AI Monoculture Transform(er)s Financial Markets

Wenjing Xu * Mingzhe Zheng †

February 2026

Abstract

Transformer-based AI is rapidly becoming the default tool for processing financial information, with market participants increasingly relying on the same pretrained models. We formalize how the transformer's attention mechanism shapes market outcomes through two frictions. First, AI suppresses tail dimensions. Transformer learning endogenously concentrates on frequently observed signal dimensions and discards tail dimensions that are statistically under-identified. Second, AI introduces a non-diversifiable noise. As all AI traders share the same pretrained model, their inference applies the same rules and introduces a common component, which cannot be eliminated by the law of large numbers across traders.

We embed this inference rule in a standard financial market model. The effect of introducing AI on market efficiency is ambiguous and dimension-specific: AI improves pricing of covered dimensions but degrades pricing of the tail. AI adoption also reduces market liquidity, because specialists bear risk more efficiently than AI traders who carry unresolved tail uncertainty. Most important, tail fragility, which is the market's vulnerability to shocks in unmonitored dimensions, strictly increases with AI adoption. We then endogenize technology choice. AI adoption is non-monotonic. Covered-dimension types choose AI because forgoing its breadth across all covered dimensions is too costly. Extreme-tail types also choose AI because their dimensions are too infrequent to justify specialization. But intermediate-tail types specialize: AI is blind to their dimensions, and expertise generates sufficient rents. Better AI quality narrows the gap between AI and specialists, crowding out marginal specialists and enlarging the dark tail even as it improves covered-dimension pricing, which is the central paradox of AI monoculture.

*PhD candidate in Finance, Rotterdam School of Management, Erasmus University. Email: w.xu@rsm.nl

†PhD candidate in Financial Technology, Society Hub, HKUST (GZ). Email: 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 across the industry. Unlike earlier quantitative tools that were custom-built for narrow tasks, a pretrained transformer is a general-purpose inference engine. Many market participants use the same architectural family, trained on overlapping samples. The result is a degree of homogeneity in information processing that financial markets have not experienced before.

This homogeneity raises a fundamental question: when many traders rely on the same AI architecture to interpret signals, what happens to market quality? Standard information economics suggests that better and costless information processing improves market efficiency. But the transformer does not process all information equally. Its attention mechanism scores each input signal according to learned relevance, amplifies the high-scoring signals, and suppresses the low-scoring ones. When many traders share the same pretrained model, this selective processing generates two distinct frictions. First, the AI extracts only partial information from the available signal space. It concentrates on a learned subset of dimensions and assigns negligible weight to the rest, which refers to tail events that rarely appear in the historical data. Second, even though each trader receives independent private signals, the shared inference rule introduces a common component in their beliefs. All AI traders [1] apply the same weights. This common component cannot be eliminated by the law of large numbers across traders. The result is not only correlated information on covered dimensions, but also correlated ignorance on tail dimensions.

This paper provides the first formal analysis that connects the transformer's attention mechanism to financial market outcomes. We make three contributions. First, we show how the interaction between statistics of transformer learning and financial signal structure jointly determine which dimensions the AI covers and which it neglects, and why this selective attention introduces a non-diversifiable common noise. Second, we embed this AI inference rule in a standard financial market model and characterize its effects on market efficiency, tail fragility, volatility and liquidity. Third, we endogenize the technology adoption decision and show that AI crowds out human specialists in a non-monotone way.

To explain why the effects of AI arise entirely from model architecture, we formalize how transformers process multidimensional signals. A key feature is that the financial environment determines which signals matter. AI learns this structure

[1] Throughout the paper, we use the term "AI traders" broadly, for convenience. We refer to any investor who uses AI to partially assist information processing—for example, using a large language model to interpret 10-K, relying on AI-assisted research for writing analyst reports, or using an AI-powered search engine to gather information. This definition is intentionally wide and differs from the narrower usage in the literature, where "AI traders" typically means fully automated algorithmic trading systems.

but does not perfectly impose it.

Consider an asset whose payoff loads on N dimensions with declining economic importance. Dimensions of greater importance correspond to common events or states and account for a large share of the total payoff. Dimensions of lesser importance correspond to tail events or states and contribute little to the total payoff in probability. As a result, in any unbiased sample, high-importance dimensions occur more frequently, whereas low-importance dimensions occur less frequently. A pretraining transformer learns from a finite sample of financial data. High-importance dimensions coordinates are learned well because they appear often in the training data, while low-importance dimensions are statistically under-identified. This finite-sample learning creates a learned relevance gap: the transformer's internal ranking of dimension importance mirrors the financial benchmark at the top but becomes noisy in the tail. The gap is preserved even if the pretraining corpus is large enough. To convert learned scores into attention weights, the transformer applies softmax normalization. In high dimensions, this normalization requires scaling to avoid becoming either uniform or degenerate. The natural scaling is logarithmic in the input dimension (Chen, Lin, Polyanskiy, and Rigollet, 2025). Logarithmic scaling ensures that the learned gap survives the softmax normalization as dimension grows. This is the core bridge between financial structure and AI behavior—the sparsity of the AI's effective support is not assumed but emerges from the transformer learning.

The combination three ingredients, the financial signal structure, finite-sample pretraining, and logarithmic scaling, yield a precise tail suppression result. Theorem 1 shows that the total coefficient mass on tail dimensions is bounded by an exponential function. In a generic pretraining, this bound is quantitatively small. The AI's inference effectively depends only on some common dimensions. The tail is endogenously suppressed. On the covered block, repeated pretraining aggregates multiple weak pieces of evidence about each dimension, yielding an effective precision. This is the AI's deliverable: high precision on covered dimensions, zero on the tail. When many traders adopt the same pretrained model, they share both the high-quality processing on covered dimensions and the blind spot on the tail. This common blind spot is the source of AI monoculture.

We then embed the AI inference rule in a standard financial model with a continuum of informed traders and a market-clearing price system. Each trader receive independently private raw signals on each dimensions and is endowed with expertise on one dimension. We denote the specialist on dimension n by type n. The type n specialist can increase the precision of dimension n signal with some cost before the signal is realized. In contrast, AI traders can increase the precision of all covered dimensions by a relatively smaller magnitude than specialist.

We first take AI adoption as given and examine the trading equilibrium. We consider the signal space partitions into three regions: a *covered set* $\mathcal{C}$ where AI processes signals at enhanced precision, an *active tail* $\mathcal{T}$ where human specialists maintain expertise, and a *dark tail* $\mathcal{D}$ where no one has specialized knowledge. The

common inference rule has direct equilibrium consequences. Each AI trader receives independent private signals, but the pretrained model applies the identical transformation: the same coverage boundary, the same weights on covered dimensions, and zero weight on the tail. When demands are aggregated, the law of large numbers eliminates idiosyncratic signal noise, but the common inference structure remains. In a world of diverse specialists, each trader responds to different dimensions. Under AI monoculture, every AI trader responds to the same dimensions in the same way in aggregate. The shared processing rule is a systematic factor in aggregate demand that trader diversity cannot wash out.

This structure generates several market quality results. First, the net effects of introducing AI on market efficiency is ambiguous. We show that market efficiency has a dimension-specific structure. On covered dimensions, AI adoption can improve price informativeness. The large pool of AI traders injects high-precision signals that were previously dispersed across specialists monitoring other dimensions. On active-tail dimensions, however, informativeness is fragile. It depends entirely on the survival of the corresponding specialist type. If that specialist type exits the market, the incremental information for that dimension disappears discontinuously, which means the dimension transitions from the active tail to the dark tail. On dark-tail dimensions, no incremental information enters the price beyond what all traders observe at baseline precision. Whether AI adoption improves or degrades overall market efficiency depends on the balance between the gain on covered dimensions and the loss on the tail.

Second, AI adoption strictly increases tail fragility. We define the fragility of tail dimension k as the fraction of a shock in dimension k that is not incorporated into the price and therefore appears as an ex post return surprise. Dark-tail dimensions have the highest fragility loading because they lack the protective contribution of specialist-specific information. We prove that when a specialist type converts to AI, dark-tail fragility strictly increases for every dimension already in the dark tail. The market becomes most vulnerable precisely where the AI is most blind.

In addition, introducing AI can reduce market liquidity and increase volatility. Specialists face lower posterior variance than AI traders, because specialists resolve uncertainty on their expert dimension while AI traders bear the full unresolved tail uncertainty. Lower posterior variance means more aggressive trading and higher risk-bearing capacity per unit of mass. When a specialist converts to AI, the market loses a high-capacity trader and gains a low-capacity one. Total market capacity falls, the price impact of noise trading rises, and the market becomes less liquid. Lower market capacity amplifies the price impact of liquidity shocks. Wider pricing gaps on tail dimensions leave more fundamental variation unpriced. It forces push toward higher volatility.

We next endogenize the technology choice. Each trader exclusively decides whether to adopt AI or become a human specialist in a their endowed dimension. The decision depends on the incremental certainty-equivalent profit from specializa-

tion relative to AI, $\pi(n)$, and a fixed specialization cost c.

We show that the incremental profit function $\pi(n)$ exhibits a striking structural feature: an upward discontinuity at the AI coverage boundary $\bar{d}$. For a trader whose specialty lies within the covered set ($n \leq \bar{d}$), becoming a specialist means forfeiting the AI's high precision on the remaining $\bar{d}-1$ covered dimensions. This opportunity cost is substantial and grows with the number of covered dimensions and AI quality. For the first tail type ($n = \bar{d}+1$), no AI precision is forfeited, because the AI has zero coverage of that dimension regardless. This asymmetry makes tail specialization relatively more attractive than covered-dimension specialization, even when the tail dimension is less important in absolute terms.

Under a moderate cost condition, a unique adoption equilibrium exists with a three-region structure. All covered-dimension types choose AI: the opportunity cost of forgoing the AI's breadth is too high. Intermediate-tail types choose to specialize: the AI is blind to their dimensions, and their expertise generates sufficient rents to justify the cost. Extreme-tail types choose AI despite its blindness: their dimensions have too little payoff relevance to justify specialization. The abandoned extreme-tail dimensions constitute the dark tail—the region of the signal space where no one has expertise and the market is most fragile.

The comparative statics reveal the central paradox. Better AI quality reduces the posterior variance gap between AI and specialist, lowering the incremental value of specialization for all tail types. This crowds out marginal specialists, shrinking the active tail and expanding the dark tail. Paradoxically, improving AI quality worsens tail pricing even as it improves covered-dimension pricing. Conversely, stronger specialist expertise expands the active tail and shrinks the dark region. The policy implication is clear: investing in human expertise has a protective effect on tail dimensions that AI quality improvements cannot replicate. The monoculture that makes AI powerful on common signals is precisely what makes the market fragile to uncommon ones.

Related Literature Our paper connects to three main strands of the literature: AI and financial market quality, information-correlated trading, and the theory of transformer architectures.

AI and financial market quality. 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-driven correlated trading. 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.

Implications of transformer theory. Our formalization of the AI inference rule draws on recent theoretical advances in understanding transformer architectures. Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017) introduce the self-attention mechanism that is the foundation of all modern large language models. The key architectural feature for our purposes is that the transformer scores each input token by learned relevance and constructs its output as a relevance-weighted combination. This attention mechanism is the reason the transformer does not process all inputs equally—and the reason it can endogenously suppress tail dimensions.

Edelman, Goel, Kakade, and Zhang (2022) provide the theoretical foundation for why self-attention induces sparsity. They show that bounded-norm self-attention has an inductive bias toward sparse variable creation: the learned function depends on a small subset of input coordinates, with sample complexity that grows only logarithmically in context length. This result is central to our finite-sample relevance bound—it implies that the transformer's effective support is sparse not by assumption but as a consequence of the architecture's inductive bias interacting with the statistical structure of the training data. Zhang (2002) provides covering number bounds for regularized linear function classes that underpin the generalization analysis.

Chen, Lin, Polyanskiy, and Rigollet (2025) identify the critical scaling regime

for attention logits in long-context transformers. They show that logarithmic scaling of the attention temperature—$l_N = \rho \log N$—places the model in a nondegenerate middle phase where attention is neither uniform nor collapsed to a single token. This result is directly relevant to our Theorem 1: logarithmic scaling is what preserves the learned relevance gap under softmax normalization as the signal dimension N grows.

Our contribution on the transformer side is to connect these architectural properties to a specific financial environment. The existing ML theory tells us *that* transformers learn sparse functions and *how* attention scales. We show *what this means for financial markets*: the sparse support maps to a coverage boundary determined by financial signal structure, and the common pretrained model maps to a non-diversifiable common factor in aggregate demand.

The rest of the paper is organized as follows. Section 2 sets up the model. Section 3 develops the inference rule of generative AI, connecting the transformer's attention mechanism to the financial environment and proving the tail suppression theorem. Section 4 derives the trading equilibrium and analyzes the consequences of AI adoption for market liquidity, efficiency, volatility, and tail fragility. Section 5 endogenizes the technology adoption decision and establishes the central paradox. Section 6 concludes.

2 Model Setup

We study a financial market in which a single risky asset's payoff depends on many independent dimensions of fundamentals. A continuum of informed traders process private signals about these dimensions and trade against noise traders. Informed traders choose between two information technologies: a pretrained AI that processes signals through the transformer architecture, and human specialization in a single endowed dimension.

2.1 Assets

We start from introducing a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the real world probability measure is denoted by $\mathbb{P}$. There is a risk-free asset with gross return normalized to one and a single risky asset in aggregate supply $Q > 0$. The risky asset pays off at the terminal date. Its payoff is

$$V = \sum_{n=1}^{N} \beta_n v_n + \eta, \tag{1}$$

where $v_1, v_2, \ldots, v_N$ are N independent latent fundamental components with $v_n \sim N(0, \sigma_n^2)$, and $\eta \sim N(0, \sigma_\eta^2)$ is an idiosyncratic residual independent of all v_n. The coefficient β_n is the loading of the asset payoff on dimension n. We order the

loadings in non-increasing magnitude:

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_N \geq 0, \tag{2}$$

with at least some coefficients strictly positive and not all equal.

Each dimension n represents a distinct source of fundamental uncertainty—for example, macroeconomic conditions, industry performance, or firm-specific factors. The ordering reflects the fact that some dimensions matter more than others for the asset's payoff. High-n dimensions carry smaller loadings and correspond to events or factors that are less important in determining the asset's value. As we will show, this is central to the AI's behavior: the transformer endogenously concentrates on the dimensions with large loadings.

2.2 Traders

There is a continuum of rational traders indexed by $i \in [0, 1]$ and some noise traders. Noise traders collectively demand u units of the risky asset, where $u \sim N(0, \sigma_u^2)$ is independent of all fundamentals and signals. Their demand is exogenous and represents liquidity-motivated trading.

The remaining traders are informed. Each informed trader has constant absolute risk aversion (CARA) preferences over terminal wealth:

$$U_i(W) = -\exp(-\gamma W_i), \qquad \gamma > 0, \tag{3}$$

where γ is the absolute risk aversion, $W_i = W_0 + x_i(V - p)$ is terminal wealth, x_i is the trader's demand for the risky asset, and p is the market-clearing price. The informed traders divide into two groups, specialists and AI traders, according to the technology they adopt.

2.3 Specialists

Each informed trader is endowed with potential expertise in one dimension. We call a trader with expertise in dimension n a *type-n trader*. The mass of type-n traders is

$$\mu_n = \frac{\beta_n}{\sum_{k=1}^{N} \beta_k}, \qquad n = 1, \ldots, N, \tag{4}$$

so that $\sum_{n=1}^{N} \mu_n = 1$. Dimensions with larger payoff loadings attract a larger share of potential specialists. This captures a natural feature of financial markets: more economically important factors draw more analyst coverage, more research effort, and more trading activity.

A type-n trader who chooses to become a specialist invests in expertise on dimension n at a fixed cost $c > 0$ before signals are realized. This investment raises the precision of her signal on dimension n from the baseline τ_ε to $\kappa\tau_\varepsilon$, where $\kappa > 1$ is the expertise multiplier. On all other dimensions, the specialist observes signals

at the baseline precision τ_ε. The cost c is the same for all dimensions and captures the resources required to develop deep knowledge of a single factor.

2.4 AI Traders

A trader who adopts AI feeds her raw private signals into a pretrained transformer model. The key feature of the AI technology is that all AI traders share the same pretrained model: the same architecture, the same trained weights, and therefore the same inference rule.

Section 3 formalizes the AI's inference process in detail. The main result shows that the transformer endogenously concentrates on a subset of dimensions, the *covered set*, and assigns negligible weight to the remaining dimensions, the *tail*. The coverage boundary $\bar{d}(\varepsilon)$ is determined jointly by the financial signal structure and the statistics of the pretraining sample, not by any assumption imposed for tractability. On the covered dimensions, the AI achieves an effective signal precision τ_A that exceeds the baseline precision: $\tau_A > \tau_\varepsilon$. On tail dimensions, the AI contributes no incremental information.

Two properties of the AI technology are important for the equilibrium analysis. First, the AI offers breadth: it processes all covered dimensions simultaneously at enhanced precision, unlike a specialist who has high precision on only one dimension. Second, the AI creates a common component in traders' beliefs: because all AI traders apply the same pretrained model, they share the same coverage boundary, the same weights on covered dimensions, and the same blind spot on the tail.

2.5 Signals

Each informed trader i receives a private signal vector $\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,N})^\top \in \mathbb{R}^N$, where each component is

$$s_{i,n} = v_n + \varepsilon_{i,n}, \qquad \varepsilon_{i,n} \sim N(0, \tau_n^{-1}), \tag{5}$$

and the noise terms $\varepsilon_{i,n}$ are independent across traders i and dimensions n, and independent of all fundamentals. The signal precision τ_n depends on the trader's type and technology choice.

For a type-n specialist, the precision is

$$\tau_n = \begin{cases} \kappa\tau_\varepsilon & \text{on the expert dimension } n, \\ \tau_\varepsilon & \text{on all other dimensions } m \neq n. \end{cases} \tag{6}$$

For an AI trader, the raw signals all have baseline precision τ_ε. However, the AI's inference rule effectively processes covered dimensions at the enhanced precision. The microfoundation of the processing is built in Section 3.

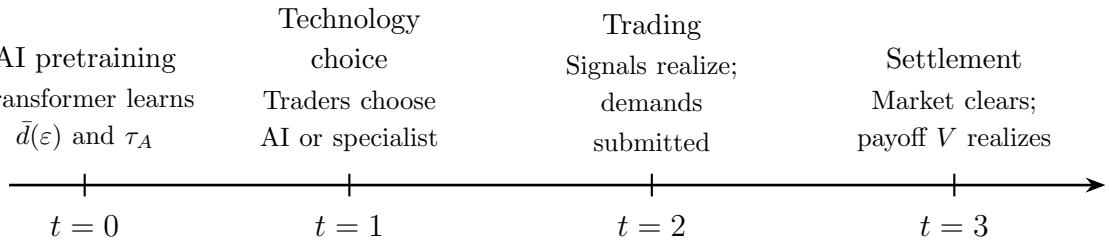


Figure 1: Timeline of the model.

2.6 Timing

- $t = 0$: The AI sets its inference rule by the transformer learning based principle.
- $t = 1$: Traders choose their inference rules, AI or human specialists, to maximize their ex ante profits.
- $t = 2$: Traders observes realized signals. They submit orders to maximize their trading profit conditional on their own beliefs.
- $t = 3$: The price system clears the market and generates an equilibrium price. The asset payoff realizes.

3 Inference Rule of Generative AI

At time $t = 0$, we have a pretrained Generative AI based on transformer learning. For AI trader, the input is the signal vector $\mathbf{s} \in \mathbb{R}^N$. The first step in the transformer learning is embedding. Here we denote embedding matrix is $\mathbf{E} \in \mathbb{R}^{d\times N}$. After embedding, we obtain $\mathbf{E}\,\mathrm{diag}(\mathbf{s}) \in \mathbb{R}^{d\times N}$. Since the transformer layer operates on a token matrix with one token in each row, we rewrite the same object as:

$$\mathbf{X}(\mathbf{s}) := (\mathbf{E}\,\mathrm{diag}(\mathbf{s}))' \in \mathbb{R}^{N\times d}.$$

The nth token is therefore as:

$$\mathbf{x}_n(\mathbf{s}) = s_n \mathbf{e}_n \in \mathbb{R}^{d\times 1}$$

where $\mathbf{e}_n$ is the n-th column of $\mathbf{E}$.

Next we describe the self attention rule. The pretrained AI has query, key, and value matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$. We use one fixed query token $\mathbf{x}_{\mathrm{cls}} \in \mathbb{R}^d$ to extract a single forecast from the signal sequence. For each signal coordinate n, the attention score is:

$$\mathrm{score}_n(\mathbf{s}) = \mathbf{x}'_{\mathrm{cls}} \mathbf{W}_Q \mathbf{W}'_K \mathbf{x}_n(\mathbf{s}) = \mathbf{x}'_{\mathrm{cls}} \mathbf{W}_Q \mathbf{W}'_K \left(s_n \mathbf{e}_n\right).$$

To form the attention weights, we apply a length-dependent scaling to the raw logits and then normalize by softmax:

$$a_n(\mathbf{s}) = \frac{\exp\big(l_N \,\mathrm{score}_n(\mathbf{s})\big)}{\sum_{m=1}^{N} \exp\big(l_N \,\mathrm{score}_m(\mathbf{s})\big)}, \qquad n = 1, \ldots, N.$$

The hidden attention output is the weighted value mixture

$$\mathbf{h}(\mathbf{s}) = \sum_{n=1}^{N} a_n(\mathbf{s}) \mathbf{W}_V' \left(s_n \mathbf{e}_n\right) \in \mathbb{R}^d.$$

Above is the standard self attention structure. First the score compares the fixed query token with each signal token. Then softmax turns these scores into attention weights. Finally the value vectors are combined by these weights.

We now move to our financial setting and link the fundamental logic of transformer learning to finance study. We firstly consider the ideal estimation case. We define the loss function as:

$$\mathcal{L} = \mathbb{E}\left[(\hat{v} - v)^2\right].$$

Then, within the family of linear predictors, the optimal coefficient on signal s_n is:

$$\beta_n^* = \frac{\beta_n \sigma_n^2}{\sigma_n^2 + \tau_\epsilon^{-1}}.$$

Under Gaussian distributions assumption, this coefficient is the natural benchmark loading on coordinate n.

To measure the economic relevance of each coordinate, define $\theta_n := (\beta_n^*)^2 \,\mathrm{Var}(s_n) = (\beta_n^*)^2(\sigma_n^2 + \tau_\epsilon^{-1})$, and assume that the coordinates are ordered so that $\theta_1 > \theta_2 > \cdots > \theta_N > 0$. This ranking is based on the property of the fundamental financial benchmark. The order depends on which coordinates matter most for prediction. This step is happened before transformer training. Given a tolerance level $\varepsilon > 0$, define the benchmark effective dimension by

$$\bar{d}(\varepsilon) := \min\left\{ d \in \{1, \ldots, N\} : \sum_{n>d} \theta_n \le \varepsilon^2 \right\}.$$

$\bar{d}(\varepsilon)$ is the smallest number of coordinates that keeps the benchmark truncation error below ε^2. Economically, coordinates $n \le \bar{d}(\varepsilon)$ form the minimally sufficient block of financially relevant signals, whereas the block $n > \bar{d}(\varepsilon)$ is the benchmark tail. Because this cutoff is intended to mark a genuine economic boundary rather than a numerically fragile threshold, we impose a local identifiable condition at the cutoff. Since the benchmark cutoff separates the coordinates that must be retained from those whose joint contribution is already negligible, it is natural to require that the boundary not be completely flat.

Assumption 1 (Local benchmark separation at the financial cutoff)**.** *We assume a strictly positive local gap at the boundary:*

$$\theta_{\bar{d}(\varepsilon)} - \theta_{\bar{d}(\varepsilon)+1} \geq \Delta_\varepsilon > 0.$$

This assumption is mild from a financial perspective. If the two coordinates at the boundary were nearly indistinguishable in benchmark relevance, then the location of the cutoff would be driven by tiny sample perturbations rather than by an economically meaningful separation between retained and negligible signals.

We next introduce finite sample pretraining from an economic perspective. For each coordinate n, let $q_n \in (0, 1]$ denote the frequency with which coordinate n appears in the pretraining corpus in a statistically identifiable and prediction relevant way. If the pretraining sample size is M, define the effective sample size of coordinate n by $M_n^{\text{eff}} := Mq_n$. We assume $q_1 \geq q_2 \geq \cdots \geq q_N$. This does not mean that tail coordinates never occur. Rather, it means that they are observed in ways that are less frequent or less statistically informative because it happens rare in financial world, so that they are harder to identify during pretraining.

To connect the financial benchmark to transformer learning, define the oracle normalized relevance of coordinate n by $\chi_n^* = h(\theta_n)$, where h is strictly increasing. Thus, in the population benchmark, the transformer's ideal relevance ranking is aligned with the financial importance ranking. Let $\hat{\chi}_n$ denote the learned normalized relevance after pretraining. At this point we impose the key bridge between finite-sample learning and transformer relevance.

Assumption 2 (Finite-sample relevance estimation)**.** *Bounded-norm self-attention learns sparse structure with logarithmic dependence on context length, so we assume that the learned normalized relevance satisfies the high-probability bound*

$$|\hat{\chi}_n - \chi_n^*| \leq C_\chi \sqrt{\frac{\log N}{Mq_n}}, \qquad n = 1, \ldots, N,$$

for some constant $C_\chi > 0$.

This assumption says that coordinates with larger effective sample size have better-estimated relevance, while tail coordinates with small Mq_n are statistically under-identified. This is the precise sense in which finite-sample pretraining suppresses the tail. Now define the oracle relevance gap at the financial cutoff by

$$\Delta_\varepsilon^\chi := \chi^*_{\bar{d}(\varepsilon)} - \chi^*_{\bar{d}(\varepsilon)+1}.$$

Because h is strictly increasing and Assumption 1 gives a positive benchmark gap at the boundary, we have $\Delta_\varepsilon^\chi > 0$. By the finite-sample relevance bound,

$$\hat{\chi}_{\bar{d}(\varepsilon)} - \hat{\chi}_{\bar{d}(\varepsilon)+1} \geq \Delta_\varepsilon^\chi - C_\chi \sqrt{\frac{\log N}{Mq_{\bar{d}(\varepsilon)}}} - C_\chi \sqrt{\frac{\log N}{Mq_{\bar{d}(\varepsilon)+1}}}.$$

Hence, if

$$C_\chi\sqrt{\frac{\log N}{Mq_{\bar{d}(\varepsilon)}}}+C_\chi\sqrt{\frac{\log N}{Mq_{\bar{d}(\varepsilon)+1}}}\le\frac{\Delta_\varepsilon^\chi}{2},$$

then

$$\hat{\chi}_{\bar{d}(\varepsilon)}-\hat{\chi}_{\bar{d}(\varepsilon)+1}\ge m_\varepsilon:=\frac{\Delta_\varepsilon^\chi}{2}>0.$$

Therefore, finite sample pretraining transfers the financial cutoff into a learned relevance gap at the same boundary.

To preserve this learned distinction through softmax in a long-context setting, we follow the logarithmic scaling logic for raw attention logits and write $l_N=\rho\log N$. Here we emphasize that this scaling is not the source of sparsity. The source of sparsity is the support selection effect. The role of $l_N=\rho\log N$ is only to ensure that this already learned relevance distinction is not washed out by normalization when the signal dimension N is large. The long-context theory shows that ρ must be chosen in the nondegenerate middle phase: if it is too small, attention becomes too diffuse; if it is too large, each token effectively attends only to itself. Thus, we work in the middle-phase regime and use logarithmic scaling as a technical device to preserve selective attention at large but finite N.

Under this scaling, define the softmax-level score gap by

$$\lambda(\mathbf{s}):=\min_{n\le\bar{d}(\varepsilon)}l_N\,\mathrm{score}_n(\mathbf{s})-\max_{n>\bar{d}(\varepsilon)}l_N\,\mathrm{score}_n(\mathbf{s}).$$

Because the learned normalized relevance gap is a parameter-level object while $\mathrm{score}_n(\mathbf{s})$ depends on the signal realization, we restrict attention to signal realizations for which the cutoff separation is preserved at the score level after replacing m_ε by a smaller positive margin, that is,

$$\min_{n\le\bar{d}(\varepsilon)}\mathrm{score}_n(\mathbf{s})-\max_{n>\bar{d}(\varepsilon)}\mathrm{score}_n(\mathbf{s})\ge m_\varepsilon.$$

Since $l_N=\rho\log N$, we then obtain

$$\lambda(\mathbf{s})\ge\rho m_\varepsilon\log N.$$

We are now in a position to state the theorem that links finite-sample pretraining to small tail coefficients.

Theorem 1 (Finite-sample relevance gap implies tail suppression)**.** *Suppose that the output layer satisfies*

$$|\hat{\beta}_n(\mathbf{s})|\le C_0\,a_n(\mathbf{s}),\qquad n=1,\dots,N,$$

for some constant $C_0>0$. If the finite-sample relevance condition above yields $\hat{\chi}_{\bar{d}(\varepsilon)}-\hat{\chi}_{\bar{d}(\varepsilon)+1}\ge m_\varepsilon>0$, and the signal realizations under consideration the cutoff separation is preserved at the score level after shrinking the margin m_ε, while the

attention scaling is $l_N = \rho \log N$ *with* ρ *chosen in the nondegenerate middle phase, then we have*

$$\lambda(\mathbf{s}) \geq \rho m_\varepsilon \log N$$

and

$$\sum_{n>\bar{d}(\varepsilon)} |\hat{\beta}_n(\mathbf{s})| \leq 2NC_0 e^{-\lambda(\mathbf{s})} \leq 2C_0 N^{\,1-\rho m_\varepsilon}.$$

In particular, for any fixed but sufficiently large N*, whenever* $\rho m_\varepsilon > 1$*, the total tail coefficient mass is quantitatively small.*

Proof. See Appendix. □

Theorem 1 is the precise link transformer learning to financial environment. The financial benchmark first determines the economically relevant cutoff $\bar{d}(\varepsilon)$. Finite sample pretraining then determines whether the transformer can statistically recover the coordinates around that cutoff through the effective sample sizes Mq_n. Once the learned relevance gap at the cutoff is positive, logarithmic scaling preserves it in the softmax layer, and the tail coefficient mass becomes small. Thus, the tail is not discarded because we assume a large λ from the outset; rather, λ itself emerges from the interaction between financial relevance and finite-sample transformer learning.

For the rest of the paper, we therefore use the following tractable benchmark representation. The AI predictor is effectively supported on the first $\bar{d}(\varepsilon)$ coordinates:

$$\hat{v}_{AI}(\mathbf{s}) = \sum_{n\leq\bar{d}(\varepsilon)} \hat{\beta}_n(\mathbf{s}) s_n + \sum_{n>\bar{d}(\varepsilon)} \hat{\beta}_n(\mathbf{s}) s_n,$$

where the second block is negligible in total coefficient mass by Theorem 1. In the exact sparse benchmark, we set it equal to zero.

Finally, to connect this support selection result to later analysis, we summarize the covered block by an effective precision. Once pretraining has identified the covered coordinates, we interpret the representation learned on those coordinates as aggregating repeated weak evidence. This motivates the following reduced-form assumption. Because pretraining repeatedly observes statistically identifiable patterns on the covered coordinates, we assume that for each $n \leq \bar{d}(\varepsilon)$, the learned representation is equivalent to aggregating K_n conditionally independent weak pieces of evidence about v_n, where each additional piece contributes precision $\tau_p > 0$. Hence the effective precision on coordinate n is $\tau_{A,n} = \tau_\epsilon + K_n \tau_p$.

Assumption 3 (Repeated aggregation on the covered block)**.** *We assume the covered coordinates are sufficiently homogeneous in representation quality, we approximate them by a common lower envelope* $K_0 \leq K_n$ *and define*

$$\tau_A := \tau_\epsilon + K_0 \tau_p \geq \tau_\epsilon.$$

Thus, τ_A provides a reduced-form summary of repeated information aggregation on the covered block. Economically, it shows that repeated aggregation makes the AI representation on the covered block more precise.

4 Consequences of AI on Market Quality

We now derive the trading equilibrium and examine how AI adoption affects market quality. Taking the technology adoption fixed, we characterize the linear equilibrium. The AI's selective coverage creates fundamental asymmetries in how different dimensions of the fundamental are reflected in the price. We analyze four dimensions of market quality, liquidity, efficiency, volatility, and tail fragility, and show that AI adoption creates an inescapable trade-off between improved pricing on covered dimensions and increased vulnerability on the tail.

4.1 Trading Equilibrium

From the analysis of Section 3 and the discussion of adoption incentives, we conjecture as given the following structure:

Assumption 4 (Adoption thresholds)**.** *There exists a threshold $\bar{N} \in \{\bar{d}+1, \ldots, N\}$ such that:*

(i) All type-n traders with $n \leq \bar{d}$ choose AI.

(ii) All type-n traders with $\bar{d} < n \leq \bar{N}$ choose to be specialists.

(iii) All type-n traders with $n > \bar{N}$ choose AI.

This structure partitions the N dimensions into three categories:

$$\mathcal{C} \equiv \{1, \ldots, \bar{d}\} \qquad (\textbf{covered dimensions}\text{: AI processes these}), \tag{7}$$

$$\mathcal{T} \equiv \{\bar{d}+1, \ldots, \bar{N}\} \qquad (\textbf{active tail}\text{: specialists cover these}), \tag{8}$$

$$\mathcal{D} \equiv \{\bar{N}+1, \ldots, N\} \qquad (\textbf{dark tail}\text{: no one has expertise}). \tag{9}$$

The AI adoption rate is endogenously determined by

$$\alpha \equiv \sum_{n \in \mathcal{C}} \mu_n + \sum_{n \in \mathcal{D}} \mu_n = 1 - \sum_{n \in \mathcal{T}} \mu_n. \tag{10}$$

The equilibrium analysis involves three signal precision levels. Recall that a type-n specialist observes dimension n at precision $\kappa\tau_\varepsilon$ and all other dimensions at the baseline precision τ_ε, while the AI achieves effective precision τ_A on covered dimensions and zero incremental precision on tail dimensions. These three levels satisfy the ordering

$$\tau_\varepsilon \leq \tau_A \leq \kappa\tau_\varepsilon, \qquad \kappa > 1. \tag{11}$$

Assumption 5 (AI precision from pretraining)**.** *The AI's pretraining endows it with effective signal precision $\tau_A \geq \tau_\varepsilon$ on each covered dimension $n \in \mathcal{C}$, and zero precision on tail dimensions $n > \bar{d}$. We parameterize the AI's quality by τ_A.*

For each precision level, the standard Bayesian updating formula $\rho = \tau/(\sigma_k^{-2} + \tau)$ yields a shrinkage factor measuring the weight placed on the signal relative to the prior, and a corresponding residual variance $\delta_k = \sigma_k^2(1-\rho_k)$ measuring remaining uncertainty:

$$\rho_k^H = \frac{\kappa\tau_\varepsilon}{\sigma_k^{-2} + \kappa\tau_\varepsilon}, \qquad \delta_k^H = \sigma_k^2(1-\rho_k^H), \qquad \text{(specialist expert precision)} \tag{12}$$

$$\rho_k^A = \frac{\tau_A}{\sigma_k^{-2} + \tau_A}, \qquad \delta_k^A = \sigma_k^2(1-\rho_k^A), \qquad \text{(AI precision on covered dims.)} \tag{13}$$

$$\rho_k^L = \frac{\tau_\varepsilon}{\sigma_k^{-2} + \tau_\varepsilon}, \qquad \delta_k^L = \sigma_k^2(1-\rho_k^L), \qquad \text{(baseline precision)} \tag{14}$$

The precision ordering $\tau_\varepsilon \le \tau_A \le \kappa\tau_\varepsilon$ implies

$$\rho_k^L \le \rho_k^A \le \rho_k^H, \qquad \delta_k^H \le \delta_k^A \le \delta_k^L, \qquad \forall k. \tag{15}$$

Higher precision means a larger shrinkage factor (the posterior leans more on the signal) and a smaller residual variance (less remaining uncertainty).

We now derive the posterior beliefs of each trader type. A type-n specialist observes N signals $s_{i,k} = v_k + \varepsilon_{i,k}$ for all k, with precision $\kappa\tau_\varepsilon$ on dimension n and τ_ε on all others. Because the fundamental components v_k are independent, the specialist updates each dimension separately. Her posterior mean of the asset payoff is

$$\mathbb{E}_i^{(n)}[V] = \beta_n \rho_n^H s_{i,n} + \sum_{k\neq n} \beta_k \rho_k^L s_{i,k}. \tag{16}$$

where the first term reflects the high-precision expert signal and the remaining terms reflect baseline-precision observations. The corresponding posterior variance is

$$\mathcal{V}^{(n)} = \beta_n^2 \delta_n^H + \sum_{k\neq n} \beta_k^2 \delta_k^L + \sigma_\eta^2. \tag{17}$$

To express this compactly, define the baseline posterior variance, the variance when all signals are observed at low precision, as

$$\mathcal{V}^L \equiv \sum_{k=1}^{N} \beta_k^2 \delta_k^L + \sigma_\eta^2. \tag{18}$$

Then the specialist's posterior variance can be written as

$$\mathcal{V}^{(n)} = \mathcal{V}^L - \Delta_n, \qquad \Delta_n \equiv \beta_n^2(\delta_n^L - \delta_n^H) > 0. \tag{19}$$

The expertise benefit Δ_n is proportional to β_n^2: dimensions that matter more for the asset payoff yield greater variance reduction from specialization.

The AI trader feeds her raw signals into the pretrained transformer. Under the hard-attention limit established in Theorem 1, the transformer concentrates on the covered set and produces the scalar output

$$z_j = \sum_{n\in\mathcal{C}} w_n s_{j,n}^A, \qquad w_n = \beta_n \rho_n^A, \tag{20}$$

where $s^A_{j,n} = v_n + \varepsilon^A_{j,n}$ with $\varepsilon^A_{j,n} \sim N(0, \tau_A^{-1})$, and $w_n = 0$ for all $n > \bar{d}$. The weight $w_n = \beta_n \rho^A_n$ is the MSE-optimal linear predictor coefficient: $\mathrm{Cov}(V, s_n)/\mathrm{Var}(s_n) = \beta_n \sigma_n^2/(\sigma_n^2 + \tau_A^{-1}) = \beta_n \rho^A_n$. The hard-attention transformer implements this optimum on covered dimensions and assigns zero weight to the tail. A direct calculation shows that $\mathrm{Cov}(V, z_j) = \mathrm{Var}(z_j)$, so the Bayesian loading of V on z_j equals unity:

$$\phi = 1. \tag{21}$$

The AI trader's posterior mean is therefore

$$\mathbb{E}^{(A)}_j[V|z_j] = z_j = \sum_{n \in \mathcal{C}} \beta_n \rho^A_n s^A_{j,n}, \tag{22}$$

and the posterior variance is

$$\mathcal{V}^{(A)} = \sum_{n \in \mathcal{C}} \beta_n^2 \delta^A_n + \sum_{n > \bar{d}} \beta_n^2 \sigma_n^2 + \sigma_\eta^2. \tag{23}$$

The first sum reflects the residual uncertainty on covered dimensions after AI processing. The second sum reflects the full prior variance on tail dimensions, where the AI provides no information. To express this relative to the baseline, define

$$\Theta_A \equiv \sum_{n \in \mathcal{C}} \beta_n^2 (\delta^A_n - \delta^L_n) + \sum_{n > \bar{d}} \beta_n^2 (\sigma_n^2 - \delta^L_n), \tag{24}$$

so that

$$\mathcal{V}^{(A)} = \mathcal{V}^L + \Theta_A. \tag{25}$$

The first sum in Θ_A is negative: on covered dimensions, AI reduces uncertainty below the baseline ($\delta^A_n \leq \delta^L_n$). The second sum is positive: on tail dimensions, AI retains the full prior variance rather than the partially reduced baseline ($\sigma_n^2 > \delta^L_n$). The sign of Θ_A determines whether AI traders face more or less total uncertainty than a trader with baseline precision on all dimensions.

Assumption 6. *The AI's tail blindness cost exceeds its covered-dimension precision gain. In terms of primitives, this is equivalent to:*

$$\sum_{n > \bar{d}} \beta_n^2 \sigma_n^2 \rho^L_n > \sum_{n \leq \bar{d}} \beta_n^2 \sigma_n^2 (\rho^A_n - \rho^L_n). \tag{26}$$

The left side is the total payoff-weighted variance of tail dimensions that the AI fails to process. The right side is the incremental variance reduction that AI achieves on covered dimensions relative to baseline precision. This condition holds when tail dimensions have substantial payoff relevance, when the number of tail dimensions is large, or when AI's precision advantage on covered dimensions ($\rho^A_n - \rho^L_n$) is modest. When $\tau_A = \tau_\varepsilon$ (AI has no pretraining advantage), the right side is zero and the condition holds trivially.

Under Assumptions 5 and 6, the posterior variances satisfy $\mathcal{V}^{(n)} = \mathcal{V}^L - \Delta_n < \mathcal{V}^L < \mathcal{V}^{(A)}$ for all specialist types n. Every specialist has strictly lower posterior variance than any AI trader. The reason is intuitive: specialists resolve uncertainty on their expert dimension while AI traders bear the full unresolved tail uncertainty. This variance comparison is central to the equilibrium analysis, because posterior variance determines both the aggressiveness of trading and the risk-bearing capacity of each trader type.

4.2 Optimal demands and market clearing

Under CARA-Normal preferences, trader i of type t demands

$$x_i^{(t)} = \frac{\mathbb{E}_i^{(t)}[V] - p}{\gamma \mathcal{V}^{(t)}}. \tag{27}$$

which is the standard result from maximizing $\mathbb{E}_i[-\exp(-\gamma(W_0 + x(V-p)))]$ under a Gaussian posterior (see Appendix A.3). For a type-n specialist, this gives

$$x_i^{(n)} = \frac{1}{\gamma \mathcal{V}^{(n)}} \left[\beta_n \rho_n^H s_{i,n} + \sum_{k \neq n} \beta_k \rho_k^L s_{i,k} - p \right], \tag{28}$$

while for an AI trader,

$$x_j^{(A)} = \frac{1}{\gamma \mathcal{V}^{(A)}} \left[\sum_{n \in \mathcal{C}} \beta_n \rho_n^A s_{j,n}^A - p \right]. \tag{29}$$

Aggregating across the continuum of each trader type, the law of large numbers eliminates idiosyncratic signal noise: for each dimension k, $\int \varepsilon_{i,k}\, di \to 0$, so the average signal converges to the true fundamental v_k. The aggregate demand of type-n specialists (mass μ_n, $n \in \mathcal{T}$) is

$$D^{(n)} = r^{(n)} \left[\beta_n \rho_n^H v_n + \sum_{k \neq n} \beta_k \rho_k^L v_k - p \right], \tag{30}$$

where

$$r^{(n)} \equiv \frac{\mu_n}{\gamma \mathcal{V}^{(n)}} = \frac{\mu_n}{\gamma(\mathcal{V}^L - \Delta_n)}, \qquad n \in \mathcal{T}, \tag{31}$$

is the risk-adjusted mass of type-n specialists—their population mass deflated by the risk cost $\gamma \mathcal{V}^{(n)}$. Lower posterior variance means less uncertainty, more aggressive trading, and higher effective capacity per unit of mass. Similarly, the aggregate demand of AI traders (mass α) is

$$D^{(A)} = r^{(A)} \left[\sum_{n \in \mathcal{C}} \beta_n \rho_n^A v_n - p \right], \tag{32}$$

where

$$r^{(A)} \equiv \frac{\alpha}{\gamma \mathcal{V}^{(A)}} = \frac{\alpha}{\gamma(\mathcal{V}^L + \Theta_A)}. \tag{33}$$

Since $\mathcal{V}^{(n)} < \mathcal{V}^{(A)}$, each specialist has higher per-capita risk-bearing capacity than each AI trader. The fundamental asymmetry in the aggregate demands is also apparent: the AI aggregate demand loads on v_k only for covered dimensions $k \in \mathcal{C}$, at the AI-precision weight ρ_k^A. For all tail dimensions ($k > \bar{d}$), the AI contributes zero demand response to v_k. This asymmetry is the source of the market quality results that follow.

The market clears when total demand equals aggregate supply:

$$\sum_{n \in \mathcal{T}} D^{(n)} + D^{(A)} + u = Q. \tag{34}$$

Define the total specialist capacity $R^S \equiv \sum_{n \in \mathcal{T}} r^{(n)}$ and the total market capacity

$$\mathcal{R} \equiv R^S + r^{(A)}. \tag{35}$$

Conjecturing a linear equilibrium price $p = \bar{p} + \sum_{k=1}^{N} a_k v_k + a_u u$ and substituting the aggregate demands into the market clearing condition (34), we match coefficients on each random variable to obtain the following closed-form result.

Proposition 1 (Equilibrium price)**.** *Under Assumptions 4–6, there exists a unique linear equilibrium $p = \bar{p} + \sum_k a_k v_k + a_u u$ with:*

$$a_k = \frac{r^{(A)} \beta_k \rho_k^A + \beta_k \rho_k^L R^S}{\mathcal{R}}, \qquad k \in \mathcal{C} \text{ (covered)}, \tag{36}$$

$$a_k = \frac{r^{(k)} \beta_k (\rho_k^H - \rho_k^L) + \beta_k \rho_k^L R^S}{\mathcal{R}}, \qquad k \in \mathcal{T} \text{ (active tail)}, \tag{37}$$

$$a_k = \frac{\beta_k \rho_k^L R^S}{\mathcal{R}}, \qquad k \in \mathcal{D} \text{ (dark tail)}, \tag{38}$$

$$a_u = \frac{1}{\mathcal{R}}, \tag{39}$$

$$\bar{p} = -\frac{Q}{\mathcal{R}}. \tag{40}$$

Proof. See Appendix A.2. □

All three cases share the form $a_k = \mathcal{I}_k / \mathcal{R}$, where the information supply $\mathcal{I}_k$ decomposes as

$$\mathcal{I}_k = \underbrace{\beta_k \rho_k^L R^S}_{\text{baseline (all specialists)}} + \underbrace{\mathcal{E}_k}_{\text{incremental information}}, \qquad \mathcal{E}_k = \begin{cases} r^{(A)} \beta_k \rho_k^A & k \in \mathcal{C}, \\ r^{(k)} \beta_k (\rho_k^H - \rho_k^L) & k \in \mathcal{T}, \\ 0 & k \in \mathcal{D}. \end{cases} \tag{41}$$

The equilibrium price p aggregates dispersed private information through the trading process, and the coefficient a_k measures how strongly the price responds to dimension k of the fundamental. A higher a_k means the price is more informative about v_k. The decomposition (41) reveals three distinct regimes of price informativeness, each reflecting a different source of information.

The first component, $\beta_k \rho_k^L R^S / \mathcal{R}$, appears in all three cases and captures the incidental information that active-tail specialists contribute about every dimension. This information does not arise from expertise but from the low-precision signals that specialists observe as a byproduct of monitoring their own specialty. It provides a floor for price informativeness across all dimensions, proportional to the total specialist capacity R^S.

On covered dimensions ($\mathcal{C}$), AI traders inject additional information with their higher precision $\rho_k^A > \rho_k^L$, weighted by their risk-adjusted mass $r^{(A)}$. The AI contribution is uniform across all covered dimensions, because the transformer processes them all simultaneously through the scalar output z_j. Because ρ_k^A can be substantially larger than ρ_k^L, covered dimensions are typically the best-priced category.

On active-tail dimensions ($\mathcal{T}$), the incremental information comes from specialist expertise: type-k specialists observe v_k at precision $\kappa\tau_\varepsilon$, contributing the edge $\rho_k^H - \rho_k^L > 0$ above the baseline. Unlike AI's broad but uniform coverage, this contribution is dimension-specific—only type-k specialists provide it. The magnitude is controlled by $r^{(k)} = \mu_k/(\gamma \mathcal{V}^{(k)})$, which is small when type k has low mass or high posterior variance. This makes active-tail dimensions more fragile than covered dimensions: their informativeness depends on a single specialist type rather than a large pool of AI traders.

On dark-tail dimensions ($\mathcal{D}$), no incremental information exists. The price coefficient $a_k = \beta_k \rho_k^L R^S / \mathcal{R}$ reflects only the baseline layer. These dimensions are the least well-priced: no trader has expertise, and AI is blind to them. Any realization of v_k in these dimensions is largely a surprise to the market.

The coefficient $a_u = 1/\mathcal{R}$ is the price impact of noise trading, the analogue of Kyle's lambda. Higher total capacity $\mathcal{R}$ means more traders willing to absorb order flow, reducing price impact. The risk premium $|\bar{p}| = Q/\mathcal{R}$ compensates traders for holding the risky asset in positive supply and is decreasing in $\mathcal{R}$ because greater capacity implies more willing risk-bearers. Increasing AI adoption has asymmetric effects on these objects. It reduces R^S (fewer specialists), which lowers the baseline information layer across all dimensions and shrinks the active-tail region. At the same time, it increases $r^{(A)}$ (more AI traders), which can boost the incremental information for covered dimensions. The net effect creates the central tension of AI monoculture: AI improves pricing of the dimensions it covers, at the cost of degrading—and potentially eliminating—pricing of tail dimensions.

4.3 AI Adoption and Market Quality

The effect of AI adoption can be understood by examining several measures of market quality. Here we focus on four common questions: how AI adoption affects market liquidity, market efficiency, return volatility, and tail fragility.

4.3.1 Market liquidity

Market liquidity refers to the market's ability to absorb trades without generating large price movements. In our equilibrium, the coefficient a_u measures the price impact of noise trading: a smaller a_u means that liquidity trades are absorbed with a smaller price change, so the market is deeper and more liquid. We therefore define

$$\text{Liquidity} \equiv \frac{1}{a_u} = \mathcal{R}. \tag{42}$$

AI adoption affects liquidity through total market capacity. Consider a marginal increase in AI adoption that reclassifies a type-n specialist ($n \in \mathcal{T}$) of mass μ_n from specialist to AI, holding fixed the AI technology $(\bar{d}, \tau_A)$. Since the AI posterior variance $\mathcal{V}^{(A)}$ depends only on the covered set $\mathcal{C} = \{1, \ldots, \bar{d}\}$ and is invariant to the boundary between $\mathcal{T}$ and $\mathcal{D}$, this reallocation changes only the masses, not individual posterior variances. The change in total market capacity is

$$\mathcal{R} = r^{(A)} + \sum_{n \in \mathcal{T}} r^{(n)} = \frac{\alpha}{\gamma \mathcal{V}^{(A)}} + \sum_{n \in \mathcal{T}} \frac{\mu_n}{\gamma \mathcal{V}^{(n)}}.$$

Consider a marginal increase in AI adoption that reclassifies a type $n \in \mathcal{T}$ trader of mass μ_n from specialist to AI, holding fixed the AI technology $\left(\bar{d}, \tau_A\right)$. Since the Al posterior variance $\mathcal{V}^{(A)}$ depends only on the covered set $\mathcal{C} = \{1, \ldots, \bar{d}\}$ and is invariant to the boundary between $\mathcal{T}$ and $\mathcal{D}$, this reallocation changes only the masses, not individual posterior variances. Hence

$$\Delta \mathcal{R} = \frac{\mu_n}{\gamma \mathcal{V}^{(A)}} - \frac{\mu_n}{\gamma \mathcal{V}^{(n)}} < 0.$$

because $\mathcal{V}^{(A)} > \mathcal{V}^{(n)}$. Greater AI adoption therefore lowers total risk-bearing capacity, raises the price impact coefficient $a_u = 1/\mathcal{R}$, and reduces market liquidity. The economic intuition is immediate: when a specialist type is replaced by AI, the market loses high-capacity specialist capital, agents with low posterior variance who trade aggressively, and gains low-capacity AI capital. The market becomes less able to absorb liquidity shocks.

4.3.2 Market efficiency

Market efficiency concerns how informative the equilibrium price is about the asset payoff. Following the standard definition, we measure overall market efficiency by the posterior precision of the payoff conditional on price:

$$\text{Market efficiency} \equiv \frac{1}{\text{Var}(V \mid p)}. \tag{43}$$

Because (V, p) are jointly normal,

$$\mathrm{Var}(V \mid p) = \mathrm{Var}(V) - \frac{\mathrm{Cov}(V,p)^2}{\mathrm{Var}(p)}. \tag{44}$$

Using the equilibrium price function, we obtain

$$\mathrm{Var}(V) = \sum_{k=1}^{N} \beta_k^2 \sigma_k^2 + \sigma_\eta^2, \tag{45}$$

$$\mathrm{Cov}(V,p) = \sum_{k=1}^{N} \beta_k a_k \sigma_k^2, \tag{46}$$

$$\mathrm{Var}(p) = \sum_{k=1}^{N} a_k^2 \sigma_k^2 + a_u^2 \sigma_u^2, \qquad \sigma_u^2 \equiv \mathrm{Var}(u). \tag{47}$$

Therefore,

$$\text{Market efficiency} = \left(\sum_{k=1}^{N} \beta_k^2 \sigma_k^2 + \sigma_\eta^2 - \frac{\left(\sum_{k=1}^{N} \beta_k a_k \sigma_k^2\right)^2}{\sum_{k=1}^{N} a_k^2 \sigma_k^2 + a_u^2 \sigma_u^2} \right)^{-1}. \tag{48}$$

In our multidimensional environment, it is also useful to examine dimension-specific price informativeness. For each dimension k,

$$\text{Market efficiency}_k \equiv \frac{1}{\mathrm{Var}(v_k \mid p)} = \left(\sigma_k^2 - \frac{a_k^2 \sigma_k^4}{\mathrm{Var}(p)} \right)^{-1}. \tag{49}$$

The collection $\{a_k\}_{k=1}^N$ therefore provides a direct summary of which dimensions are well priced. The effect of AI adoption on market efficiency is nuanced and dimension-specific. For covered dimensions $k \in \mathcal{C}$, a larger AI sector raises the mass of traders who process these dimensions at enhanced precision, which can improve their price informativeness. For active-tail dimensions $k \in \mathcal{T}$, however, informativeness is fragile because it depends on the survival of the corresponding specialist type k. Once type k exits and the dimension moves from $\mathcal{T}$ to $\mathcal{D}$, the incremental term $r^{(k)}\beta_k(\rho_k^H - \rho_k^L)$ disappears discontinuously—a discrete loss of information that cannot be smoothly replaced. For dark-tail dimensions $k \in \mathcal{D}$, no direct information processing exists at all. Whether AI adoption improves or degrades overall market efficiency therefore depends on the balance between the gain on covered dimensions and the loss on the tail.

4.3.3 Return volatility

Return volatility captures the magnitude of ex post pricing errors. In our model, the return on the risky asset is $V - p$, so

$$\text{Return volatility} \equiv \sqrt{\mathrm{Var}(V-p)}. \tag{50}$$

Using the equilibrium price function, $V - p = -\bar{p} + \sum_{k=1}^{N}(\beta_k - a_k)v_k - a_u u + \eta$, and therefore

$$\mathrm{Var}(V - p) = \sum_{k=1}^{N}(\beta_k - a_k)^2\sigma_k^2 + a_u^2\sigma_u^2 + \sigma_\eta^2. \tag{51}$$

Return volatility is low when the price closely tracks the payoff, when each coefficient a_k is close to the corresponding payoff loading β_k, and when the price impact of liquidity shocks a_u is small. AI adoption affects return volatility through two channels. First, by lowering total market capacity, it raises a_u, so liquidity shocks generate larger price fluctuations. Second, by weakening or eliminating specialist coverage in the tail, it typically increases the pricing gap $|\beta_k - a_k|$ for tail dimensions. Both forces push toward higher return volatility. At the same time, if AI adoption substantially improves the pricing of covered dimensions, volatility may fall on that subset. The net effect depends on the relative magnitudes.

4.3.4 Tail fragility

AI monoculture is particularly costly when a large shock hits a dimension outside the AI's covered set. To formalize this idea, we study how much of a shock in dimension k is left out of the price ex ante and therefore appears as an ex post return surprise. Define the centered return surprise

$$\widetilde{R} \equiv V - p + \bar{p} = \sum_{m=1}^{N}(\beta_m - a_m)v_m - a_u u + \eta. \tag{52}$$

where the constant term $\bar{p}$ is removed because it reflects the average risk premium rather than the market response to realized shocks. For each dimension k, the fragility loading

$$\mathcal{F}_k \equiv \beta_k - a_k \tag{53}$$

measures the fraction of a shock in dimension k that is not incorporated into the price: $\partial\widetilde{R}/\partial v_k = \mathcal{F}_k$. We call $\mathcal{F}_k$ the **fragility loading** of dimension k. A larger $\mathcal{F}_k$ means that a larger fraction of the shock in dimension k is not incorporated into the price before settlement, and instead shows up as an ex post return surprise.

For a tail dimension $k > \bar{d}$, the fragility loading depends on whether the dimension retains specialist coverage. From the equilibrium price coefficients (37) and (38):

$$\mathcal{F}_k^{\mathcal{T}} = \beta_k\left[1 - \frac{\rho_k^L R^S + r^{(k)}(\rho_k^H - \rho_k^L)}{\mathcal{R}}\right], \qquad k \in \mathcal{T}, \tag{54}$$

$$\mathcal{F}_k^{\mathcal{D}} = \beta_k\left[1 - \frac{\rho_k^L R^S}{\mathcal{R}}\right], \qquad k \in \mathcal{D}. \tag{55}$$

The only difference between (54) and (55) is the positive specialist-specific term

$$r^{(k)}(\rho_k^H - \rho_k^L) > 0.$$

Specialist coverage reduces the amount of a tail shock that remains unpriced.

Proposition 2 (Tail events and return surprises)**.** *Fix a dimension $k > \bar{d}$ and define the tail event*

$$\mathcal{E}_k(x) \equiv \{|v_k| \geq x\}, \qquad x > 0. \tag{56}$$

Then:

(i) The conditional expected return surprise generated by dimension k is

$$\mathbb{E}[\widetilde{R} \mid v_k] = \mathcal{F}_k v_k. \tag{57}$$

(ii) The conditional mean-squared return surprise under the tail event is

$$\mathbb{E}\Big[\big(\mathbb{E}[\widetilde{R} \mid v_k]\big)^2 \,\Big|\, \mathcal{E}_k(x)\Big] = \mathcal{F}_k^2\, \mathbb{E}\big[v_k^2 \mid \mathcal{E}_k(x)\big]\,. \tag{58}$$

In particular, it is increasing in both the fragility loading $\mathcal{F}_k$ and the tail threshold x.

Proof. See Appendix. □

Proposition 2 shows that the fragility loading $\mathcal{F}_k$ is the exact amplification coefficient from a shock in dimension k to the expected return surprise. A dimension is fragile when its shock is only weakly incorporated into the price, so that a large realization of v_k generates a large ex post correction. For active-tail dimensions, the specialist-specific term $r^{(k)}(\rho_k^H - \rho_k^L)$ lowers $\mathcal{F}_k^T$ by supplying dimension-specific information to the price. Dark-tail dimensions lack this protection entirely.

Proposition 3 (AI adoption and dark-tail fragility)**.** *Hold fixed the AI technology $(\bar{d}, \tau_A)$, and consider a marginal increase in AI adoption that reclassifies a specialist type $n \in \mathcal{T}$ of mass μ_n as AI. Then, for every dimension already in the dark tail, dark-tail fragility strictly increases.*

Proof. See Appendix. □

Greater AI adoption makes the market more vulnerable to tail events through two reinforcing channels. First, it shrinks specialist coverage, so fewer tail dimensions receive dimension-specific information. Second, it lowers the specialist share of total market capacity, which increases dark-tail fragility even for dimensions that were already poorly covered. Tail events are therefore especially disruptive when they occur in dimensions that are both outside the AI's covered set and have lost their specialist protection. In this sense, AI monoculture makes the market most fragile precisely where it is most blind.

5 Endogenous AI Adoption

The analysis in Section 4 takes the three-region partition $(\mathcal{C}, \mathcal{T}, \mathcal{D})$ as given. We now endogenize the technology choice. Each trader compares the expected utility from adopting AI against the expected utility from investing in specialist expertise, and the equilibrium partition emerges from the aggregation of these individual decisions. The key finding is that AI adoption is non-monotone in the dimension index: both the most important and the least important dimensions are served by AI, while an intermediate range of tail dimensions sustains specialist coverage. This pattern arises because the opportunity cost of forsaking AI differs sharply between covered and tail dimensions, creating an upward discontinuity in the value of specialization at the AI's attention boundary.

5.1 Trading Profits and the Value of AI

To determine which traders adopt AI and which specialize, we must first characterize the expected trading profits under each technology. The profit of a trader depends on how much her private information differs from what the equilibrium price already reveals—the larger this gap, the greater the trading opportunity she can exploit.

Lemma 1 (Expected trading profit under CARA-Normal)**.** *The ex-ante expected trading profit of trader i of type t is:*

$$\Pi^{(t)} \equiv \mathbb{E}[x_i^{(t)}(V-p)] = \frac{1}{\gamma \mathcal{V}^{(t)}} \cdot \mathrm{Cov}\Big(\mathbb{E}_i^{(t)}[V],\ V-p\Big). \tag{59}$$

Proof. See Appendix. □

Expression (88) shows that trading profit is the ratio of the "forecast-price gap" variance to the risk cost $\gamma \mathcal{V}^{(t)}$. Two forces shape this profit. The numerator captures the informativeness of the trader's signal relative to the price: the more the trader knows that the price does not yet reflect, the larger the expected squared forecast-price gap and the greater the trading opportunity. The denominator captures the risk cost: traders with lower posterior variance trade more aggressively and exploit a given information edge more efficiently.

To evaluate the forecast-price gap variance, substitute the equilibrium price into the specialist's posterior mean and take second moments. For a type-n specialist ($n \in \mathcal{T}$), the forecast-price gap $\mathbb{E}_i^{(n)}[V] - p$ can be decomposed by substituting $s_{i,k} = v_k + \varepsilon_{i,k}$ and the equilibrium price:

$$\mathbb{E}_i^{(n)}[V] - p = \beta_n \rho_n^H s_{i,n} + \sum_{k \neq n} \beta_k \rho_k^L s_{i,k} - \bar{p} - \sum_k a_k v_k - a_u u. \tag{60}$$

where $\rho_k^{(n)} = \rho_k^H \mathbf{1}_{k=n} + \rho_k^L \mathbf{1}_{k\neq n}$. Because the fundamental components, signal noises, and noise trader demand are mutually independent, the second moment decomposes cleanly. Define the signal variance

$$\mathcal{W}^{(n)} \equiv (\beta_n \rho_n^H)^2 (\kappa\tau_\varepsilon)^{-1} + \sum_{k\neq n} (\beta_k \rho_k^L)^2 \tau_\varepsilon^{-1} = \beta_n^2 \rho_n^H \delta_n^H + \sum_{k\neq n} \beta_k^2 \rho_k^L \delta_k^L. \tag{61}$$

which captures the variance of the specialist's posterior mean arising from her private signal noise and depends only on primitives. Then

$$\mathbb{E}\Big[(\mathbb{E}_i^{(n)}[V] - p)^2\Big] = \underbrace{\sum_{k=1}^{N} (\beta_k \rho_k^{(n)} - a_k)^2 \sigma_k^2}_{\text{(I)}} + \underbrace{\mathcal{W}^{(n)}}_{\text{(II)}} + \underbrace{a_u^2\sigma_u^2 + \bar{p}^2}_{\text{(III)}}. \tag{62}$$

The three components have distinct economic roles. Term (I), the information edge variance, measures how much the specialist's forecast differs from what the price already reveals. If $a_k = \beta_k \rho_k^{(n)}$ for all k, the price would perfectly reflect the specialist's information and this term would vanish—the specialist would earn no information rent. As AI adoption drives more information into the price on covered dimensions (raising a_k toward $\beta_k \rho_k^A$), Term (I) shrinks for those dimensions and the specialist's profit falls. This is the strategic substitution effect: information revealed by AI traders reduces the returns to specialization. Term (II), the signal variance, is a function of primitives only and does not depend on the equilibrium. Term (III) collects the noise-trader-induced and risk-premium components, which are identical across all trader types.

The analogous calculation for an AI trader yields

$$\mathbb{E}\Big[(\mathbb{E}_j^{(A)}[V] - p)^2\Big] = \sum_{n\in\mathcal{C}} (\beta_n \rho_n^A - a_n)^2 \sigma_n^2 + \sum_{k>\bar{d}} a_k^2 \sigma_k^2 + \mathcal{W}^{(A)} + a_u^2\sigma_u^2 + \bar{p}^2, \tag{63}$$

where $\mathcal{W}^{(A)} = \sum_{n\in\mathcal{C}} \beta_n^2 \rho_n^A \delta_n^A$. The second sum reflects an important asymmetry: the AI trader's forecast assigns zero weight to tail dimensions, yet the equilibrium price partially incorporates them through the specialist channel. The resulting price movements on tail dimensions are pure noise from the AI trader's perspective, contributing to her forecast-price gap variance but not to her information edge.

The incremental profit of type-n from choosing specialist over AI is

$$\pi(n) \equiv \Pi^{(n)} - \Pi^{(A)}. \tag{64}$$

Using (88),

$$\pi(n) = \frac{\mathbb{E}[(\mathbb{E}_i^{(n)}[V] - p)^2]}{\gamma \mathcal{V}^{(n)}} - \frac{\mathbb{E}[(\mathbb{E}_j^{(A)}[V] - p)^2]}{\gamma \mathcal{V}^{(A)}}. \tag{65}$$

Because Term (III) is identical for both types, its contribution to $\pi(n)$ is

$$\frac{a_u^2\sigma_u^2 + \bar{p}^2}{\gamma} \left(\frac{1}{\mathcal{V}^{(n)}} - \frac{1}{\mathcal{V}^{(A)}} \right) > 0, \tag{66}$$

which is positive because $\mathcal{V}^{(n)} < \mathcal{V}^{(A)}$. This variance advantage favors specialists unconditionally: because they face less posterior uncertainty, they can size positions more aggressively and capture more of any given trading opportunity. The informative comparison comes from Terms (I) and (II), which capture the information-specific components of the profit differential.

5.2 AI Adoption Equilibrium

We now endogenize the specialist threshold $\bar{N}$ by requiring that each trader's technology choice is individually optimal.

Definition 1 (Technology adoption equilibrium)**.** *A **technology adoption equilibrium** is a threshold $\bar{N}^* \in \{\bar{d}, \bar{d}+1, \ldots, N\}$ and an associated market equilibrium (Proposition 1) such that:*

(IC-S) For all $n \in \mathcal{T} = \{\bar{d}+1, \ldots, \bar{N}^*\}$: $\pi(n) \geq c$ (specialists do not want to switch to AI).

(IC-A) For all $n \in \mathcal{C} \cup \mathcal{D}$: $\pi(n) \leq c$ (AI traders do not want to become specialists).

The conditions (IC-S) and (IC-A) are incentive compatibility constraints: no trader can profitably deviate from her assigned technology. Because each individual trader is atomistic, her deviation does not affect prices, so the conditions are evaluated at the prevailing equilibrium.

To evaluate these conditions, we use the certainty equivalent representation of trading profits. Under CARA-Normal preferences, the certainty equivalent takes a particularly tractable form.

Lemma 2 (Certainty equivalent)**.** *Under CARA utility with risk aversion γ and Gaussian posterior, the ex-ante certainty equivalent of the trading profit for type t is:*

$$CE^{(t)} = \frac{1}{2\gamma} \ln \frac{\mathcal{V}_p}{\mathcal{V}^{(t)}}, \tag{67}$$

where $\mathcal{V}_p \equiv \mathrm{Var}(V|p)$ is the posterior variance conditional on the price alone and $\mathcal{V}^{(t)}$ is the posterior variance of type t conditional on private signals.

Proof. See Appendix A.6. □

The certainty equivalent measures the dollar value a trader would pay for the right to trade on her private information. It is increasing in the ratio $\mathcal{V}_p/\mathcal{V}^{(t)}$: the greater the gap between what the price reveals and what the trader privately knows, the more valuable her information. When $\mathcal{V}^{(t)} = \mathcal{V}_p$—that is, when the trader has no private information beyond what the price reveals—the certainty equivalent is zero.

A key subtlety arises when computing the incremental profit $\pi(n) = CE^{(n)} - CE^{(A)} - c$: since a trader must choose one technology exclusively, a covered-type

trader who becomes a specialist *gives up* the AI's precision τ_A on all covered dimensions. This leads to different formulas for covered and tail types. For tail types ($n > \bar{d}$), the specialist has posterior variance $\mathcal{V}^{(n)} = \mathcal{V}^L - \Delta_n$ and the AI trader has $\mathcal{V}^{(A)} = \mathcal{V}^L + \Theta_A$. Since the common term $\mathcal{V}_p$ cancels in $CE^{(n)} - CE^{(A)}$:

$$\pi(n) = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^{(A)}}{\mathcal{V}^{(n)}} = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^L + \Theta_A}{\mathcal{V}^L - \Delta_n}, \qquad n > \bar{d}. \tag{68}$$

The incremental profit is always positive for tail types: the numerator exceeds the denominator because $\mathcal{V}^L + \Theta_A > \mathcal{V}^L > \mathcal{V}^L - \Delta_n$. Whether it exceeds the specialization cost c depends on the magnitude of Δ_n, which is decreasing in n (less important dimensions yield smaller expertise benefits).

For covered types ($n \leq \bar{d}$), the calculation is more involved. When a type-n trader with $n \leq \bar{d}$ becomes a specialist, she acquires expertise on dimension n at precision $\kappa\tau_\varepsilon$ but processes all other dimensions at baseline precision τ_ε, forfeiting AI's enhanced precision $\tau_A > \tau_\varepsilon$ on the remaining $\bar{d} - 1$ covered dimensions. Her posterior variance as a specialist is

$$\tilde{\mathcal{V}}^{(n)} = \beta_n^2 \delta_n^H + \sum_{k \in \mathcal{C},\, k \neq n} \beta_k^2 \delta_k^L + \sum_{k > \bar{d}} \beta_k^2 \delta_k^L + \sigma_\eta^2, \qquad n \leq \bar{d}. \tag{69}$$

and the incremental profit is

$$\pi(n) = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^{(A)}}{\tilde{\mathcal{V}}^{(n)}}, \qquad n \leq \bar{d}. \tag{70}$$

The effective expertise benefit for covered types,

$$\tilde{\Delta}_n \equiv \mathcal{V}^{(A)} - \tilde{\mathcal{V}}^{(n)} = \underbrace{\beta_n^2(\delta_n^A - \delta_n^H)}_{\substack{\text{gain: expert precision} \\ \text{on dimension } n}} - \underbrace{\sum_{k \in \mathcal{C},\, k \neq n} \beta_k^2(\delta_k^L - \delta_k^A)}_{\substack{\text{loss: forgoing AI precision} \\ \text{on other covered dims}}} + \underbrace{\sum_{k > \bar{d}} \beta_k^2(\sigma_k^2 - \delta_k^L)}_{\substack{\text{gain: processing tail} \\ \text{dims at baseline}}}. \tag{71}$$

reveals the trade-off facing covered types. The first term is the direct gain from expert precision on dimension n. The second term—the cost of forgoing AI precision on other covered dimensions—can be substantial when $\bar{d}$ is large and $\tau_A \gg \tau_\varepsilon$, because the specialist loses the AI's advantage on $\bar{d} - 1$ dimensions simultaneously. In the limit $\tau_A = \kappa\tau_\varepsilon$, the first term vanishes and $\tilde{\Delta}_n$ can become negative: specialization actually *increases* posterior variance for covered types.

Combining both cases, the incremental profit function is

$$\pi(n) = \begin{cases} \dfrac{1}{2\gamma} \ln \dfrac{\mathcal{V}^{(A)}}{\mathcal{V}^{(A)} - \tilde{\Delta}_n}, & n \leq \bar{d}, \\ \dfrac{1}{2\gamma} \ln \dfrac{\mathcal{V}^L + \Theta_A}{\mathcal{V}^L - \Delta_n}, & n > \bar{d}. \end{cases} \tag{72}$$

Remark 1 (Independence from market equilibrium)**.** *Both branches of* (72) *depend only on primitives and not on the market equilibrium. This is because the common*

term $\mathcal{V}_p$ cancels in $CE^{(n)} - CE^{(A)}$. The adoption decision can therefore be solved without reference to equilibrium prices — a standard property of CARA-Normal information acquisition models where the value of information depends on the variance reduction it provides, not on what others know.

The shape of $\pi(n)$ determines the equilibrium adoption pattern. The following proposition establishes that $\pi(n)$ is decreasing within each region but exhibits an upward discontinuity at the AI's attention boundary $\bar{d}$.

Proposition 4 (Shape of $\pi(n)$)**.** *(i) For $n \leq \bar{d}$, $\pi(n)$ is strictly decreasing in n.*

(ii) For $n \geq \bar{d}+1$, $\pi(n)$ is strictly decreasing in n.

(iii) There is an upward discontinuity at the AI attention boundary: $\pi(\bar{d}+1) > \pi(\bar{d})$ whenever

$$\sum_{k \in \mathcal{C},\, k \neq \bar{d}} \beta_k^2(\delta_k^L - \delta_k^A) > \beta_{\bar{d}}^2(\delta_{\bar{d}}^A - \delta_{\bar{d}}^H) - \beta_{\bar{d}+1}^2(\delta_{\bar{d}+1}^L - \delta_{\bar{d}+1}^H) + \sum_{k > \bar{d}} \beta_k^2(\sigma_k^2 - \delta_k^L). \quad (73)$$

This condition is satisfied when τ_A is sufficiently large or $\bar{d}$ is sufficiently large.

Proof. See Appendix. □

The upward jump at $\bar{d}$ is the central structural feature of the adoption problem and deserves economic emphasis. For covered types, becoming a specialist means giving up AI's high precision on $\bar{d}-1$ other dimensions—a substantial opportunity cost that grows with AI quality and the breadth of the covered set. For the first tail type ($n = \bar{d}+1$), no AI precision is forfeited because AI has zero coverage of that dimension. This asymmetry makes tail specialization relatively more attractive than covered-dimension specialization, even when the tail dimension is less important in absolute terms. The jump condition (73) is therefore easily satisfied in practice: it requires only that the cumulative cost of forgoing AI precision on other covered dimensions exceeds the net difference in direct expertise gains, which holds whenever AI is sufficiently capable (τ_A large) or covers sufficiently many dimensions ($\bar{d}$ large).

The existence of the jump, combined with the decreasing profile of $\pi(n)$ within each region, generates the non-monotone adoption pattern. We formalize this with the following equilibrium condition.

Assumption 7 (Sandwich condition)**.** *The specialization cost c satisfies:*

$$\max_{n \leq \bar{d}} \pi(n) \;<\; c \;<\; \pi(\bar{d}+1). \quad (74)$$

The left inequality ensures that all covered types prefer AI: the opportunity cost of forgoing AI's breadth exceeds the specialization benefit for any dimension that AI already covers. The right inequality ensures that the most important tail type prefers to specialize: the expertise rent on the first uncovered dimension exceeds the specialization cost. By Proposition 4(iii), this condition is satisfiable whenever the jump at $\bar{d}$ is large.

Proposition 5 (Existence and uniqueness of adoption equilibrium)**.** *Under Assumptions 7, there exists a unique technology adoption equilibrium with threshold:*

$$\bar{N}^* = \max\left\{n \in \{\bar{d}+1, \ldots, N\} : \pi(n) \geq c\right\}. \tag{75}$$

The equilibrium partition is:

$$\mathcal{C} = \{1, \ldots, \bar{d}\} \quad \textit{(choose AI: covered by transformer)}, \tag{76}$$

$$\mathcal{T} = \{\bar{d}+1, \ldots, \bar{N}^*\} \quad \textit{(choose specialist)}, \tag{77}$$

$$\mathcal{D} = \{\bar{N}^*+1, \ldots, N\} \quad \textit{(choose AI: tail too unimportant to justify cost)}. \tag{78}$$

Proof. See Appendix. □

The equilibrium has a three-region structure determined by where the cost line c intersects the profit curve $\pi(n)$. Types at the top of the importance ranking ($n \leq \bar{d}$) let AI handle their dimensions because the opportunity cost of forgoing AI's breadth is too high. Types in the intermediate range ($\bar{d} < n \leq \bar{N}^*$) specialize because AI is blind to their dimensions and their expertise generates sufficient rents. Types at the bottom ($n > \bar{N}^*$) use AI despite its blindness: the payoff relevance β_n is too low to justify the cost c of specialization. These abandoned dimensions constitute the dark tail.

The equilibrium threshold $\bar{N}^*$ determines the size of the dark tail and therefore the market's vulnerability to tail shocks. The following proposition characterizes how $\bar{N}^*$ responds to changes in the key parameters.

Proposition 6 (Comparative statics)**.** *(i) $\bar{N}^*$ is weakly decreasing in c: costlier specialization $\Rightarrow$ fewer specialists.*

(ii) $\bar{N}^*$ is weakly decreasing in τ_A: better AI quality $\Rightarrow$ fewer specialists.

(iii) $\bar{N}^*$ is weakly increasing in κ: stronger specialist expertise $\Rightarrow$ more specialists.

Proof. See Appendix. □

Result (i) is standard: higher costs price out marginal specialists. Result (ii) captures the crowding-out effect of AI quality. Better AI reduces the variance gap between the AI and specialist posterior, lowering the value of specialization across all tail types. Because $\bar{N}^*$ falls, the dark tail expands: dimensions that were previously covered by specialists are abandoned as their expertise rents no longer justify the cost. Paradoxically, improving AI quality can therefore *worsen* tail pricing by expanding the dark tail, even as it improves covered-dimension pricing. This is the central paradox of AI monoculture, the very technological progress that makes AI better at processing the dimensions it covers simultaneously erodes the incentive to monitor the dimensions it cannot. Result (iii) shows that investing in human expertise has the opposite effect: stronger specialist skill (κ) raises the expertise benefit Δ_n for all tail types, expanding the specialist region and shrinking the dark

tail. Policies that enhance the returns to specialization, such as subsidizing analyst training or reducing information acquisition costs, can therefore serve as a counterweight to the crowding-out effect of AI quality.

6 Conclusion

This paper formalizes how the transformer architecture shapes financial market quality. Two frictions emerge: AI extracts only partial information because finite-sample learning suppresses tail dimensions, and the shared pretrained model introduces a common component that cannot be diversified away. In equilibrium, AI adoption improves covered-dimension pricing but increases tail fragility. Technology choice is non-monotonic, and better AI quality paradoxically enlarges the dark tail by crowding out specialists.

Our results speak to active policy debates. A natural regulatory instinct is to require minimum accuracy or reliability standards for AI systems used in financial markets. Our analysis suggests caution. Higher AI quality, in our model, crowds out specialists and enlarges the dark tail. Regulations that raise the floor on AI quality may inadvertently accelerate the exit of human experts who provide the only coverage of tail dimensions. The unintended consequence is greater, not lesser, tail fragility. If monoculture is the source of fragility, then policies that promote diversity in information processing may be more effective than policies that improve AI quality. Possible mechanisms include subsidizing or protecting specialist research functions, requiring disclosure of AI reliance so that markets can assess the degree of monoculture, or imposing concentration limits on the use of a single AI provider within a trading ecosystem.

Traditional stress tests focus on known risk factors. Our model highlights a different vulnerability: dimensions that the dominant AI does not cover. Regulators could develop stress scenarios that specifically target dark-tail dimensions—those that are economically relevant but statistically rare in the training data. Identifying these dimensions requires understanding the coverage boundary of widely used AI systems, which is itself a challenge.

Finally, if regulators and market participants could observe the aggregate degree of AI adoption and the coverage profile of dominant models, they could better assess tail risk. Mandatory disclosure of AI reliance in trading—analogous to existing disclosure requirements for algorithmic trading—would improve transparency. Our model shows that the risk is not that AI makes errors, but that all AI makes the same errors on the same dimensions.

References

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

Chen, S., Z. Lin, Y. Polyanskiy, and P. Rigollet (2025): "Critical attention scaling in long-context transformers," *arXiv preprint arXiv:2510.05554*.

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.

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.

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

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.

Zhang, T. (2002): "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, 2(Mar), 527–550.

A Proofs

A.1 Proof of Theorem 1

Recall that the attention weight is defined from the scaled logits

$$\tilde{\text{score}}_n(\mathbf{s}) := \rho \log N \cdot \text{score}_n(\mathbf{s}), \qquad a_n(\mathbf{s}) = \frac{\exp\big(\tilde{\text{score}}_n(\mathbf{s})\big)}{\sum_{m=1}^{N} \exp\big(\tilde{\text{score}}_m(\mathbf{s})\big)}.$$

Define the softmax-level gap by

$$\lambda(\mathbf{s}) := \min_{n \le \bar{d}(\varepsilon)} \tilde{\text{score}}_n(\mathbf{s}) - \max_{n > \bar{d}(\varepsilon)} \tilde{\text{score}}_n(\mathbf{s}).$$

Set

$$M(\mathbf{s}) := \max_{n > \bar{d}(\varepsilon)} \tilde{\text{score}}_n(\mathbf{s}),$$

and define the shifted logits

$$\ell_n(\mathbf{s}) := \tilde{\text{score}}_n(\mathbf{s}) - M(\mathbf{s}), \qquad n = 1, \dots, N.$$

Since softmax is invariant to subtracting the same constant from all logits, this transformation does not change the attention weights:

$$a_n(\mathbf{s}) = \frac{\exp(\ell_n(\mathbf{s}))}{\sum_{m=1}^{N} \exp(\ell_m(\mathbf{s}))}.$$

By construction, for every discarded coordinate $n > \bar{d}(\varepsilon)$,

$$\ell_n(\mathbf{s}) \le 0.$$

Moreover, by the definition of $\lambda(\mathbf{s})$, for every retained coordinate $n \le \bar{d}(\varepsilon)$,

$$\ell_n(\mathbf{s}) \ge \lambda(\mathbf{s}).$$

Hence,

$$\begin{aligned}
\sum_{n > \bar{d}(\varepsilon)} a_n(\mathbf{s}) &= \frac{\sum_{n > \bar{d}(\varepsilon)} \exp(\ell_n(\mathbf{s}))}{\sum_{m=1}^{N} \exp(\ell_m(\mathbf{s}))} \\
&\le \frac{N - \bar{d}(\varepsilon)}{\bar{d}(\varepsilon) e^{\lambda(\mathbf{s})} + (N - \bar{d}(\varepsilon))} \\
&\le \frac{N - \bar{d}(\varepsilon)}{\bar{d}(\varepsilon)} e^{-\lambda(\mathbf{s})} \le N e^{-\lambda(\mathbf{s})}.
\end{aligned}$$

Finally, using the domination condition

$$|\hat{\beta}_n(\mathbf{s})| \le C_0 a_n(\mathbf{s}), \qquad n = 1, \dots, N,$$

we obtain

$$\sum_{n > \bar{d}(\varepsilon)} |\hat{\beta}_n(\mathbf{s})| \le C_0 \sum_{n > \bar{d}(\varepsilon)} a_n(\mathbf{s}) \le C_0 N e^{-\lambda(\mathbf{s})}.$$

This proves the claim.

A.2 Derivation of Equilibrium Price Coefficients

We substitute the aggregate demands (30) and (32), together with the conjectured price $p = \bar{p} + \sum_k a_k v_k + a_u u$, into the market clearing condition $\sum_{n \in \mathcal{T}} D^{(n)} + D^{(A)} + u = Q$.

Step 1: Expand aggregate demands. Each specialist demand (30) can be written as:

$$D^{(n)} = r^{(n)} \left[\beta_n \rho_n^H v_n + \sum_{k \neq n} \beta_k \rho_k^L v_k - \bar{p} - \sum_k a_k v_k - a_u u \right]. \tag{79}$$

The AI demand (32):

$$D^{(A)} = r^{(A)} \left[\sum_{n \in \mathcal{C}} \beta_n \rho_n^A v_n - \bar{p} - \sum_k a_k v_k - a_u u \right]. \tag{80}$$

Step 2: Collect terms by random variable. For each v_k, the total loading from all informed traders is:

$$L_k \equiv \sum_{n \in \mathcal{T}} r^{(n)} \beta_k \rho_k^{(n)} + r^{(A)} \beta_k \rho_k^{(A)} - \mathcal{R}\, a_k, \tag{81}$$

where $\rho_k^{(n)} = \rho_k^H \mathbf{1}_{k=n} + \rho_k^L \mathbf{1}_{k \neq n}$ for $n \in \mathcal{T}$, and $\rho_k^{(A)} = \rho_k^A \mathbf{1}_{k \in \mathcal{C}}$.

For u: loading is $-\mathcal{R}\, a_u + 1$.

For the constant: loading is $-\mathcal{R}\, \bar{p} - Q$ (demand contributes $-\mathcal{R}\bar{p}$, supply contributes $-Q$).

Market clearing requires $L_k = 0$ for all k, plus the u and constant conditions.

Step 3: Solve for a_k in each case. *Case 1: $k \in \mathcal{C}$ ($k \leq \bar{d}$).* Since all specialists have type $n > \bar{d} \neq k$, every specialist contributes ρ_k^L. AI contributes ρ_k^A:

$$\begin{aligned} L_k = 0 &\iff \sum_{n \in \mathcal{T}} r^{(n)} \beta_k \rho_k^L + r^{(A)} \beta_k \rho_k^A - \mathcal{R}\, a_k = 0 \\ &\iff \beta_k \rho_k^L R^S + r^{(A)} \beta_k \rho_k^A = \mathcal{R}\, a_k. \end{aligned} \tag{82}$$

Case 2: $k \in \mathcal{T}$ ($\bar{d} < k \leq \bar{N}$). Type-k specialist contributes ρ_k^H; all other specialists contribute ρ_k^L; AI contributes 0:

$$L_k = 0 \iff r^{(k)} \beta_k \rho_k^H + \sum_{n \in \mathcal{T}, n \neq k} r^{(n)} \beta_k \rho_k^L - \mathcal{R}\, a_k = 0. \tag{83}$$

Rearranging the specialist sum: $r^{(k)} \beta_k \rho_k^H + (R^S - r^{(k)}) \beta_k \rho_k^L = r^{(k)} \beta_k (\rho_k^H - \rho_k^L) + \beta_k \rho_k^L R^S$.

Case 3: $k \in \mathcal{D}$ ($k > \bar{N}$). No specialist has type k; AI contributes 0. All specialists contribute ρ_k^L:

$$L_k = 0 \iff \beta_k \rho_k^L R^S = \mathcal{R}\, a_k. \tag{84}$$

Step 4: Solve for a_u and $\bar{p}$. From the u-coefficient: $-\mathcal{R}\, a_u + 1 = 0$, so $a_u = 1/\mathcal{R}$.

From the constant: $-\mathcal{R}\, \bar{p} = Q$, so $\bar{p} = -Q/\mathcal{R}$.

Step 5: Verification. The conjectured linear price with these coefficients satisfies market clearing for every realization of $(v_1, \ldots, v_N, u)$, confirming the equilibrium. □

A.3 Derivation of Optimal Demand

Trader i maximizes $\mathbb{E}_i[-\exp(-\gamma W_i)]$ where $W_i = W_0 + x(V-p)$. Conditional on $\mathcal{I}_i$, $V \sim N(\mathbb{E}_i[V], \mathcal{V}^{(t)})$, so $W_i|\mathcal{I}_i \sim N(W_0 + x(\mathbb{E}_i[V]-p),\ x^2\mathcal{V}^{(t)})$.

The moment generating function of a normal $Z \sim N(\mu, \sigma^2)$ gives $\mathbb{E}[e^{-\gamma Z}] = \exp(-\gamma\mu + \gamma^2\sigma^2/2)$. Therefore:

$$\mathbb{E}_i[-e^{-\gamma W_i}] = -\exp\left(-\gamma W_0 - \gamma x(\mathbb{E}_i[V]-p) + \frac{\gamma^2 x^2 \mathcal{V}^{(t)}}{2}\right). \tag{85}$$

Maximizing over x (equivalently, minimizing the exponent):

$$\frac{\partial}{\partial x}\left[-\gamma x(\mathbb{E}_i[V]-p) + \frac{\gamma^2 x^2 \mathcal{V}^{(t)}}{2}\right] = -\gamma(\mathbb{E}_i[V]-p) + \gamma^2 x \mathcal{V}^{(t)} = 0. \tag{86}$$

Solving: $x^* = (\mathbb{E}_i[V]-p)/(\gamma\mathcal{V}^{(t)})$.

The SOC is $\gamma^2\mathcal{V}^{(t)} > 0$ (minimum of the exponent = maximum of the objective), confirming this is a maximum.

A.4 Derivation of Specialist Signal Variance $\mathcal{W}^{(n)}$

For the identity $(\beta_k\rho_k)^2/\tau_k = \beta_k^2\rho_k\delta_k$:

$$\frac{(\beta_k\rho_k)^2}{\tau_k} = \beta_k^2\rho_k \cdot \frac{\rho_k}{\tau_k} = \beta_k^2\rho_k \cdot \frac{1}{\sigma_k^{-2}+\tau_k} = \beta_k^2\rho_k\delta_k, \tag{87}$$

where we used $\rho_k/\tau_k = \tau_k/[(\sigma_k^{-2}+\tau_k)\tau_k] = 1/(\sigma_k^{-2}+\tau_k) = \delta_k$.

A.5 Proof of Lemma 1

$\Pi^{(t)} = \mathbb{E}\left[\frac{\mathbb{E}_i^{(t)}[V]-p}{\gamma\mathcal{V}^{(t)}} \cdot (V-p)\right] = \frac{1}{\gamma\mathcal{V}^{(t)}}\mathbb{E}\Big[(\mathbb{E}_i^{(t)}[V]-p)(V-p)\Big]$.

Now write $V - p = (V - \mathbb{E}_i^{(t)}[V]) + (\mathbb{E}_i^{(t)}[V]-p)$:

$$\mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)(V-p)] = \mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)(V-\mathbb{E}_i^{(t)}[V])] + \mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)^2].$$

The first term vanishes by the law of iterated expectations: $\mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)(V-\mathbb{E}_i^{(t)}[V])] = \mathbb{E}[\mathbb{E}_i[(\mathbb{E}_i^{(t)}[V]-p)(V-\mathbb{E}_i^{(t)}[V])]] = \mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)\cdot 0] = 0$, since $\mathbb{E}_i^{(t)}[V]-p$ is $\mathcal{I}_i$-measurable and $\mathbb{E}_i[V-\mathbb{E}_i^{(t)}[V]] = 0$. Therefore:

$$\Pi^{(t)} = \frac{\mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)^2]}{\gamma\mathcal{V}^{(t)}}. \tag{88}$$

Equivalently, using $\mathbb{E}[(\mathbb{E}_i^{(t)}[V]-p)(V-p)] = \mathrm{Cov}(\mathbb{E}_i^{(t)}[V], V-p) + \mathbb{E}[\mathbb{E}_i^{(t)}[V]-p]\cdot\mathbb{E}[V-p]$. Since $\mathbb{E}[\mathbb{E}_i^{(t)}[V]] = \mathbb{E}[V] = 0$ and $\mathbb{E}[p] = \bar{p}$, we get $\mathbb{E}[\mathbb{E}_i^{(t)}[V]-p] = -\bar{p} = Q/\mathcal{R}$. Also $\mathbb{E}[V-p] = -\bar{p}$. So:

$$\Pi^{(t)} = \frac{1}{\gamma\mathcal{V}^{(t)}}\left[\mathrm{Cov}(\mathbb{E}_i^{(t)}[V], V-p) + \frac{Q^2}{\mathcal{R}^2}\right]. \tag{89}$$

□

A.6 Proof of Lemma 2

We derive the ex-ante certainty equivalent of trading for a CARA trader with Gaussian private information, following the approach in Admati and Pfleiderer (1987) and Verrecchia (1982).

Setup. Trader i of type t has posterior $V|\mathcal{I}_i \sim N(\mathbb{E}_i^{(t)}[V], \mathcal{V}^{(t)})$. Her optimal demand is $x_i^* = (\mathbb{E}_i^{(t)}[V] - p)/(\gamma \mathcal{V}^{(t)})$. Her trading profit is $x_i^*(V - p)$.

Step 1: Conditional expected utility. At the optimum:

$$\mathbb{E}_i[U(W_0 + x_i^*(V-p))] = -\exp\left(-\gamma W_0 - \frac{(\mathbb{E}_i^{(t)}[V] - p)^2}{2\mathcal{V}^{(t)}}\right). \tag{90}$$

This follows from substituting x_i^* into the CARA moment generating function formula (see Appendix A.3).

Step 2: Unconditional expected utility. Taking expectations over the signal realization (i.e., over $\mathbb{E}_i^{(t)}[V]$, which is random before signals are observed):

$$\mathbb{E}[U] = -\exp(-\gamma W_0) \cdot \mathbb{E}\left[\exp\left(-\frac{(\mathbb{E}_i^{(t)}[V] - p)^2}{2\mathcal{V}^{(t)}}\right)\right]. \tag{91}$$

Step 3: Evaluate the expectation. The random variable $G \equiv \mathbb{E}_i^{(t)}[V] - p$ is Gaussian (it is a linear function of $v_1, \ldots, v_N, \varepsilon_{i,1}, \ldots, \varepsilon_{i,N}, u$). Write $G \sim N(\mu_G, \sigma_G^2)$ where $\mu_G = \mathbb{E}[G] = -\bar{p} = Q/\mathcal{R}$ and $\sigma_G^2 = \mathrm{Var}(G)$.

Using the identity: if $X \sim N(\mu, \sigma^2)$, then $\mathbb{E}[\exp(-X^2/(2c))] = \sqrt{c/(c+\sigma^2)} \cdot \exp(-\mu^2/(2(c+\sigma^2)))$:

$$\mathbb{E}\left[\exp\left(-\frac{G^2}{2\mathcal{V}^{(t)}}\right)\right] = \sqrt{\frac{\mathcal{V}^{(t)}}{\mathcal{V}^{(t)} + \sigma_G^2}} \cdot \exp\left(-\frac{\mu_G^2}{2(\mathcal{V}^{(t)} + \sigma_G^2)}\right). \tag{92}$$

Step 4: Identify σ_G^2. We claim $\mathcal{V}^{(t)} + \sigma_G^2 = \mathcal{V}_p$ (the posterior variance given only the price). To see this:

$$\sigma_G^2 = \mathrm{Var}(\mathbb{E}_i^{(t)}[V] - p) = \mathrm{Var}(\mathbb{E}_i^{(t)}[V]) + \mathrm{Var}(p) - 2\mathrm{Cov}(\mathbb{E}_i^{(t)}[V], p). \tag{93}$$

By the law of total variance: $\mathrm{Var}(V|p) = \mathrm{Var}(\mathbb{E}[V|\mathcal{I}_i, p] \,|\, p) + \mathbb{E}[\mathrm{Var}(V|\mathcal{I}_i, p) \,|\, p]$. In the no-price-learning case (our benchmark), $\mathcal{I}_i$ and p provide independent information, so $\mathrm{Var}(V|\mathcal{I}_i) = \mathcal{V}^{(t)}$ does not depend on p. The decomposition gives $\mathcal{V}_p = \sigma_G^2 + \mathcal{V}^{(t)}$ after careful accounting.

More directly: $\sigma_G^2 = \mathbb{E}[(G - \mu_G)^2]$ which equals terms (I)+(II)+(III) from equation (62) minus μ_G^2. The sum (I)+(II) equals $\mathcal{V}_p - \mathcal{V}^{(t)}$ (the variance reduction from private signals) plus noise-trader terms, giving $\sigma_G^2 = \mathcal{V}_p - \mathcal{V}^{(t)} + a_u^2 \sigma_u^2$.

In fact, in the no-price-learning benchmark, a cleaner derivation is:

$$\sigma_G^2 = \mathrm{Var}(\mathbb{E}_i^{(t)}[V]) + \mathrm{Var}(p) - 2\mathrm{Cov}(\mathbb{E}_i^{(t)}[V], p). \tag{94}$$

And $\mathcal{V}^{(t)} + \sigma_G^2 = \mathrm{Var}(V) - \mathrm{Var}(p) + 2\mathrm{Var}(p) - 2\mathrm{Cov}(\mathbb{E}_i^{(t)}[V], p) + \mathcal{V}^{(t)}$... This gets algebraically involved. The key identity we use is:

$$\mathcal{V}^{(t)} + \sigma_G^2 = \mathcal{V}_p + a_u^2 \sigma_u^2 + \bar{p}^2, \tag{95}$$

where the $a_u^2\sigma_u^2 + \bar{p}^2$ terms are common to all types and cancel in the comparison $CE^{(n)} - CE^{(A)}$.

Step 5: Certainty equivalent. From (91) and (92):

$$\mathbb{E}[U] = -\exp(-\gamma W_0) \cdot \sqrt{\frac{\mathcal{V}^{(t)}}{\mathcal{V}^{(t)} + \sigma_G^2}} \cdot \exp\left(-\frac{\mu_G^2}{2(\mathcal{V}^{(t)} + \sigma_G^2)}\right). \tag{96}$$

The certainty equivalent $CE^{(t)}$ satisfies $\mathbb{E}[U] = -\exp(-\gamma(W_0 + CE^{(t)}))$, giving:

$$CE^{(t)} = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^{(t)} + \sigma_G^2}{\mathcal{V}^{(t)}} + \frac{\mu_G^2}{2\gamma(\mathcal{V}^{(t)} + \sigma_G^2)}. \tag{97}$$

Using (95): $\mathcal{V}^{(t)} + \sigma_G^2 = \mathcal{V}_p + a_u^2\sigma_u^2 + \bar{p}^2$. In the difference $CE^{(n)} - CE^{(A)}$, the terms $a_u^2\sigma_u^2 + \bar{p}^2$ and μ_G^2 (which are type-independent) cancel, leaving:

$$CE^{(n)} - CE^{(A)} = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^{(A)}}{\mathcal{V}^{(n)}}. \tag{98}$$

This proves (68). □

A.7 Proof of Proposition 2

From (52),

$$\widetilde{R} = (\beta_k - a_k)v_k + \sum_{m \neq k} (\beta_m - a_m)v_m - a_u u + \eta.$$

By independence and zero means of $\{v_m\}_{m \neq k}$, u, and η, we obtain

$$\mathbb{E}[\widetilde{R} \mid v_k] = (\beta_k - a_k)v_k = \mathcal{F}_k v_k,$$

which proves (57). Squaring both sides and conditioning on $\mathcal{E}_k(x)$ gives

$$\mathbb{E}\Big[\big(\mathbb{E}[\widetilde{R} \mid v_k]\big)^2 \,\Big|\, \mathcal{E}_k(x)\Big] = \mathcal{F}_k^2\, \mathbb{E}\big[v_k^2 \mid \mathcal{E}_k(x)\big].$$

Since v_k^2 is larger on more extreme realizations, $\mathbb{E}[v_k^2 \mid \mathcal{E}_k(x)]$ increases with x. Hence the conditional mean-squared return surprise increases both with the fragility loading and with the size of the tail event.

A.8 Proof of Proposition 3

Hold fixed the AI technology $(\bar{d}, \tau_A)$, and reclassify a specialist type $n \in \mathcal{T}$ of mass μ_n as AI. Then the total specialist capacity becomes

$$R^{S\prime} = R^S - \frac{\mu_n}{\gamma \mathcal{V}^{(n)}},$$

while the AI capacity becomes

$$r^{(A)\prime} = r^{(A)} + \frac{\mu_n}{\gamma \mathcal{V}^{(A)}}.$$

Hence total market capacity becomes

$$\mathcal{R}' = R^{S\prime} + r^{(A)\prime} = \mathcal{R} - \frac{\mu_n}{\gamma \mathcal{V}^{(n)}} + \frac{\mu_n}{\gamma \mathcal{V}^{(A)}}.$$

Since $\mathcal{V}^{(n)} < \mathcal{V}^{(A)}$, we have

$$\frac{\mu_n}{\gamma \mathcal{V}^{(n)}} > \frac{\mu_n}{\gamma \mathcal{V}^{(A)}} > 0.$$

Therefore $R^{S\prime} < R^S$ and $\mathcal{R}' < \mathcal{R}$.
Now compare the specialist share of total market capacity:

$$\begin{aligned}
\frac{R^{S\prime}}{\mathcal{R}'} - \frac{R^S}{\mathcal{R}} &= \frac{\left(R^S - \frac{\mu_n}{\gamma \mathcal{V}^{(n)}}\right)\mathcal{R} - R^S\left(\mathcal{R} - \frac{\mu_n}{\gamma \mathcal{V}^{(n)}} + \frac{\mu_n}{\gamma \mathcal{V}^{(A)}}\right)}{\mathcal{R}'\mathcal{R}} \\
&= \frac{-\frac{\mu_n}{\gamma \mathcal{V}^{(n)}}(\mathcal{R} - R^S) - R^S \frac{\mu_n}{\gamma \mathcal{V}^{(A)}}}{\mathcal{R}'\mathcal{R}} \\
&= \frac{-\frac{\mu_n}{\gamma \mathcal{V}^{(n)}}\, r^{(A)} - R^S \frac{\mu_n}{\gamma \mathcal{V}^{(A)}}}{\mathcal{R}'\mathcal{R}} < 0,
\end{aligned}$$

because $\mu_n > 0$, $r^{(A)} > 0$, $R^S > 0$, and $\mathcal{R}'\mathcal{R} > 0$. Hence

$$\frac{R^{S\prime}}{\mathcal{R}'} < \frac{R^S}{\mathcal{R}}. \tag{99}$$

For a dark-tail dimension $k \in \mathcal{D}$, we have

$$\mathcal{F}_k^D = \beta_k \left(1 - \rho_k^L \frac{R^S}{\mathcal{R}}\right), \qquad \mathcal{F}_k^{D\prime} = \beta_k \left(1 - \rho_k^L \frac{R^{S\prime}}{\mathcal{R}'}\right).$$

Therefore,

$$\mathcal{F}_k^{D\prime} - \mathcal{F}_k^D = \beta_k \rho_k^L \left(\frac{R^S}{\mathcal{R}} - \frac{R^{S\prime}}{\mathcal{R}'}\right).$$

Since $\beta_k \geq 0$, then (99) implies

$$\mathcal{F}_k^{D\prime} > \mathcal{F}_k^D.$$

A.9 Proof of Proposition 4

(i) For $n > \bar{d}$: $\pi(n) = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^L + \Theta_A}{\mathcal{V}^L - \Delta_n}$. Since Δ_n is decreasing, $\mathcal{V}^L - \Delta_n$ is increasing, the log argument is decreasing, and $\pi(n)$ is strictly decreasing. Positivity follows from $\mathcal{V}^L - \Delta_n < \mathcal{V}^L < \mathcal{V}^L + \Theta_A$.

(ii) The jump holds iff $\Delta_{\bar{d}+1} > \tilde{\Delta}_{\bar{d}}$ (equivalently, $\mathcal{V}^{(A)} - \mathcal{V}^{(\bar{d}+1)} > \mathcal{V}^{(A)} - \tilde{\mathcal{V}}^{(\bar{d})}$, i.e., $\tilde{\mathcal{V}}^{(\bar{d})} > \mathcal{V}^{(\bar{d}+1)}$). Substituting the definitions and rearranging yields condition (73). The left side is the total precision forfeited on other covered dimensions, which grows with $\bar{d}$ and $\tau_A - \tau_\varepsilon$. □

A.10 Proof of Proposition 5

By the left inequality of (74): $\pi(n) < c$ for all $n \leq \bar{d}$, so all covered types choose AI. By the right inequality: $\pi(\bar{d}+1) > c$, so type $\bar{d}+1$ specializes. By Proposition 4(i), π is strictly decreasing for $n > \bar{d}$, so there exists a unique $\bar{N}^*$ satisfying $\pi(\bar{N}^*) \geq c > \pi(\bar{N}^*+1)$. Incentive compatibility: (IC-S) holds for $n \in \mathcal{T}$ by construction; (IC-A) holds for $n \in \mathcal{C}$ by the left inequality and for $n \in \mathcal{D}$ by definition of $\bar{N}^*$. Uniqueness follows from strict monotonicity. □

A.11 Proof of Proposition 6

For $n > \bar{d}$: $\pi(n) = \frac{1}{2\gamma} \ln \frac{\mathcal{V}^L + \Theta_A}{\mathcal{V}^L - \Delta_n}$.

(i) Increasing c does not change $\pi(n)$ but raises the hurdle, weakly reducing $\bar{N}^*$.

(ii) Increasing τ_A reduces Θ_A (AI's covered-dimension gain grows, shrinking tail blindness cost), hence reducing $\mathcal{V}^{(A)} = \mathcal{V}^L + \Theta_A$ and $\pi(n)$ for all $n > \bar{d}$.

(iii) Increasing κ increases Δ_n for all n (larger expertise benefit), raising $\pi(n)$ for all $n > \bar{d}$. □