of these N sequence classes. Here we index from the highest pretraining frequency to the lowest pretraining frequency.

Assumption 2 (Pretraining class frequencies)**.** *For the fixed query token m, let p_n^{tr} denote the conditional probability*

$$p_n^{\mathrm{tr}} = P\left(x_{T+1} = n \mid x_T = m\right), \qquad n = 1, \ldots, N,$$

where

$$p_1^{\mathrm{tr}} > p_2^{\mathrm{tr}} > \cdots \; p_N^{\mathrm{tr}} > 0, \qquad \sum_{n=1}^{N} p_n^{\mathrm{tr}} = 1.$$

The class frequency p_n^{tr} determines how often the entire evidence environment $\mathfrak{C}_{m,n}$ appears in the training sample. Consequently, at global training time t, class n has received an amount of pretraining proportional to $p_n^{\mathrm{tr}} t$.

Definition 1 (Local learning clock)**.** *The local learning clock of class n is*

$$\ell_n(t) := p_n^{\mathrm{tr}} t.$$

It measures the expected amount of pretraining received by class n at global training time t.

All classes share the same global training time. A class with a higher pretraining frequency therefore runs a faster local clock and moves further along its learning path at every moment,

$$p_i^{\mathrm{tr}} > p_j^{\mathrm{tr}} \qquad \Longrightarrow \qquad \ell_i(t) > \ell_j(t) \qquad \text{for every } t > 0.$$

Each class represents the evidence environment associated with one payoff dimension. The frequency distribution of the pretraining corpus therefore generates uneven learning across the payoff dimensions of the risky asset. Class frequency determines the pace of learning, and we next study the consequences of these non-uniform data.

3.2 Endogenous Learning Cutoff

We characterize how the differences in local exposure determine the learning outcome of each payoff dimension. Following Tian, Wang, Chen, and Du (2023),

we hold the self-attention matrix Z fixed and study the effective logit dynamics generated by Y. The complete derivation is provided in Appendix A.1.

Let $\alpha_{nn}(t)$ denote the probability assigned to the correct next token for class n at global training time t. The effective logit dynamics show that this probability depends on class n only through its local exposure $p_n^{\mathrm{tr}}t$. At the beginning of training, the model assigns probability $1/M$ to every candidate token. The probability $\alpha_{nn}(t)$ then increases toward one as the local exposure of class n increases.

Every class follows the same relationship between local exposure and the correct token probability. The differences across classes arise from their pretraining frequencies. At the same global training time, a more frequent class has received more local exposure and therefore has a higher correct token probability.

The same effective logit dynamics determine the pretraining loss. Let $\mathcal{L}(t)$ denote the aggregate population loss at global training time t. Appendix A.1 shows that the loss is strictly decreasing and strictly convex. Additional training therefore continues to reduce prediction errors, while the marginal reduction becomes smaller over time.

In practice, the trainer chooses the training time by trading off the aggregate pretraining loss against the cost of computation. Let $c > 0$ denote the normalized marginal computation cost. The optimal training time is

$$T^\star = \arg\min_{t\geq 0} \left\{\mathcal{L}(t) + \eta_Y c t\right\}.$$

Proposition 12 shows that $T^\star$ is finite and unique. So when the solution is positive, training stops at the time that the marginal reduction in the aggregate loss equals the marginal computation cost.

The pretraining frequency of class n enters this marginal reduction twice. The class receives weight p_n^{tr} in the aggregate loss, and its local exposure increases at rate p_n^{tr}. Its contribution to the marginal reduction in the population loss is therefore proportional to $(p_n^{\mathrm{tr}})^2$. A rare class receives fewer updates and improvements in that class receive less weight in the training objective.

Now we can evaluate the class specific correct token probabilities at the optimal training time. Perfect learning would require the model to assign probability one to the correct token. This value can be approached as training continues indefinitely. Here we consider a conservative threshold(=one half) as a sufficient criteria. At this threshold, the correct token receives at least as much probability as the incorrect tokens combined.

Theorem 1 shows that the corresponding local learning threshold is $\ell_{1/2} = \frac{M-2+(M-1)\log(M-1)}{M\eta_Y}$, which depends only on the vocabulary size M and the learning rate η_Y. At the optimal training time, class n reaches this threshold if and only if $p_n^{\mathrm{tr}}T^\star \geq \ell_{1/2}$. We define the endogenous learning cutoff by

$$\bar{d} = \#\left\{n \in [N] \colon p_n^{\mathrm{tr}}T^\star \geq \ell_{1/2}\right\}.$$

Because the classes are indexed from the highest pretraining frequency to the lowest, the dimensions that reach the threshold form an initial segment of $[N]$. The

first $\bar{d}$ dimensions are learned to the threshold, while the remaining dimensions are undertrained at $T^\star$. Figure 2 provides a numerical illustration of how nonuniform pretraining frequencies generate different learning progress across payoff dimensions.

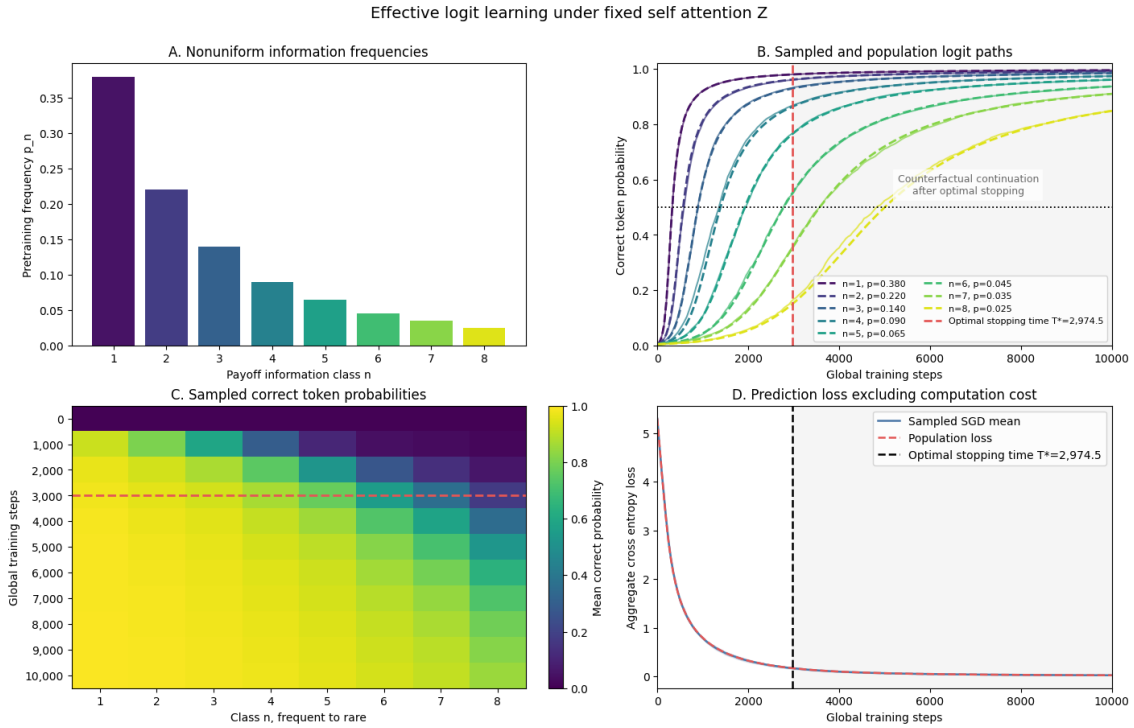


Figure 2: Effective logit learning under nonuniform pretraining frequencies. Panel A presents the original frequencies. Panel B compares the continuous population paths with the means across the sampled paths. Panel C shows the sampled correct token probabilities across classes and training times. Panel D presents the prediction loss without the computation cost. The parameters are $N = 8$, $M = 200$, $\eta_Y = 0.05$, and $c = 0.002$.

At the optimal stopping time, the first six classes have crossed the learning threshold, while the last two classes remain below it. The figure clarifies that the resulting cutoff reflects finite computation. One thing needs to clarify is that the correct token probabilities of the rare classes can continue to increase in the counterfactual region after $T^\star$. This is because the undertrained phenomenon is from the developer stopping choice rather than these classes can not be learned.

Furthermore, Theorem 1 shows that $\bar{d}$ is weakly decreasing in the computation cost c. A higher cost reduces the optimal training time and allows fewer local clocks to reach the common threshold. Figure 3 connects the class specific learning paths to the trainer's stopping decision and the resulting cutoff.

Panels A and C show why the developer stops with an interior cutoff. Continuing beyond $T^\star$ would allow additional dimensions to cross the learning threshold, but it would also increase the developer's total objective. Panel B illustrates the comparative static in Theorem 1. A higher computation cost moves the optimal

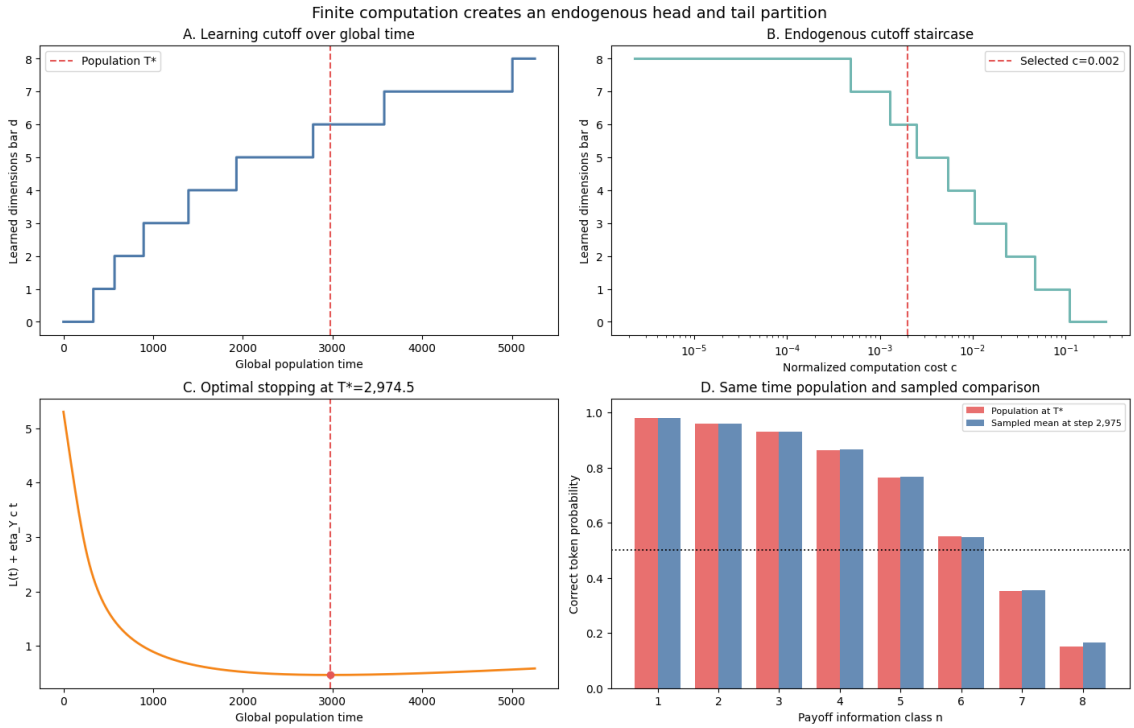


Figure 3: Optimal stopping and the endogenous learning cutoff. Panel A shows the number of payoff dimensions that reach the learning threshold as global training time increases. Panel B shows the cutoff as a function of the normalized computation cost. Panel C shows the developer's objective, including both the aggregate pretraining loss and computation cost. Its unique minimum determines $T^\star$. Panel D compares the population probabilities at $T^\star$ with the sampled mean at the nearest integer training step.

stopping time earlier and weakly reduces $\bar{d}$. Panel D shows that the continuous population result closely matches the sampled effective logit updates at the same stopping time.

Finally, since the cutoff theorem allows the corner cases $\bar{d} = 0$ and $\bar{d} = N$, we next give conditions under which the optimal training time produces a nondegenerate cutoff. Appendix A.3 derives these conditions from the pretraining frequencies, the vocabulary size, and the computation cost.

Fix $k \in \{0, \ldots, N-2\}$. Corollary 11 shows that if

$$\frac{p_N^{\mathrm{tr}} \log M}{\eta_Y \ell_{1/2}} < c \leq \frac{M}{4(M-1)} \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2,$$

then

$$k+1 \leq \bar{d} \leq N-1.$$

The lower bound on c ensures that training stops before the least frequent class reaches the learning threshold. The upper bound on c ensures that the squared frequency mass beyond class k remains large enough for training to continue until at least class $k+1$ reaches the threshold. Together, these conditions state that some payoff dimensions are learned to the threshold while at least one dimension remains undertrained.

The endogenous cutoff therefore divides the N payoff information dimensions into two groups. Dimensions $\{1, \ldots, \bar{d}\}$ have reached the common learning threshold by the optimal training time, while left dimensions $\{\bar{d}+1, \ldots, N\}$ remain below this threshold. Motivated by this difference in pretraining outcomes, the next section formally states that the AI processes the first group with higher precision and the second group with lower precision.

4 Trading Equilibrium and Endogenous Specialization

4.1 Trading equilibrium

We solve backward for a fixed $x \in [0,1]^N$. Suppose aggregate flow takes the form

$$Y = \sum_{n=1}^{N} g_n v_n + u. \tag{13}$$

Because the fundamentals and noise order are independent normal variables, the market maker infers each dimension from the same scalar flow according to

$$\mathbb{E}[v_n \mid Y] = \frac{g_n/\tau_n}{\sigma_u^2 + \sum_{k=1}^{N} g_k^2/\tau_k} Y, \qquad n \in \mathcal{N}. \tag{14}$$

Aggregating these dimension-level posterior means gives

$$p(Y) = \lambda Y, \qquad \lambda = \frac{\sum_{n=1}^{N} \beta_n g_n/\tau_n}{\sigma_u^2 + \sum_{n=1}^{N} g_n^2/\tau_n}. \tag{15}$$

The zero intercept follows from zero unconditional means. Define the residual payoff exposure of dimension n by

$$a_n \equiv \beta_n - \lambda g_n. \tag{16}$$

Thus the pricing error is

$$V - p(Y) = \sum_{n=1}^{N} a_n v_n - \lambda u. \tag{17}$$

The conditional pricing-error variances of an AI-only investor and a type-n specialist are

$$\Omega^A \equiv \lambda^2 \sigma_u^2 + \sum_{k=1}^{N} \frac{a_k^2}{\tau_k + q_k}, \tag{18}$$

$$\Omega^{S,n} \equiv \Omega^A - \frac{a_n^2 \tau_H}{(\tau_n + q_n)(\tau_n + q_n + \tau_H)}. \tag{19}$$

These are equilibrium objects, not additional primitives.

Proposition 1 (Trading equilibrium)**.** *Fix $x \in [0,1]^N$. A linear trading equilibrium exists. Every such equilibrium has $\lambda > 0$, $a_n > 0$, and $g_n > 0$ for all n. Its unique individual best responses are*

$$X_i^A = \frac{\sum_{k=1}^{N} a_k \widehat{v}_{i,k}^A}{\gamma \Omega^A}, \tag{20}$$

$$X_i^{S,n} = \frac{\sum_{k \neq n} a_k \widehat{v}_{i,k}^A + a_n \widehat{v}_{i,n}^{S,n}}{\gamma \Omega^{S,n}}. \tag{21}$$

The $N+1$ equilibrium coefficients solve (15) *and*

$$g_n = \frac{a_n}{\gamma} \left[\frac{q_n}{\tau_n + q_n} \left(\frac{\mu_A}{\Omega^A} + \sum_{\ell \neq n} \frac{m_\ell}{\Omega^{S,\ell}} \right) + \frac{m_n}{\Omega^{S,n}} \frac{q_n + \tau_H}{\tau_n + q_n + \tau_H} \right], \quad n \in \mathcal{N}, \tag{22}$$

where a_n, Ω^A, and $\Omega^{S,n}$ are given by (16)–(19)*. Conversely, every positive solution of this explicit system generates a linear trading equilibrium through* (20)–(21)*.*

Some global results below use a sufficient condition that selects a single trading-equilibrium branch. Define

$$\bar{t}(\Theta^{FM}) \equiv \left[2\gamma \sigma_u \sqrt{\sum_{k=1}^{N} \frac{\beta_k^2}{\tau_k + q_k + \tau_H}} - 1 \right]^{-1}, \tag{23}$$

whenever the bracket is positive. The sufficient condition is

$$0 < \bar{t}(\Theta^{FM}) < \min\left\{\frac{1}{3}, \min_{k\in\mathcal{N}} \frac{\tau_k}{\tau_k + q_k + \tau_H}\right\}. \tag{24}$$

Under (24), the positive trading equilibrium is unique for every $x \in [0,1]^N$. Appendix Lemma 5 provides the proof. We invoke this condition only for results that require global rather than branch-specific comparisons.

Equation (22) displays how market makers learn. By the exact law of large numbers, the average posterior mean among AI-only investors loads on v_n by $q_n/(\tau_n+q_n)$. The corresponding specialist loading is $(q_n+\tau_H)/(\tau_n+q_n+\tau_H)$. These slopes, risk-adjusted trading intensities, and group masses determine g_n. The market maker then uses (15) to project V on the scalar statistic Y. Substituting that regression coefficient back into investors' residual exposures closes the fixed point.

The regression formula is equivalently the orthogonality condition

$$\lambda\sigma_u^2 = \sum_{n=1}^{N} \frac{a_n g_n}{\tau_n}. \tag{25}$$

This identity is often more useful than (15). It shows immediately why a negative-price-impact equilibrium cannot occur: (22) implies that $a_n g_n > 0$ whenever $a_n \neq 0$.

4.2 The value of human expertise

Define pricing inefficiency as the unconditional pricing-error variance:

$$\mathcal{I} \equiv \mathrm{Var}(V - p(Y)) = \lambda^2\sigma_u^2 + \sum_{k=1}^{N} \frac{a_k^2}{\tau_k}. \tag{26}$$

This notation abbreviates a derived expression and introduces no primitive. The next result gives both expected profits and the decision-relevant value of information.

Define the value of human expertise as the certainty equivalent of the difference between the ex ante utility with human expertise and without expertise. Given the equilibrium allocation x, the value of type-n specialization is

$$H_n(x) = \frac{1}{\gamma}\log\left(\frac{\mathbb{E}\left[\mathbb{E}\left[\exp\left\{-\gamma X_i^*\left(V - p(x)\right)\right\}|F_i^A\right]\right]}{\mathbb{E}\left[\mathbb{E}\left[\exp\left\{-\gamma X_i^*\left(V - p(x)\right)\right\}|F_i^{S,n}\right]\right]}\right). \tag{27}$$

Proposition 2 (The value of human expertise)**.** *At the trading equilibrium associated with x, the value of human expertise is*

$$H_n(x) = \frac{1}{2\gamma}\log\left(\frac{\Omega^A}{\Omega^{S,n}}\right) = -\frac{1}{2\gamma}\log\left[1 - \frac{a_n^2\tau_H}{(\tau_n+q_n)(\tau_n+q_n+\tau_H)}\frac{1}{\lambda^2\sigma_u^2 + \sum_k a_k^2/(\tau_k+q_k)}\right]. \tag{28}$$

A type-n investor specializes if and only if $H_n(x) > c$, remains AI-only if and only if $H_n(x) < c$, and is indifferent if and only if $H_n(x) = c$.

Equation (28) converts the improvement in the investor's information set into a monetary willingness to pay. The financial-market feedback is already included because λ, g, and the residual exposures are evaluated at the trading equilibrium associated with x. Define the equilibrium attenuation of dimension n by

$$t_n \equiv \frac{\lambda g_n}{a_n} > 0. \tag{29}$$

The identity $a_n = \beta_n/(1 + t_n)$ follows from (16). It separates primitive payoff exposure from endogenous price inference.

Corollary 1 (Cross-sectional ranking)**.** *At any trading equilibrium,*

$$H_n(x) > H_\ell(x) \quad \Longleftrightarrow \quad \frac{\Delta_n^H}{(1+t_n)^2} > \frac{\Delta_\ell^H}{(1+t_\ell)^2}. \tag{30}$$

The same ordering applies to gross expected-profit gains. If dimensions n and ℓ have the same prior and signal precisions, then

$$H_n(x) > H_\ell(x) \quad \Longleftrightarrow \quad \frac{\beta_n}{1+t_n} > \frac{\beta_\ell}{1+t_\ell}. \tag{31}$$

Human expertise is therefore most valuable where it removes a large share of the remaining payoff risk. A high primitive loading β_n raises this value, but informative order flow attenuates it through t_n; payoff importance alone does not rank specialization incentives.

Corollary 2 (Ordered exposure and the AI boundary)**.** *Suppose $0 < \bar{d} < N$. At any trading equilibrium, the value of expertise jumps upward at the coverage boundary if and only if*

$$\left(\frac{\beta_{\bar{d}+1}}{\beta_{\bar{d}}}\right)^2 \left(\frac{1+t_{\bar{d}}}{1+t_{\bar{d}+1}}\right)^2 \frac{(\tau+\tau_A)(\tau+\tau_A+\tau_H)}{(\tau+\tau_L)(\tau+\tau_L+\tau_H)} > 1. \tag{32}$$

For a closed-form quality threshold, let $B \equiv \tau+\tau_L$ and let $\widetilde{t}$ be any valid uniform upper bound on t_n over the precision range under study. Set

$$R_{\bar{d}} \equiv (1+\widetilde{t})^2 \left(\frac{\beta_{\bar{d}}}{\beta_{\bar{d}+1}}\right)^2. \tag{33}$$

Then a boundary jump is guaranteed whenever

$$\kappa_A > \kappa_{\bar{d}}^* \equiv \frac{1}{2}\left[\sqrt{\tau_H^2 + 4R_{\bar{d}}B(B+\tau_H)} - (2B+\tau_H)\right]. \tag{34}$$

Equation (32) displays the central tradeoff. The first uncovered dimension has a smaller payoff exposure, but AI no longer compresses its information gap. The attenuation ratio records the equilibrium price response. Equation (34) gives the AI improvement that dominates both the exposure decline and the largest admissible inference response.

Corollary 3 (Strategic interaction in human expertise)**.** *For any types n and ℓ on a locally unique branch,*

$$\operatorname{sgn}\left(\frac{\partial H_n}{\partial x_\ell}\right) = \operatorname{sgn}\left(2\frac{\partial \log a_n}{\partial x_\ell} - \frac{\partial \log \Omega^A}{\partial x_\ell}\right). \tag{35}$$

Specialization by type ℓ is a strategic complement to type n when the percentage decline in total residual risk exceeds twice the percentage decline in dimension-n residual exposure. It is a strategic substitute when the reverse inequality holds.

Proposition 3 (The trading externality of specialization)**.** *Suppose condition* (24) *holds. At the unique trading equilibrium, increasing any specialization share x_ℓ strictly lowers every residual payoff exposure, raises price impact, and improves price efficiency:*

$$\frac{\partial a_n}{\partial x_\ell} < 0 \quad \textit{for every } n, \qquad \frac{\partial \lambda}{\partial x_\ell} > 0, \qquad \frac{\partial \mathcal{I}}{\partial x_\ell} < 0. \tag{36}$$

The derivatives are one-sided at a boundary share. Hence, for two share profiles $x' \geq x$ with at least one strict inequality, with the trading equilibrium re-solved at each profile,

$$a_n(x') < a_n(x) \quad \textit{for every } n, \qquad \lambda(x') > \lambda(x), \qquad \mathcal{I}(x') < \mathcal{I}(x). \tag{37}$$

Specialization consequently has a positive information externality: research in one dimension improves the price as a signal about the entire payoff. This does not by itself make research choices complements, because the more informative price can also reduce the private return to additional expertise.

4.3 Endogenous specialization

Combining Proposition 1 with Proposition 2 yields a finite complementarity problem. For every n,

$$\begin{cases} x_n = 1, & H_n(x) > c, \\ x_n = 0, & H_n(x) < c, \\ x_n \in [0,1], & H_n(x) = c. \end{cases} \tag{38}$$

Equivalently, for any fixed positive scaling number, x equals the projection onto $[0,1]^N$ of x plus that number times $H(x) - c\mathbf{1}$. This projected equation is convenient computationally but does not alter the economics.

Proposition 4 (Existence and uniqueness of specialization equilibrium)**.**

1. *A joint equilibrium exists on every globally unique trading branch. More generally, existence holds for any continuous selection of trading equilibria.*

2. *If H is continuously differentiable, the joint equilibrium is globally unique if, for every $x \in [0,1]^N$ and every n,*

$$-\frac{\partial H_n}{\partial x_n} > \frac{1}{2}\sum_{\ell \neq n}\left|\frac{\partial H_n}{\partial x_\ell} + \frac{\partial H_\ell}{\partial x_n}\right|. \tag{39}$$

3. *If the Jacobian of the active indifference equations is nonsingular at a joint equilibrium, the active shares are locally unique.*

Condition (39) is a global dominance test, not a new primitive. Every derivative is obtained from the primitive trading system as in Proposition 3. Economically, the absolute own-share response must dominate the symmetrized spillovers from all other specialization choices. The condition permits either strategic substitutes or complements and is stronger than the local nonsingularity test. Because cost does not enter these derivatives, it establishes uniqueness for every c once it holds uniformly on the share cube.

The local comparative statics have a compact closed form. Let $\mathcal{A}$ denote the set of interior types and define

$$\mathcal{G}(g, \lambda, x_{\mathcal{A}}; \theta) \equiv \begin{pmatrix} F(g, \lambda; q(\theta), x) \\ H_{\mathcal{A}}(q(\theta), x) - c(\theta)\mathbf{1} \end{pmatrix}. \tag{40}$$

Corner shares are fixed while their inequalities remain strict.

Proposition 5 (Closed-form local joint comparative statics)**.** *Suppose the Jacobian $\mathcal{G}_{(g,\lambda,x_{\mathcal{A}})}$ is nonsingular. For any scalar parameter θ,*

$$\begin{pmatrix} dg/d\theta \\ d\lambda/d\theta \\ dx_{\mathcal{A}}/d\theta \end{pmatrix} = -\mathcal{G}^{-1}_{(g,\lambda,x_{\mathcal{A}})}\mathcal{G}_{\theta}. \tag{41}$$

In particular,

$$\begin{pmatrix} dg/dc \\ d\lambda/dc \\ dx_{\mathcal{A}}/dc \end{pmatrix} = \mathcal{G}^{-1}_{(g,\lambda,x_{\mathcal{A}})} \begin{pmatrix} 0_{N+1} \\ \mathbf{1} \end{pmatrix}. \tag{42}$$

Equations (80) *and* (41) *then give every response of a, H, price efficiency, and depth. No equilibrium coefficient is held fixed.*

Proposition 6 (Closed-form joint equilibrium in one dimension)**.** *Suppose $N = 1$ and suppress subscripts. The joint market-specialization equilibrium is globally unique for every primitive vector. If*

$$c \geq \frac{1}{2\gamma} \log\left(\frac{\tau + q + \tau_H}{\tau + q}\right), \tag{43}$$

then $x^* = 0$. Otherwise define the derived number

$$z_c = \frac{\tau}{e^{2\gamma c} - 1}\left[\frac{1}{\tau + q} - \frac{e^{2\gamma c}}{\tau + q + \tau_H}\right] > 0. \tag{44}$$

Then the specialization share is the following closed-form function of the primitives:

$$x^* = \min\left\{1, \max\left\{0, \frac{\dfrac{\gamma\beta\sigma_u\sqrt{z_c}}{\sqrt{\tau}(1+z_c)} - \dfrac{q}{(\tau+q)\left[z_c + \tau/(\tau+q)\right]}}{\dfrac{q+\tau_H}{(\tau+q+\tau_H)\left[z_c + \tau/(\tau+q+\tau_H)\right]} - \dfrac{q}{(\tau+q)\left[z_c+\tau/(\tau+q)\right]}}\right\}\right\}. \tag{45}$$

If (45) *is interior, the trading coefficient is* $z = z_c$ *and* (86) *gives* (g, λ, a)*. At a corner,* z *is the unique root of* (85) *evaluated at* x^**. Whenever the share is interior,* $dx^*/dc < 0$*.*

The formula provides a closed-form benchmark for the multidimensional complementarity system. The exponential term converts the certainty-equivalent cost into the unique residual-risk ratio at which the type is indifferent. The remaining fraction in (45) is the specialist mass required to generate that ratio in order flow. Clipping occurs only when the indifference point lies outside the trading outcomes attainable between the two corner allocations.

Interior specialization is not generically restricted to a single dimension in a fully asymmetric N-dimensional system: several equations $H_n(x) = c$ can determine several shares. What is generic is that an interior share must satisfy an equality rather than an inequality. A claim that only one type mixes requires an additional rank or single-crossing restriction.

The next corollary states exactly when the economically useful three-region pattern emerges.

Corollary 4 (Covered AI-only region, active human tail, and dark tail)**.** *Consider a joint equilibrium. Suppose*

$$\frac{a_n^2 \tau_H}{(\tau_n + q_n)(\tau_n + q_n + \tau_H)} < \Omega^A\big(1 - e^{-2\gamma c}\big) \quad \text{for every } n \le \bar{d}, \tag{46}$$

the reverse strict inequality holds at $n = \bar{d} + 1$*, and the left side of* (46) *is strictly decreasing for* $n > \bar{d}$*. Then there is an index* $r \in \{\bar{d} + 2, \dots, N + 1\}$ *such that*

$$x_n = 0 \quad (n \le \bar{d}), \qquad x_n = 1 \quad (\bar{d} < n < r), \qquad x_n = 0 \quad (r < n \le N), \tag{47}$$

If $r \le N$*, then* $x_r = 0$ *when* $H_r < c$*, while* $x_r \in [0, 1]$ *is possible only when* $H_r = c$*. If no uncovered type is exactly indifferent, all shares are corners.*

Corollary 5 (Unique boundary mixing)**.** *Fix a candidate uncovered boundary type* $r \le N$*, set all other shares to the three-region corners in* (47)*, and let* $H_r(s)$ *denote its equilibrium value when* $x_r = s \in [0, 1]$*. If* $H_r(s)$ *is strictly decreasing, then its unique equilibrium share is*

$$x_r^* = \begin{cases} 0, & c \ge H_r(0), \\ 1, & c \le H_r(1), \\ \text{the unique } s \in (0,1) \text{ satisfying } H_r(s) = c, & H_r(1) < c < H_r(0). \end{cases} \tag{48}$$

At an interior solution that preserves the surrounding corner inequalities,

$$\frac{dx_r^*}{dc} = \left(\frac{\partial H_r}{\partial x_r}\right)^{-1} < 0. \tag{49}$$

Thus strict single crossing restricts mixing to at most one type, and strategic substitutability makes that remaining scalar fixed point unique.

Without the AI coverage discontinuity, the specialization pattern inherits the cross-sectional ranking of residual information value. Selective AI changes this benchmark by creating the upward jump at $\bar{d}+1$.

The "dark" tail is not uninformed: its investors retain the baseline signals. It is dark in the narrower sense that no costly human overlay is produced there. The result also clarifies the boundary jump. Raising precision on covered dimensions lowers their direct human value, while the first uncovered dimension retains the low baseline precision. If its residual payoff exposure is sufficiently large, specialization value jumps upward exactly where AI coverage ends.

No-AI benchmark. The no-AI economy is obtained by setting $\bar{d}=0$, so $q_n=\tau_L$ for every n. No strategy or equilibrium definition changes. Let a superscript 0 denote the solution of (15), (22), and (38) under this precision profile. The no-AI specialization value is

$$H_n^0(x)=\frac{1}{2\gamma}\log\left[\frac{(\lambda^0)^2\sigma_u^2+\sum_k(a_k^0)^2/(\tau_k+\tau_L)}{(\lambda^0)^2\sigma_u^2+\sum_k(a_k^0)^2/(\tau_k+\tau_L)-(a_n^0)^2\tau_H/((\tau_n+\tau_L)(\tau_n+\tau_L+\tau_H))}\right]. \tag{50}$$

Corollary 6 (No-AI cutoff)**.** *If*

$$\frac{(a_n^0)^2\tau_H}{(\tau_n+\tau_L)(\tau_n+\tau_L+\tau_H)} \tag{51}$$

is weakly decreasing in n at a joint no-AI equilibrium, the sets of types with $H_n^0>c$, $H_n^0=c$, and $H_n^0<c$ are consecutive, possibly empty blocks. Their shares are, respectively, one, unrestricted in $[0,1]$, and zero. If the displayed expression is strictly decreasing, the equality block contains at most one type and there is a single cutoff $r^0\in\{1,\dots,N+1\}$: $x_n^0=1$ for $n<r^0$, $x_n^0=0$ for $n>r^0$, and a boundary share can be interior only at equality.

The heterogeneous model also gives primitive corner bounds.

Corollary 7 (Primitive cost bounds under heterogeneity)**.** *Define*

$$\underline{c}(q)\equiv\min_n -\frac{1}{2\gamma}\log\left[1-\frac{\Delta_n^H}{[1+\bar{t}(\Theta)]^2\left[D^A+\sum_k\beta_k^2/(4\tau_k)\right]}\right], \tag{52}$$

$$\overline{c}(q)\equiv\max_n \frac{1}{2\gamma}\log\left(\frac{\tau_n+q_n+\tau_H}{\tau_n+q_n}\right). \tag{53}$$

Every type specializes when $c<\underline{c}(q)$. No type specializes when $c>\overline{c}(q)$. These bounds are sufficient and use the heterogeneous primitives directly.

Figure 4 plots the equilibrium value of expertise against its cost. In the selective-AI economy, the value jumps at the coverage boundary: covered dimensions remain AI-only, an adjacent uncovered block specializes, and the lowest-exposure dimensions form a dark tail. Without AI, the coverage jump disappears and the value declines to a single research cutoff. The figure illustrates Corollaries 4 and 6.

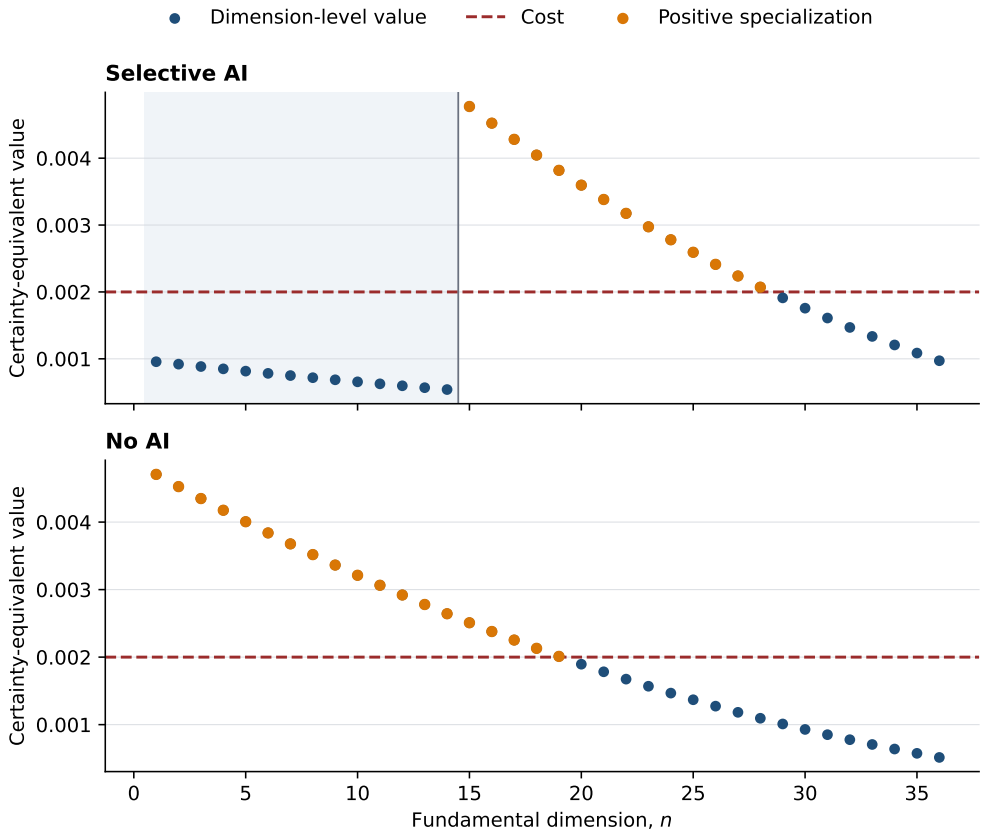


Figure 4: **The value and location of human specialization.** Each blue point is the separately computed certainty-equivalent value in (28) for one discrete dimension. There is no smoothing or interpolation. The dashed line is c, orange markers identify dimensions with positive specialization, and shading identifies the AI-covered set. We set $N = 36$, $\mu_n = 1/36$, $\tau_n = 1$, and let β_n decline linearly from 1.65 to 0.55. The remaining primitives are $\bar{d} = 14, \tau_L = 0.55, \kappa_A = 5, \tau_H = 2.8, \gamma = 4, c = 0.002, \sigma_u^2 = 1.2$.

5 The Effects of AI: AI versus no AI

5.1 AI and human expertise

Let x^{AI} and x^0 be joint equilibria under selective AI and no AI, respectively. Dimension-level and aggregate changes in human expertise are

$$\Delta x_n = x_n^{AI} - x_n^0, \qquad \Delta X^H = \sum_{n=1}^{N} \mu_n (x_n^{AI} - x_n^0). \tag{54}$$

These differences need not have a common sign. AI makes the human signal more redundant on a covered dimension. It also changes background risk and the price exposure of every dimension. The full response of each value uses the joint system in (41). Ordered exposures then turn a common technology change into different entry and exit thresholds across types.

Acquisition cost activates this heterogeneity. The corner bounds in (52) and (53) imply that both AI and no-AI economies have full research at sufficiently low costs and no research at sufficiently high costs. Between those regions, types move at different thresholds. AI can therefore change the level and location of research only on the intermediate part of the cost path. No common share or representative dimension is imposed.

Proposition 7 (AI reallocation under cutoff equilibria)**.** *Suppose the selective-AI equilibrium has the strict corner structure in Corollary 4, with active set* $\{\bar{d}+1,\dots,r^{AI}-1\}$*, and the no-AI equilibrium has the strict corner structure in Corollary 6, with active set* $\{1,\dots,r^0-1\}$*. Then*

$$\Delta x_n = \mathbf{1}_{\{\bar{d}<n<r^{AI}\}} - \mathbf{1}_{\{n<r^0\}}, \tag{55}$$

$$\Delta X^H = \sum_{\bar{d}<n<r^{AI}} \mu_n - \sum_{n<r^0} \mu_n. \tag{56}$$

If $r^{AI} \le r^0$*, AI weakly reduces specialization in every dimension and weakly reduces aggregate expertise. The aggregate reduction is strict if and only if the no-AI active set differs from the selective-AI active set. In particular, it is strict if* $r^{AI} < r^0$*, or if* $\bar{d} > 0$ *and the no-AI active set is nonempty. If* $r^{AI} > r^0$*, AI raises aggregate expertise if and only if*

$$\sum_{\max\{\bar{d}+1,r^0\}\le n<r^{AI}} \mu_n > \sum_{n\le\min\{\bar{d},r^0-1\}} \mu_n. \tag{57}$$

The left side is newly created expertise in the uncovered tail. The right side is expertise displaced from AI-covered dimensions.

Proposition 8 (One-step expansion of AI coverage)**.** *Fix all primitives other than the integer coverage boundary. Suppose the equilibria at coverage* $\bar{d}$ *and* $\bar{d}+1$ *both have the strict three-region structure in Corollary 4, with no indifferent boundary type and active sets* $\{\bar{d}+1,\dots,r_{\bar{d}}-1\}$ *and* $\{\bar{d}+2,\dots,r_{\bar{d}+1}-1\}$*, respectively. Then the dimension-level change is exactly*

$$x_n^{\bar{d}+1} - x_n^{\bar{d}} = \mathbf{1}_{\{\bar{d}+1<n<r_{\bar{d}+1}\}} - \mathbf{1}_{\{\bar{d}<n<r_{\bar{d}}\}}, \tag{58}$$

and the aggregate change is

$$X^H(\bar{d}+1) - X^H(\bar{d}) = \begin{cases} -\mu_{\bar{d}+1} + \displaystyle\sum_{r_{\bar{d}}\le n<r_{\bar{d}+1}} \mu_n, & r_{\bar{d}+1} \ge r_{\bar{d}}, \\ -\mu_{\bar{d}+1} - \displaystyle\sum_{r_{\bar{d}+1}\le n<r_{\bar{d}}} \mu_n, & r_{\bar{d}+1} < r_{\bar{d}}. \end{cases} \tag{59}$$

Thus broader coverage raises aggregate human expertise if and only if $r_{\bar{d}+1} > r_{\bar{d}}$ *and the mass of newly active dark-tail types in the first line of* (59) *exceeds* $\mu_{\bar{d}+1}$*. If* $\mu_n = 1/N$ *for every type, the expression simplifies for every cutoff movement to*

$$X^H(\bar{d}+1) - X^H(\bar{d}) = \frac{r_{\bar{d}+1} - r_{\bar{d}} - 1}{N}. \tag{60}$$

Aggregate expertise therefore rises exactly when the dark-tail cutoff moves outward by at least two dimensions, is unchanged when it moves outward by one, and falls when it does not move outward.

Covering the next dimension mechanically removes the human specialist who previously occupied the left edge of the active tail. Cross-dimension complementarity can replace that researcher only through entry at the dark-tail boundary. Proposition 8 measures the two forces without treating the dimension label as continuous. The first term in (59) is displaced boundary expertise. The sum records entry or exit among discrete tail types. A one-position outward movement merely relocates an equal-mass researcher. Only a larger movement raises total expertise.

On a locally unique branch, Proposition 5 gives the exact response to AI quality or cost. Corner shares remain fixed until a value inequality binds. Neither parameter has an unconditional aggregate sign because the covered and uncovered types face different terms in (80).

Increasing coverage by one dimension is the discrete counterpart. Both the old and new joint equilibria are solved in full. Proposition 8 then shows when entry at the dark-tail boundary is large enough to dominate the researcher displaced from the new covered dimension.

Figure 5 varies AI quality and specialization cost. Human-specialist mass is nonmonotone in AI quality: a small improvement initially redirects expertise toward uncovered dimensions, stronger AI then crowds out covered expertise, and further improvements increase the relative value of uncovered dimensions enough to induce partial recovery. Specialization declines with cost along the selected stable branch. Price-error variance falls with AI quality while λ rises, illustrating that more informative prices need not be more liquid.

5.2 Financial market quality

We use three market-quality measures. First, pricing inefficiency was defined in (26). Normal projection gives the equivalent explained-variance representation

$$\mathcal{I} = \sum_{n=1}^{N} \frac{\beta_n^2}{\tau_n} - \frac{\left(\sum_n \beta_n g_n/\tau_n\right)^2}{\sigma_u^2 + \sum_n g_n^2/\tau_n}. \tag{61}$$

Lower $\mathcal{I}$ means a more informative price. Second, market depth is

$$\mathcal{L} \equiv \frac{1}{\lambda}. \tag{62}$$

Because $\lambda > 0$, a larger $\mathcal{L}$ means that a given order imbalance moves prices less.

Third, the Gaussian-linear structure makes tail risk exact. For any fixed loss threshold $z > 0$,

$$\Pr\{|V - p(Y)| > z\} = 2\Phi\left(-\frac{z}{\sqrt{\mathcal{I}}}\right), \tag{63}$$

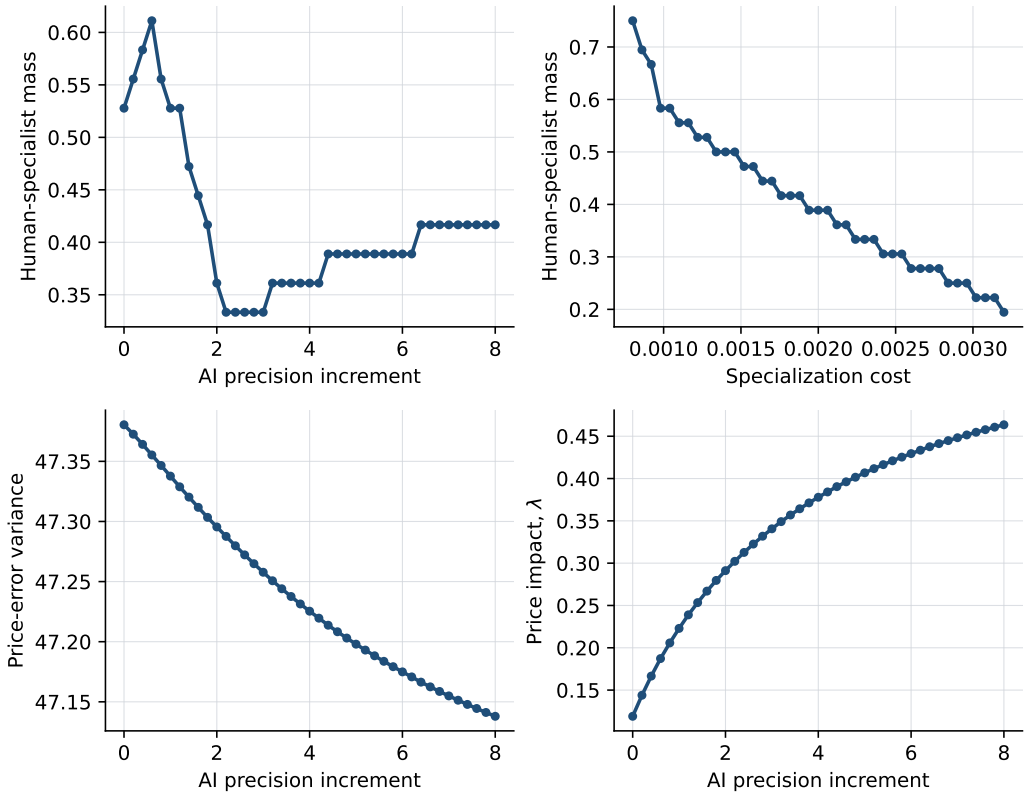


Figure 5: **AI quality, human expertise, and market quality.** Each point resolves both the trading equilibrium and specialization complementarity problem. The curves are numerical illustrations of the direct substitution, cross-dimension complementarity, and endogenous price-feedback channels.

and for a fixed confidence level $\alpha \in (1/2, 1)$, one-sided expected shortfall is

$$ES_\alpha(V - p) = \frac{\phi(\Phi^{-1}(\alpha))}{1 - \alpha}\sqrt{\mathcal{I}}. \tag{64}$$

Therefore every Gaussian tail measure in (63)–(64) ranks equilibria exactly as $\mathcal{I}$ does. A tail-risk result distinct from price efficiency would require changing a primitive, for example by introducing non-Gaussian fundamentals or state-dependent liquidity. We do not add such an assumption to manufacture a separate result.

Proposition 9 (The direct technology channel)**.** *Fix the specialization vector x. A marginal increase in the signal precision of any dimension j strictly lowers every residual payoff exposure and price-error variance, raises price impact, and lowers depth:*

$$\frac{\partial a_n}{\partial q_j} < 0 \quad \textit{for every } n, \qquad \frac{\partial \mathcal{I}}{\partial q_j} < 0, \qquad \frac{\partial \lambda}{\partial q_j} > 0, \qquad \frac{\partial \mathcal{L}}{\partial q_j} < 0. \tag{65}$$

For two precision profiles $q' \geq q$ with at least one strict inequality, if (24) *holds*

along the segment joining them, then

$$\begin{aligned} a_n(q',x) < a_n(q,x) \quad \textit{for every } n, \qquad \mathcal{I}(q',x) < \mathcal{I}(q,x), \\ \lambda(q',x) > \lambda(q,x), \qquad \mathcal{L}(q',x) < \mathcal{L}(q,x). \end{aligned} \tag{66}$$

The direct technology channel therefore has no efficiency–depth ambiguity. Better signals make order flow more revealing, so market makers optimally respond more aggressively to it. The ambiguity in the total AI effect comes from endogenous human expertise: AI may crowd out, create, or relocate specialization while it improves the underlying signals.

The AI-minus-no-AI market-quality effects are

$$\Delta\mathcal{I} = \mathcal{I}^{AI} - \mathcal{I}^{0}, \qquad \Delta\mathcal{L} = \frac{1}{\lambda^{AI}} - \frac{1}{\lambda^{0}}, \qquad \Delta ES_\alpha = \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}\left(\sqrt{\mathcal{I}^{AI}} - \sqrt{\mathcal{I}^{0}}\right). \tag{67}$$

These expressions show why efficiency and liquidity can separate. Reweighting order flow toward payoff-relevant fundamentals lowers $\mathcal{I}$, while the regression slope λ can rise because market makers rationally react more strongly to the now more informative order. A higher λ is lower depth even when price discovery improves.

Corollary 8 (Finite efficiency–depth accounting)**.** *At every trading equilibrium,*

$$\lambda^2\left(\sigma_u^2 + \sum_{n=1}^{N}\frac{g_n^2}{\tau_n}\right) = \sum_{n=1}^{N}\frac{\beta_n^2}{\tau_n} - \mathcal{I}. \tag{68}$$

Consequently, any finite AI–no-AI comparison satisfies the exact ratio

$$\frac{\mathcal{L}^{AI}}{\mathcal{L}^{0}} = \left[\frac{\sigma_u^2 + \sum_n (g_n^{AI})^2/\tau_n}{\sigma_u^2 + \sum_n (g_n^{0})^2/\tau_n}\right]^{1/2}\left[\frac{\sum_n \beta_n^2/\tau_n - \mathcal{I}^0}{\sum_n \beta_n^2/\tau_n - \mathcal{I}^{AI}}\right]^{1/2}. \tag{69}$$

If AI improves price efficiency and weakly reduces total order-flow variance, then it strictly reduces depth. If AI worsens price efficiency and weakly raises total order-flow variance, then it strictly raises depth. In every other case, (69) *gives the exact tradeoff.*

The first factor in (69) is the change in total order-flow variance. The second is the inverse change in the payoff variance explained by order flow. Thus liquidity can improve when expertise is displaced because less informed trading reduces adverse selection, even if prices become less informative. Conversely, AI can make prices more informative yet reduce depth when market makers respond to a more payoff-relevant flow. The identity uses equilibrium loadings rather than specialist counts, so it remains valid under mixed cross-dimensional reallocation.

The effect of AI can be decomposed into a technology channel and an expertise-mediated channel. Write the arguments (q, x) only here to expose which derived equilibrium is being evaluated. Then

$$\Delta\mathcal{I} = \left[\mathcal{I}(q^{AI}, x^{AI}) - \mathcal{I}(q^{AI}, x^{0})\right] + \left[\mathcal{I}(q^{AI}, x^{0}) - \mathcal{I}(q^{0}, x^{0})\right], \tag{70}$$

$$\Delta\mathcal{L} = \left[\mathcal{L}(q^{AI}, x^{AI}) - \mathcal{L}(q^{AI}, x^{0})\right] + \left[\mathcal{L}(q^{AI}, x^{0}) - \mathcal{L}(q^{0}, x^{0})\right]. \tag{71}$$

The first bracket in each line changes human specialization while holding the selective-AI precision profile fixed. The second changes information technology while holding the no-AI specialization allocation fixed.

Corollary 9 (Technology versus endogenous expertise)**.** *Suppose condition* (24) *holds under* q^{AI} *and along the segment from* q^0 *to* q^{AI}*. The second bracket in each of* (70) *and* (71) *is strictly negative whenever AI covers at least one dimension.*

1. *If* $x^{AI} \geq x^0$ *coordinatewise, both expertise-mediated brackets are nonpositive and*

$$\Delta\mathcal{I} < 0, \qquad \Delta\mathcal{L} < 0. \tag{72}$$

   Expertise expansion reinforces AI's direct improvement in price discovery and further reduces depth.

2. *If* $x^{AI} \leq x^0$ *coordinatewise, both expertise-mediated brackets are nonnegative, strictly if at least one share falls. Expertise displacement therefore offsets AI's price-discovery improvement and restores depth. Either total sign is possible according to which bracket dominates.*

3. *If AI reallocates expertise so that the share vectors are not coordinatewise ordered, neither expertise-mediated bracket has a general sign.*

The same decomposition applies to AI-quality, coverage, or cost comparisons.

Corollary 10 (Market quality under one-step coverage expansion)**.** *Suppose the premises of Proposition 8 hold and condition* (24) *holds under the* $\bar{d}+1$ *precision profile and along the one-coordinate precision segment from coverage* $\bar{d}$ *to* $\bar{d}+1$*. Holding the old specialization allocation fixed, covering dimension* $\bar{d}+1$ *strictly lowers both pricing inefficiency and market depth.*

1. *If* $r_{\bar{d}+1} \leq r_{\bar{d}}$*, then* $x^{\bar{d}+1} \leq x^{\bar{d}}$ *coordinatewise, with a strict decline for dimension* $\bar{d}+1$*. The endogenous-expertise channel therefore strictly raises both pricing inefficiency and depth, opposing the direct coverage channel. Either total sign is possible.*

2. *If* $r_{\bar{d}+1} > r_{\bar{d}}$*, the share vectors are not coordinatewise ordered: dimension* $\bar{d}+1$ *exits while dimensions* $r_{\bar{d}}, \ldots, r_{\bar{d}+1}-1$ *enter. Neither expertise-mediated market-quality effect has a general sign.*

In either case, the total efficiency sign is given exactly by (75) *after replacing the superscripts AI and* 0 *with* $\bar{d}+1$ *and* $\bar{d}$*.*

The mixed case is precisely the reallocation isolated by Propositions 7 and 8: newly active uncovered types and displaced covered types exert opposing market-quality effects. Aggregate specialist mass alone is insufficient because the entering

and exiting dimensions can have different payoff exposures and prior precisions. Equations (70)–(71) prevent that reallocation from being conflated with AI's direct technological effect.

Ordered exposure changes the shape of the adjustment. A common precision increase need not move all research shares together. High and low exposure types reach their indifference conditions at different values of AI quality. Between those thresholds, the direct technology effect and the loss of one type's research can alternate in importance. Proposition 11 below gives the exact local sign at every point.

The cost comparison nevertheless has a general structure. It does not require equal exposures.

Proposition 10 (Cost-induced market-quality reversals under heterogeneity)**.** *Let $q^{AI} \geq q^0$ with at least one strict inequality. Suppose* (24) *holds along the segment joining the two precision profiles for every specialization vector. Suppose also that* (39) *holds in both economies. Define*

$$c_F \equiv \min\{\underline{c}(q^{AI}), \underline{c}(q^0)\}, \qquad c_0 \equiv \max\{\overline{c}(q^{AI}), \overline{c}(q^0)\}. \tag{73}$$

Then both economies have $x = \mathbf{1}$ for $c < c_F$ and $x = \mathbf{0}$ for $c > c_0$. In both regions,

$$\Delta\mathcal{I}(c) < 0, \qquad \Delta\mathcal{L}(c) < 0. \tag{74}$$

If some $\widehat{c} \in (c_F, c_0)$ satisfies $\Delta\mathcal{I}(\widehat{c}) > 0$, then $\Delta\mathcal{I}(c)$ changes sign at least twice. Expected shortfall and every fixed-threshold Gaussian tail probability have the same sign changes. If instead $\Delta\mathcal{L}(\widehat{c}) > 0$, the depth difference changes sign at least twice.

The proposition separates a robust boundary fact from the interior allocation. At low and high costs, the two economies have the same research allocation. Better technology then improves price discovery and lowers depth. In the middle, ordered types enter and exit at different costs. A reversal occurs exactly when that endogenous reallocation outweighs the technology effect. Heterogeneity can generate more than two crossings because each type adds a distinct margin.

For finite changes, (61) gives the exact AI-efficiency criterion

$$\mathcal{I}^{AI} < \mathcal{I}^0 \quad \Longleftrightarrow \quad \frac{\left(\sum_n \beta_n g_n^{AI}/\tau_n\right)^2}{\sigma_u^2 + \sum_n (g_n^{AI})^2/\tau_n} > \frac{\left(\sum_n \beta_n g_n^0/\tau_n\right)^2}{\sigma_u^2 + \sum_n (g_n^0)^2/\tau_n}. \tag{75}$$

AI improves price efficiency precisely when its equilibrium order flow has greater payoff-explaining power after accounting for noise and the total variance of informed orders. The condition does not require that every dimension receive more weight.

Proposition 11 (Exact local market-quality effects)**.** *Let θ be any scalar parameter that varies along a differentiable equilibrium branch while $(\{\beta_n, \tau_n\}_{n=1}^N, \sigma_u^2)$ remain*

fixed, including κ_A, c, or a continuous perturbation used to compare two precision profiles. Then

$$\frac{d\mathcal{I}}{d\theta} = -2\lambda \sum_{n=1}^{N} \frac{a_n}{\tau_n} \frac{dg_n}{d\theta}, \tag{76}$$

$$\frac{d\lambda}{d\theta} = \frac{1}{\sigma_u^2 + \sum_n g_n^2/\tau_n} \sum_{n=1}^{N} \frac{\beta_n - 2\lambda g_n}{\tau_n} \frac{dg_n}{d\theta}, \tag{77}$$

$$\frac{d\mathcal{L}}{d\theta} = -\frac{1}{\lambda^2} \frac{d\lambda}{d\theta}, \qquad \frac{dES_\alpha}{d\theta} = \frac{\phi(\Phi^{-1}(\alpha))}{2(1-\alpha)\sqrt{\mathcal{I}}} \frac{d\mathcal{I}}{d\theta}. \tag{78}$$

In particular, any perturbation that raises every g_n weakly and at least one strictly improves price efficiency because $\lambda > 0$ and $a_n > 0$. The same perturbation can reduce depth: price impact rises exactly when its loading changes have a positive weighted inner product with $\beta_n - 2\lambda g_n$ as displayed in (77).

All loading derivatives in Proposition 11 are obtained by differentiating the primitive coefficient system and, when specialization is endogenous, the active indifference equations. Equation (76) shows that the apparent noise-price and residual-exposure terms cancel by market-maker orthogonality. Endogenous specialization can make the g_n responses change sign across dimensions, providing a disciplined route to hump-shaped or U-shaped outcomes without manufacturing a separate mechanism.

Figure 6 reports the difference indices in (54) and (67). At low AI quality, uncovered entry can initially exceed covered displacement, so AI raises specialist mass relative to the no-AI benchmark. As quality increases, the difference turns negative and later partially recovers. This nonmonotonicity reflects changes in the two endogenous research cutoffs rather than a smooth representative-agent response.

AI lowers both price-error variance and market depth at every positive quality or coverage point shown. The total signs are calibration specific because expertise is reallocated across dimensions with different β_n. Varying cost changes the expertise difference nonmonotonically as the AI and no-AI cutoffs cross cost at different values. The corner bounds show why this channel is concentrated at intermediate costs.

Figure 7 applies Proposition 10. At the low- and high-cost corners, the AI and no-AI economies have the same research allocation, so the direct technology channel determines the comparison. In the shaded intermediate region, staggered type-level exits change the location of expertise and can temporarily overturn that comparison.

6 Conclusion

This paper studies how transformer pretraining changes the production of financial information. Nonuniform representation of payoff-relevant topics makes some learning clocks run faster than others. Because additional training is costly and its

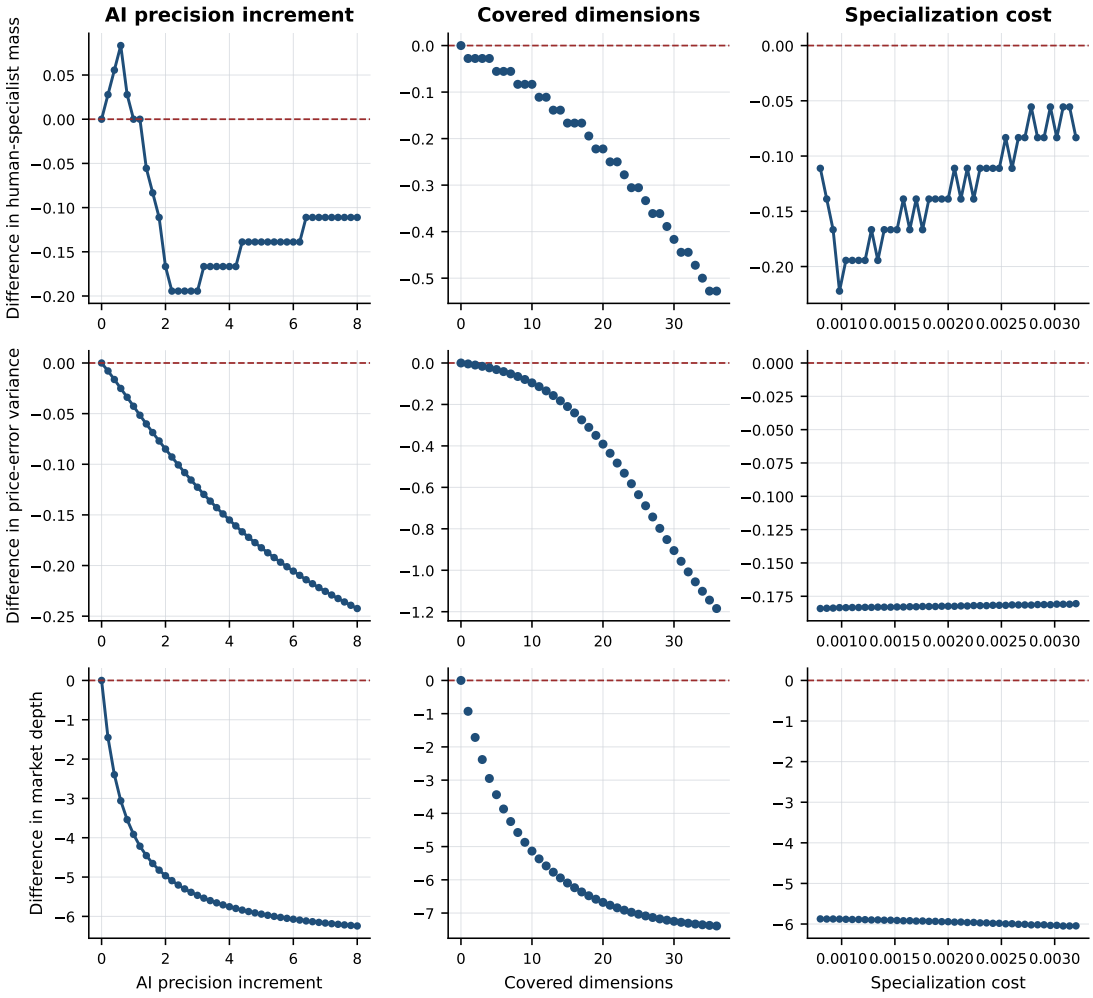


Figure 6: **AI-minus-no-AI differences.** Rows report the change in human-specialist mass, price-error variance, and market depth. Columns vary AI precision, the integer number of covered dimensions, and specialization cost. Every marker re-solves both the trading and specialization equilibria. Coverage is discrete. The middle-column markers correspond to all integers $\bar{d} = 0, \ldots, 36$ and are not connected or interpolated. In the cost column, the no-AI benchmark is separately re-solved at every cost.

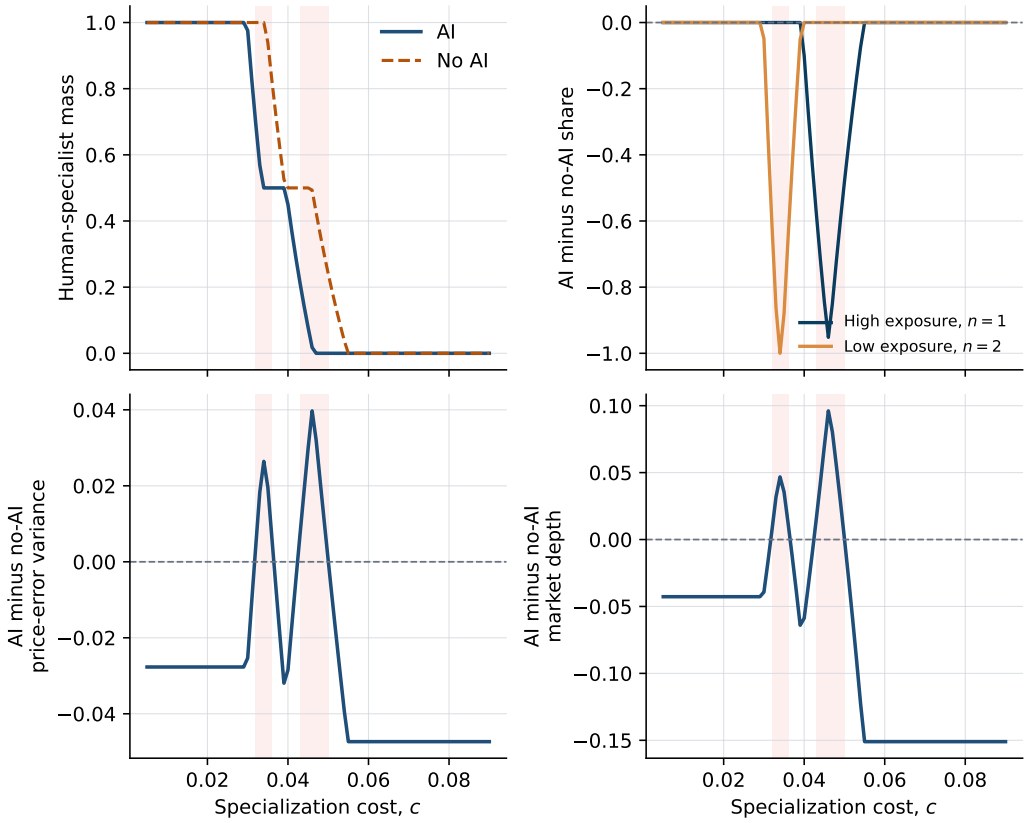


Figure 7: **Acquisition cost with heterogeneous payoff exposure.** The top panels show aggregate and type-level specialization responses. The bottom panels report AI-minus-no-AI market outcomes. Red shading marks costs at which AI raises price-error variance. Better technology dominates at the common corners. Type-specific research exits generate the interior reversals.

average benefit declines, the model developer optimally stops with an endogenous coverage boundary. Investors who use the pretrained technology share this uneven precision profile even though their private signal realizations remain independent. AI monoculture in the model is therefore a common pattern of attention, not a mechanical assumption of identical forecasts.

Embedding this technology in a financial market yields two main economic forces. First, AI substitutes directly for human research on covered dimensions but can raise the relative value of expertise just outside its coverage boundary. Under single crossing, human specialization is concentrated in an intermediate uncovered region, between an AI-covered region and a low-exposure dark tail. Second, AI affects market quality through both technology and endogenous information production. Holding research fixed, higher AI precision improves price discovery and lowers market depth because market makers respond more strongly to informative order flow. Entry, exit, and relocation of specialists can reinforce or offset this direct effect. These reallocations matter most at intermediate research costs, where the AI

and no-AI economies support different research allocations.

The analysis highlights why average predictive performance is insufficient for evaluating AI in financial markets. The location of model coverage and the identity of the specialists displaced or induced by AI determine which risks enter prices. This observation also points to empirically relevant objects: topic frequencies in training data, category-level AI performance, changes in analyst specialization around coverage boundaries, and the payoff-weighted information content of order flow. Extensions could allow multiple competing models, endogenous corpus design, or non-Gaussian tail risks. Those additions may diversify coverage or create distinct stability effects, but the basic mechanism remains: a technology that improves information processing unevenly also changes where private incentives sustain human expertise.

References

Brunnermeier, M. K., and L. H. Pedersen (2009): "Market liquidity and funding liquidity," *The review of financial studies*, 22(6), 2201–2238.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse signals in the cross-section of returns," *The Journal of Finance*, 74(1), 449–492.

Colliard, J.-E., T. Foucault, and S. Lovo (2022): "Algorithmic pricing and liquidity in securities markets," *Working paper, HEC Paris*.

Danielsson, J., H. S. Shin, and J.-P. Zigrand (2012): "Endogenous and systemic risk," in *Quantifying systemic risk*, pp. 73–94. University of Chicago Press.

Dou, W. W., I. Goldstein, and Y. Ji (2025a): "AI-powered trading, algorithmic collusion, and price efficiency," *NBER Working Paper No. 34054*.

——— (2025b): "Financial Market Fragility in the Era of AI Planning," *Available at SSRN 5763222*.

Dugast, J., and T. Foucault (2018): "Data abundance and asset price informativeness," *Journal of Financial Economics*, 130(2), 367–391.

Edelman, B. L., S. Goel, S. Kakade, and C. Zhang (2022): "Inductive biases and variable creation in self-attention mechanisms," in *International Conference on Machine Learning*, pp. 5793–5831. PMLR.

Farboodi, M., A. Matray, L. Veldkamp, and V. Venkateswaran (2022): "Where has all the data gone?," *The Review of Financial Studies*, 35(7), 3101–3138.

Farboodi, M., and L. Veldkamp (2020): "Long-run growth of financial data technology," *American Economic Review*, 110(8), 2485–2523.

Goldstein, I., C. S. Spatt, and M. Ye (2021): "Big data in finance," *Review of Financial Studies*, 34(7), 3213–3225.

Grossman, S. J., and J. E. Stiglitz (1980): "On the impossibility of informationally efficient markets," *The American economic review*, 70(3), 393–408.

Kelly, B. T., B. Kuznetsov, S. Malamud, and T. A. Xu (2025): "Artificial intelligence asset pricing models," *NBER Working Paper No. 33351*.

Khandani, A. E., and A. W. Lo (2011): "What happened to the quants in August 2007? Evidence from factors and transactions data," *Journal of Financial Markets*, 14(1), 1–46.

Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2026): "Trading in crowded markets," *Journal of Financial and Quantitative Analysis*, 61(1), 137–175.

Li, Y., Y. Li, and A. Risteski (2023): "How do transformers learn topic structure: Towards a mechanistic understanding," in *International Conference on Machine Learning*, pp. 19689–19729. PMLR.

Lopez-Lira, A., and Y. Tang (2023): "Can chatgpt forecast stock price movements? return predictability and large language models," *arXiv preprint arXiv:2304.07619*.

Mondria, J. (2010): "Portfolio choice, attention allocation, and price comovement," *Journal of Economic Theory*, 145(5), 1837–1864.

Snell, C., R. Zhong, D. Klein, and J. Steinhardt (2021): "Approximating how single head attention learns," *arXiv preprint arXiv:2103.07601*.

Stein, J. C. (2009): "Presidential address: Sophisticated investors and market efficiency," *The Journal of Finance*, 64(4), 1517–1548.

Tian, Y., Y. Wang, B. Chen, and S. S. Du (2023): "Scan and snap: Understanding training dynamics and token composition in 1-layer transformer," *Advances in neural information processing systems*, 36, 71911–71947.

Tian, Y., Y. Wang, Z. Zhang, B. Chen, and S. Du (2024): "Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention," in *International Conference on Learning Representations*, vol. 2024, pp. 36365–36390.

Van Nieuwerburgh, S., and L. Veldkamp (2009): "Information immobility and the home bias puzzle," *The Journal of Finance*, 64(3), 1187–1215.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017): "Attention is all you need," *Advances in neural information processing systems*, 30.

Veldkamp, L. L. (2006): "Media frenzies in markets for financial information," *American Economic Review*, 96(3), 577–601.

Verrecchia, R. E. (1982): "Information acquisition in a noisy rational expectations economy," *Econometrica: Journal of the Econometric Society*, pp. 1415–1430.

A Proof of Section 3

The appendix is organized as follows. Appendix A.1 collects the lemmas used throughout. Appendix A.2 characterizes the endogenous dimension cutoff point. Appendix A.3 gives sufficient conditions for an interior cutoff $\bar{d}$ in terms of the pretraining frequencies and the vocabulary size.

A.1 Technical Lemmas

We follow the decoder analysis and most notation of Tian, Wang, Chen, and Du (2023) and extend their setup by allowing sequence class n with different frequency $p_n^{\mathrm{tr}} > 0$, where $\sum_{n=1}^{N} p_n^{\mathrm{tr}} = 1$. As in their analysis, we take the long sequence limit and hold the self attention matrix Z fixed while studying the decoder matrix Y. The class representations f_n and the matrix F are therefore fixed.

Define $W = F^\top Y$. The nth row of W is $w_n^\top = f_n^\top Y$. Equivalently, $w_n = Y^\top f_n$ is the vector of decoder logits produced for class n. Under the modified decoder gradient of Tian, Wang, Chen, and Du (2023), the effective logit vectors are separable across classes and measured by its local exposure. Also, every class follows the same logit dynamics.

We study the deterministic population path. At global training time t, the local exposure of class n is $\ell_n(t) = p_n^{\mathrm{tr}} t$. Let ℓ denote a generic value of this local population clock. The effective logit dynamics are

$$\frac{dw_n(\ell)}{d\ell} = \eta_Y \left[e_n - \frac{\exp\big(w_n(\ell)\big)}{\mathbf{1}^\top \exp\big(w_n(\ell)\big)} \right], \qquad w_n(0) = 0.$$

Lemma 1 (Effective logit dynamics)**.** *Define*

$$\zeta_n = \frac{M}{M-1}\left(e_n - \frac{1}{M}\mathbf{1}\right).$$

Then the effective logit vector for class n satisfies

$$w_n(\ell) = (M-1)h(\ell)\zeta_n,$$

where

$$\frac{dh(\ell)}{d\ell} = \frac{\eta_Y}{(M-1) + \exp\big(Mh(\ell)\big)}, \qquad h(0) = 0.$$

Equivalently,

$$\exp\big(Mh(\ell)\big) + (M-1)Mh(\ell) = M\eta_Y \ell + 1.$$

Proof. The vector ζ_n has entry 1 in coordinate n and entry $-1/(M-1)$ in every coordinate $j \neq n$. Therefore, if $w_n(\ell) = (M-1)h(\ell)\zeta_n$, then $w_{nn}(\ell) = (M-1)h(\ell), j = n$ and $w_{nj}(\ell) = -h(\ell), j \neq n$. It follows that

$$\mathbf{1}^\top \exp\big(w_n(\ell)\big) = \exp\big((M-1)h(\ell)\big) + (M-1)\exp\big(-h(\ell)\big).$$

Factoring out $\exp\big(-h(\ell)\big)$ gives

$$\mathbf{1}^\top \exp\big(w_n(\ell)\big) = \exp\big(-h(\ell)\big)\left[(M-1) + \exp\big(Mh(\ell)\big)\right].$$

Hence the nth entry of $e_n - \frac{\exp\big(w_n(\ell)\big)}{\mathbf{1}^\top \exp\big(w_n(\ell)\big)}$ is $\frac{M-1}{(M-1)+\exp\big(Mh(\ell)\big)}$, while every entry $j \neq n$ is $-\frac{1}{(M-1)+\exp\big(Mh(\ell)\big)}$. Using the entries of ζ_n, the entire vector can therefore be written as

$$e_n - \frac{\exp\big(w_n(\ell)\big)}{\mathbf{1}^\top \exp\big(w_n(\ell)\big)} = \frac{M-1}{(M-1)+\exp\big(Mh(\ell)\big)}\zeta_n.$$

On the other hand, differentiating $w_n(\ell) = (M-1)h(\ell)\zeta_n$ gives $\frac{dw_n(\ell)}{d\ell} = (M-1)\frac{dh(\ell)}{d\ell}\zeta_n$. Substitution into the decoder dynamics yields

$$\eta_Y \frac{M-1}{(M-1)+\exp\big(Mh(\ell)\big)}\zeta_n = (M-1)\frac{dh(\ell)}{d\ell}\zeta_n.$$

Canceling the common factor gives

$$\frac{dh(\ell)}{d\ell} = \frac{\eta_Y}{(M-1)+\exp\big(Mh(\ell)\big)}.$$

Since $w_n(0) = 0$, we have $h(0) = 0$. The decoder dynamics have a unique solution, so the row remains in the direction ζ_n for all $\ell \geq 0$. Finally,

$$\begin{aligned}\frac{d}{d\ell}\left[\exp\big(Mh(\ell)\big) + (M-1)Mh(\ell)\right] &= M\left[\exp\big(Mh(\ell)\big) + (M-1)\right]\frac{dh(\ell)}{d\ell} \\ &= M\eta_Y.\end{aligned}$$

Integrating and using $h(0) = 0$ gives

$$\exp\big(Mh(\ell)\big) + (M-1)Mh(\ell) = M\eta_Y\ell + 1.$$

□

For later use, the probability assigned to the correct next token for class n at global training time t is $\alpha_{nn}(t) = \frac{\exp\big(Mh(p_n^{\mathrm{tr}}t)\big)}{(M-1)+\exp(Mh(p_n^{\mathrm{tr}}t))}$. This expression follows by evaluating the effective logit path in Lemma 1 at the population local exposure $\ell = p_n^{\mathrm{tr}}t$.

Remark 1 (Continuous population path)**.** *The discrete decoder updates studied by Tian, Wang, Chen, and Du (2023) admit a continuous approximation with a controlled approximation error. We therefore formulate our analysis in continuous local exposure and treat global training time as continuous. All subsequent dynamics, prediction probabilities, loss functions, and stopping results are defined on this continuous population path.*

Lemma 2 (Class specific and aggregate loss)**.** *Under the effective logit dynamics in Lemma 1, the cross entropy loss of class n is*

$$\mathcal{L}_n(\ell) = \log\left(1 + (M-1)\exp\big(-Mh(\ell)\big)\right).$$

Its derivative with respect to the local clock is

$$\frac{d\mathcal{L}_n(\ell)}{d\ell} = -\frac{M(M-1)\eta_Y}{\left[(M-1)+\exp\big(Mh(\ell)\big)\right]^2} < 0.$$

At global training time t, the aggregate population loss is

$$\mathcal{L}(t) = \sum_{n=1}^{N} p_n^{\mathrm{tr}} \log\left(1 + (M-1)\exp\left[-Mh\left(p_n^{\mathrm{tr}} t\right)\right]\right).$$

Its first derivative is

$$\frac{d\mathcal{L}(t)}{dt} = -M(M-1)\eta_Y \sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1) + \exp\left(Mh\left(p_n^{\mathrm{tr}} t\right)\right)\right]^2} < 0.$$

Its second derivative is

$$\frac{d^2\mathcal{L}(t)}{dt^2} = 2M^2(M-1)\eta_Y^2 \sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^3 \exp\left(Mh\left(p_n^{\mathrm{tr}} t\right)\right)}{\left[(M-1) + \exp\left(Mh\left(p_n^{\mathrm{tr}} t\right)\right)\right]^4} > 0.$$

The class specific loss is strictly decreasing in its local clock, while the aggregate population loss is strictly decreasing and strictly convex in global training time.

Proof. By Lemma 1, the decoder row associated with class n satisfies $w_n(\ell) = (M-1)h(\ell)\zeta_n$. Its correct token entry is $w_{nn}(\ell) = (M-1)h(\ell)$, and every other entry is $w_{nj}(\ell) = -h(\ell), j \neq n$. The logit probability assigned to the correct next token is therefore

$$\frac{\exp\big((M-1)h(\ell)\big)}{\exp\big((M-1)h(\ell)\big) + (M-1)\exp\big(-h(\ell)\big)}.$$

Multiplying the numerator and denominator by $\exp\big(h(\ell)\big)$ gives and taking the negative logarithm yields

$$\begin{aligned}\mathcal{L}_n(\ell) &= -\log\left(\frac{\exp\big(Mh(\ell)\big)}{(M-1) + \exp\big(Mh(\ell)\big)}\right) \\ &= \log\left(1 + (M-1)\exp\big(-Mh(\ell)\big)\right).\end{aligned}$$

Differentiating with respect to ℓ gives

$$\frac{d\mathcal{L}_n(\ell)}{d\ell} = -\frac{M(M-1)}{(M-1) + \exp\big(Mh(\ell)\big)} \frac{dh(\ell)}{d\ell}.$$

Lemma 1 gives

$$\frac{dh(\ell)}{d\ell} = \frac{\eta_Y}{(M-1) + \exp\big(Mh(\ell)\big)}.$$

It follows that

$$\frac{d\mathcal{L}_n(\ell)}{d\ell} = -\frac{M(M-1)\eta_Y}{\big[(M-1) + \exp\big(Mh(\ell)\big)\big]^2} < 0.$$

At global training time t, the population local clock of class n is $p_n^{\mathrm{tr}} t$. Since class n enters the pretraining population with frequency p_n^{tr}, the aggregate population loss is $\mathcal{L}(t) = \sum_{n=1}^{N} p_n^{\mathrm{tr}} \mathcal{L}_n\left(p_n^{\mathrm{tr}} t\right)$. Substituting the class specific loss gives

$$\mathcal{L}(t) = \sum_{n=1}^{N} p_n^{\mathrm{tr}} \log\left(1 + (M-1)\exp\left[-Mh\left(p_n^{\mathrm{tr}} t\right)\right]\right).$$

Applying the chain rule gives

$$\frac{d\mathcal{L}(t)}{dt} = \sum_{n=1}^{N} p_n^{\mathrm{tr}} \frac{d\mathcal{L}_n(\ell)}{d\ell}\Big|_{\ell=p_n^{\mathrm{tr}}t} p_n^{\mathrm{tr}}$$
$$= -M(M-1)\eta_Y \sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)\right]^2}.$$

Since $p_n^{\mathrm{tr}} > 0$ for every $n \in [N]$, the derivative is strictly negative. Differentiating once more gives

$$\frac{d^2\mathcal{L}(t)}{dt^2} = 2M(M-1)\eta_Y \sum_{n=1}^{N} \left(p_n^{\mathrm{tr}}\right)^2$$
$$\times \frac{M\exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)\frac{d}{dt}h\left(p_n^{\mathrm{tr}}t\right)}{\left[(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)\right]^3}.$$

The local clock dynamics imply

$$\frac{d}{dt}h\left(p_n^{\mathrm{tr}}t\right) = \frac{\eta_Y p_n^{\mathrm{tr}}}{(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)}.$$

Substitution yields

$$\frac{d^2\mathcal{L}(t)}{dt^2} = 2M^2(M-1)\eta_Y^2 \sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^3 \exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)}{\left[(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}t\right)\right)\right]^4} > 0.$$

Therefore, the aggregate population loss is strictly decreasing and strictly convex in global training time.

□

A.2 Endogenous learning cutoff $\bar{d}$

This subsection derives the optimal training time and the resulting learning cutoff $\bar{d}$. The developer chooses how long to train by trading off the aggregate pretraining loss against the cost of computation. We first show that this stopping problem has a unique solution. Then we evaluate the class specific correct token probabilities at the optimal training time and define $\bar{d}$ as the number of payoff information dimensions that reach the learning threshold.

A.2.1 Optimal training time

Proposition 12 (Optimal training time)**.** *The developer chooses*

$$T^\star = \arg\min_{t\geq 0} \left\{\mathcal{L}(t) + \eta_Y ct\right\},$$

where $c > 0$ is the normalized marginal computation cost. The optimal training time is finite and unique. When $c \geq \frac{M-1}{M}\sum_{n=1}^{N}\left(p_n^{\mathrm{tr}}\right)^2$, the optimal time $T^\star = 0$.

When $0 < c < \frac{M-1}{M}\sum_{n=1}^{N}\left(p_n^{\mathrm{tr}}\right)^2$, the optimal training time is strictly positive and is uniquely characterized by

$$M(M-1)\sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}T^\star\right)\right)\right]^2} = c.$$

Proof. By Lemma 2, $\mathcal{L}(t)$ is strictly convex. Therefore, $\mathcal{L}(t)+\eta_Y ct$ is also strictly convex.

Since $\mathcal{L}(t)\geq 0$ and $c>0$, $\mathcal{L}(t)+\eta_Y ct\longrightarrow\infty$ as $t\longrightarrow\infty$. Hence, a finite minimizer exists. Strict convexity implies that this minimizer is unique.

At $t=0$, $h(0)=0$. Lemma 2 gives

$$\begin{aligned}\frac{d}{dt}\left[\mathcal{L}(t)+\eta_Y ct\right]\Big|_{t=0} &= -M(M-1)\eta_Y\sum_{n=1}^{N}\frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1)+1\right]^2}+\eta_Y c\\ &= \eta_Y\left[c-\frac{M-1}{M}\sum_{n=1}^{N}\left(p_n^{\mathrm{tr}}\right)^2\right].\end{aligned}$$

If $c\geq\frac{M-1}{M}\sum_{n=1}^{N}\left(p_n^{\mathrm{tr}}\right)^2$, the derivative of the objective is nonnegative at $t=0$. Since the objective is strictly convex, its derivative is strictly increasing. The unique minimizer is therefore $T^\star=0$.

Now suppose $0<c<\frac{M-1}{M}\sum_{n=1}^{N}\left(p_n^{\mathrm{tr}}\right)^2$. The derivative of the objective is negative at $t=0$. Since $p_n^{\mathrm{tr}}>0$ for every $n\in[N]$, the implicit logit solution implies $h\left(p_n^{\mathrm{tr}}t\right)\longrightarrow\infty$ as $t\longrightarrow\infty$. Hence, $d\mathcal{L}(t)/dt\longrightarrow 0$, and therefore $\frac{d}{dt}\left[\mathcal{L}(t)+\eta_Y ct\right]\longrightarrow\eta_Y c>0$. Furthermore, the continuity gives an interior solution $T^\star>0$ to

$$\frac{d\mathcal{L}(T^\star)}{dt}+\eta_Y c=0.$$

Strict convexity implies that this solution is unique. Using the derivative in Lemma 2 and cancelling the common factor η_Y gives

$$M(M-1)\sum_{n=1}^{N}\frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1)+\exp\left(Mh\left(p_n^{\mathrm{tr}}T^\star\right)\right)\right]^2}=c.$$

□

A.2.2 Endogenous cutoff

Theorem 1 (Endogenous learnedness cutoff)**.** *Define the local learning threshold by*

$$\ell_{1/2}=\frac{M-2+(M-1)\log(M-1)}{M\eta_Y}.$$

At the optimal training time $T^\star$, define

$$\bar{d}=\#\left\{n\in[N]\colon p_n^{\mathrm{tr}}T^\star\geq\ell_{1/2}\right\}.$$

Then $\bar{d}\in\{0,\ldots,N\}$. The logit assigns at least one half of its probability to the correct next token for exactly the first $\bar{d}$ payoff information dimensions. Thus, the learned dimensions are $\{1,\ldots,\bar{d}\}$ and the undertrained dimensions are $\{\bar{d}+1,\ldots,N\}$. The corner cases satisfy

$$\bar{d}=0\quad\Longleftrightarrow\quad p_1^{\mathrm{tr}}T^\star<\ell_{1/2}$$

and

$$\bar{d}=N\quad\Longleftrightarrow\quad p_N^{\mathrm{tr}}T^\star\geq\ell_{1/2}.$$

If $p_{\bar{d}}^{\mathrm{tr}}T^\star\geq\ell_{1/2}>p_{\bar{d}+1}^{\mathrm{tr}}T^\star$, then the cutoff is

$$1\leq\bar{d}\leq N-1.$$

Moreover, $\bar{d}$ is weakly decreasing in the normalized computation cost c.

Proof. By Lemma 1, after local exposure ℓ the probability assigned to the correct next token is $\frac{\exp\big(Mh(\ell)\big)}{(M-1)+\exp\big(Mh(\ell)\big)}$. This probability is at least one half if and only if $\exp\big(Mh(\ell)\big) \geq M-1$, or equivalently $Mh(\ell) \geq \log(M-1)$. The logit dynamics satisfy

$$\exp\big(Mh(\ell)\big) + (M-1)Mh(\ell) = M\eta_Y \ell + 1.$$

At the point where $Mh(\ell) = \log(M-1)$, this relation becomes

$$(M-1) + (M-1)\log(M-1) = M\eta_Y \ell + 1.$$

Solving for ℓ gives

$$\ell = \frac{M - 2 + (M-1)\log(M-1)}{M\eta_Y} = \ell_{1/2}.$$

Since $dh(\ell)/d\ell > 0$, the decoder assigns at least one half of its probability to the correct next token if and only if $\ell \geq \ell_{1/2}$.

At global training time $T^\star$, dimension n follows the population local clock $p_n^{\mathrm{tr}} T^\star$. Hence, dimension n is learned to the threshold if and only if $p_n^{\mathrm{tr}} T^\star \geq \ell_{1/2}$. Because $p_1^{\mathrm{tr}} T^\star > p_2^{\mathrm{tr}} T^\star > \cdots > p_N^{\mathrm{tr}} T^\star$ whenever $T^\star > 0$, the dimensions satisfying this inequality form an initial segment of $[N]$. For an interior optimal training time,

$$\mathcal{L}'(T^\star) + \eta_Y c = 0.$$

Differentiating with respect to c gives

$$\mathcal{L}''(T^\star)\frac{dT^\star}{dc} + \eta_Y = 0,$$

so

$$\frac{dT^\star}{dc} = -\frac{\eta_Y}{\mathcal{L}''(T^\star)} < 0.$$

For costs above the interior region, Proposition 12 gives $T^\star = 0$. Hence, $T^\star$ is weakly decreasing in c over the full parameter range. Every population local clock $p_n^{\mathrm{tr}} T^\star$ is therefore weakly decreasing in c, so the number of dimensions that cross $\ell_{1/2}$ can only decrease. □

A.3 Frequency based conditions for an interior $\bar{d}$

This subsection gives sufficient conditions for an nontrivial interior cutoff in terms of the pretraining frequencies and the vocabulary size.

Lemma 3 (Frequency based lower bound for the learned cutoff)**.** *Under the conditions of Theorem 1, the squared frequency mass beyond the learned cutoff satisfies*

$$\sum_{n>\bar{d}} \big(p_n^{\mathrm{tr}}\big)^2 < \frac{4(M-1)}{M} c.$$

Consequently, for every $k \in \{0, \ldots, N-1\}$,

$$\frac{M}{4(M-1)} \sum_{n>k} \big(p_n^{\mathrm{tr}}\big)^2 \geq c \quad \Longrightarrow \quad \bar{d} \geq k+1.$$

Proof. The lower bound condition states that when the squared frequency mass beyond class k is sufficiently large relative to the marginal computation cost, training continues until at least class $k+1$ enters the learned region.

We first establish the bound on the squared frequency mass beyond the learned cutoff. Consider first the case $T^\star = 0$. Proposition 12 implies

$$c \geq \frac{M-1}{M} \sum_{n=1}^{N} \left(p_n^{\mathrm{tr}}\right)^2.$$

Since M is large, the transition clock satisfies $\ell_{1/2} > 0$. Theorem 1 therefore gives $\bar{d} = 0$. Moreover, since $\frac{M-1}{M} > \frac{M}{4(M-1)}$. So it follows that

$$c > \frac{M}{4(M-1)} \sum_{n>\bar{d}} \left(p_n^{\mathrm{tr}}\right)^2.$$

Now consider the case $T^\star > 0$. The first order condition for the optimal training time gives

$$c = M(M-1) \sum_{n=1}^{N} \frac{\left(p_n^{\mathrm{tr}}\right)^2}{\left[(M-1) + \exp\left(Mh\left(p_n^{\mathrm{tr}} T^\star\right)\right)\right]^2}.$$

For every class beyond the learned cutoff, Theorem 1 gives

$$p_n^{\mathrm{tr}} T^\star < \ell_{1/2}, \qquad n > \bar{d}.$$

Since h is strictly increasing and $Mh(\ell_{1/2}) = \log(M-1)$, we have

$$\exp\left(Mh\left(p_n^{\mathrm{tr}} T^\star\right)\right) < M-1, \qquad n > \bar{d}.$$

Hence, for every $n > \bar{d}$,

$$\left[(M-1) + \exp\left(Mh\left(p_n^{\mathrm{tr}} T^\star\right)\right)\right]^2 < 4(M-1)^2.$$

Each tail class therefore contributes more than $\frac{M}{4(M-1)} \left(p_n^{\mathrm{tr}}\right)^2$ to the marginal benefit of training. Summing these contributions gives

$$\begin{aligned} c &> M(M-1) \sum_{n>\bar{d}} \frac{\left(p_n^{\mathrm{tr}}\right)^2}{4(M-1)^2} \\ &= \frac{M}{4(M-1)} \sum_{n>\bar{d}} \left(p_n^{\mathrm{tr}}\right)^2. \end{aligned}$$

When $\bar{d} = N$, the tail sum equals zero, and the same result follows from $c > 0$. Thus, in every case,

$$c > \frac{M}{4(M-1)} \sum_{n>\bar{d}} \left(p_n^{\mathrm{tr}}\right)^2.$$

We now prove the lower bound for the learned cutoff. Fix $k \in \{0, \ldots, N-1\}$ and suppose that

$$c \leq \frac{M}{4(M-1)} \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2.$$

Assume for contradiction that $\bar{d} \leq k$. The tail beyond $\bar{d}$ then contains every class beyond k. Since all class frequencies are strictly positive,

$$\sum_{n>\bar{d}} \left(p_n^{\mathrm{tr}}\right)^2 \geq \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2.$$

Combining this inequality with the tail bound gives

$$\begin{aligned} c &> \frac{M}{4(M-1)} \sum_{n>\bar{d}} \left(p_n^{\mathrm{tr}}\right)^2 \\ &\geq \frac{M}{4(M-1)} \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2 \\ &\geq c. \end{aligned}$$

This contradiction gives $\bar{d} > k$. Since $\bar{d}$ is an integer, $\bar{d} \geq k+1$.

□

Lemma 4 (Upper bound condition)**.** *For every $c > 0$, the optimal training time satisfies*

$$T^\star \leq \frac{\log M}{\eta_Y c}.$$

Consequently,

$$c > \frac{p_N^{\mathrm{tr}} \log M}{\eta_Y \ell_{1/2}} \quad \Longrightarrow \quad \bar{d} \leq N-1.$$

Proof. If $T^\star = 0$, the first inequality is immediate. Suppose therefore that $T^\star > 0$. Since $\mathcal{L}(t)$ is convex, $-\mathcal{L}'(t)$ is decreasing. Hence,

$$\begin{aligned} \mathcal{L}(0) - \mathcal{L}(T^\star) &= \int_0^{T^\star} -\mathcal{L}'(t)\, dt \\ &\geq T^\star \left[-\mathcal{L}'(T^\star)\right]. \end{aligned}$$

The interior first order condition gives $-\mathcal{L}'(T^\star) = \eta_Y c$. Moreover, $\mathcal{L}(0) = \log M$ and $\mathcal{L}(T^\star) \geq 0$. Therefore,

$$\log M \geq \mathcal{L}(0) - \mathcal{L}(T^\star) \geq T^\star \eta_Y c,$$

which gives $T^\star \leq \frac{\log M}{\eta_Y c}$. If

$$c > \frac{p_N^{\mathrm{tr}} \log M}{\eta_Y \ell_{1/2}},$$

then

$$p_N^{\mathrm{tr}} T^\star \leq \frac{p_N^{\mathrm{tr}} \log M}{\eta_Y c} < \ell_{1/2}.$$

Theorem 1 therefore implies $\bar{d} \leq N-1$. □

Corollary 11 (Primitive nondegenerate cutoff region)**.** *Fix $k \in \{0, \ldots, N-2\}$. If*

$$\frac{p_N^{\mathrm{tr}} \log M}{\eta_Y \ell_{1/2}} < c \leq \frac{M}{4(M-1)} \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2,$$

then

$$k+1 \leq \bar{d} \leq N-1.$$

The stated cost region is nonempty whenever

$$\frac{p_N^{\mathrm{tr}} \log M}{\eta_Y \ell_{1/2}} < \frac{M}{4(M-1)} \sum_{n>k} \left(p_n^{\mathrm{tr}}\right)^2 .$$

Because $\eta_Y \ell_{1/2}$ depends only on M, this condition is a direct restriction on the pretraining frequencies and the vocabulary size.

Proof. The upper bound on c and Proposition 3 give $\bar{d} \geq k+1$. The lower bound on c and Proposition 4 give $\bar{d} \leq N-1$. The cost region is nonempty whenever its lower endpoint is strictly smaller than its upper endpoint. □

We next bounds the cutoff using the final prediction performance of the trained model. The bound uses observable quantities and requires no numerical value for the cost of computation.

Corollary 12 (Performance based lower bound)**.** *Define the average probability assigned to the correct next token by*

$$\pi^\star = \sum_{n=1}^{N} p_n^{\mathrm{tr}} \alpha_{nn}(T^\star).$$

This quantity is an average correct token probability rather than a top one accuracy. Define the cumulative pretraining frequencies by

$$S_0 = 0, \qquad S_k = \sum_{n=1}^{k} p_n^{\mathrm{tr}}, \qquad k \in [N].$$

Then

$$S_{\bar{d}} > 2\pi^\star - 1.$$

Consequently,

$$\bar{d} \geq \min\left\{k \in \{0, \ldots, N\} \colon S_k > 2\pi^\star - 1\right\}.$$

In particular,

$$\pi^\star \geq \frac{1}{2} \quad \Longrightarrow \quad \bar{d} \geq 1,$$

and

$$\pi^\star \geq \frac{1 + p_1^{\mathrm{tr}}}{2} \quad \Longrightarrow \quad \bar{d} \geq 2.$$

More generally, for every $k \in \{0, \ldots, N-1\}$,

$$\pi^\star \geq \frac{1 + S_k}{2} \quad \Longrightarrow \quad \bar{d} \geq k+1.$$

Proof. By Theorem 1, $\alpha_{nn}(T^\star) < 1/2$ for every $n > \bar{d}$. Since $T^\star$ is finite, each decoder probability is also strictly below one. Therefore,

$$\begin{aligned}
\pi^\star &= \sum_{n \leq \bar{d}} p_n^{\mathrm{tr}} \alpha_{nn}(T^\star) + \sum_{n > \bar{d}} p_n^{\mathrm{tr}} \alpha_{nn}(T^\star) \\
&< \sum_{n \leq \bar{d}} p_n^{\mathrm{tr}} + \frac{1}{2} \sum_{n > \bar{d}} p_n^{\mathrm{tr}} \\
&= S_{\bar{d}} + \frac{1}{2}\left(1 - S_{\bar{d}}\right) \\
&= \frac{1 + S_{\bar{d}}}{2}.
\end{aligned}$$

Hence, $S_{\bar{d}} > 2\pi^\star - 1$.

Since S_k is strictly increasing in k and $S_N = 1$, the set

$$\{k \in \{0, \ldots, N\} \colon S_k > 2\pi^\star - 1\}$$

is nonempty. Its smallest element cannot exceed $\bar{d}$, which gives the lower bound.

If $\pi^\star \geq (1 + S_k)/2$, then $2\pi^\star - 1 \geq S_k$. No index $j \leq k$ can satisfy $S_j > 2\pi^\star - 1$. The smallest index that satisfies the inequality is therefore at least $k + 1$, which proves the final implication. The first two statements follow from $S_0 = 0$ and $S_1 = p_1^{\mathrm{tr}}$. □

B Proof of Section 4

Proof of Proposition 1. Conjecture (13) and a linear price. Conditional on either information set, $V - p(Y)$ is normal. Its conditional mean is $\sum_k a_k \widehat{v}^A_{i,k}$ for an AI-only investor and $\sum_{k \neq n} a_k \widehat{v}^A_{i,k} + a_n \widehat{v}^{S,n}_{i,n}$ for a specialist. Independence gives conditional variances (18) and (19). If a normal payoff Z has conditional mean M and variance $\Omega > 0$, then

$$\mathbb{E}[-e^{-\gamma X Z} \mid \mathcal{F}] = -\exp\{-\gamma X M + \tfrac{1}{2}\gamma^2 X^2 \Omega\}.$$

The exponent is strictly convex and has the unique minimizer $X = M/(\gamma\Omega)$. Because utility is the negative exponential of that exponent, this minimizer is the unique utility-maximizing order. This proves (20) and (21).

For every dimension, the exact law of large numbers implies

$$\int \widehat{v}^A_{i,n}\, di = \frac{q_n}{\tau_n + q_n} v_n \quad \text{within each AI-information group,}$$

and the corresponding specialist average is $(q_n + \tau_H)v_n/(\tau_n + q_n + \tau_H)$. Weighting (20) and (21) by group masses gives (22). Normal projection gives (15).

Combining $a_n = \beta_n - \lambda g_n$ with (22) shows that $g_n = a_n K_n/\gamma$ for a strictly positive finite K_n. The projection condition is equivalent to (25). Hence

$$\lambda\sigma_u^2 = \sum_n \frac{a_n^2 K_n}{\gamma\tau_n} > 0.$$

Thus $\lambda > 0$. Moreover,

$$a_n = \frac{\beta_n}{1 + \lambda K_n/\gamma} > 0, \qquad g_n = \frac{a_n K_n}{\gamma} > 0.$$

Existence follows by a cone fixed-point argument that also rules out an artificial zero-trade solution. Let $K_n(a, \lambda, x) > 0$ denote the bracket multiplying a_n/γ in (22), and make the proof-coordinate change

$$t_n = \frac{\lambda K_n(a, \lambda, x)}{\gamma}.$$

This is not an additional primitive. At any positive fixed point,

$$a_n = \frac{\beta_n}{1 + t_n}, \qquad g_n = \frac{\beta_n t_n}{\lambda(1 + t_n)}, \qquad \lambda^2\sigma_u^2 = \sum_n \frac{\beta_n^2 t_n}{\tau_n(1 + t_n)^2}.$$

Substitute these expressions into (18)–(19) and hence into K_n. With

$$S(t) = \sum_k \frac{\beta_k^2 t_k}{\tau_k(1 + t_k)^2},$$

this defines a continuous map of the nonnegative cone into itself,

$$T_n(t) = \frac{K_n(t)}{\gamma\sigma_u}\sqrt{S(t)}.$$

We verify the two bounds uniformly. Use the sup norm and write $M = \|t\|_\infty$. If $M \le 1$, the coordinate attaining M implies

$$S(t) \ge \min_k \frac{\beta_k^2}{\tau_k}\frac{M}{(1+M)^2} \ge \frac{M}{4}\min_k \frac{\beta_k^2}{\tau_k}.$$

At $t = 0$, every conditional variance is strictly positive and every K_n is strictly positive. Continuity therefore bounds all K_n away from zero on a sufficiently small cube. Consequently $\|T(t)\|_\infty$ is bounded below by a positive primitive constant times $\sqrt{M}$, and hence exceeds M on a sufficiently small nonzero sphere.

For the large-sphere bound, both (18) and every (19) are at least $S(t)$ after the substitution above. All posterior-mean slopes lie in $(0, 1)$ and group masses sum to one, so $K_n(t) \le 1/S(t)$ for every n. For $M \ge 1$, the coordinate attaining M gives

$$S(t) \ge \min_k \frac{\beta_k^2}{\tau_k}\frac{M}{(1+M)^2} \ge \frac{1}{4M}\min_k \frac{\beta_k^2}{\tau_k}.$$

It follows that

$$\|T(t)\|_\infty \le \frac{2\sqrt{M}}{\gamma\sigma_u\sqrt{\min_k \beta_k^2/\tau_k}},$$

which is smaller than M on every sufficiently large sphere. The cone compression-expansion fixed-point theorem now yields a nonzero fixed point between the two spheres. If one coordinate of a nonzero t were zero, then its image under T would be strictly positive, so every coordinate of the fixed point is positive. The displayed transformations recover finite positive (a, g, λ) satisfying the coefficient system. The converse follows by reversing the calculations. □

Lemma 5 (Uniform trading uniqueness)**.** *If* (24) *holds, the positive trading equilibrium is globally unique for every* $x \in [0, 1]^N$.

Proof. Use the proof coordinates from Proposition 1 and write

$$S(t) = \sum_k \frac{\beta_k^2 t_k}{\tau_k(1+t_k)^2}, \qquad M = \max_k t_k.$$

After substituting $a_k = \beta_k/(1+t_k)$ and $\lambda^2\sigma_u^2 = S(t)$, the conditional residual variance of any information group r is

$$\Omega^r(t) = S(t) + \sum_k \frac{\beta_k^2}{(\tau_k + q_k + \delta_k^r)(1+t_k)^2},$$

where $\delta_k^A = 0$ and $\delta_k^{S,\ell} = \tau_H \mathbf{1}_{\{k=\ell\}}$. Each coordinate of the fixed-point map is a positive weighted sum of

$$\frac{\sqrt{S(t)}}{\gamma\sigma_u\Omega^r(t)},$$

where the weights are group masses multiplied by posterior-mean slopes and sum to at most one. Consequently,

$$\Omega^r(t) \geq S(t) + \frac{1}{(1+M)^2} \sum_k \frac{\beta_k^2}{\tau_k + q_k + \tau_H}.$$

Maximizing $\sqrt{S}/(S+C)$ over $S > 0$ gives $1/(2\sqrt{C})$, so every positive fixed point satisfies

$$M \leq \frac{1+M}{2\gamma\sigma_u \sqrt{\sum_k \beta_k^2/(\tau_k + q_k + \tau_H)}}.$$

Under (24), rearrangement gives

$$M \leq \bar{t}(\Theta^{FM}). \tag{79}$$

It remains to rule out two fixed points inside this bound. Put $w_k = \log t_k$. For one information group, suppress the superscript and write

$$\frac{\sqrt{S(t)}}{\Omega(t)}, \quad S(t) = \sum_k \frac{b_k t_k}{(1+t_k)^2}, \quad \Omega(t) = S(t) + \sum_k \frac{c_k}{(1+t_k)^2},$$

where $b_k = \beta_k^2/\tau_k$ and $c_k = \beta_k^2/(\tau_k + q_k + \delta_k^r)$. Direct differentiation shows that the sign of its log derivative with respect to w_j is the sign of

$$b_j(1-t_j)\left[\sum_k \frac{c_k}{(1+t_k)^2} - S(t)\right] + 4S(t)c_j.$$

By (24) and (79),

$$t_k \leq \bar{t}(\Theta^{FM}) < \frac{\tau_k}{\tau_k + q_k + \tau_H} \leq \frac{c_k}{b_k}.$$

Thus every coordinate derivative is nonnegative. Under a proportional scaling of all t_k, the logarithmic elasticity of $S(t)$ is strictly below one. Each summand $(b_k t_k + c_k)/(1+t_k)^2$ of $\Omega(t)$ has logarithmic elasticity greater than $-1/2$ because $t_k < 1/3$. Hence every row sum of the log-Jacobian is strictly below one. A positive weighted sum across information groups preserves these two properties.

If t and t' were distinct fixed points, the line segment joining $\log t$ and $\log t'$ would remain inside (79). The mean-value inequality would imply

$$\|\log t - \log t'\|_\infty < \|\log t - \log t'\|_\infty,$$

a contradiction. The transformations in the proof of Proposition 1 are one-to-one, so (g, λ) is globally unique. □

Proof of Proposition 2. Let $Z = V - p(Y)$, $M_j = \mathbb{E}[Z \mid \mathcal{F}_i^j]$, and $\Omega_j = \mathrm{Var}(Z \mid \mathcal{F}_i^j)$ for $j \in \{A, (S,n)\}$. Proposition 1 gives the optimal order $X_j = M_j/(\gamma\Omega_j)$ and conditional optimized utility $-\exp\{-M_j^2/(2\Omega_j)\}$. The law of total variance implies that M_j is normal with variance $\mathcal{I} - \Omega_j$. Hence

$$\mathbb{E}\left[e^{-M_j^2/(2\Omega_j)}\right] = \left(1 + \frac{\mathcal{I} - \Omega_j}{\Omega_j}\right)^{-1/2} = \sqrt{\frac{\Omega_j}{\mathcal{I}}}.$$

It follows that the gross ex ante indirect utility is $\mathcal{V}^j(x) = -\sqrt{\Omega_j/\mathcal{I}}$. Definition (**??**) therefore yields $H_n(x) = (2\gamma)^{-1}\log(\Omega^A/\Omega^{S,n})$. The acquisition cost multiplies specialist expected utility by $e^{\gamma c}$. Because utility is negative, specialization is preferred exactly when $e^{\gamma c}\sqrt{\Omega^{S,n}/\mathcal{I}} < \sqrt{\Omega^A/\mathcal{I}}$, or $c < H_n(x)$. Substitution of (18)–(19) gives (28). □

Proof of Corollary 1. Equation (28) is strictly increasing in $\Omega^A - \Omega^{S,n}$ because Ω^A is common across types at a given equilibrium. Equations (19) and (29) give

$$\Omega^A - \Omega^{S,n} = \frac{\Delta_n^H}{(1+t_n)^2}.$$

This proves (30). When the two dimensions have common prior and signal precisions, all terms other than $\beta_n/(1+t_n)$ cancel. □

Proof of Corollary 2. Corollary 1 implies that the boundary value ratio is ordered by

$$\frac{\Delta_{\bar{d}+1}^H/(1+t_{\bar{d}+1})^2}{\Delta_{\bar{d}}^H/(1+t_{\bar{d}})^2}.$$

Substitute $q_{\bar{d}} = \tau_A$, $q_{\bar{d}+1} = \tau_L$, and the common prior precision. This gives (32).

For the final statement, write $B = \tau + \tau_L$. Given the uniform attenuation bound $\widetilde{t}$, it is sufficient that

$$\frac{(B+\kappa_A)(B+\kappa_A+\tau_H)}{B(B+\tau_H)} > R_{\bar{d}}.$$

The left side is strictly increasing in κ_A. Solving the associated quadratic equality for its nonnegative root gives (34). □

Proposition 13 (Equilibrium response of specialization value)**.** *Let θ vary along a differentiable trading-equilibrium branch. The specialization profile x, the precision profile q, or both may change with θ. For every type n,*

$$\frac{dH_n}{d\theta} = \frac{\rho_n}{2\gamma(1-\rho_n)}\left[2\frac{d\log a_n}{d\theta} - \frac{d\log\Omega^A}{d\theta} - \left(\frac{1}{\tau_n+q_n} + \frac{1}{\tau_n+q_n+\tau_H}\right)\frac{dq_n}{d\theta}\right], \tag{80}$$

$$\frac{d\Omega^A}{d\theta} = 2\lambda\sigma_u^2\frac{d\lambda}{d\theta} + 2\sum_{k=1}^{N}\frac{a_k}{\tau_k+q_k}\frac{da_k}{d\theta} - \sum_{k=1}^{N}\frac{a_k^2}{(\tau_k+q_k)^2}\frac{dq_k}{d\theta}, \tag{81}$$

$$\frac{da_k}{d\theta} = -g_k\frac{d\lambda}{d\theta} - \lambda\frac{dg_k}{d\theta}. \tag{82}$$

The market responses are

$$\begin{pmatrix} dg/d\theta \\ d\lambda/d\theta \end{pmatrix} = -F_{(g,\lambda)}^{-1}\left(F_q\frac{dq}{d\theta} + F_x\frac{dx}{d\theta}\right). \tag{83}$$

All derivatives are evaluated at the same equilibrium.

Proof of Proposition 13. Let $F(g,\lambda;q,x) = 0$ collect (15) and (22). On a locally unique branch, write $H_n(q,x)$ for the value in (28) after solving this trading system. Thus a, g, and λ are functions of the objects being varied.

Define

$$\rho_n \equiv \frac{a_n^2\tau_H}{(\tau_n+q_n)(\tau_n+q_n+\tau_H)\Omega^A} \in (0,1). \tag{84}$$

Then $H_n = -(2\gamma)^{-1}\log(1-\rho_n)$.

Differentiate $H_n = -(2\gamma)^{-1}\log(1-\rho_n)$. This gives

$$\frac{dH_n}{d\theta} = \frac{\rho_n}{2\gamma(1-\rho_n)}\frac{d\log\rho_n}{d\theta}.$$

Taking the logarithmic derivative of (84) yields (80). Direct differentiation of (18) and $a_k = \beta_k - \lambda g_k$ yields (81) and (82). Finally, differentiating $F(g,\lambda;q,x) = 0$ and using the nonsingularity of $F_{(g,\lambda)}$ gives (83). □

Proof of Corollary 3. Set $\theta = x_\ell$ in (80). Signal precisions do not change, so the last term is zero. The factor multiplying the bracket is strictly positive. This proves (35). All market responses are those in (83). □

Proof of Proposition 3. Under (24), the proof of Lemma 5 shows that the log-Jacobian of the fixed-point map is entrywise nonnegative with row sums below one. At a fixed t, differentiating the bracket K_n in the proof of Proposition 1 with respect to x_ℓ gives

$$\frac{\partial K_n}{\partial x_\ell} = \mu_\ell \frac{q_n}{\tau_n + q_n}\left(\frac{1}{\Omega^{S,\ell}} - \frac{1}{\Omega^A}\right) > 0, \qquad n \neq \ell,$$

and

$$\frac{\partial K_\ell}{\partial x_\ell} = \mu_\ell \left[\frac{q_\ell + \tau_H}{(\tau_\ell + q_\ell + \tau_H)\Omega^{S,\ell}} - \frac{q_\ell}{(\tau_\ell + q_\ell)\Omega^A}\right] > 0.$$

Thus the log fixed-point map shifts strictly upward in every coordinate. If its log-Jacobian is denoted by D, implicit differentiation gives

$$\frac{\partial \log t}{\partial x_\ell} = (I - D)^{-1}\frac{\partial \log T}{\partial x_\ell} > 0,$$

because $(I - D)^{-1} = \sum_{j=0}^{\infty} D^j$ is entrywise nonnegative and the vector on the right is strictly positive. Hence every t_n strictly rises with x_ℓ, and $a_n = \beta_n/(1 + t_n)$ strictly falls.

The same transformation gives

$$\lambda^2\sigma_u^2 = \sum_n \frac{\beta_n^2 t_n}{\tau_n(1 + t_n)^2}, \qquad \mathcal{I} = \sum_n \frac{\beta_n^2}{\tau_n(1 + t_n)}.$$

The first expression is strictly increasing in every t_n because $t_n < 1/3$, while the second is strictly decreasing. This proves (36). Integrating the derivatives along the segment from x to x' proves (37). □

Proof of Proposition 4. We first justify continuity when the positive trading equilibrium is globally unique. The small- and large-sphere inequalities in the proof of Proposition 1 are uniform over $x \in [0, 1]^N$: group masses range over a compact set, sum to one, and every baseline posterior-mean slope is strictly positive. Hence every positive fixed point in the proof coordinates lies in one common compact annulus bounded away from zero and infinity. If $x^j \to x$, any subsequence of the associated unique fixed points has a further convergent subsequence. Continuity of the fixed-point map makes its limit a positive fixed point at x, and uniqueness identifies that limit with the equilibrium at x. Thus the unique trading equilibrium, and therefore H, is continuous in x. The same conclusion holds for any continuous selection when trading is not globally unique.

For each x, the specialization best-response correspondence is nonempty and convex: it is $\{0\}$ below cost, $\{1\}$ above cost, and $[0, 1]$ at equality. Continuity of H makes the correspondence upper hemicontinuous. Kakutani's theorem yields a fixed point.

For global uniqueness, write $G(x) = c\mathbf{1} - H(x)$. The complementarity conditions are equivalent to the variational inequality

$$G(x)'(z - x) \geq 0 \qquad \text{for every } z \in [0, 1]^N.$$

Under (39), the symmetric part of the Jacobian of G has positive diagonal entries and is strictly diagonally dominant, and is therefore positive definite. For distinct x and y,

$$(x - y)'[G(x) - G(y)] = \int_0^1 (x - y)'\frac{\partial G(y + s(x - y))}{\partial x}(x - y)\, ds > 0.$$

If both x and y solved the variational inequality, evaluating each at the other point and adding would instead give $(x-y)'[G(x)-G(y)] \leq 0$, a contradiction. The local statement follows from the implicit-function theorem applied to the active equalities together with the strict corner inequalities. □

Proof of Proposition 5. On a branch with active set $\mathcal{A}$, the equilibrium is the square system $\mathcal{G}(g, \lambda, x_{\mathcal{A}}; \theta) = 0$. Strict corner inequalities keep all other shares locally fixed. The implicit-function theorem gives

$$\frac{d(g, \lambda, x_{\mathcal{A}})}{d\theta} = -\mathcal{G}^{-1}_{(g,\lambda,x_{\mathcal{A}})}\mathcal{G}_{\theta},$$

which is (41). Cost enters only the active equations $H_{\mathcal{A}} - c\mathbf{1} = 0$, so $\mathcal{G}_c = (0'_{N+1}, -\mathbf{1}')'$. Substitution gives (42). The remaining responses follow by differentiating their equilibrium definitions. □

Proof of Proposition 6. For a fixed share x, define $z = g^2/(\tau\sigma_u^2)$. The scalar trading equilibrium is characterized by

$$\frac{\sqrt{z}}{(1+z)\left[(1-x)\dfrac{q}{\tau+q}\Big/\left(z+\dfrac{\tau}{\tau+q}\right)+x\dfrac{q+\tau_H}{\tau+q+\tau_H}\Big/\left(z+\dfrac{\tau}{\tau+q+\tau_H}\right)\right]} = \frac{\sqrt{\tau}}{\gamma\beta\sigma_u}. \quad (85)$$

Once z is obtained,

$$g = \sqrt{\tau\sigma_u^2 z}, \qquad \lambda = \frac{\beta g}{\tau\sigma_u^2 + g^2}, \qquad a = \frac{\beta}{1+z}. \quad (86)$$

The left side of (85) is strictly increasing in z and, at a fixed z, strictly decreasing in x. Hence the unique solution $z(x)$ is strictly increasing. Moreover,

$$H(z) = \frac{1}{2\gamma}\log\left[\frac{z/\tau + 1/(\tau+q)}{z/\tau + 1/(\tau+q+\tau_H)}\right]$$

is strictly decreasing in z. Thus $H(x)$ is strictly decreasing and the scalar complementarity problem has a unique solution. Its largest feasible value is below its limit at $z = 0$, which is the right side of (43); that cost condition therefore implies $x^* = 0$.

Below the threshold, solving $H(z) = c$ gives (44). Rearranging (85) at $z = z_c$ gives

$$(1-x)\frac{q/(\tau+q)}{z_c + \tau/(\tau+q)} + x\frac{(q+\tau_H)/(\tau+q+\tau_H)}{z_c + \tau/(\tau+q+\tau_H)} = \frac{\gamma\beta\sigma_u\sqrt{z_c}}{\sqrt{\tau}(1+z_c)}.$$

Solving this linear equation for x yields the inner fraction in (45). Truncating it to $[0, 1]$ selects the appropriate corner whenever z_c lies outside the attainable interval. At an interior solution, implicit differentiation gives $dx^*/dc = 1/H'(x^*) < 0$. □

Proof of Corollary 4. Equation (28) implies

$$H_n \gtreqless c \quad \Longleftrightarrow \quad \frac{a_n^2\tau_H}{(\tau_n+q_n)(\tau_n+q_n+\tau_H)\Omega^A} \gtreqless 1 - e^{-2\gamma c}.$$

Thus (46) puts every covered type at the zero corner, while the reverse inequality at $\bar{d}+1$ puts the first uncovered type at the one corner. Strictly decreasing uncovered variance reduction makes H_n strictly decreasing on that set because Ω^A is common across types. It can therefore cross cost at most once, which gives (47) and the stated boundary treatment. □

Proof of Corollary 5. Continuity of the selected trading equilibrium makes $H_r(s)$ continuous. If it is strictly decreasing, it crosses the cost line at most once. The endpoint inequalities give the two corners, and the intermediate value theorem gives the unique interior solution in (48). Differentiating $H_r(x_r^*) = c$ while the surrounding strict corner inequalities remain slack gives (49). □

Proof of Corollary 6. Under the stated weak ranking, Corollary 1 makes H_n^0 weakly decreasing. The above-cost, equal-to-cost, and below-cost sets are therefore consecutive blocks. Complementarity sets shares to one in the first block and zero in the last, while any share is optimal in the equality block. Under strict ranking, the equality block contains at most one type, giving the stated cutoff and boundary treatment. □

Proof of Corollary 7. The bounds established above imply

$$\frac{\Omega^A - \Omega^{S,n}}{\Omega^A} \geq \frac{\Delta_n^H}{[1+\bar{t}(\Theta^{FM})]^2\left[D^A + \sum_k \beta_k^2/(4\tau_k)\right]}.$$

Because $-(2\gamma)^{-1}\log(1-r)$ is strictly increasing in r, $c < \underline{c}(q)$ makes $H_n > c$ for every type and every share vector, so complementarity requires $x = \mathbf{1}$.

For the upper bound, let $d_n = \tau_n + q_n$ and collect every component of $\Omega^{S,n}$ other than the residual variance of dimension n in $B_n \geq 0$. Then

$$e^{2\gamma H_n} = \frac{B_n + a_n^2/d_n}{B_n + a_n^2/(d_n + \tau_H)} \leq \frac{d_n + \tau_H}{d_n}.$$

Thus $c > \bar{c}(q)$ makes $H_n < c$ for every type and share vector, so complementarity requires $x = \mathbf{0}$. □

C Proofs of Sections 5

Proof of Proposition 7. Under the two strict corner structures, $x_n^{AI} = \mathbf{1}_{\{\bar{d}<n<r^{AI}\}}$ and $x_n^0 = \mathbf{1}_{\{n<r^0\}}$. Subtraction and aggregation give (55) and (56). If $r^{AI} \leq r^0$, the selective-AI active set is a subset of the no-AI active set. Their aggregate masses are equal only if the sets coincide because every μ_n is positive. If $r^{AI} > r^0$, canceling the overlap leaves newly active uncovered types $\max\{\bar{d}+1, r^0\} \leq n < r^{AI}$ and displaced covered types $n \leq \min\{\bar{d}, r^0 - 1\}$. Comparing their masses gives (57). □

Proof of Proposition 8. The two strict three-region allocations give $x_n^{\bar{d}} = \mathbf{1}_{\{\bar{d}<n<r_{\bar{d}}\}}$ and $x_n^{\bar{d}+1} = \mathbf{1}_{\{\bar{d}+1<n<r_{\bar{d}+1}\}}$, proving (58). If $r_{\bar{d}+1} \geq r_{\bar{d}}$, canceling the common active types leaves the exit of dimension $\bar{d}+1$ and entry of dimensions $r_{\bar{d}}, \ldots, r_{\bar{d}+1} - 1$. If $r_{\bar{d}+1} < r_{\bar{d}}$, it leaves the exit of dimension $\bar{d}+1$ together with dimensions $r_{\bar{d}+1}, \ldots, r_{\bar{d}} - 1$. Aggregating gives (59). With equal masses, counting the entering and exiting indices gives (60). □

Proof of Proposition 9. Fix the proof coordinates t and use the information-group notation in the proof of Lemma 5. Increasing q_j strictly lowers every group-specific conditional residual variance

$$\Omega^r(t) = S(t) + \sum_k \frac{\beta_k^2}{(\tau_k + q_k + \delta_k^r)(1+t_k)^2}.$$

For $n \neq j$, every posterior-mean slope multiplying the group terms in $K_n(t)$ is unchanged, so the decline in all denominators strictly raises $K_n(t)$. For $n = j$, this effect is reinforced

because both $q_j/(\tau_j + q_j)$ and $(q_j + \tau_H)/(\tau_j + q_j + \tau_H)$ increase. Hence the log fixed-point map shifts upward in every coordinate. Under (24), its log-Jacobian D is entrywise nonnegative with spectral radius below one, and therefore

$$\frac{\partial \log t}{\partial q_j} = (I - D)^{-1} \frac{\partial \log T}{\partial q_j} > 0.$$

The formulas in the proof of Proposition 3 then imply that every a_n and $\mathcal{I}$ falls and λ rises. Since $\mathcal{L} = 1/\lambda$, depth falls. Integrating along the precision segment proves (66). □

Proof of Corollary 8. Squaring (15) and using (61) gives

$$\lambda^2 \left(\sigma_u^2 + \sum_n \frac{g_n^2}{\tau_n} \right) = \frac{\left(\sum_n \beta_n g_n / \tau_n \right)^2}{\sigma_u^2 + \sum_n g_n^2/\tau_n} = \sum_n \frac{\beta_n^2}{\tau_n} - \mathcal{I},$$

which proves (68). Positivity of λ implies

$$\mathcal{L} = \left[\frac{\sigma_u^2 + \sum_n g_n^2/\tau_n}{\sum_n \beta_n^2/\tau_n - \mathcal{I}} \right]^{1/2}.$$

Taking the AI-to-no-AI ratio proves (69); the two sign statements follow by comparing its factors. □

Proof of Corollary 9. Proposition 9 makes the technology bracket in each decomposition strictly negative. Holding q^{AI} fixed, Proposition 3 implies that a componentwise reduction from x^0 to x^{AI} weakly raises $\mathcal{I}$ and weakly lowers λ, strictly if a component changes. Because $\mathcal{L} = 1/\lambda$, it weakly raises depth. A componentwise increase reverses these signs. Adding the two brackets proves parts (i) and (ii). Without a componentwise ordering, the specialization externality does not sign the net change, proving part (iii). □

Proof of Corollary 10. Proposition 9 signs the direct effect of the one-coordinate precision increase. If $r_{\bar{d}+1} \leq r_{\bar{d}}$, equation (58) makes the new active set a strict subset of the old one, so Proposition 3 signs the expertise-mediated brackets in the opposite direction. If $r_{\bar{d}+1} > r_{\bar{d}}$, the same equation records both exit at the coverage boundary and entry at the dark-tail boundary. The share vectors are then not coordinatewise ordered and no aggregate sign follows. The finite criterion is (75) applied to the two coverage equilibria. □

Proof of Proposition 10. Corollary 7 gives the common allocations $x = \mathbf{1}$ for $c < c_F$ and $x = \mathbf{0}$ for $c > c_0$. At either allocation, Proposition 9 compares the two precision profiles while re-solving every trading coefficient, giving $\Delta\mathcal{I} < 0$ and $\Delta\mathcal{L} < 0$.

Joint uniqueness and continuity imply that equilibrium shares, price inefficiency, and depth are continuous in cost. If $\Delta\mathcal{I}(\widehat{c}) > 0$ at an intermediate cost, the intermediate value theorem gives one zero between the low-cost region and $\widehat{c}$ and another between $\widehat{c}$ and the high-cost region. Equations (63) and (64) transfer these sign changes to the Gaussian tail measures. The same argument applies to $\Delta\mathcal{L}$. □

Proof of Proposition 11. Let

$$A = \sum_n \frac{\beta_n g_n}{\tau_n}, \qquad B = \sigma_u^2 + \sum_n \frac{g_n^2}{\tau_n}.$$

From (15) and (61), $\lambda = A/B$ and $\mathcal{I} = \sum_n \beta_n^2/\tau_n - A^2/B$. Differentiation gives

$$d\mathcal{I} = -2\lambda \sum_n \frac{\beta_n}{\tau_n}\, dg_n + 2\lambda^2 \sum_n \frac{g_n}{\tau_n}\, dg_n = -2\lambda \sum_n \frac{a_n}{\tau_n}\, dg_n,$$

which proves (76). Direct differentiation of A/B gives

$$d\lambda = \frac{1}{B} \sum_n \frac{\beta_n - 2\lambda g_n}{\tau_n} \, dg_n,$$

proving (77). Equation (78) follows by differentiating (62) and (64). Positivity of λ and every a_n proves the componentwise-loading claim. □