# Neural Attention in Heterogeneous-Agent Economies

Alessandro T. Villa*

August 19, 2026

**Abstract**

Transformers, the core architecture behind modern large language models, combine flexible nonlinear approximation with attention, a mechanism that learns how information is aggregated across inputs. Can attention reveal which parts of an agent distribution matter for aggregate dynamics? I represent cross-sectional distributions as economic tokens and embed a transformer in a recursive Krusell–Smith-style solution method. The transformer forecasts the continuation distribution entering individual Bellman equations. I establish an analytical link between transformer attention and the sensitivity of aggregate policy functions to distributional components, with attention recovering their normalized marginal sensitivities under a restricted representation. In Krusell–Smith benchmarks, the transformer rediscovers approximate aggregation and assigns nearly uniform attention when aggregate capital is nearly sufficient. In a heterogeneous-firm economy with endogenous default, attention concentrates near the firm default margin and becomes state dependent: during downturns attention shifts toward financially fragile firms, whose balance sheets become more consequential for leverage and aggregate dynamics.

**Keywords:** Heterogeneous-agent economies; heterogeneous firms; firm dynamics; financial frictions; endogenous default; transformer attention; computational economics; distributional dynamics; neural networks; machine learning.

**JEL Codes:** C45; C63; C68; D22; E21; E32; E44; G33.

*Federal Reserve Bank of Chicago, 230 South LaSalle Street, Chicago, IL 60604, USA. Email: alessandro.tenzin.villa@gmail.com. Acknowledgments: I thank Nicolò Ceneri, Christian Hellwig, Giuseppe Lopomo, Walker Ray, and Nicolas Werquin for helpful comments and encouragement. Disclaimer: The views expressed in this paper do not represent the views of the Chicago Fed or the Federal Reserve System.

# 1 Introduction

Transformers (Vaswani et al., 2017) are a core architecture behind modern artificial intelligence. They combine flexible nonlinear approximation with attention, a mechanism that learns how to combine information across inputs. This paper explores a natural point of contact with heterogeneous agent models, where a central problem is also one of aggregation: which parts of a high-dimensional cross-sectional distribution matter for aggregate dynamics, and when? I represent economic states as tokens—such as wealth bins or capital–debt–productivity cells—and use a transformer to learn how these distributional objects are combined when forecasting equilibrium dynamics. To my knowledge, this is the first application of transformer attention as a recursive state-reduction device in heterogeneous agent and heterogeneous firm economies. I ask whether attention can both help solve these models and provide an economically interpretable map of how distributional information enters aggregate outcomes. The contribution is threefold.

First, I show analytically that, under a simple restricted attention representation, attention coincides with the normalized marginal effect of each distributional component on the aggregate policy function.[1] A two-period mean-sufficient economy provides the simplest illustration. Because every wealth-bin contribution has the same marginal effect on next-period aggregate capital, the analytical mapping implies uniform attention across the distribution. I then use a second analytical benchmark to illustrate how the mapping operates away from mean sufficiency. With production at the individual-producer level and diminishing returns, the marginal contribution of capital to aggregate accumulation is larger for smaller producers. The analytical mapping therefore implies a non-uniform attention profile, with normalized marginal relevance declining with producer capital. I evaluate both benchmarks using a richer full self-attention transformer in which the restrictions used to derive the analytical mapping are relaxed. Although attention weights and value representations are learned from the state, the resulting attention profiles closely track the analytical benchmarks: nearly uniform in the mean-sufficient economy and systematically declining with capital in the individual-production economy.

I then generalize the analytical result beyond the restricted transformer attention representation underlying the two examples above. Those benchmarks isolate the value channel: attention weights are held fixed under the local perturbation, so a change in one distributional component affects the prediction only through the economic content being aggregated. In a more general attention representation, the same perturbation can also change the attention

[1] A distributional component can be, for example, the contribution of a wealth bin to aggregate capital or, in the firm application, a capital–debt–productivity cell.

weights themselves, creating an attention-reallocation channel. For a general differentiable aggregate policy function, I show that local sensitivity to the distribution decomposes into these two components: a value channel, which captures changes in the information being aggregated holding attention weights fixed, and an attention-reallocation channel, which captures how the perturbation changes the weights assigned across the cross section. The exact mapping between attention and normalized policy sensitivity derived in the two analytical benchmarks is recovered as the special case in which attention directly weights the distributional components and those weights are held fixed under the local perturbation.

Second, methodologically, I develop a recursive solution method, the Distributional State Transformer (DST), in the spirit of Krusell and Smith (1998) and building on recent machine-learning approaches to heterogeneous-agent models (Fernandez-Villaverde et al., 2023; Han et al., 2025; Kase et al., 2025; Gu et al., 2026). Rather than committing ex ante to a small set of aggregate moments, I represent the cross-sectional distribution through economic tokens: wealth-productivity bins in a household economy, or capital–debt–productivity cells in a firm economy. The underlying economy retains a continuum-agent representation: individual states are discretized on a standard grid, while aggregate states are represented by a sparse library of complete distributions over that grid rather than by a finite simulated population. A transformer forecasts the next tokenized distribution, which enters individual Bellman equations through continuation values. The neural-network component provides a flexible approximation to the equilibrium law of motion, while attention determines how information from different regions of the distribution is combined in a state-contingent fashion. I validate the method in a standard one-asset heterogeneous-household economy against a one-moment Krusell–Smith solution. Although the DST observes and forecasts the cross-sectional distribution rather than aggregate capital, it rediscovers the Krusell–Smith approximate-aggregation structure: once aggregate capital is recovered from the forecast distribution, the implied policy function is nearly identical to the one-moment Krusell–Smith benchmark, while attention remains nearly flat across wealth bins.

Third, I apply the method to a heterogeneous-firm economy in which endogenous firm default makes the aggregation of firm balance sheets economically important. While recent machine-learning methods for heterogeneous-agent macroeconomics have been applied primarily to household economies, firm balance sheets provide a natural setting for distributional attention. The model builds on Khan and Thomas (2013), combining partial capital irreversibility with costly equity issuance and defaultable debt priced under endogenous default. It is also closely related to Khan et al. (2021), who study defaultable debt and endogenous default and exit while solving their stochastic economy with a Krusell–Smith approximation based on aggregate capital. These papers show that rich firm heterogeneity

can coexist with parsimonious aggregate forecasting rules. Rather than imposing such a low-dimensional representation ex ante, I let the transformer observe the joint distribution of capital, debt, and productivity directly. In the quantitative model, attention is highly nonuniform: it concentrates near the endogenous default margin and becomes state dependent.

The policy-sensitivity decomposition then shows why this attention pattern is economically relevant. The transformer forecasts the next cross-sectional distribution, from which aggregate objects can be recovered. I illustrate the decomposition using next-period capital and leverage, neither of which is a separate prediction target. The exercise delivers two findings. First, different aggregate outcomes depend on the firm distribution through different margins. Next-period capital is driven primarily by the persistence of firms' existing capital stocks, with investment adjusting this stock at the margin, whereas leverage is more sensitive to how heterogeneity in firms' balance sheets is aggregated across the cross section. Second, this aggregation margin becomes especially important near the endogenous default boundary. For leverage, the value channel—the effect of a redistribution through the information carried by firms' balance-sheet regions, holding aggregation weights fixed—and the attention-reallocation channel—the additional effect arising because those weights themselves change—have similar average magnitudes, but attention reallocation becomes particularly important near default. Shifting firm mass toward financially fragile regions therefore changes not only the composition of the cross section, but also which other regions become relevant for implied aggregate leverage. Because default risk and debt prices vary sharply near the default boundary, the mapping from the firm distribution into aggregate outcomes becomes itself sensitive to the distribution's composition. Financial fragility therefore makes economic aggregation endogenous to the cross-sectional state.

Finally, transformer attention is state dependent. A negative aggregate productivity shock raises firm default and temporarily shifts attention toward financially fragile firms. A positive shock of the same size lowers default risk and generates a much weaker reallocation of attention. The aggregate responses mirror this asymmetry. In busts, default rises and investment and hours contract, while attention moves toward financially fragile firms; in booms, default risk falls and the attention response is substantially smaller. Thus, the part of the distribution that matters for aggregate policy sensitivity changes over the business cycle: financially fragile balance sheets become more consequential for aggregate dynamics precisely in adverse states.

**Related Literature.** This paper contributes to two main strands of the literature: (i) computational methods for heterogeneous-agent economies, especially recent machine-learning

approaches that build on the Krusell–Smith approach, and (ii) firm dynamics with financial frictions. A key contribution is to connect these strands through a recursive algorithm that represents the cross-sectional distribution as economic tokens, learns its law of motion with a transformer, and uses the resulting attention weights as an interpretable diagnostic of how distributional information is aggregated across aggregate states.

*Machine Learning and Computational Methods in Heterogeneous-Agent Models.* An increasingly rich literature uses machine-learning tools to solve high-dimensional dynamic economic models, including Scheidegger and Bilionis (2019), Maliar et al. (2021), Azinovic et al. (2022), Valaitis and Villa (2024), Payne et al. (2025), and Gopalakrishna et al. (2026). Within this literature, the closest methodological papers for heterogeneous-agent economies are Ebrahimi Kahou et al. (2021), Fernandez-Villaverde et al. (2023), Han et al. (2025), Kase et al. (2025), and Gu et al. (2026).

Ebrahimi Kahou et al. (2021) show that permutation symmetries across heterogeneous agents can be exploited in neural architectures to make high-dimensional dynamic programs and recursive competitive equilibria tractable. Their approach is complementary to the one taken here. The DST works directly with a continuum distribution, where individual agent labels have already been integrated out, and represents the remaining cross-sectional state through economically meaningful regions whose aggregation is learned by transformer attention. Mean sufficiency provides a useful special case, in which these regions have equal marginal relevance and the corresponding attention benchmark is uniform.[2] Fernandez-Villaverde et al. (2023) develop a global nonlinear solution and estimation method for a continuous-time heterogeneous-agent economy with financial frictions. In the spirit of Krusell and Smith, they summarize the cross-sectional distribution with a finite set of features and use neural networks to approximate nonlinear perceived laws of motion. Their approach shows how flexible neural-network approximations can extend Krusell–Smith methods to environments with rich nonlinear aggregate dynamics.

Han et al. (2025) develop DeepHAM, a global solution method that uses neural networks to approximate value and policy functions and learned generalized moments to summarize the cross-sectional distribution. Kase et al. (2025) develop a neural-network framework for globally solving and estimating nonlinear heterogeneous-agent models, representing the cross section through a finite population of agents.[3] By contrast, Gu et al. (2026) ex-

[2] Section 2 develops this benchmark analytically and evaluates it computationally.

[3] These approaches also relate to simulation-based solution methods. Duffy and McNelis (2001) introduced neural networks into parameterized-expectations algorithms, Judd et al. (2011) developed simulation-based techniques for high-dimensional dynamic economies, and Valaitis and Villa (2024) use supervised learning to approximate expectation terms in macro-finance models.

ploit the continuous-time master-equation formulation, represent the distribution by a high-dimensional finite-dimensional state, and use neural networks to obtain a global solution.

I complement these approaches in three ways. First, I use a transformer architecture that produces pairwise attention weights across distributional tokens. These weights are part of the forecasting rule itself: they determine how information from different regions of the cross section is combined before predicting the next state. Attention is therefore part of the learned aggregation mechanism and also provides an economic diagnostic: after solving the model, the same weights can be used to study which regions of the state space receive greater relevance in equilibrium forecasts and how this relevance changes across aggregate states. I provide analytical conditions under which received attention coincides with normalized marginal policy sensitivity and, more generally, characterize how state-dependent attention enters the local sensitivity of aggregate policy functions to the distribution.

Second, I represent the cross-sectional distribution through economic tokens without committing ex ante to a small set of aggregate moments, while retaining a continuum-agent representation in the underlying numerical solution. Each token describes one region of the cross section, and the transformer forecasts the next tokenized distribution, which enters directly into the individual Bellman equation through continuation values. Because storing equilibrium objects over the full high-dimensional distribution space is infeasible, I construct a sparse library from model-generated distributions that cover the region of the state space relevant in equilibrium. The continuous transformer forecast is mapped to nearby library distributions to evaluate continuation values, while the underlying distribution is updated using Young (2010)'s non-stochastic transition rather than by simulating a finite population of agents.

Third, another distinction is the domain of the quantitative application. I apply the method to a heterogeneous-firm economy that builds on Khan and Thomas (2013), adding costly equity issuance and defaultable debt with endogenous firm default. These financial frictions create financially fragile regions of the firm distribution, providing a setting in which the aggregate relevance of the joint distribution of capital, debt, and productivity can be studied directly.

*Firm Dynamics and Financial Frictions.* A large literature studies firm dynamics with investment frictions and financial frictions, including Hopenhayn (1992), Hopenhayn and Rogerson (1993), Cooley and Quadrini (2001), Gomes (2001), Hennessy and Whited (2005), Clementi and Hopenhayn (2006), Cooper and Haltiwanger (2006), Hennessy and Whited (2007), Jermann and Quadrini (2012), Clementi and Palazzo (2016), Lanteri (2018), Ottonello and Winberry (2020), Villa (2026), and Ippolito et al. (2026). These papers show

that firm-level balance sheets, financing frictions, and default risk can shape financing, investment, and aggregate dynamics. The quantitative application in this paper belongs to this tradition: firms choose capital and defaultable debt, face costly equity issuance, and default endogenously. The application builds directly on Khan and Thomas (2008) and Khan and Thomas (2013), and is closely related to Khan et al. (2021). Khan and Thomas (2013) study a production-heterogeneity economy with partial capital irreversibility and collateralized borrowing. Khan et al. (2021) instead study defaultable debt with endogenous default and exit and equilibrium loan schedules that depend on firm and aggregate states, while abstracting from partial capital irreversibility and external equity issuance. My application combines partial capital irreversibility with defaultable debt and endogenous default, while allowing firms to raise costly external equity.

A useful lesson from this literature is that the economic importance of the firm distribution need not map directly into the dimension of the aggregate forecasting rule. Khan et al. (2021), for example, solve their stochastic economy using a Krusell–Smith approximation based on the first moment of the capital distribution, while showing that features of the firm age–size distribution can have first-order effects on aggregate responses to financial shocks. My application takes this aggregation question as a starting point. Rather than imposing a low-dimensional representation of the distribution ex ante, I let the transformer forecast the distribution itself and use attention to identify which regions of the cross section matter for aggregate dynamics and how their relevance varies with the aggregate state.

In the quantitative model considered here, this relevance is highly uneven. Firms near the endogenous default margin are particularly consequential for aggregate leverage, and financial fragility creates a state-dependent aggregation margin: changes in the composition of fragile balance-sheet regions also change how cross-sectional information is aggregated. During downturns, when default risk rises, attention shifts toward firms near the default margin. Thus, the contribution of the quantitative application is to identify where distributional heterogeneity matters, through which aggregation channel, and how its relevance changes with aggregate conditions.

**Outline.** Section 2 develops the economic interpretation of transformer attention through two analytical aggregation benchmarks and a general link between attention and aggregate policy sensitivity. Section 3 presents the Distributional State Transformer algorithm and validates it in a standard heterogeneous-household economy. Section 4 introduces the heterogeneous-firm economy with endogenous default, and Section 5 presents the quantitative results. Section 6 concludes.

# 2 Economic Aggregation and Transformer Attention

This section develops the economic interpretation of transformer attention using two analytically tractable aggregation problems. In the first, individual capital is aggregated before production, so only mean capital matters for aggregate accumulation and all regions of the cross section have the same marginal relevance. In the second, production takes place before aggregation, so diminishing returns at the individual-producer level make marginal relevance vary systematically with capital. The two economies therefore deliver sharp analytical benchmarks—uniform in the first case and non-uniform in the second—which I then confront with a richer full self-attention transformer.

The final subsection nests both benchmarks in a general local sensitivity result. It characterizes when transformer attention measures marginal economic relevance directly and, more generally, how a change in one region of the distribution can affect aggregate outcomes both through its own contribution and by changing the relevance of the rest of the cross section.

## 2.1 Mean-Sufficient Attention in a Two-Period Economy

I begin with a benchmark in which the cross-sectional distribution matters only through its mean. The environment is deliberately simple, so that the economic aggregation result is known analytically and the attention diagnostic can be evaluated against a sharp benchmark.

Consider a two-period economy with a continuum of households indexed by $i \in [0, 1]$. Household $i$ enters the period with capital $k_i$, and aggregate capital is $K = \int_0^1 k_i \, di$. There are no idiosyncratic productivity shocks. A representative firm rents aggregate capital and produces final output according to $Y = ZK^\alpha$ with $0 < \alpha < 1$, where $Z$ is aggregate productivity. The competitive return on capital is $R_K = \alpha Z K^{\alpha-1}$.

Household $i$ chooses next-period capital $k_i'$ to solve

$$\max_{k_i'} \left\{ \log c_i + \beta \log c_i' \right\},$$

subject to

$$c_i + k_i' = R_K k_i, \qquad c_i' = R_K' k_i'.$$

The next-period return $R_K'$ is taken as exogenous to the household and drops out of the saving first-order condition under logarithmic utility. Substituting the constraints into the objective yields

$$\max_{k_i'} \left\{ \log(R_K k_i - k_i') + \beta \log(R_K' k_i') \right\}.$$

The first-order condition gives the closed-form saving rule

$$k_i' = \kappa R_K k_i, \qquad \kappa \equiv \frac{\beta}{1+\beta}. \tag{1}$$

Thus individual saving is linear in individual capital. Aggregating equation (1) across households gives

$$K' = \int_0^1 k_i' \, di = \kappa R_K \int_0^1 k_i \, di = \kappa R_K K.$$

Using the firm first-order condition,

$$K' = \kappa \alpha Z K^{\alpha}. \tag{2}$$

Hence the entire distribution of capital is summarized by the scalar $K$. Conditional on $K$ and $Z$, reallocating capital across households has no effect on aggregate capital accumulation.

**The prediction problem.** The objective is to approximate the aggregation rule (2) mapping the current aggregate state into next-period aggregate capital. Let $X$ denote a collection of economic inputs describing the aggregate state. The transformer is a parameterized mapping

$$\widehat{y} = \mathcal{T}_\theta(X),$$

where $\theta$ denotes the network parameters. In the present example, $\widehat{y} = \widehat{K}'$ although the same architecture can be used to predict multiple aggregate objects.

The network is trained using observations generated from the analytical law

$$K' = \kappa \alpha Z \left( \sum_{j=1}^{J} \mu_j k_j \right)^{\alpha},$$

which serves only to construct the training targets and is not supplied to the model.

**Mapping the model to the transformer.** The input $X$ consists of a collection of vectors, called *tokens*, where each token represents one economic object. Representing the aggregate state as a collection of tokens allows the transformer to learn mappings over complex distributional states, where tokens may encode capital, debt, productivity, or other characteristics of the cross-sectional distribution.

In this example, I approximate the wealth distribution with $J$ wealth bins. Bin $j$ has representative capital $k_j$ and mass $\mu_j$. Together with aggregate productivity, these define

the input sequence

$$X = \{ \underbrace{x_1, \ldots, x_J}_{\text{wealth-bin tokens}} , \underbrace{x_Z}_{\text{aggregate-productivity token}} \}.$$

For $j = 1, \ldots, J$, the wealth-bin token is

$$x_j = \left( \underbrace{k_j, \mu_j}_{\text{wealth-bin state}} \middle| \underbrace{0}_{\text{aggregate productivity}} \middle| \underbrace{0}_{\text{token type}} \right),$$

while the aggregate-productivity token is

$$x_Z = \left( \underbrace{0, 0}_{\text{wealth-bin state}} \middle| \underbrace{Z}_{\text{aggregate productivity}} \middle| \underbrace{1}_{\text{token type}} \right).$$

Thus each token has dimension $d_x = 4$. The first two entries encode wealth-bin information, the third aggregate productivity, and the last identifies the token type.

**Processing the tokenized state.** The transformer processes the tokenized state through three successive operations. First, it constructs informative features for each economic object. Second, it learns how information should be exchanged across those objects to form a representation of the aggregate state. Finally, it maps this learned state representation into the desired economic prediction using a flexible nonlinear approximator. Figure 1 summarizes these three stages.

First, an *embedding layer* maps each raw token into a learned feature vector,

$$h_j = E_{\theta_E}(x_j), \qquad h_Z = E_{\theta_E}(x_Z).$$

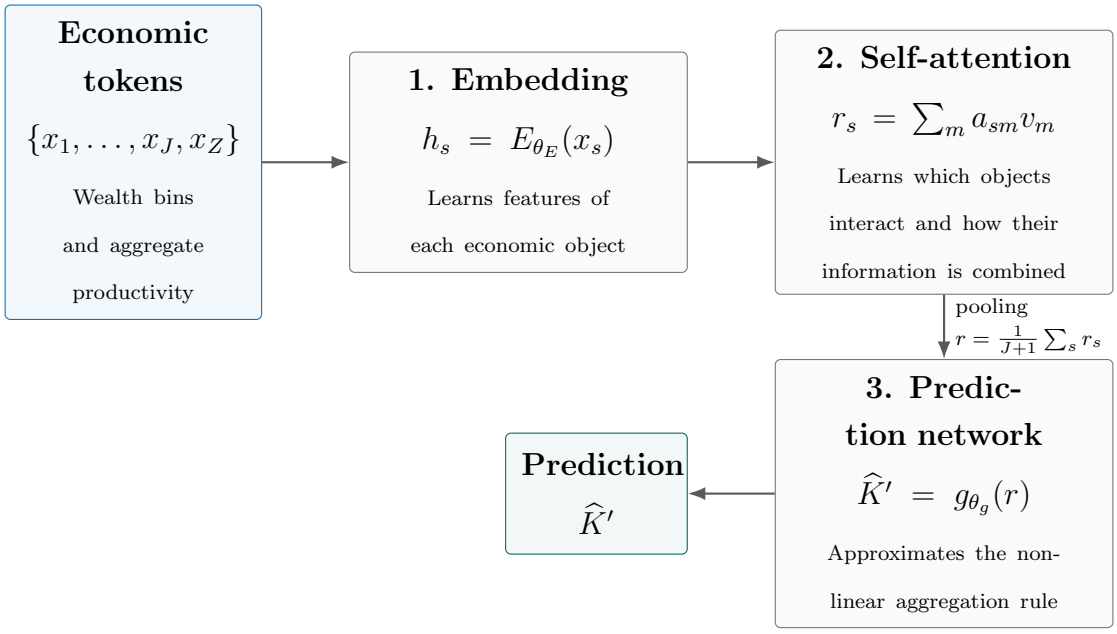
The raw wealth-bin token contains only $k_j$ and $\mu_j$, together with placeholders and a type indicator. These coordinates describe the bin, but need not be the most informative features for prediction or for comparing the token with the remaining economic objects. For example, in richer environments a token may contain capital, debt, and productivity, while prediction depends on nonlinear combinations such as leverage or net worth. The embedding therefore constructs a learned representation of each token before interactions across tokens are evaluated. A simple embedding is

$$E_{\theta_E}(x) = \sigma(W_E x + b_E),$$

where the researcher specifies the functional form and embedding dimension, while $W_E$ and

$b_E$ are estimated jointly with the remaining network parameters. Thus, the researcher does not specify which transformations of the raw token variables are economically relevant; these features are learned directly from the prediction objective.

Figure 1: Transformer Processing of the Tokenized Economic State




*Notes:* The embedding layer constructs learned features from each raw economic token. Self-attention then updates each token using information from the remaining wealth-bin and aggregate-productivity tokens. After pooling, a feedforward neural network maps the resulting representation into the prediction of next-period aggregate capital.

Second, a *self-attention layer* allows each embedded token to interact with every other token. In this benchmark, it learns which regions of the wealth distribution are most informative for predicting $K'$, how information should be aggregated across wealth bins, and how the distribution interacts with aggregate productivity. More generally, the attention mechanism learns both which economic objects matter for prediction and how information should flow across them. The output is therefore a collection of context-dependent token representations: each wealth-bin token is represented not in isolation, but conditional on the rest of the distribution and on $Z$.

Finally, the context-dependent token representations are pooled into a single aggregate representation and passed to a standard feedforward neural network,

$$\widehat{K}' = g_{\theta_g}(r).$$

At this stage, the relevant features have already been constructed and the economically important interactions identified by the attention layer. The role of the prediction network

is therefore to provide a flexible nonlinear approximation of the equilibrium mapping from the learned aggregate-state representation to the desired prediction. All network parameters are estimated jointly to minimize prediction errors.

The notation below makes the self-attention block in Figure 1 explicit. Starting from the embedded features $h_s$ produced in the first stage, the attention layer constructs query, key, and value vectors. Queries and keys determine the attention weights $a_{sm}$, while values contain the information combined according to those weights to form $r_s = \sum_m a_{sm} v_m$. Pooling the resulting token representations produces $r$, which is then passed to the prediction network in the third stage. The complete architecture is summarized in the box below.

**Self-attention architecture**

Let $\mathcal{I} = \{1, \ldots, J, Z\}$ index the $J$ wealth-bin tokens and the aggregate-productivity token. Given $x_s \in \mathbb{R}^{d_x}$, $s \in \mathcal{I}$, the transformer computes:

**1. Embedding**

$$h_s = E_{\theta_E}(x_s) \in \mathbb{R}^d.$$

**2. Self-attention**

$$\begin{aligned}
q_s &= W_Q h_s \in \mathbb{R}^d, \qquad k_s = W_K h_s \in \mathbb{R}^d, \qquad v_s = W_V h_s \in \mathbb{R}^d, \\
e_{sm} &= \frac{q_s^\top k_m}{\sqrt{d}} \in \mathbb{R}, \qquad m \in \mathcal{I}, \\
a_{sm} &= \frac{\exp(e_{sm})}{\sum_{\ell \in \mathcal{I}} \exp(e_{s\ell})}, \qquad a_{sm} \geq 0, \qquad \sum_{m \in \mathcal{I}} a_{sm} = 1, \\
r_s &= \sum_{m \in \mathcal{I}} a_{sm} v_m \in \mathbb{R}^d.
\end{aligned}$$

*Pooling*

$$r = \frac{1}{J+1} \sum_{s \in \mathcal{I}} r_s \in \mathbb{R}^d.$$

**3. Prediction network**

$$\widehat{K}' = g_{\theta_g}(r).$$

**Results.** Remark 1 provides the economic interpretation of the attention diagnostic. Attention weights capture the normalized marginal relevance of each distributional component for the aggregate policy function. In the mean-sufficient economy, an additional unit of capital has the same effect on $K'$ regardless of where it is located in the wealth distribution. In this example, mean sufficiency implies that each wealth bin has the same marginal effect on next-period aggregate capital. Therefore, the analytical aggregation result translates into a sharp prediction for the transformer: received attention is uniform across wealth bins.

**Mean Sufficiency and Uniform Attention**

*Remark* 1. Let $s_j = \mu_j k_j$ denote the contribution of wealth bin $j$ to aggregate capital, so that $K = \sum_{j=1}^{J} s_j$. The aggregate policy function (2) is $K' = \kappa\alpha Z \left(\sum_{j=1}^{J} s_j\right)^{\alpha}$. The normalized marginal contribution of wealth bin $j$ to aggregate capital is

$$\omega_j^E \equiv \frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \frac{1}{J}.$$

To isolate the attention channel, consider a restricted specialization of the transformer containing only the $J$ wealth-bin tokens. Set $v_j = s_j$ and let queries and keys be fixed bin embeddings, so that the attention weights

$$a_{qj} = \frac{\exp(e_{qj})}{\sum_{\ell=1}^{J} \exp(e_{q\ell})}, \qquad \sum_{j=1}^{J} a_{qj} = 1,$$

are independent of $(s_1, \ldots, s_J, Z)$. The mean-pooled representation is $r^W = \sum_{j=1}^{J} \omega_j^T s_j$, where normalized received attention is $\omega_j^T = \frac{1}{J}\sum_{q=1}^{J} a_{qj}$, with $\sum_{j=1}^{J} \omega_j^T = 1$.
Let $\widehat{K}' = g(r^W, Z)$, where $g$ is differentiable and $g_r(r^W, Z) \neq 0$. If the restricted transformer represents the aggregate policy exactly on an open admissible set, $\widehat{K}' = K'$, then

$$\omega_j^T = \frac{1}{J}\sum_{q=1}^{J} a_{qj} = \frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \omega_j^E = \frac{1}{J}, \qquad j = 1, \ldots, J. \tag{3}$$

*Proof.* See Appendix B.1. □

## 2.2 Non-Uniform Marginal Relevance with Individual Production

The previous example places production in a representative firm, so that individual capital stocks are aggregated before entering production. I now make one change to that environment: production takes place at the individual level. Preferences, timing, and aggregate productivity are otherwise unchanged.

Household $i$ operates a production unit with capital $k_i$ and technology $y_i = Zk_i^\alpha$, with $0 < \alpha < 1$. Capital is the only input, so $\alpha$ governs decreasing returns at the individual producer. Household $i$ chooses next-period capital $k_i'$ to solve

$$\max_{k_i'} \left\{ \log c_i + \beta \log c_i' \right\},$$

subject to

$$c_i + k_i' = Zk_i^\alpha, \qquad c_i' = Z'(k_i')^\alpha.$$

Substituting the constraints into the objective gives

$$\max_{k_i'} \left\{ \log(Zk_i^\alpha - k_i') + \beta \log(Z'(k_i')^\alpha) \right\}.$$

The first-order condition is

$$\frac{1}{Zk_i^\alpha - k_i'} = \frac{\beta\alpha}{k_i'},$$

which yields

$$k_i' = \kappa_\alpha Z k_i^\alpha, \qquad \kappa_\alpha \equiv \frac{\beta\alpha}{1+\beta\alpha}. \tag{4}$$

Unlike equation (1), individual saving is now nonlinear in current capital. Aggregating equation (4) across households gives

$$K' = \kappa_\alpha Z \int_0^1 k_i^\alpha \, di. \tag{5}$$

The contrast with the mean-sufficient benchmark is immediate:

$$\underbrace{K' \propto Z \left( \int_0^1 k_i \, di \right)^\alpha}_{\text{capital aggregated before production}} \qquad \text{versus} \qquad \underbrace{K' \propto Z \int_0^1 k_i^\alpha \, di}_{\text{capital transformed before aggregation}} .$$

Mean capital therefore no longer summarizes the cross section. With production at the individual level, decreasing returns make the marginal product of capital depend on producer size, with smaller producers having a higher marginal product of capital. The nonlinear

aggregation rule in (5) implies a sharp non-uniform benchmark for marginal relevance. Remark 2 formalizes this mapping between producer capital and transformer attention.

**Individual Production and Non-Uniform Attention**

*Remark* 2. As in Remark 1, let $s_j = \mu_j k_j$ denote the capital contribution of bin $j$. Using a $J$-bin approximation to the distribution, the aggregate policy function (5) is

$$K' = \kappa_\alpha Z \sum_{j=1}^{J} \mu_j k_j^\alpha = \kappa_\alpha Z \sum_{j=1}^{J} \mu_j^{1-\alpha} s_j^\alpha.$$

The normalized marginal contribution of capital bin $j$ to aggregate capital accumulation is

$$\omega_j^E \equiv \frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \frac{k_j^{\alpha-1}}{\sum_{\ell=1}^{J} k_\ell^{\alpha-1}}.$$

To isolate the attention channel, consider a restricted specialization of the transformer containing only the $J$ capital-bin tokens. Set $v_j = s_j$ and let queries and keys generate attention weights

$$a_{qj}(s) = \frac{\exp(e_{qj}(s))}{\sum_{\ell=1}^{J} \exp(e_{q\ell}(s))}, \qquad \sum_{j=1}^{J} a_{qj}(s) = 1,$$

where $s = (s_1, \ldots, s_J)$. Normalized received attention is $\omega_j^T(s) = J^{-1} \sum_{q=1}^{J} a_{qj}(s)$, with $\sum_{j=1}^{J} \omega_j^T(s) = 1$. At a given state $s$, hold these attention weights fixed under a local perturbation $\widetilde{s}$, and define the mean-pooled representation $r^W(\widetilde{s}; s) = \sum_{j=1}^{J} \omega_j^T(s) \widetilde{s}_j$. Let $\widehat{K}' = g(r^W(\widetilde{s}; s), Z)$, where $g$ is differentiable and $g_r(r^W, Z) \neq 0$. If the restricted transformer represents the aggregate policy to first order at $\widetilde{s} = s$, then

$$\omega_j^T(s) = \frac{1}{J} \sum_{q=1}^{J} a_{qj}(s) = \frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \omega_j^E(s) = \frac{k_j^{\alpha-1}}{\sum_{\ell=1}^{J} k_\ell^{\alpha-1}}, \qquad j = 1, \ldots, J. \tag{6}$$

*Proof.* See Appendix B.2. □

Because $0 < \alpha < 1$, normalized marginal relevance declines with capital: smaller producers have a higher marginal contribution to aggregate capital accumulation and therefore receive greater attention in the restricted representation.

## 2.3 Computational Validation

Remarks 1 and 2 provide analytical ground truth for the cross-sectional pattern of marginal relevance. Such a benchmark is typically unavailable in language applications, where attention can be informative about input relevance but need not coincide with alternative measures of importance (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). The two economies above instead make marginal relevance known analytically, allowing learned attention to be evaluated directly against an economic benchmark.

I evaluate these predictions using the same tokenization and network structure across both benchmarks: a full self-attention transformer trained only to predict $K'$. The numerical architecture is richer than the restricted representations in the remarks: embeddings, queries, keys, and values are learned from the state, attention may vary across realized states, and the resulting representation is passed through a nonlinear prediction network.[4] The analytical results therefore provide exact benchmarks for the attention–sensitivity mapping, while the numerical exercise asks whether this economic structure remains visible under a richer representation.

I simulate economies with $J = 20$ equal-mass capital bins, ordered from low to high capital within each economy. The transformer observes the full cross section and aggregate productivity $Z$, but not the analytical aggregation rule or its marginal sensitivities. Figure 2 reports received attention across capital bins on held-out economies.

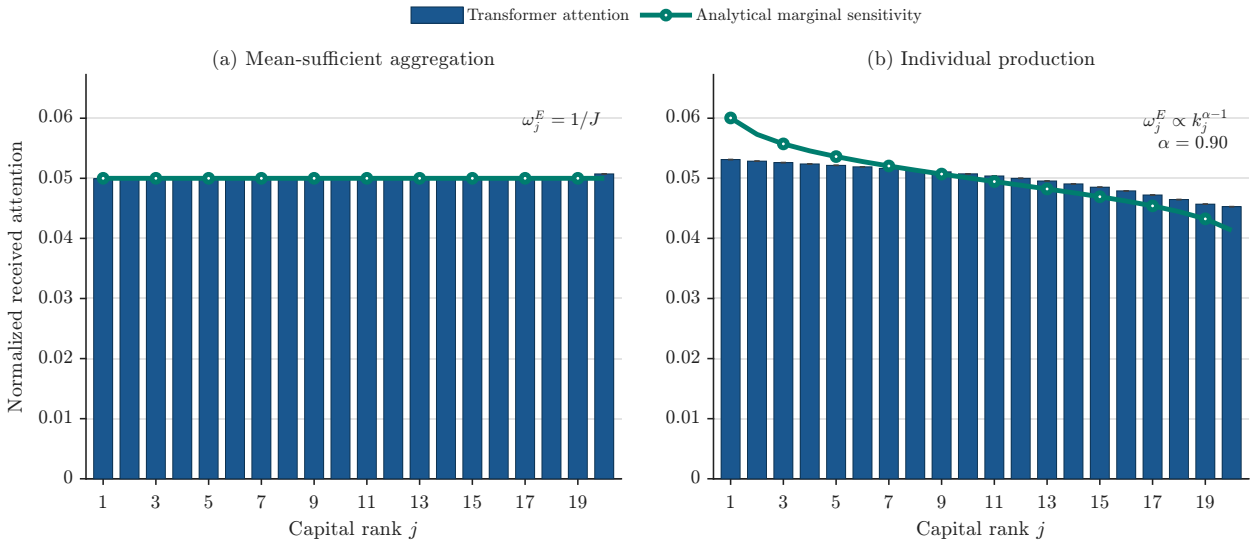
Figure 2 shows that learned attention closely tracks the analytical benchmarks for economic marginal relevance. In Panel (a), attention is essentially uniform across capital ranks, matching the $1/J$ prediction implied by mean sufficiency despite systematic differences in capital across bins. In Panel (b), because production takes place at the individual-producer level, decreasing returns make the marginal product of capital higher at smaller producers, so analytical marginal relevance declines with capital. Learned attention reproduces this non-uniform pattern closely, assigning greater weight to low-capital producers and progressively less weight to larger producers. Thus, when the economic structure of aggregation changes from uniform to heterogeneous marginal relevance, the learned attention profile changes accordingly and remains aligned with the corresponding analytical benchmark.

## 2.4 Policy Sensitivity and Attention Weights

The two examples above establish the attention–sensitivity link in environments where marginal relevance is known analytically. I now generalize that link. A perturbation to

[4] Architecture, training details, and out-of-sample fit are reported in Appendix A.1. Both transformers fit their respective analytical aggregation rules essentially exactly.

Figure 2: Economic Marginal Relevance and Transformer Attention




*Notes:* Bars report normalized received attention across capital-bin tokens; error bars denote 95% confidence intervals. Teal lines report normalized analytical marginal sensitivities. In Panel (a), mean sufficiency implies $\omega_j^E = 1/J$. In Panel (b), decreasing returns in individual production imply $\omega_j^E = k_j^{\alpha-1} / \sum_\ell k_\ell^{\alpha-1}$. Bins have equal mass and are ordered from low to high capital. The transformer is trained only to predict $K'$.

one component of the distribution can affect an aggregate prediction in two ways: by changing the information carried by the token representations and by changing how attention is allocated across tokens. Proposition 1 characterizes these two channels for a general differentiable aggregate policy function.

## Transformer Attention and Distributional Policy Sensitivity

**Proposition 1.** *Let $s = (s_1, \ldots, s_J)$ denote the distributional components entering a scalar aggregate policy function*

$$Y = F(s, z),$$

*where $z$ collects aggregate conditions. Consider an attention aggregation of the form*

$$r(s,z) = \sum_{m=1}^{J} \omega_m^T(s,z) v_m(s,z), \qquad \sum_{m=1}^{J} \omega_m^T(s,z) = 1,$$

*where $v_m(s,z) \in \mathbb{R}^d$ is the value representation associated with component $m$, and*

$\omega_m^T(s,z)$ *is its attention weight. Let the transformer prediction be*

$$\widehat{y} = g(r(s,z),z).$$

*If $g$, the attention weights, and the value representations are differentiable, and the transformer represents the equilibrium mapping exactly on an open set, $\widehat{y} = F(s,z)$, then*

$$\frac{\partial F(s,z)}{\partial s_j} = \nabla_r g(r,z)^\top \left[ \underbrace{\sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} v_m(s,z)}_{\text{attention-reallocation channel}} + \underbrace{\sum_{m=1}^{J} \omega_m^T(s,z) \frac{\partial v_m(s,z)}{\partial s_j}}_{\text{value channel}} \right], \qquad j = 1,\dots,J. \tag{7}$$

*Proof.* See Appendix B.3. □

Equation (7) separates two sources of policy sensitivity. The attention-reallocation channel captures changes in how the network weights the cross section when $s_j$ changes; the value channel captures changes in the economic information carried by the token representations. The gradient $\nabla_r g$ maps both responses into units of the aggregate outcome. The result is local and does not depend on the time horizon of the underlying economy, so the same decomposition applies to differentiable policy functions in recursive environments.

**Restricted attention and normalized policy sensitivity.** The two remarks above arise when these additional channels are restricted. Consider a local representation in which $d = 1$, values coincide with the underlying economic components, $v_m = s_m$, and attention weights are held fixed when evaluating a local perturbation of the distribution. Then $\frac{\partial \omega_m^T}{\partial s_j} = 0$ for the perturbation under consideration, and Proposition 1 reduces to

$$\frac{\partial F(s,z)}{\partial s_j} = g_r(r,z)\omega_j^T \quad \text{and} \quad \sum_{\ell=1}^{J} \frac{\partial F(s,z)}{\partial s_\ell} = g_r(r,z),$$

with $\sum_j \omega_j^T = 1$. Hence, whenever $g_r(r,z) \neq 0$,

$$\omega_j^T = \omega_j^E(s,z) \equiv \frac{\partial F(s,z)/\partial s_j}{\sum_{\ell=1}^{J} \partial F(s,z)/\partial s_\ell}. \tag{8}$$

Under this restricted transformer representation, attention therefore recovers normalized marginal policy sensitivity exactly.

**The two-period benchmarks as special cases.** Remark 1 follows immediately. With $s_j = \mu_j k_j$, the mean-sufficient policy is

$$K' = \kappa \alpha Z \left( \sum_{j=1}^{J} s_j \right)^{\alpha},$$

so that $\partial K'/\partial s_j$ is identical across bins. Equation (8) therefore gives $\omega_j^T = \omega_j^E = \frac{1}{J}$.

Remark 2 gives the corresponding non-uniform case. With individual production,

$$K' = \kappa_\alpha Z \sum_{j=1}^{J} \mu_j^{1-\alpha} s_j^{\alpha} \quad \text{and hence} \quad \frac{\partial K'}{\partial s_j} = \kappa_\alpha \alpha Z k_j^{\alpha-1}.$$

Normalized marginal relevance is therefore

$$\omega_j^T = \omega_j^E = \frac{k_j^{\alpha-1}}{\sum_{\ell=1}^{J} k_\ell^{\alpha-1}}.$$

Thus, uniform attention in the first benchmark and greater attention to smaller producers in the second are two special cases of the same attention–sensitivity mapping.

**From attention to the DST.** Proposition 1 characterizes how a self-attention aggregation transmits distributional changes to an aggregate prediction. The DST developed in the next section embeds this mechanism in a richer transformer architecture. In particular, residual connections and tokenwise feed-forward blocks provide additional paths through which the economic state can affect the forecast without operating through the self-attention update itself. The proposition therefore isolates the two economic channels operating within attention—changes in token values and the reallocation of attention—while the full architecture also permits a remaining-network channel. Section 3 turns this architecture into a recursive solution method, and the quantitative application later measures the relative importance of these channels in the heterogeneous-firm economy.

# 3 The Distributional State Transformer

The examples in Section 2 illustrate how attention organizes distributional information, while Proposition 1 links this aggregation to local policy sensitivity. This section turns the same architecture into a general solution method. The *Distributional State Transformer* (DST) is an iterative algorithm for recursive heterogeneous-agent economies in which the cross-sectional distribution is part of the aggregate state. It represents the distribution and aggregate conditions as tokens, learns a perceived law of motion for the distribution from model-consistent simulations, and inserts that law directly into the Bellman equation.

## 3.1 Recursive Economy and Tokenized State

Let $a \in \mathcal{X}$ denote an individual state and let $z_t \in \mathcal{Z}$ collect the aggregate shocks. The cross-sectional distribution of individual states at date $t$ is denoted by $\mu_t$. The recursive aggregate state is

$$\Omega_t = (\mu_t, z_t)\,.$$

Given individual policy rules $g$, the model implies the distributional transition

$$\mu_{t+1} = \Gamma\left(\mu_t, z_t, z_{t+1}; g\right),$$

where $\Gamma$ is the transition operator induced by individual policy rules, idiosyncratic shocks, and the transition of the aggregate shocks from $z_t$ to $z_{t+1}$.

The purpose of the DST is to approximate the equilibrium transition of this distribution without reducing it to a small, prespecified set of aggregate moments. To do so, the method represents the current aggregate state as a fixed-length sequence of economic tokens,

$$X_t = \tau(\Omega_t) = \{x_{1,t}, \ldots, x_{J,t}\}\,, \qquad x_{j,t} \in \mathbb{R}^{d_x},$$

where $\tau$ denotes the tokenization map and $J$ is the fixed number of tokens.

A token is an economic object written in numerical format. Distribution tokens represent regions of the cross section, such as wealth bins or firm balance-sheet cells. For example, in the two-period economy of Section 2.1, each distribution token represents a wealth bin with its capital and mass, while aggregate productivity is included as a separate token. In the recursive Krusell–Smith economy of Section 3.7, each wealth-bin token contains the bin mass and conditional moments of assets and idiosyncratic productivity.

Tokens may represent also aggregate shocks, prices, policy variables, or structural parameters if the purpose is also to estimate the model, as in Kase et al. (2025). A generic

token has the form

$$x_{j,t} = (\text{economic content} \mid \text{placeholders} \mid \text{type indicators}) .$$

Placeholders allow economically different objects to have the same numerical dimension, while the type indicators identify the economic role of each token.

Let $m_{j,t}$ denote the distributional coordinates associated with region $j$ whose transition is forecast by the transformer. These coordinates need not coincide with the economic content supplied to the corresponding input token $x_{j,t}$. Depending on the application, the input token may contain cell masses and conditional moments, while the forecast coordinates may use mass-weighted moments or other statistics describing the same region of the distribution. Stacking these coordinates gives

$$m_t \equiv \text{vec}\,(m_{1,t}, \ldots, m_{J,t}) .$$

The vector $m_t$ is the distributional representation of $\mu_t$ whose transition is learned by the transformer. Aggregate moments can instead be computed from $m_t$.

## 3.2 The Perceived Distributional Law

The transformer maps the current tokenized state $X_t$ into a perceived law of motion for the distribution. Because distributional states are persistent, the network predicts the change in the distributional vector,

$$\Delta m_{t+1} \equiv m_{t+1} - m_t,$$

rather than its next-period level. This residual formulation asks the network to learn the typically smaller and less variable adjustment to the current distribution. The resulting forecast is

$$\widehat{m}_{t+1} = m_t + \mathcal{T}_{\theta,m}(X_t), \tag{9}$$

where $\mathcal{T}_{\theta,m}(X_t)$ denotes the transformer prediction of $\Delta m_{t+1}$.

Some applications also require auxiliary equilibrium objects inside the recursive problem. Let $\psi_t$ collect these objects, with prediction $\widehat{\psi}_t = \mathcal{T}_{\theta,\psi}(X_t)$. For example, in the heterogeneous-firm economy of Section 4, it is convenient to track the log marginal utility as an auxiliary target, $\psi_t = \ell_t \equiv \log U_C(C_t)$, from which consumption, wages, and stochastic discount factors are recovered.

Training targets are generated by applying the distributional transition operator to the current distribution under the current policy functions. The parameters $\theta$ of the transformer

are found to minimize the following loss

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{T}_{\text{tr}}|} \sum_{t \in \mathcal{T}_{\text{tr}}} \left[ \|\Delta m_{t+1} - \mathcal{T}_{\theta,m}(X_t)\|_2^2 + \lambda_\psi \|\psi_t - \mathcal{T}_{\theta,\psi}(X_t)\|_2^2 \right],$$

where $\lambda_\psi = 0$ when no auxiliary target is used. Inputs and targets are standardized during training and transformed back into economic units before entering the recursive problem.

## 3.3 Sparse Library of Distributional States

A Cartesian grid over all coordinates of the tokenized distribution is infeasible, as even a small number of grid points along each coordinate would generate an exponentially large aggregate state space. A complementary approach is to solve the model directly over the ergodic set generated by stochastic simulation. For example, Valaitis and Villa (2024) combine stochastic simulation with parameterized expectations in a representative-agent setting. Han et al. (2025) solve heterogeneous-agent models by training neural networks over simulated equilibrium trajectories, while Kase et al. (2025) approximate the distribution with a finite simulated population and train over the resulting simulated state space.

Here, instead, I retain complete distributions over the underlying individual-state grid and use the non-stochastic simulation method of Young (2010), which preserves the continuum-agent representation. The remaining challenge is to approximate equilibrium objects over the resulting high-dimensional distribution space. I do so by replacing the Cartesian grid with a sparse library of feasible distributions, $\mathcal{L} = \{\mu^1, \ldots, \mu^S\}$, where each node $\mu^s$ is an entire cross-sectional distribution over the underlying individual-state grid. The transformer operates on the tokenized representation and produces a continuous distributional forecast, while the sparse library is used to interpolate stored equilibrium objects at that forecast. This sparse representation creates two challenges: selecting a small set of distributions that adequately covers the relevant region of the aggregate state space, and interpolating equilibrium objects between sparse library nodes.

**Selecting the library nodes.** The library $\mathcal{L}$ may be fixed at initialization or enriched during the outer fixed-point iteration with newly simulated distributions that expand its coverage of the aggregate state space. To construct the initial library, I first generate a candidate set of feasible distributions,

$$\mathcal{C}_{\text{sim}} = \{\nu^1, \ldots, \nu^N\}, \qquad N \geq S,$$

from a bootstrap simulation under a preliminary perceived law of motion. The candidate set may be augmented with anchor distributions chosen to improve coverage of economically relevant regions that are rarely visited in the bootstrap simulation,

$$\mathcal{C} = \mathcal{C}_{\text{sim}} \cup \mathcal{C}_{\text{anc}}.$$

The library $\mathcal{L} = \{\mu^1, \ldots, \mu^S\}$ is then selected as a space-filling subset of $\mathcal{C}$. Anchor distributions therefore enter the candidate set but are retained only when they improve coverage.

To compare candidate distributions, let

$$m(\nu) = \text{vec}\big(m_1(\nu), \ldots, m_J(\nu)\big)$$

denote the stacked distributional coordinates associated with $\nu$. The selection metric may augment these distributional coordinates with aggregate moments that are particularly relevant for continuation values or prices. For concreteness, in the heterogeneous-firm application of Sections 4–5, cell $j$ corresponds to a capital–debt–productivity region $\mathcal{B}_j$, and its forecast coordinates are the mass-weighted moments

$$m_j(\nu) = \big(\omega_j(\nu)\bar{k}_j(\nu), \omega_j(\nu), \omega_j(\nu)\bar{b}_j(\nu), \omega_j(\nu)\bar{\varepsilon}_j(\nu)\big), \qquad \omega_j(\nu) = \int_{\mathcal{B}_j} d\nu.$$

The preliminary perceived law in this application is initialized with a low-dimensional Krusell–Smith approximation, and the selection coordinates additionally include aggregate capital ($K(\nu)$), debt ($B(\nu)$), and leverage ($L(\nu)$). After standardization across the candidate set, the resulting coordinate vector is

$$c(\nu) = \left[\frac{m(\nu) - \bar{m}}{s_m}, \lambda_K \frac{K(\nu) - \bar{K}}{s_K}, \lambda_B \frac{B(\nu) - \bar{B}}{s_B}, \lambda_L \frac{L(\nu) - \bar{L}}{s_L}\right],$$

where the $\lambda$'s determine the relative weight placed on the aggregate coordinates.

The $S$ library nodes are selected from $\mathcal{C}$ using a farthest-point rule. Define

$$d(\nu, \nu') = \frac{1}{\sqrt{d_c}} \left\| c(\nu) - c(\nu') \right\|,$$

where $\|\cdot\|$ denotes the Euclidean norm and $d_c$ is the dimension of $c(\nu)$. The first node is the candidate closest to the center of the candidate cloud,

$$\mu^1 \in \arg\min_{\nu \in \mathcal{C}} \left\| c(\nu) - \bar{c} \right\|, \qquad \bar{c} = \frac{1}{|\mathcal{C}|} \sum_{\nu \in \mathcal{C}} c(\nu).$$

Given $\mathcal{L}_{r-1} = \{\mu^1, \ldots, \mu^{r-1}\}$, subsequent nodes satisfy

$$\mu^r \in \arg\max_{\nu \in \mathcal{C}} \min_{\mu \in \mathcal{L}_{r-1}} d(\nu, \mu), \qquad r = 2, \ldots, S.$$

Thus, each new node is the candidate farthest from the portion of the state space already represented in the library. If the library is enriched during the outer iteration, the same criterion is applied to newly simulated distributions.

**Interpolating over the sparse library.** For a current state, the transformer produces a continuous forecast $\widehat{m}'$. Because $\widehat{m}'$ will generally lie between library nodes, the method approximates equilibrium objects using nearby stored distributions. Let $\Xi(m)$ denote the coordinate vector obtained from a distributional moment vector $m$ using the same standardization and additional aggregate coordinates used to represent the library nodes through $c(\mu^r)$. Thus, $\Xi(\widehat{m}')$ expresses the continuous transformer forecast in the same coordinate system as the stored library distributions, allowing distances between the forecast and each library node to be computed directly. The distance between the forecast and library node $r$ is

$$d_r(\widehat{m}') = \frac{1}{\sqrt{d_c}} \left\| \Xi(\widehat{m}') - c(\mu^r) \right\|,$$

where $d_c$ is the dimension of the standardized coordinate vector. Let $\mathcal{N}(\widehat{m}')$ denote the set of nearest library nodes. In the absence of an exact match, the interpolation weights are

$$\omega_r(\widehat{m}') = \frac{d_r(\widehat{m}')^{-2}}{\sum_{\ell \in \mathcal{N}(\widehat{m}')} d_\ell(\widehat{m}')^{-2}}, \qquad r \in \mathcal{N}(\widehat{m}'), \tag{10}$$

so that $\sum_{r \in \mathcal{N}(\widehat{m}')} \omega_r(\widehat{m}') = 1$. An exact match receives unit weight. For any equilibrium object $F$ stored at the library nodes, its value at the predicted distribution is approximated by

$$\widehat{F}(\widehat{m}') = \sum_{r \in \mathcal{N}(\widehat{m}')} \omega_r(\widehat{m}') F(\mu^r).$$

Thus, the continuous transformer forecast determines which stored distributions enter the approximation and how much weight each receives. The same interpolation weights can be applied to continuation values, policies, default probabilities, and other equilibrium objects stored on the library.

## 3.4 Embedding the Law in the Bellman Equation

The perceived distributional law enters the recursive problem. At library node $\mu^s$ and aggregate state $z$, the transformer produces the distributional forecast

$$\widehat{m}'_{sz} = m(\mu^s) + \mathcal{T}_{\theta,m}(X_{sz}).$$

As described in Section 3.3, this forecast determines a set of nearby library nodes $\mathcal{N}(s,z)$ and interpolation weights $\{\omega_{sz,r}\}_{r\in\mathcal{N}(s,z)}$.

Let $a$ denote the current individual state and let $\alpha$ denote an individual choice. Suppressing application-specific constraints, the continuation value associated with $(a,\alpha)$ is

$$\mathcal{C}_{sz}(a,\alpha) = \sum_{z'} P_z(z,z') \sum_{r\in\mathcal{N}(s,z)} \omega_{sz,r}\, \mathbb{E}\left[M_{sz,rz'} V_r(a',z') \mid a,\alpha,z,z'\right]. \tag{11}$$

Here $V_r(a',z')$ is the value function stored at library node $\mu^r$, and the expectation integrates over the transition of the individual state. The stochastic discount factor $M_{sz,rz'}$ is determined by the economic model and may depend on auxiliary objects predicted by the transformer.

The Bellman equation at $(\mu^s,z)$ is then

$$V_s(a,z) = \max_{\alpha\in\mathcal{A}(a,z,\mu^s)} \left\{ u\left(a,\alpha,z,\mu^s;\widehat{\psi}_{sz}\right) + \mathcal{C}_{sz}(a,\alpha) \right\}.$$

The learned law therefore affects current decisions through changes in the predicted library nodes and interpolation weights entering the continuation value.

## 3.5 Equilibrium Iteration

The equilibrium is a fixed point between the perceived distributional law and the policy functions, as in Krusell and Smith (1998). Generally, the sparse library $\mathcal{L}$ may be enriched during the outer iteration. In the implementation used below, it is constructed before the final equilibrium iteration and then held fixed.

At outer iteration $k$, the transformer is trained on transitions generated by the previous policy functions. It is then evaluated at each current aggregate state $(\mu^s,z)$, producing the forecast $\widehat{m}'_{sz}$, the neighboring library nodes $\mathcal{N}(s,z)$, the interpolation weights $\omega_{sz,r}$, and any auxiliary predictions $\widehat{\psi}_{sz}$. Holding these objects fixed, the individual Bellman equation is solved to convergence. The resulting policy functions are then simulated using the non-stochastic distributional transition of Young (2010), generating the training sample for the

next outer iteration. Computationally, the transformer training step is parallelized on GPU, while the Bellman solves are parallelized across library nodes and aggregate states on CPU.[5]

**Algorithm 1** DST Equilibrium Iteration

Construct a bootstrap policy under a simple perceived law of motion.
Simulate the policy using the non-stochastic distributional transition operator of Young (2010).
Select and fix the sparse library $\mathcal{L} = \{\mu^1, \ldots, \mu^S\}$.
Initialize value and policy functions at all library nodes.
Construct the initial training sample $\{X_t, \Delta m_{t+1}, \psi_t\}_{t \in \mathcal{I}_{\text{tr}}}$.
**for** outer iteration $k = 1, 2, \ldots, K_{\max}$ **do**
    Train or fine-tune $\mathcal{T}_{\theta^{(k)}}$ on the current sample. ▷ **Parallelized on GPU**
    Evaluate $\mathcal{T}_{\theta^{(k)}}$ at every $(\mu^s, z)$ and compute $\widehat{m}'_{sz}$, $\mathcal{N}(s, z)$, $\omega_{sz,r}$, and $\widehat{\psi}_{sz}$.
    Solve the individual problem at all library nodes. ▷ **Parallelized on CPU**
    Simulate the resulting policy functions using the non-stochastic distributional transition operator.
    Update the training sample with the simulated transitions.
    Compute the policy change $\Delta_g^{(k)}$ and the perceived-law change $\Delta_{\mathcal{T}}^{(k)}$.
    **if** $\Delta_g^{(k)} < \tau_g$ and $\Delta_{\mathcal{T}}^{(k)} < \tau_{\mathcal{T}}$ **then**
        Stop.
    **end if**
**end for**

The policy criterion $\Delta_g^{(k)}$ measures the change in policy functions across successive outer iterations on the fixed library. The law criterion $\Delta_{\mathcal{T}}^{(k)}$ measures the corresponding change in transformer forecasts at the same states.

## 3.6 Attention Diagnostics

The perceived law in Section 3.2 is produced by repeatedly combining the input tokens through self-attention. The same attention weights used to construct the transformer outputs can therefore be inspected to study how the network organizes the economic state.

For observation $t$, layer $\ell$, and head $h$, let

$$A_t^{\ell h} = \left(a_{jm,t}^{\ell h}\right)_{j,m=1}^{J}, \qquad \sum_{m=1}^{J} a_{jm,t}^{\ell h} = 1. \tag{12}$$

[5] Transformer training is parallelized on GPU, using CUDA or Apple's Metal backend; see Vaswani et al. (2017) for the parallel attention architecture and Paszke et al. (2019) for the automatic-differentiation framework used in the implementation.

The row index $j$ denotes the query token whose representation is being updated, while the column index $m$ denotes the key-value token used in that update. Thus, $a^{\ell h}_{jm,t}$ measures the weight placed on token $m$ when updating token $j$ in layer $\ell$ and head $h$.

To summarize attention across test observations, layers, and heads, define the average attention matrix

$$\bar{A} = (\bar{a}_{jm})_{j,m=1}^{J}, \qquad \bar{a}_{jm} = \frac{1}{|\mathcal{I}_{\text{te}}|LH} \sum_{t \in \mathcal{I}_{\text{te}}} \sum_{\ell=1}^{L} \sum_{h=1}^{H} a^{\ell h}_{jm,t}.$$

I use two attention diagnostics. The first diagnostic measures *received attention*. For token $m$ in a group of economic tokens $\mathcal{G}$, define

$$\text{Rec}_m(\mathcal{G}) = \frac{\sum_{j \in \mathcal{Q}} \bar{a}_{jm}}{\sum_{m' \in \mathcal{G}} \sum_{j \in \mathcal{Q}} \bar{a}_{jm'}}, \qquad m \in \mathcal{G}, \tag{13}$$

where $\mathcal{Q}$ is the set of query tokens included in the diagnostic. This statistic measures how the attention received by group $\mathcal{G}$ is distributed across its members. It satisfies $\sum_{m \in \mathcal{G}} \text{Rec}_m(\mathcal{G}) = 1$. If the $J_{\mathcal{G}}$ tokens in the group enter symmetrically, the corresponding benchmark is $1/J_{\mathcal{G}}$.

The second diagnostic measures *directional group attention*. For two token groups $\mathcal{A}$ and $\mathcal{B}$, define

$$\text{Attn}(\mathcal{A} \to \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \sum_{m \in \mathcal{B}} \bar{a}_{jm}. \tag{14}$$

This statistic measures the average share of attention that query tokens in $\mathcal{A}$ assign to key-value tokens in $\mathcal{B}$. Because the first group supplies the queries and the second supplies the key-value tokens,

$$\text{Attn}(\mathcal{A} \to \mathcal{B}) \neq \text{Attn}(\mathcal{B} \to \mathcal{A})$$

in general. In the heterogeneous-firm application that follows, these attention measures show how the transformer combines regions of the distribution associated with different levels of capital, debt, productivity, and default risk. In this sense, attention could be interpreted as a description of the network's learned aggregation rule.

## 3.7 Validation in a One-Asset Household Economy

I validate the DST in a standard one-asset heterogeneous-household economy for which a one-moment Krusell–Smith approximation provides a natural reference benchmark. Households differ in assets $a$ and idiosyncratic labor productivity $\varepsilon$, while aggregate productivity is $Z$.

Labor supply is fixed, and the household problem is

$$V(a, \varepsilon, Z, \mu) = \max_{a' \geq \underline{a}} \{u(c) + \beta \mathbb{E}\left[V(a', \varepsilon', Z', \mu') \mid \varepsilon, Z\right]\},$$

subject to

$$c + a' = R(Z, \mu)a + w(Z, \mu)\varepsilon.$$

A representative firm produces

$$Y = ZK^{\alpha}N^{1-\alpha}, \qquad K = \int a \, d\mu, \qquad N = \int \varepsilon \, d\mu,$$

with factor prices given by marginal products. The distribution $\mu$ evolves endogenously under household saving policies.

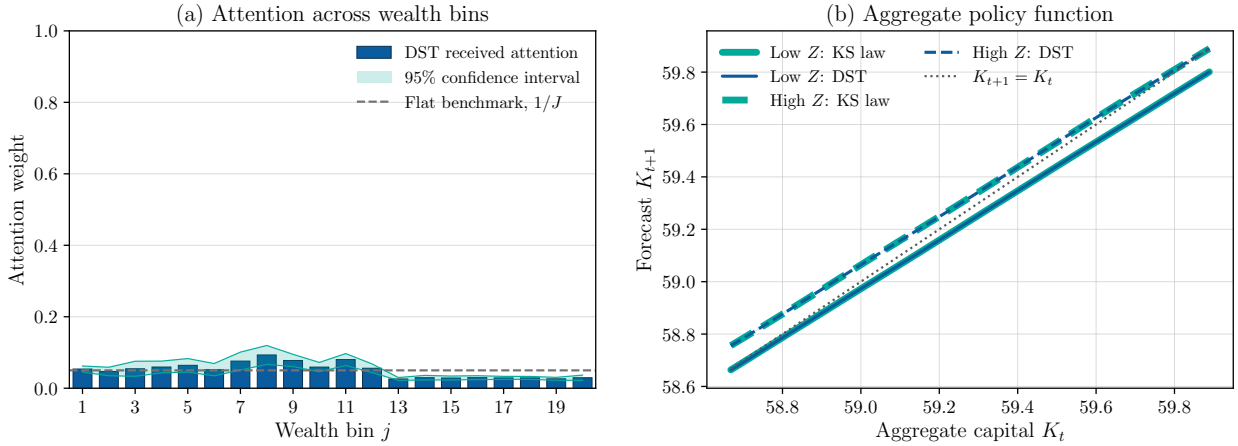
I solve the same economy in two ways. Implementation details are reported in Appendix A.2. The reference solution uses a standard Krusell–Smith perceived law based on aggregate capital $K$. The DST instead takes the cross-sectional distribution as the aggregate state and applies the algorithm described above. The distribution is represented by $J = 20$ wealth-bin tokens containing bin mass and conditional moments of assets and productivity, and the transformer forecasts the change in the tokenized distribution according to (9). Aggregate capital is not a direct prediction target; it is recovered by aggregating the predicted distribution.[6]

The comparison asks whether a method that conditions on and forecasts the distribution itself recovers the approximate aggregation captured by the one-moment Krusell–Smith solution. To assess the robustness of this comparison across simulated aggregate histories, I solve the economy ten times, each time using a distinct realization of aggregate productivity shocks and a 50,000-period simulation. Figure 3 summarizes the resulting validation exercise. The aggregate-capital forecast implied by the DST distributional law achieves $R^2 = 1.00000$ to five decimal places, and the implied aggregate policy functions are nearly identical to those from the one-moment Krusell–Smith solution. At the same time, received attention across wealth bins remains broadly close to the uniform benchmark across runs, with only modestly greater attention over intermediate-low wealth bins and somewhat lower attention in the upper tail. Thus, despite conditioning on and forecasting the full cross-sectional distribution, the DST recovers the familiar approximate-aggregation structure without imposing aggregate capital as the state variable ex ante.

The full self-attention architecture also constructs directional interactions between every

[6]The transformer predicts all 80 distributional coordinates. The discretization, tokenization, admissibility projection, library construction, and numerical implementation are described in Appendix A.2.

Figure 3: DST Validation in a One-Asset Household Economy




*Notes:* Panel (a) reports received attention across wealth-bin tokens. Bars denote mean received attention across ten independent model solutions, each based on a distinct realization of aggregate productivity shocks and a 50,000-period simulation; shaded regions report 95% confidence intervals across runs. The dashed horizontal line denotes the flat-attention benchmark, $1/J$. Panel (b) compares the aggregate-capital policy functions implied by the DST-predicted distribution with the corresponding one-moment Krusell–Smith laws, separately for low and high aggregate productivity. The dotted line is $K_{t+1} = K_t$.

pair of tokens. In particular, each wealth-bin token can attend to aggregate productivity, and the aggregate-productivity token can attend to each wealth bin. Appendix Figure 7 in Appendix C reports these two directional attention profiles. Both remain broadly flat across wealth bins, indicating that no particular region of the wealth distribution interacts with aggregate productivity substantially more than the others, in either direction.[7]

# 4 Heterogeneous Firms and Endogenous Default

I next apply the DST to a heterogeneous-firm economy with capital and debt. The environment builds on the production-heterogeneity framework of Khan and Thomas (2013): firms differ in idiosyncratic productivity, choose physical capital subject to partial irreversibility, face borrowing constraints, and operate under aggregate productivity shocks. I extend that benchmark by adding costly equity issuance and endogenous firm default. This makes the joint distribution of capital, debt, and productivity relevant for aggregate dynamics. Appendix A.3 gives the computational details for this application.

[7]Directional attention is not symmetric. $Z \to W_j$ measures the weight placed on wealth-bin token $W_j$ when updating the representation of aggregate productivity $Z$, whereas $W_j \to Z$ measures the weight placed on $Z$ when updating wealth-bin token $W_j$. Because attention is normalized separately over the key-value tokens available to each query, these two weights need not coincide.

## 4.1 Model

**Environment.** Time is discrete and one model period is one year. Aggregate productivity $Z_t$ follows an AR(1) process in logs, i.e. $\log Z_t = \rho_Z \log Z_{t-1} + \sigma_Z \eta_t^Z$, with $\eta_t^Z \sim \mathcal{N}(0,1)$. A continuum of firms is indexed by idiosyncratic productivity $\varepsilon_t$, physical capital $k_t$, and one-period debt $b_t$. Idiosyncratic productivity $\varepsilon_t$ also follows an AR(1) process in logs, i.e. $\log \varepsilon_t = \rho_\varepsilon \log \varepsilon_{t-1} + \sigma_\varepsilon \eta_t^\varepsilon$, with $\eta_t^\varepsilon \sim \mathcal{N}(0,1)$. The production technology of an active firm is

$$y_t(\varepsilon, k, n; Z) = Z_t \varepsilon_t k_t^\alpha n_t^\nu, \qquad \alpha + \nu < 1. \tag{15}$$

Operating profits, net of a fixed operating cost $f$, are

$$\pi_t(\varepsilon, k; Z, w) = Z_t \varepsilon_t k_t^\alpha n_t^\nu - w_t n_t - f. \tag{16}$$

Given the real wage $w_t$, labor demand is given by the marginal product of labor:

$$n_t(\varepsilon, k; Z, w) = \left( \frac{\nu Z_t \varepsilon_t k_t^\alpha}{w_t} \right)^{1/(1-\nu)}. \tag{17}$$

The representative household owns both firms and competitive lenders. Its preferences are

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[ \frac{C_t^{1-\gamma}}{1-\gamma} - \chi N_t \right],$$

with the logarithmic case obtained as $\gamma \to 1$. Its budget constraint is

$$C_t = w_t N_t + \int \mathcal{D}_t(\varepsilon, k, b)\, d\mu_t(\varepsilon, k, b) + \Pi_t^L, \tag{18}$$

where $\mu_t$ denotes the time-$t$ distribution of firms over capital, debt, and idiosyncratic productivity, $\mathcal{D}_t(\varepsilon, k, b)$ denotes the net payout to shareholders by a firm in state $(\varepsilon, k, b)$, and $\Pi_t^L$ denotes the realized net payout from competitive lending. Perfect competition among lenders implies that newly issued debt is priced to satisfy lenders' expected discounted zero-profit condition. Because the household owns both firms and lenders, debt issuance and repayment are internal financial transfers and cancel upon consolidation. Combining (18) with firm and lender cash flows therefore yields the aggregate resource constraint (26).

The representative household's labor-supply first-order condition is

$$w_t = \chi C_t^\gamma, \tag{19}$$

so aggregate consumption determines the wage. Because the household owns both firms and lenders, the stochastic discount factor relevant for their intertemporal payoffs is

$$M_{t,t+1} = \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\gamma}. \tag{20}$$

**Firm finance and default.** A firm enters period $t$ with idiosyncratic productivity, capital, and debt $(\varepsilon, k, b)$. Upon entering the period, equity holders choose whether to default or continue. A defaulting firm exits before producing; after physical depreciation, its remaining capital $(1-\delta)k$ is liquidated at the same resale price $\theta_k$ that applies to disinvestment. Conditional on continuation, the firm operates, repays current debt $b$, chooses next-period capital $k'$, and issues one-period debt $b'$. New debt sells at price $q_t(\varepsilon, k', b')$. Let net investment be

$$i(k, k') \equiv k' - (1 - \delta)k.$$

Following Khan and Thomas (2013), capital is subject to partial irreversibility. The real resource cost of investing from $k$ to $k'$ is

$$\mathcal{I}(k, k') = \begin{cases} i(k, k'), & i(k, k') \geq 0, \\ \theta_k i(k, k'), & i(k, k') < 0, \end{cases} \tag{21}$$

with $\theta_k < 1$. Thus, the gross resale value of fully liquidated capital is $\theta_k(1-\delta)k$. The payout before external-equity costs is therefore

$$d_t = \pi_t(\varepsilon, k; Z, w) + q_t(\varepsilon, k', b')b' - b - \mathcal{I}(k, k'). \tag{22}$$

Debt is subject to a borrowing constraint,

$$b' \leq \bar{b}(k') \equiv \theta_b \theta_k k'.$$

I keep this constraint to preserve comparability with the debt-capacity margin in Khan and Thomas (2013). The constraint limits debt issuance ex ante, whereas the recovery rate defined below determines lenders' payoff ex post in default. The two need not coincide: debt permitted by the contractual limit may be only partially collateralized and is priced accordingly. In the baseline calibration with endogenous default, the constraint is effectively slack: the stationary share of firms at or near the borrowing limit is numerically zero. Debt pricing therefore disciplines leverage before the borrowing limit becomes active, because firms approaching states with high default risk face sharply lower bond prices.

If the payout in (22) is negative, the firm raises external equity and pays a quadratic issuance cost. Equity holders receive

$$\mathcal{D}(d_t) = \begin{cases} d_t, & d_t \geq 0, \\ d_t - \dfrac{\phi_e}{2}(-d_t)^2, & d_t < 0. \end{cases}$$

This friction makes internal finance valuable and gives leverage a role beyond the borrowing constraint.

Let $V_t^c(\varepsilon, k, b)$ denote the value of continuing, after optimizing over $(k', b')$, and let $V^d$ denote the equity value of default. I normalize $V^d = 0$. Upon default, equity holders are wiped out and creditors recover up to the face value of outstanding debt from the liquidation of the firm's capital. Any residual resale value after creditor claims have been satisfied is absorbed by bankruptcy and liquidation costs, keeping the equity payoff from default equal to zero. Equity holders default whenever the value of continuing is below the value of default. Hence, the default-adjusted firm value and the associated default rule are[8]

$$V_t(\varepsilon, k, b) = \max\left\{V_t^c(\varepsilon, k, b), V^d\right\}, \qquad \mathbf{1}_t^d(\varepsilon, k, b) = \mathbf{1}\left\{V_t^c(\varepsilon, k, b) < V^d\right\}.$$

The default rule is indexed by the state at which the decision is taken. From the perspective of period $t$, debt issued today pays its face value at $t+1$ if the firm continues upon entering next period with state $(\varepsilon_{t+1}, k', b')$. If the firm defaults instead, creditors recover up to the face value of debt from the liquidation of the firm's capital, as specified below.

Defaulting firms exit and are replaced by entrants so that the mass of firms remains constant. Entrants start with capital $k^e = K^{ss}$, where $K^{ss}$ is the stationary aggregate capital stock of the deterministic economy, zero debt, and idiosyncratic productivity drawn from the stationary idiosyncratic distribution. Entrants enter the next-period distribution with newly installed capital. The capital of defaulting firms is liquidated, and creditor recoveries offset part of the resources required to install the replacement entrants' capital.

Define the amount recovered by creditors upon default as

$$\Lambda(k, b) = \min\left\{b,\, \theta_k(1-\delta)k\right\}. \tag{23}$$

For $b > 0$, the corresponding recovery rate is

$$\mathcal{R}(k, b) = \frac{\Lambda(k, b)}{b} = \min\left\{1,\, \frac{\theta_k(1-\delta)k}{b}\right\}. \tag{24}$$

[8] In the numerical implementation, I use a smooth approximation to the default rule. Appendix B.4 describes the approximation and shows that it is quantitatively close to the hard-threshold rule.

Thus, for debt $b'$ issued in period $t$, the relevant recovery rate in period $t+1$ is $\mathcal{R}(k', b')$. A continuum of identical competitive lenders price debt using the household stochastic discount factor. For $b' > 0$, the one-period debt price satisfies

$$q_t(\varepsilon, k', b') = \mathbb{E}_t \left[ M_{t,t+1} \left( 1 - p^d_{t+1}(\varepsilon_{t+1}, k', b') + p^d_{t+1}(\varepsilon_{t+1}, k', b') \mathcal{R}(k', b') \right) \right]. \tag{25}$$

The expectation in (25) is over next-period aggregate productivity, idiosyncratic productivity, and the transformer-implied distributional transition described below. Thus default pricing is one of the channels through which the forecasted aggregate distribution affects current firm decisions.

The continuing-firm Bellman equation is

$$V^c(\varepsilon, k, b; \Omega_t) = \max_{k', b'} \left\{ \mathcal{D}\big(d_t(\varepsilon, k, b, k', b'; \Omega_t)\big) + \mathbb{E}_t \left[ M_{t,t+1} V(\varepsilon', k', b'; \Omega_{t+1}) \right] \right\},$$

subject to $b' \leq \bar{b}(k')$. The default-adjusted value summarizes the transition decision: firms with low continuation value exit and are replaced in the next-period distribution, while continuing firms carry their chosen $(k', b')$ forward.

**Aggregation and market clearing.** Recall that $\mu_t(\varepsilon, k, b)$ denotes the cross-sectional distribution of firms entering period $t$, and define $p^c_t(\varepsilon, k, b) \equiv 1 - p^d_t(\varepsilon, k, b)$. Aggregate capital and debt at the beginning of the period are $K_t = \int k \, d\mu_t$ and $B_t = \int b \, d\mu_t$.

Default occurs before production, hence only continuing firms produce and hire labor:

$$Y_t = \int p^c_t(\varepsilon, k, b) \, y_t(\varepsilon, k, n; Z) \, d\mu_t, \qquad N_t = \int p^c_t(\varepsilon, k, b) \, n_t(\varepsilon, k; Z, w) \, d\mu_t.$$

Aggregate capital evolves according to

$$K_{t+1} = \int p^c_t(\varepsilon, k, b) \, k'_t(\varepsilon, k, b) \, d\mu_t + k^e \int p^d_t(\varepsilon, k, b) \, d\mu_t,$$

where defaulting firms are replaced one-for-one by entrants with capital $k^e$. Define aggregate investment as $I_t \equiv K_{t+1} - (1-\delta)K_t$. The aggregate resource constraint is

$$\begin{aligned} C_t + I_t + \int p^c_t(\varepsilon, k, b) \, \Psi(k, k') \, d\mu_t + \int p^d_t(\varepsilon, k, b) \, \Psi^d(k, b) \, d\mu_t \\ + \int p^c_t(\varepsilon, k, b) \, f \, d\mu_t + \int p^c_t(\varepsilon, k, b) \, \Phi_e(d_t) \, d\mu_t = Y_t, \end{aligned} \tag{26}$$

where

$$\Psi(k,k') = (1-\theta_k)[-i(k,k')]_+, \qquad \Psi^d(k,b) = (1-\delta)k - \Lambda(k,b), \qquad \Phi_e(d) = \frac{\phi_e}{2}[-d]_+^2.$$

Appendix A.3.5 details the associated resource accounting and shows how liquidation recoveries offset the resource cost of replacement entry. The equilibrium wage is the fixed point satisfying (19) and (26).

## 4.2 Perceived Distributional Law

I now specialize the DST of Section 3 to the firm economy. The aggregate state is the joint distribution of firms over productivity, capital, and debt, together with aggregate productivity $Z_t$. I partition the firm state space into capital–debt–productivity regions and represent each region by a token containing its probability mass and conditional means of capital, debt, and productivity. The baseline uses 15 capital bins, 8 debt bins, and 9 productivity groups, yielding $J = 1{,}080$ distributional tokens.

Let $\mu_{j,t}$ denote the mass in region $j$, and let $\bar{k}_{j,t}$, $\bar{b}_{j,t}$, and $\bar{\varepsilon}_{j,t}$ denote the corresponding conditional means. The distributional coordinates whose transition is forecast are the mass-weighted moments

$$m_{j,t} = \mu_{j,t}\left(\bar{k}_{j,t}, 1, \bar{b}_{j,t}, \bar{\varepsilon}_{j,t}\right), \qquad m_t = \operatorname{vec}\left(m_{1,t}, \ldots, m_{J,t}\right).$$

This representation preserves economically important aggregates directly. In particular,

$$K_t = \sum_{j=1}^{J} \mu_{j,t}\bar{k}_{j,t}, \qquad B_t = \sum_{j=1}^{J} \mu_{j,t}\bar{b}_{j,t}.$$

Aggregate capital and debt are therefore implications of the predicted distribution rather than separate transformer targets.

The transformer learns the perceived distributional law

$$\widehat{m}_{t+1} = m_t + \mathcal{T}_{\theta,m}(X_t), \qquad \widehat{\ell}_t = \mathcal{T}_{\theta,\ell}(X_t), \tag{27}$$

where $X_t$ contains the distributional tokens and aggregate productivity, and

$$\ell_t \equiv \log U_C(C_t) = -\gamma \log C_t$$

is included as an auxiliary equilibrium object. The latter implies

$$\widehat{C}_t = \exp\left(-\frac{\widehat{\ell}_t}{\gamma}\right), \qquad \widehat{w}_t = \chi \widehat{C}_t^{\gamma}.$$

As in Section 3, the continuous forecast $\widehat{m}'_{sZ}$ at a library state $(\mu^s, Z)$ determines neighboring future distributions $\mu^r$ and interpolation weights $\omega_{sZ,r}$. Together with the marginal-utility prediction, these weights determine the perceived continuation value

$$\mathcal{C}_{sZ}(\varepsilon, k', b') = \sum_{Z',\varepsilon'} P_Z(Z, Z') P_\varepsilon(\varepsilon, \varepsilon') \sum_{r \in \mathcal{N}(s,Z)} \omega_{sZ,r} \widehat{M}_{sZ,rZ'} V_r(\varepsilon', k', b'; Z'), \tag{28}$$

where

$$\widehat{M}_{sZ,rZ'} = \beta \left(\frac{\widehat{C}_{rZ'}}{\widehat{C}_{sZ}}\right)^{-\gamma}.$$

Competitive lenders use the same perceived transition. The price of one unit of debt issued by a firm choosing $(k', b')$ is

$$\begin{aligned} q_{sZ}(\varepsilon, k', b') = \sum_{Z',\varepsilon'} P_Z(Z, Z') P_\varepsilon(\varepsilon, \varepsilon') \sum_{r \in \mathcal{N}(s,Z)} \omega_{sZ,r} \widehat{M}_{sZ,rZ'} \\ \times \left[1 - p^{d,r}(Z', \varepsilon', k', b') + p^{d,r}(Z', \varepsilon', k', b') \mathcal{R}(k', b')\right]. \end{aligned} \tag{29}$$

Thus the predicted distribution affects current firm decisions through both continuation values and debt prices, while the auxiliary marginal-utility prediction determines the perceived wage and stochastic discount factor inside the recursive problem.

Given these perceived objects, the firm problem is the one described in Section 4. The resulting policies are simulated on the complete firm distribution, imposing current-period market clearing exactly. The model-generated distributional transitions and marginal utilities are then used to update the transformer until the perceived law and firm policies converge. Appendix A.3 provides the full numerical implementation, including tokenization, admissibility projection, transformer architecture, library construction, interpolation, debt pricing, simulation, and equilibrium iteration.

# 5 Quantitative Results

This section presents the quantitative analysis of the heterogeneous-firm economy. After calibrating the model and characterizing its stationary equilibrium, I use transformer attention to identify which regions of the firm distribution matter most for aggregate dynamics.

Attention concentrates on financially fragile firms near the default margin, and shifts toward these firms during downturns as their default risk becomes more consequential for aggregate dynamics.

## 5.1 Calibration

The calibration follows Khan and Thomas (2013) closely. I set $\alpha = 0.27$, $\nu = 0.60$, $\beta = 0.96$, $\delta = 0.065$, $\rho_Z = 0.852$, $\sigma_Z = 0.014$, and $\theta_b = 1.350$ as in their benchmark, and choose $\chi$ so that stationary hours are approximately one third. The parameters governing capital irreversibility and idiosyncratic productivity, $\theta_k = 0.96$, $\rho_\varepsilon = 0.7$, and $\sigma_\varepsilon = 0.112$, are disciplined by the three establishment-level investment moments used by Khan and Thomas (2013), based on Cooper and Haltiwanger (2006).[9] I add two moments to discipline the financial frictions introduced here: the default frequency disciplines the operating cost $f = 0.0735$, while the equity-issuance frequency disciplines the issuance-cost parameter $\phi_e = 1.3$. Table 1 reports the resulting fit.

Table 1: Targeted Moments

| Moment | Model | Target | Source |
|---|---|---|---|
| Mean investment rate, $i/k$ | 0.116 | 0.130 | Khan and Thomas (2013) |
| Std. investment rate, $i/k$ | 0.320 | 0.337 | Khan and Thomas (2013) |
| Serial correlation of $i/k$ | 0.060 | 0.058 | Khan and Thomas (2013) |
| Default frequency | 0.019 | 0.020 | Ippolito et al. (2026) |
| Equity-issuance frequency | 0.032 | 0.040 | Ippolito et al. (2026) |

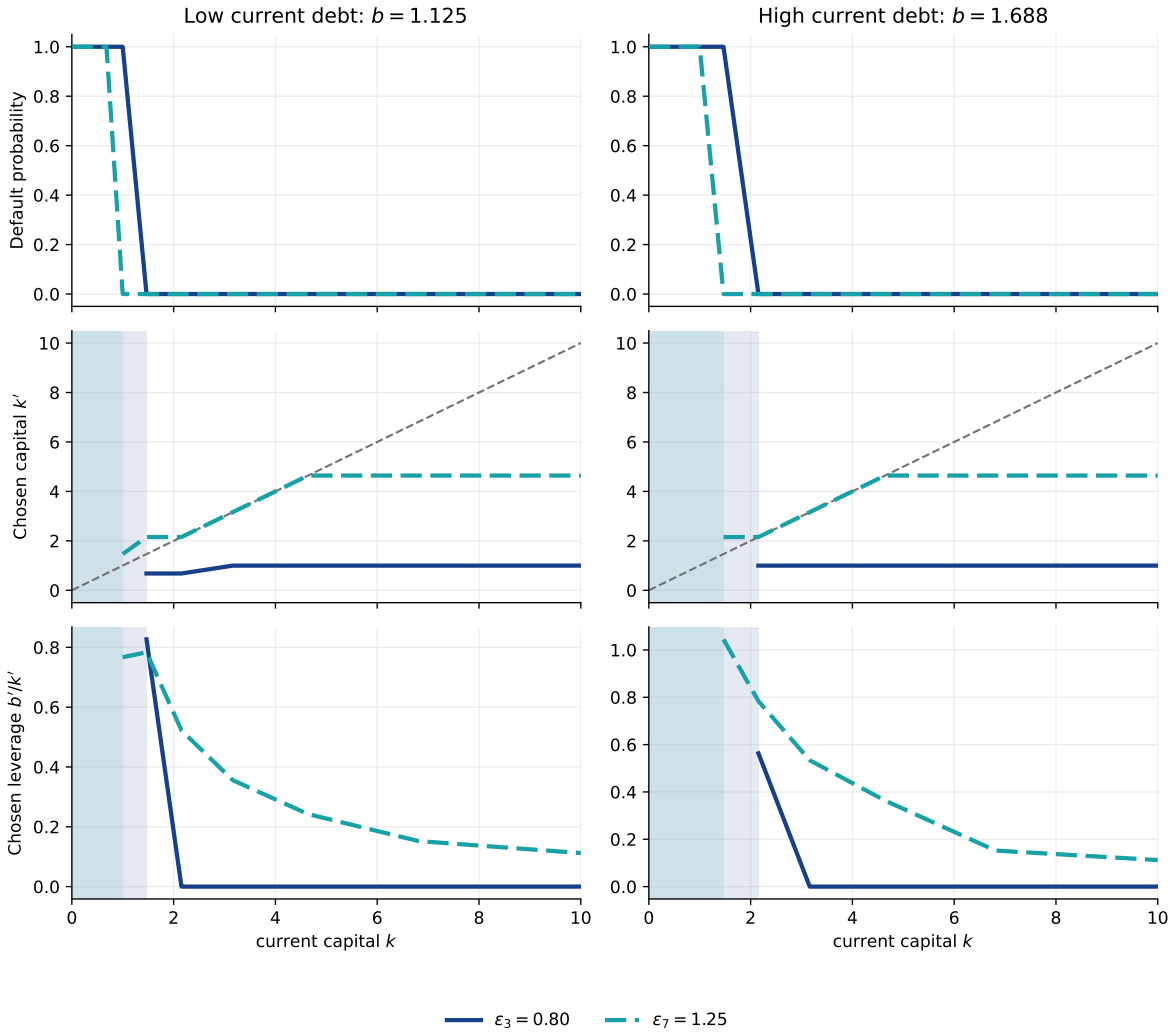
*Notes:* The first three moments discipline $\theta_k, \rho_\varepsilon, \sigma_\varepsilon$; the default and equity-issuance frequencies discipline $f$ and $\phi_e$, respectively. Investment is $i/k = (k' - (1-\delta)k)/k$; its standard deviation and serial correlation are winsorized at $[-1, 1]$. Model moments are computed from the stationary distribution. Equity issuance is counted when gross external equity exceeds 10 percent of current capital.

**Stationary policies.** Figure 4 reports slices of the stationary policy functions for two levels of inherited debt and two idiosyncratic-productivity states. As in Khan and Thomas (2013), partial capital irreversibility makes firm scale persistent and limits capital reallocation. Endogenous default adds an extensive margin: for a given productivity level, higher inherited debt shifts the default boundary toward higher capital, while higher productivity raises desired scale and moves firms away from default. Conditional on continuation, leverage generally declines with capital. Thus inherited debt mainly affects whether firms survive,

[9] Aggregate productivity is discretized with three Rouwenhorst nodes and idiosyncratic productivity with nine.

whereas productivity and capital shape the scale and financing choices of continuing firms. Appendix Figure 9 in Appendix C shows the corresponding debt-price schedules, which deteriorate sharply in financially fragile states and approach the safe-debt price as default risk vanishes.

Figure 4: Stationary Firm Policies




*Notes:* The figure reports stationary firm policies as functions of current capital $k$ for selected idiosyncratic-productivity states $\varepsilon$. Columns condition on low and high current debt, respectively; rows report default probability, chosen next-period capital $k'$, and chosen leverage $b'/k'$. The dashed line in the capital panels is $k' = k$.

## 5.2 Results

The quantitative results for the model introduced in Section 4 use the final converged DST law; the transformer architecture is reported in Appendix A.3.2. The equilibrium is solved using 50,000-period simulations, and the attention diagnostics below are computed over

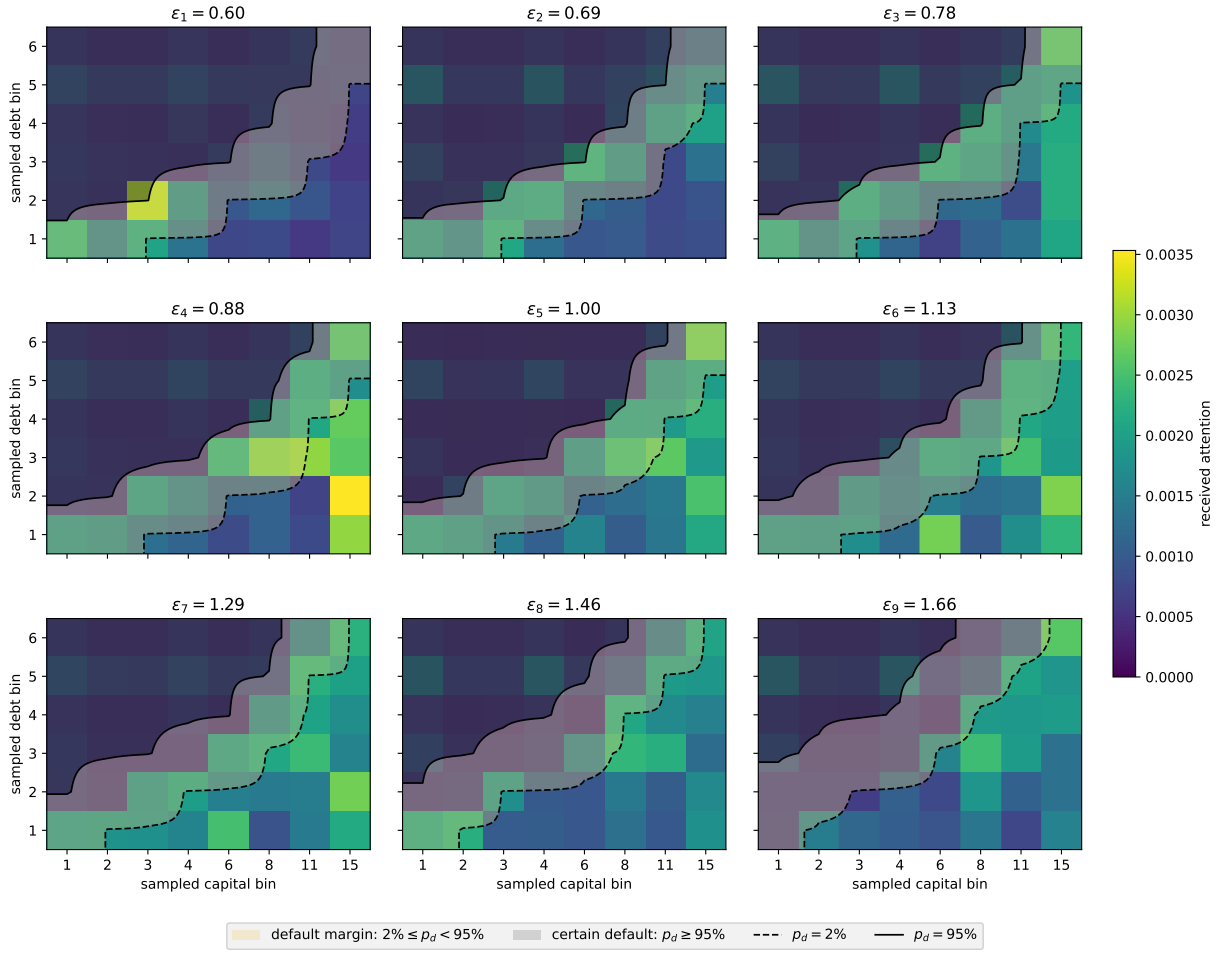
the resulting ergodic path. The learned distributional law closely reproduces the aggregate dynamics implied by the model while forecasting the distribution itself rather than separate aggregate laws.[10]

**Where does attention concentrate?** Figure 5 provides the main cross-sectional diagnostic. The heatmaps report ergodic received attention across capital–debt cells and idiosyncratic productivity states, averaged over the final simulated path. Attention is highly nonuniform. It is concentrated near the endogenous default margin, especially for low- and middle-productivity firms, where balance-sheet conditions are particularly consequential for continuation and default decisions.

The pattern changes systematically with idiosyncratic productivity. For the lowest productivity states, attention is almost entirely concentrated in the intermediate default-risk region, $2\% \leq p^d < 95\%$. Little attention is placed on states with $p^d \geq 95\%$, where default is essentially certain and the continuation decision carries little marginal information. At intermediate productivity levels, attention remains elevated near the default margin but becomes more diffuse over non-defaulting firms. For the highest productivity states, especially the two highest productivity realizations, received attention is lower in absolute terms and less concentrated, consistent with these firms being farther from the default boundary.

[10]For example, aggregating the predicted distribution yields $R^2 = 0.9993$ for next-period aggregate capital and $R^2 = 0.9925$ for next-period leverage. The auxiliary marginal-utility prediction has $R^2 = 0.9996$.

Figure 5: Attention and the Endogenous Default Region




*Notes.* The figure reports ergodic received attention across capital–debt cells for each idiosyncratic productivity state, $\epsilon$. The overlay is constructed from equilibrium default probabilities: yellow denotes firms with default risk above the calibrated stationary default rate but below near-certain default, $2\% \le p^d < 95\%$, and dark shading denotes near-certain default, $p^d \ge 95\%$. Capital and debt bins outside the ergodic support of the library are omitted.

Appendix Figure 8 in Appendix C provides a complementary model-based diagnostic by measuring how local redistributions of firm mass affect aggregate default exposure, holding local default probabilities fixed. The resulting sensitivity is concentrated around the endogenous default region, identifying the parts of the firm distribution in which small reallocations of mass have the largest direct effect on aggregate default.

**Policy-sensitivity decomposition.** To connect the quantitative results to Proposition 1, I study mass-preserving perturbations of the firm distribution. Each perturbation reallocates a small amount of firm mass toward one capital–debt–productivity region, removing it proportionally from the rest of the distribution while holding within-cell characteristics fixed. The exercise delivers two economic takeaways. First, different aggregate outcomes depend

on the distribution through different margins: next-period aggregate capital mainly reflects the persistence of firms' existing capital stocks, whereas aggregate leverage is more sensitive to how information from firms' balance sheets is aggregated. Second, this aggregation margin becomes especially important near the endogenous default boundary. Shifting firm mass toward financially fragile regions changes not only the composition of firms' balance sheets, but also the relevance of other parts of the distribution for predicted aggregate outcomes. The recession experiment in Figure 6 provides a directional counterpart to this result: as default risk rises, attention shifts toward firms near the default margin.

Because the transformer predicts changes in the distribution,

$$\widehat{m}_{t+1} = m_t + \mathcal{T}_{\theta,m}(X_t),$$

local sensitivity separates into a direct current-state component and a transformer-mediated component. Within self-attention, the latter contains the attention-reallocation and value channels in equation (7); the full quantitative decomposition also accounts for the remaining network paths.[11] Importantly, the transformer forecasts the next distribution rather than separate aggregate laws, so capital and leverage are implied by the same distributional forecast.

For next-period capital, the direct current-state response is substantially larger than the transformer-mediated response (0.521 versus 0.209), consistent with the persistence of the capital stock. Within the latter, the attention-reallocation and value channels have average magnitudes equal to about 37% and 21%, respectively. For next-period aggregate leverage, by contrast, the two attention channels have similar average magnitudes, 47% and 45% of the transformer-mediated response. Near the default margin, the transformer-mediated response exceeds the direct response (0.068 versus 0.047), and attention reallocation alone has magnitude 0.059, about 87% of the transformer-mediated response.

Economically, reallocating mass toward the default region places more weight on firms for which default risk and financing conditions are particularly sensitive to balance-sheet positions. Financially fragile regions therefore matter not only through the change in the composition of firms' balance sheets, but also because that change alters how information elsewhere in the cross section is aggregated. The state dependence of attention is thus economically relevant for predicted aggregate leverage and, through firms' financing decisions, aggregate dynamics. The next exercise shows how this mechanism changes over the business

[11] Appendix B.5 derives the full decomposition for the quantitative transformer, and Appendix Table 2 reports the corresponding results for aggregate capital and leverage. As a numerical check, the full accounting closes to numerical precision, and derivatives computed by automatic differentiation closely match central finite differences.

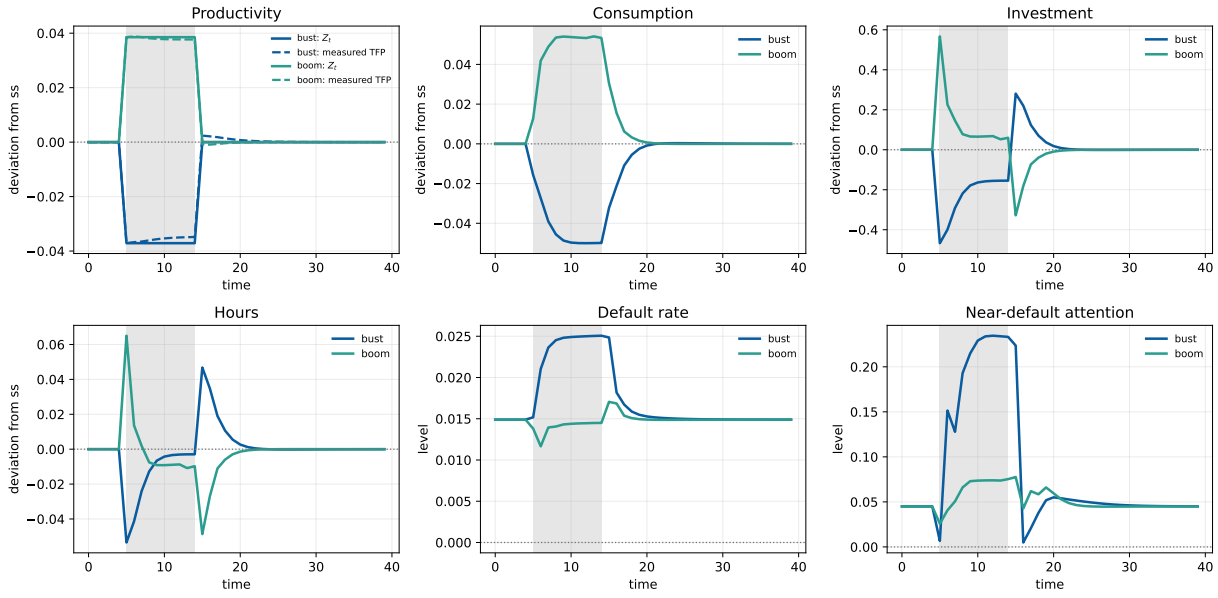
cycle.

**State-dependent attention.** The previous result identifies where attention is allocated on average. I next ask whether the importance of this region changes with aggregate conditions. Starting from the dynamic stationary distribution associated with the learned law, I subject the economy to temporary aggregate-productivity busts and booms of equal magnitude. Figure 6 reports the responses.

The adjustment is asymmetric. A downturn generates a sharp and persistent rise in default, together with larger contractions in investment and hours than the corresponding expansions generated by the boom. At the same time, attention received by firms near the default margin rises markedly during the bust, while the response in the boom is much smaller. The learned aggregation rule is therefore state dependent: adverse aggregate conditions shift attention toward financially fragile regions of the distribution precisely when default becomes more consequential for aggregate dynamics.

Figure 6: Bust–Boom Asymmetry and Transformer Attention

Bust-boom asymmetry from the dynamic stationary distribution




*Notes:* The economy starts from the dynamic stationary distribution implied by the final transformer law. The shaded region denotes the temporary aggregate-productivity shock. All variables except default and attention are deviations from their dynamic stationary values. Near-default attention is attention received by firm tokens close to the endogenous default margin.

# 6 Conclusion

Transformers have become central to modern artificial intelligence in part because they combine flexible neural-network approximation with an attention mechanism that learns how to aggregate information across complex inputs. This paper asks whether the same mechanism can shed light on a classic problem in heterogeneous agent models: which parts of a high-dimensional cross-sectional distribution matter for aggregate dynamics, and how their importance changes with the aggregate state.

I develop the Distributional State Transformer (DST), a recursive solution method that represents the cross-sectional distribution as economic tokens and learns its perceived law of motion without imposing a preselected set of aggregate moments. I show theoretically that local policy sensitivity decomposes into attention-reallocation and value channels; under a restricted aggregation structure, received attention coincides with normalized aggregate policy function sensitivities. In the mean-sufficient and Krusell–Smith exercises, the transformer recovers the underlying aggregation structure from the full distribution, with attention remaining flat or nearly flat across wealth bins.

In a heterogeneous-firm economy with endogenous default, attention instead concentrates near the default margin, and attention reallocation is quantitatively important for aggregate leverage. In this environment, financial fragility therefore makes economic aggregation endogenous to the cross-sectional state. Moreover, the relevant parts of the distribution change with aggregate conditions: during downturns, attention shifts toward financially fragile firms, whose balance sheets become more consequential for financing, investment, and aggregate dynamics. These results suggest that attention can serve simultaneously as a computational device for solving high-dimensional recursive models and as a diagnostic tool for understanding how the aggregation of cross-sectional heterogeneity changes over the business cycle.

# References

Azinovic, M., Gaegauf, L., and Scheidegger, S. (2022). Deep equilibrium nets. *International Economic Review*, 63(4):1471–1525.

Clementi, G. L. and Hopenhayn, H. A. (2006). A theory of financing constraints and firm dynamics. *Quarterly Journal of Economics*, 121(1):229–265.

Clementi, G. L. and Palazzo, B. (2016). Entry, exit, firm dynamics, and aggregate fluctuations. *American Economic Journal: Macroeconomics*, 8(3):1–41.

Cooley, T. F. and Quadrini, V. (2001). Financial markets and firm dynamics. *American Economic Review*, 91(5):1286–1310.

Cooper, R. W. and Haltiwanger, J. C. (2006). On the nature of capital adjustment costs. *Review of Economic Studies*, 73(3):611–633.

Duffy, J. and McNelis, P. D. (2001). Approximating and simulating the stochastic growth model: Parameterized expectations, neural networks, and the genetic algorithm. *Journal of Economic Dynamics and Control*, 25(9):1273–1303.

Ebrahimi Kahou, M., Fernández-Villaverde, J., Perla, J., and Sood, A. (2021). Exploiting symmetry in high-dimensional dynamic programming. Working Paper 28981, National Bureau of Economic Research.

Fernandez-Villaverde, J., Hurtado, S., and Nuno, G. (2023). Financial frictions and the wealth distribution. *Econometrica*, 91(3):869–901.

Gomes, J. F. (2001). Financing investment. *American Economic Review*, 91(5):1263–1285.

Gopalakrishna, G., Gu, Z., and Payne, J. (2026). Institutional asset pricing with segmentation and household heterogeneity. Working paper, Rotman School of Management.

Gu, Z., Laurière, M., Merkel, S., and Payne, J. (2026). Global solutions to master equations for continuous time heterogeneous agent macroeconomic models. *Journal of Political Economy Macroeconomics*. Accepted.

Han, J., Yang, Y., and E, W. (2025). Deepham: A global solution method for heterogeneous agent models with aggregate shocks. Working paper.

Hennessy, C. A. and Whited, T. M. (2005). Debt dynamics. *Journal of Finance*, 60(3):1129–1165.

Hennessy, C. A. and Whited, T. M. (2007). How costly is external financing? evidence from a structural estimation. *Journal of Finance*, 62(4):1705–1745.

Hopenhayn, H. and Rogerson, R. (1993). Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy*, 101(5):915–938.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*, 60(5):1127–1150.

Ippolito, F., Steri, R., Tebaldi, C., and Villa, A. T. (2026). Mind the gap: The market price of financial (in)flexibility. *Journal of Finance, forthcoming*.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jermann, U. and Quadrini, V. (2012). Macroeconomic effects of financial shocks. *American Economic Review*, 102(1):238–271.

Judd, K. L., Maliar, L., and Maliar, S. (2011). Numerically stable and accurate stochastic simulation approaches for solving dynamic economic models. *Quantitative Economics*, 2(2):173–210.

Kase, H., Melosi, L., and Rottner, M. (2025). Estimating nonlinear heterogeneous agent models with neural networks. Working paper.

Khan, A., Senga, T., and Thomas, J. K. (2021). Default risk and aggregate fluctuations in an economy with production heterogeneity. Working paper.

Khan, A. and Thomas, J. K. (2008). Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. *Econometrica*, 76(2):395–436.

Khan, A. and Thomas, J. K. (2013). Credit shocks and aggregate fluctuations in an economy with production heterogeneity. *Journal of Political Economy*, 121(6):1055–1107.

Krusell, P. and Smith, Jr., A. A. (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, 106(5):867–896.

Lanteri, A. (2018). The market for used capital: Endogenous irreversibility and reallocation over the business cycle. *American Economic Review*, 108(9):2383–2419.

Maliar, L., Maliar, S., and Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, 122:76–101.

Ottonello, P. and Winberry, T. (2020). Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, 88(6):2473–2502.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Payne, J., Rebei, A., and Yang, Y. (2025). Deep learning for search and matching models. Research Paper 25-05, Swiss Finance Institute.

Scheidegger, S. and Bilionis, I. (2019). Machine learning for high-dimensional dynamic stochastic economies. *Journal of Computational Science*, 33:68–82.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Valaitis, V. and Villa, A. T. (2024). A machine learning projection method for macro-finance models. *Quantitative Economics*, 15(1):145–173.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Villa, A. T. (2026). Macro shocks and firm dynamics with oligopolistic financial intermediaries. *Review of Economic Studies, forthcoming*.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Young, E. R. (2010). Solving the incomplete markets model with aggregate uncertainty using the krusell–smith algorithm and non-stochastic simulations. *Journal of Economic Dynamics and Control*, 34(1):36–41.

# Online Appendix

This Appendix is organized in three parts. Appendix A provides implementation details, Appendix B contains the mathematical derivations, and Appendix C reports additional figures.

# A Computational Appendix

Appendix A.1 describes the implementation of the two-period benchmarks, including the transformer architecture, training procedure, and attention diagnostics. Appendix A.2 provides implementation details for the infinite-horizon household benchmark, including tokenization, projection, library interpolation, and the non-stochastic distributional transition. Appendix A.3 specializes the DST algorithm to the heterogeneous-firm application, including firm tokenization, transformer architecture, library construction, debt pricing, market clearing, and equilibrium iteration.

## A.1 Implementation of the Two-Period Models

The two exercises in Section 2.3 use the same full self-attention architecture. The cross section is represented by $J = 20$ equal-mass capital-bin tokens, ordered by capital within each economy, together with a separate aggregate-productivity token $Z$. Capital-bin tokens contain the bin capital $k_j$, its mass $\mu_j$, and token-type indicators; the aggregate token contains $Z$ and the corresponding type indicator. Inputs and prediction targets are standardized using training-sample moments.

Each token is mapped through a learned nonlinear embedding with internal dimension $d = 128$. Queries, keys, and values are learned linear transformations of the embedded tokens,

$$q_j = W_Q h_j, \qquad k_j^A = W_K h_j, \qquad v_j = W_V h_j,$$

and a single full self-attention block allows every token to attend to every other token. The resulting token representations are mean pooled and passed through a prediction network with one hidden layer of 64 units and a ReLU activation to produce $\widehat{K}'$. Thus, unlike the restricted representations in Remarks 1 and 2, both attention weights and value representations can depend on the realized economic state.

Each benchmark uses 100,000 simulated economies, with 80,000 used for training and 20,000 held out for evaluation. Networks are trained for 80 epochs using Adam, batches of 256 observations, and learning rate $10^{-3}$. Forecast fit is essentially exact: the out-of-sample

$R^2$ is 0.999938 in the mean-sufficient economy and 0.999866 in the individual-production economy.

For the attention diagnostic in Figure 2, I restrict attention to the capital-bin block. For each held-out economy, received attention for bin $j$ is the average attention assigned to that bin by the $J$ capital-bin queries, normalized to sum to one across capital bins. Figure entries are test-sample averages of these normalized weights; error bars report 95% confidence intervals across held-out economies.

## A.2 Implementation of the One-Asset Household Model

This appendix describes the implementation of the one-asset household economy used to validate the DST in Section 3.7. The general perceived distributional law, sparse-library construction, interpolation operator, and equilibrium iteration are described in Section 3; here I specialize those objects to the household economy.

### A.2.1 Economic Environment and Numerical State

Households differ in assets $a$ and idiosyncratic labor productivity $\varepsilon$. Let

$$\mathcal{A} = \{a_1, \ldots, a_{N_a}\}, \qquad \mathcal{E} = \{\varepsilon_1, \ldots, \varepsilon_{N_\varepsilon}\}, \qquad \mathcal{Z} = \{Z_B, Z_G\}$$

denote the asset, idiosyncratic-productivity, and aggregate-productivity grids. Idiosyncratic productivity is approximated by a seven-state Rouwenhorst chain, while aggregate productivity follows a two-state Markov chain. The two processes are independent.

The numerical aggregate state is the complete distribution

$$\mu \in \Delta(\mathcal{A} \times \mathcal{E}).$$

Labor supply is fixed at one, so the household chooses only next-period assets. At aggregate state $(\mu, Z)$, the household problem is

$$V(a, \varepsilon, Z, \mu) = \max_{a' \geq \underline{a}} \left\{ u(R(Z, \mu)a + w(Z, \mu)\varepsilon - a') + \beta \mathbb{E}\left[ V(a', \varepsilon', Z', \mu') \mid \varepsilon, Z \right] \right\}.$$

Aggregate capital and effective labor are

$$K(\mu) = \sum_{a \in \mathcal{A}} \sum_{\varepsilon \in \mathcal{E}} a \, \mu(a, \varepsilon), \qquad N(\mu) = \sum_{a \in \mathcal{A}} \sum_{\varepsilon \in \mathcal{E}} \varepsilon \, \mu(a, \varepsilon).$$

A representative firm produces

$$Y = ZK^{\alpha}N^{1-\alpha},$$

so factor prices are

$$R(Z,\mu) = 1 - \delta + \alpha ZK(\mu)^{\alpha-1}N(\mu)^{1-\alpha},$$

and

$$w(Z,\mu) = (1-\alpha)ZK(\mu)^{\alpha}N(\mu)^{-\alpha}.$$

The complete distribution $\mu$, rather than its tokenized representation, is propagated through the economy. Tokenization is used only to construct the state supplied to the transformer's perceived distributional law. The one-moment Krusell–Smith reference solution uses the same underlying economy but summarizes the aggregate state entering its perceived law by aggregate capital $K$.

### A.2.2 Tokenization and Distributional Forecast

I partition the asset support into $J = 20$ fixed wealth bins

$$\{\mathcal{B}_j\}_{j=1}^{J}.$$

For distribution $\mu$, let the mass in bin $j$ be

$$\omega_j(\mu) = \sum_{a\in\mathcal{B}_j}\sum_{\varepsilon\in\mathcal{E}} \mu(a,\varepsilon).$$

For a positive-mass bin, define the conditional moments

$$\bar{a}_j(\mu) = \frac{1}{\omega_j(\mu)}\sum_{a\in\mathcal{B}_j}\sum_{\varepsilon\in\mathcal{E}} a\,\mu(a,\varepsilon),$$

$$\bar{\varepsilon}_j(\mu) = \frac{1}{\omega_j(\mu)}\sum_{a\in\mathcal{B}_j}\sum_{\varepsilon\in\mathcal{E}} \varepsilon\,\mu(a,\varepsilon),$$

and

$$\overline{\varepsilon^2}_j(\mu) = \frac{1}{\omega_j(\mu)}\sum_{a\in\mathcal{B}_j}\sum_{\varepsilon\in\mathcal{E}} \varepsilon^2\mu(a,\varepsilon).$$

The distributional coordinates for bin $j$ are

$$m_j(\mu) = \left(\bar{a}_j(\mu), \omega_j(\mu), \bar{\varepsilon}_j(\mu), \overline{\varepsilon^2}_j(\mu)\right),$$

and

$$m(\mu) = \operatorname{vec}\left(m_1(\mu), \ldots, m_J(\mu)\right) \in \mathbb{R}^{80}.$$

If $\omega_j(\mu) = 0$, the asset coordinate is set to the midpoint of $\mathcal{B}_j$, while the productivity coordinates are assigned fixed feasible values.

In this application, the economic coordinates supplied to the wealth-bin tokens coincide with the coordinates whose transition is forecast. The transformer input is

$$X(\mu, Z) = \{x_1, \ldots, x_J, x_Z\},$$

where

$$x_j = \left(\bar{a}_j, \omega_j, \bar{\varepsilon}_j, \overline{\varepsilon^2}_j \;\middle|\; 0 \;\middle|\; 1, 0\right), \qquad j = 1, \ldots, J,$$

and

$$x_Z = (0, 0, 0, 0 \mid Z \mid 0, 1).$$

As in equation (9), the transformer predicts changes rather than levels,

$$\Delta m_{t+1} = m_{t+1} - m_t.$$

Inputs and target changes are standardized using moments from the current training sample. Let $\bar{\Delta}_m$ and $s_{\Delta_m}$ denote the coordinate-specific mean and standard deviation of the target changes. The forecast transformed back into economic units is

$$m_{t+1}^{\text{raw}} = m_t + \bar{\Delta}_m + s_{\Delta_m} \odot \widetilde{\mathcal{T}}_{\theta,m}(\widetilde{X}_t).$$

The perceived distributional law is

$$\widehat{m}_{t+1} = \Pi_H\left(m_{t+1}^{\text{raw}}\right), \tag{30}$$

where $\Pi_H$ projects the raw forecast onto the admissible household token space. The projection imposes

$$\widehat{\omega}_j \geq 0, \qquad \sum_{j=1}^{J} \widehat{\omega}_j = 1,$$

keeps each conditional asset mean inside its associated wealth bin, and restricts the conditional productivity moments to their feasible support.

Aggregate capital is not a separate transformer target. It is implied by the projected

distributional forecast:

$$\widehat{K}_{t+1}^{\mathrm{DST}} = \sum_{j=1}^{J} \widehat{\omega}_{j,t+1} \widehat{\bar{a}}_{j,t+1}. \tag{31}$$

Thus the comparison with the one-moment Krusell–Smith solution evaluates whether forecasting the distribution itself recovers the familiar approximate aggregation through aggregate capital.

### A.2.3 Sparse Library and Household Problem

Value and policy functions are stored on a fixed library of complete distributions,

$$\mathcal{L} = \{\mu^1, \ldots, \mu^S\},$$

constructed from model-generated distributions using the procedure in Section 3.3. At node $\mu^s$, the stored objects are

$$V^s(Z, \varepsilon, a) \equiv V(a, \varepsilon, Z, \mu^s)$$

and

$$g_a^s(Z, \varepsilon, a) \equiv g_a(a, \varepsilon, Z, \mu^s).$$

For the household benchmark, distances between distributions are computed in standardized token space. At a current node $(\mu^s, Z)$, the transformer forecast is first projected according to equation (30). I then select the $M = 4$ nearest library distributions and construct the inverse-distance interpolation weights in equation (10).

These weights specialize the generic continuation value in equation (11) to

$$\mathcal{C}_{sZ}(a', \varepsilon) = \sum_{Z' \in \mathcal{Z}} P_Z(Z, Z') \sum_{\varepsilon' \in \mathcal{E}} P_\varepsilon(\varepsilon, \varepsilon') \sum_{r \in \mathcal{N}(s,Z)} \omega_{sZ,r} V^r(Z', \varepsilon', a'). \tag{32}$$

The Bellman equation at library node $\mu^s$ is therefore

$$V^s(Z, \varepsilon, a) = \max_{a' \geq \underline{a}} \left\{ u(R(Z, \mu^s)a + w(Z, \mu^s)\varepsilon - a') + \beta \mathcal{C}_{sZ}(a', \varepsilon) \right\}. \tag{33}$$

Current factor prices in equation (33) are computed from the current complete distribution $\mu^s$. The transformer affects the household problem through the perceived future distribution entering the continuation value.

Along an equilibrium simulation, the realized distribution $\mu_t$ need not coincide with a library node. I therefore locate its nearest library distributions using the same standardized

metric and interpolate the stored policies $g_a^s$. The resulting policy

$$g_a(a, \varepsilon, Z_t, \mu_t)$$

is applied to the complete current distribution.

### A.2.4 Distributional Transition and Equilibrium Iteration

Given $g_a$, $\mu_t$, and the realized aggregate transition $Z_t \to Z_{t+1}$, the next distribution is computed using the non-stochastic transition of Young (2010),

$$\mu_{t+1} = \mathcal{Y}_H \left(\mu_t, g_a, Z_t, Z_{t+1}\right).$$

For each current grid point $(a_i, \varepsilon_j)$, the chosen asset position is represented as a lottery over neighboring points of $\mathcal{A}$. Current mass is split across these points using the corresponding interpolation weights and across future idiosyncratic-productivity states according to $P_\varepsilon$. The simulation therefore propagates the complete continuum distribution rather than a finite population of households.

The equilibrium iteration follows Algorithm 1. A bootstrap perceived law is first used to simulate feasible distributions and construct the library. The library is then held fixed during the final equilibrium iteration. Given the current transformer law, I solve equation (33) at every library node, interpolate the resulting policies along the simulated path, and update the complete distribution using $\mathcal{Y}_H$. The model-generated pairs

$$\{X_t, \Delta m_{t+1}\}$$

form the supervised sample used to update the transformer. The household problem is then resolved under the updated perceived law, and the procedure is repeated until both the policy functions and transformer forecasts satisfy the convergence criteria in Section 3.5.

Received attention in Figure 3 and directional attention in Appendix Figure 7 are computed from the converged transformer using the definitions in Section 3.6. Because the network predicts all 80 coordinates of $m_{t+1}$, the aggregate-capital fit reported in the main text is entirely implied by the distributional forecast rather than by a separately estimated law for $K_{t+1}$.

### A.2.5 Numerical Implementation

The economic and machine-learning components are implemented separately. NumPy is used for array operations, Numba for JIT-compiled Bellman and distribution-transition routines, and PyTorch for transformer training, inference, and attention extraction. Bellman and Young-transition operations are executed on the CPU, while transformer training and inference use CUDA when available, Apple's `mps` backend on compatible hardware, and otherwise the CPU.

## A.3 Implementation of the Heterogeneous-Firm Application

This appendix provides implementation details for the heterogeneous-firm application in Section 4. The general DST algorithm, including the perceived distributional law, sparse library, interpolation operator, and equilibrium iteration, is described in Section 3. Here I specialize these objects to the firm economy and describe the transformer architecture, debt-pricing implementation, distributional transition, and market-clearing procedure.

A firm state is $(\varepsilon, k, b)$, and the complete aggregate distribution is

$$\mu \in \Delta(\mathcal{E} \times \mathcal{K} \times \mathcal{B}).$$

Value and policy functions are stored on a fixed sparse library of complete firm distributions. The transformer supplies a perceived law for the next tokenized distribution and an auxiliary prediction of household marginal utility. The former determines continuation values and debt prices, while the latter determines the perceived stochastic discount factor and wage inside the firm problem.

### A.3.1 Tokenization and Admissibility Projection

The firm state space is partitioned into $J = 1{,}080$ regions corresponding to 15 capital bins, 8 debt bins, and 9 idiosyncratic-productivity groups. Let $\mathcal{C}_j$ denote region $j$. Its probability mass under distribution $\mu$ is

$$\mu_j(\mu) = \sum_{(\varepsilon,k,b)\in\mathcal{C}_j} \mu(\varepsilon, k, b).$$

For cells with positive mass, let $\bar{k}_j(\mu)$, $\bar{b}_j(\mu)$, and $\bar{\varepsilon}_j(\mu)$ denote the corresponding conditional means.

The numerical input token associated with cell $j$ is

$$x_j^{\mu}(\mu) = \left(\bar{k}_j(\mu), \mu_j(\mu), \bar{b}_j(\mu), \bar{\varepsilon}_j(\mu), 0, 1, 0\right),$$

while aggregate productivity is represented by

$$x^Z = (0, 0, 0, 0, Z, 0, 1).$$

The final two entries are token-type indicators, while zero entries provide placeholders that give all tokens the same dimension. The transformer input is therefore

$$X(\mu, Z) = \left\{x_1^\mu(\mu), \ldots, x_J^\mu(\mu), x^Z\right\}.$$

The distributional forecasting target is represented by the mass-weighted moment vector

$$m_j(\mu) = \left(\mu_j(\mu)\bar{k}_j(\mu), \mu_j(\mu), \mu_j(\mu)\bar{b}_j(\mu), \mu_j(\mu)\bar{\varepsilon}_j(\mu)\right),$$

and

$$m(\mu) = \operatorname{vec}\left(m_1(\mu), \ldots, m_J(\mu)\right).$$

Mass weighting is convenient for regions with little or no probability mass and makes aggregate capital and debt additive:

$$K(m) = \sum_{j=1}^{J} m_{j,k}, \qquad B(m) = \sum_{j=1}^{J} m_{j,b},$$

where

$$m_{j,k} = \mu_j \bar{k}_j, \qquad m_{j,b} = \mu_j \bar{b}_j.$$

The perceived law predicts changes in the distributional vector. Let $\bar{\Delta}_m$ and $s_{\Delta_m}$ denote the coordinate-specific mean and standard deviation of the changes in the current training sample. If $\widetilde{\mathcal{T}}_{\theta,m}(\widetilde{X})$ is the transformer output in standardized units, the raw forecast in economic units is

$$m'^{,\text{raw}} = m(\mu) + \bar{\Delta}_m + s_{\Delta_m} \odot \widetilde{\mathcal{T}}_{\theta,m}(\widetilde{X}).$$

The forecast entering the equilibrium algorithm is

$$\widehat{m}'(\mu, Z; \theta) = \Pi_F\left(m'^{,\text{raw}}\right), \tag{34}$$

where $\Pi_F$ projects the raw prediction onto the admissible token space. Cell masses are restricted to be nonnegative and normalized to sum to one. For positive-mass cells, conditional capital, debt, and productivity means are recovered from the corresponding mass-weighted coordinates and restricted to the support of the associated cell. All mass-weighted coordinates are set to zero for cells with zero projected mass.

The auxiliary transformer target is log household marginal utility,

$$\ell \equiv \log U_C(C) = -\gamma \log C.$$

The corresponding prediction

$$\widehat{\ell}(\mu, Z) = \mathcal{T}_{\theta,\ell}(X(\mu, Z))$$

implies

$$\widehat{C}(\mu, Z) = \exp\left[-\frac{\widehat{\ell}(\mu, Z)}{\gamma}\right], \qquad \widehat{w}(\mu, Z) = \chi \widehat{C}(\mu, Z)^{\gamma}. \tag{35}$$

### A.3.2 Transformer Architecture

Stacking the $J + 1$ economic tokens gives the transformer input $X_t$. Economic coordinates are standardized using the current training sample, while token-type indicators are left unchanged. A linear embedding maps each token into $\mathbb{R}^d$, with $d = 64$. A learned CLS token is prepended and learned positional embeddings are added to the resulting sequence. Thus, if $\widetilde{X}_t$ denotes the standardized token matrix,

$$H_t^{(0)} = \begin{bmatrix} c \\ \widetilde{X}_t W_E^{\top} + \mathbf{1} b_E^{\top} \end{bmatrix} + P,$$

where $c \in \mathbb{R}^{64}$ is the learned CLS token.

The network contains $L = 2$ pre-normalized transformer blocks with $H = 4$ attention heads per block. Each head has dimension $d_h = d/H = 16$. For block $\ell$, define

$$\bar{H}_t^{(\ell)} = \mathrm{LN}_1^{(\ell)}\left(H_t^{(\ell-1)}\right).$$

For head $h$,

$$Q_h^{(\ell)} = \bar{H}_t^{(\ell)} W_{Q,h}^{(\ell)}, \qquad K_h^{(\ell)} = \bar{H}_t^{(\ell)} W_{K,h}^{(\ell)}, \qquad V_h^{(\ell)} = \bar{H}_t^{(\ell)} W_{V,h}^{(\ell)},$$

and

$$\Omega_h^{(\ell)} = \mathrm{softmax}\left(\frac{Q_h^{(\ell)} K_h^{(\ell)\top}}{\sqrt{d_h}}\right), \qquad R_h^{(\ell)} = \Omega_h^{(\ell)} V_h^{(\ell)}.$$

The multi-head attention output is

$$A_t^{(\ell)} = \left[R_1^{(\ell)}, \ldots, R_H^{(\ell)}\right] W_O^{(\ell)} + b_O^{(\ell)},$$

and enters through a residual connection,

$$\widetilde{H}_t^{(\ell)} = H_t^{(\ell-1)} + A_t^{(\ell)}.$$

Each block then applies a tokenwise feed-forward network,

$$\mathrm{FFN}^{(\ell)}(H) = W_2^{(\ell)}\,\mathrm{GELU}\left(W_1^{(\ell)}\,\mathrm{LN}_2^{(\ell)}(H) + b_1^{(\ell)}\right) + b_2^{(\ell)},$$

with hidden dimension 256, followed by a second residual connection,

$$H_t^{(\ell)} = \widetilde{H}_t^{(\ell)} + \mathrm{FFN}^{(\ell)}\left(\widetilde{H}_t^{(\ell)}\right).$$

The final quantitative specification uses zero dropout.

After the second block, a final layer normalization produces

$$\bar{H}_t = \mathrm{LN}_f(H_t^{(2)}).$$

A shared linear head applied separately to each of the $J$ firm tokens produces four standardized distributional outputs,

$$\widehat{z}_{j,t+1}^{m} = W_m \bar{h}_{j,t} + b_m \in \mathbb{R}^4, \qquad j = 1, \ldots, J,$$

while a separate head applied to the CLS representation produces the log-marginal-utility prediction,

$$\widehat{z}_t^{\ell} = W_\ell \bar{h}_{\mathrm{CLS},t} + b_\ell.$$

After undoing target standardization, the firm-token outputs give $\widehat{\Delta m}_{j,t+1}$, where

$$m_{j,t} = \mu_{j,t}\left(\bar{k}_{j,t}, 1, \bar{b}_{j,t}, \bar{\varepsilon}_{j,t}\right).$$

The perceived distributional law is therefore

$$\widehat{m}_{j,t+1} = m_{j,t} + \widehat{\Delta m}_{j,t+1},$$

while the CLS head yields

$$\widehat{\ell}_t = \log \widehat{U_C(C_t)}.$$

This residual forecasting law is distinct from the internal residual connections of the transformer blocks.

### A.3.3 Fixed Library and Interpolation

The firm application uses the fixed distributional library

$$\mathcal{L} = \{\mu^1, \ldots, \mu^S\}, \qquad S = 300.$$

The library is constructed before the final transformer fixed-point iteration. Candidate distributions include the stationary distribution, capital and leverage perturbations around the stationary state, and distributions generated by a preliminary low-dimensional Krusell–Smith-style solution. The latter provides a warm start for the library and initial policy arrays only; the perceived law used in the final equilibrium is the transformer law.

For interpolation, I augment the standardized distributional coordinates with aggregate capital, debt, and leverage implied by the same distributional moment vector. Define

$$L(m) = \frac{B(m)}{K(m)}$$

and

$$c_F(m) = (\mathsf{S}_m m,\, \lambda_K K(m),\, \lambda_B B(m),\, \lambda_L L(m))\,,$$

where $\mathsf{S}_m$ standardizes the distributional coordinates and $(\lambda_K, \lambda_B, \lambda_L)$ determine the relative scaling of the aggregate coordinates.

At library node $(\mu^s, Z)$, the transformer generates

$$\widehat{m}'_{sZ} = \widehat{m}'(\mu^s, Z; \theta).$$

Let $\mathcal{N}(s, Z)$ contain the $M = 4$ library distributions nearest to this forecast under Euclidean distance in $c_F$-space. For $r \in \mathcal{N}(s, Z)$, define

$$d_{sZ,r} = \|c_F(\widehat{m}'_{sZ}) - c_F(m(\mu^r))\|_2\,.$$

In the absence of an exact match, the interpolation weights are

$$\omega_{sZ,r} = \frac{d_{sZ,r}^{-2}}{\sum_{\ell \in \mathcal{N}(s,Z)} d_{sZ,\ell}^{-2}}, \qquad r \in \mathcal{N}(s, Z), \tag{36}$$

with unit weight assigned to an exact match.

### A.3.4 Continuation Values, Default, and Debt Pricing

At current library state $(\mu^s, Z)$, the transformer provides both $\widehat{m}'_{sZ}$ and the auxiliary marginal-utility prediction $\widehat{\ell}_{sZ}$. For a future library node $r \in \mathcal{N}(s, Z)$ and aggregate state $Z'$, the perceived stochastic discount factor is

$$\widehat{M}_{sZ,rZ'} = \beta \left( \frac{\widehat{C}_{rZ'}}{\widehat{C}_{sZ}} \right)^{-\gamma}, \tag{37}$$

where $\widehat{C}_{sZ} \equiv \widehat{C}(\mu^s, Z)$.

For a continuing firm choosing $(k', b')$, the perceived continuation value is

$$\mathcal{C}_{sZ}(\varepsilon, k', b') = \sum_{Z'} P_Z(Z, Z') \sum_{r \in \mathcal{N}(s,Z)} \omega_{sZ,r} \sum_{\varepsilon'} P_\varepsilon(\varepsilon, \varepsilon') \widehat{M}_{sZ,rZ'} V^r(Z', \varepsilon', k', b'). \tag{38}$$

The same perceived transition prices one-period defaultable debt. For a prospective choice $(k', b')$, the recovery rate in next period is

$$\mathcal{R}(k', b') = \min \left\{ 1, \frac{\theta_k (1-\delta) k'}{b'} \right\}, \qquad b' > 0.$$

Using the smoothed default probability described in Appendix B.4, the debt price is

$$\begin{aligned} q^d_{sZ}(\varepsilon, k', b') = \sum_{Z'} P_Z(Z, Z') \sum_{r \in \mathcal{N}(s,Z)} \omega_{sZ,r} \sum_{\varepsilon'} P_\varepsilon(\varepsilon, \varepsilon') \widehat{M}_{sZ,rZ'} \\ \times \left[ 1 - p^{d,r}(Z', \varepsilon', k', b') + p^{d,r}(Z', \varepsilon', k', b') \mathcal{R}(k', b') \right]. \end{aligned} \tag{39}$$

At aggregate state $(\mu^s, Z)$, the current payout associated with $(k', b')$ is

$$d_{sZ}(\varepsilon, k, b, k', b') = \pi\left(\varepsilon, k; Z, \widehat{w}_{sZ}\right) + q^d_{sZ}(\varepsilon, k', b') b' - b - \mathcal{I}(k, k').$$

The continuing-firm problem is

$$V^{c,s}(Z, \varepsilon, k, b) = \max_{k', b'} \left\{ \mathcal{D}\left( d_{sZ}(\varepsilon, k, b, k', b') \right) + \mathcal{C}_{sZ}(\varepsilon, k', b') \right\}, \tag{40}$$

subject to

$$b' \leq \theta_b \theta_k k'.$$

The equity value of default remains normalized to zero. The hard default rule is therefore

$$V^s = \max\left\{ V^{c,s}, 0 \right\}, \qquad d^{d,s} = \mathbf{1}\left\{ V^{c,s} < 0 \right\},$$

with the smooth counterpart used in the numerical implementation as described in Appendix B.4. Debt prices, default probabilities, and firm values are iterated jointly to convergence for fixed transformer forecasts and interpolation weights.

### A.3.5 Simulation and Market Clearing

The realized distribution generally does not coincide with a library node. At each simulated date, I compute $c_F(m(\mu_t))$, identify the nearest library distributions, and use the same inverse-distance interpolation scheme to recover the current capital and debt policies and default probabilities. Denote the resulting policies by

$$g_k(\varepsilon, k, b, Z_t, \mu_t), \qquad g_b(\varepsilon, k, b, Z_t, \mu_t).$$

Because default occurs before production, only continuing mass produces and hires labor. Current output and labor therefore use the continuation probability $p_t^c = 1 - p_t^d$. The next distribution is computed using the non-stochastic operator

$$\mu_{t+1} = \mathcal{Y}_F\left(\mu_t, g_k, g_b, p^d, Z_t, Z_{t+1}\right).$$

For continuing firms, chosen capital and debt are represented as lotteries over neighboring points on the numerical $(k, b)$ grid. Firm mass is then allocated across these grid points and next-period idiosyncratic-productivity states according to $P_\varepsilon$. Defaulting mass exits and is replaced one-for-one by entrants with capital $k^e = K^{ss}$, zero debt, and productivity drawn from the stationary idiosyncratic distribution. Thus the simulation propagates the complete continuum distribution rather than a finite population of firms.

For market clearing, aggregate capital evolves according to

$$K_{t+1} = \int p_t^c(\varepsilon, k, b) g_k(\varepsilon, k, b, Z_t, \mu_t)\, d\mu_t + k^e \int p_t^d(\varepsilon, k, b)\, d\mu_t.$$

Stock-based investment is

$$I_t = K_{t+1} - (1 - \delta)K_t.$$

Using $p_t^c + p_t^d = 1$, aggregate investment can equivalently be written as

$$I_t = \int p_t^c\, i(k, g_k)\, d\mu_t + \int p_t^d\, [k^e - (1 - \delta)k]\, d\mu_t.$$

Thus, $I_t$ is a stock-accounting measure rather than the resource cost of capital formation.

For continuing firms, partial irreversibility generates the resource loss

$$\Psi(k, k') = (1 - \theta_k)[-i(k, k')]_+,$$

so that

$$i(k, k') + \Psi(k, k') = \mathcal{I}(k, k').$$

For defaulting firms, creditor recovery is

$$\Lambda(k, b) = \min\left\{b,\, \theta_k(1 - \delta)k\right\},$$

and the associated resource loss is

$$\Psi^d(k, b) = (1 - \delta)k - \Lambda(k, b).$$

Hence, the resource use associated with default and replacement entry is

$$k^e - (1 - \delta)k + \Psi^d(k, b) = k^e - \Lambda(k, b).$$

Creditor recovery therefore offsets part of the resources required to install the replacement entrant's capital.

Current consumption is therefore computed from the resource constraint

$$\begin{aligned} C_t + I_t + \int p_t^c\, \Psi(k, g_k)\, d\mu_t + \int p_t^d\, \Psi^d(k, b)\, d\mu_t \\ + \int p_t^c f\, d\mu_t + \int p_t^c \Phi_e(d_t)\, d\mu_t = Y_t. \end{aligned}$$

The perceived aggregate consumption object implied by the transformer enters the recursive problem solved at the library nodes. Through the household intratemporal condition it determines the perceived wage used in current profits and continuation decisions. This perceived wage is not used to impose current-period market clearing along the simulated equilibrium path. For each realized $(\mu_t, Z_t)$, the current wage is instead obtained exactly from

$$\mathcal{R}_w(w; \mu_t, Z_t) \equiv w - \chi C(w; \mu_t, Z_t)^\gamma = 0, \tag{41}$$

where $C(w; \mu_t, Z_t)$ is the consumption level implied by the resource constraint above. Along the simulated path, firm policies are given, so this is a scalar market-clearing problem. Current-period market clearing is therefore imposed exactly in simulation, while transformer predictions determine the perceived aggregate objects entering the recursive problem.

### A.3.6 Equilibrium Iteration

The initial library and policy arrays are obtained from the stationary economy and the low-dimensional warm start described above. The library is then held fixed throughout the final DST equilibrium iteration.

At each outer iteration, the simulated path generates supervised observations of

$$(X_t, \Delta m_{t+1}, \log U_C(C_t)) .$$

I supplement these observations with one-step transitions generated at the library nodes under the current policy functions. This ensures that the transformer is trained not only along the simulated path but also at the distributional states at which the Bellman equation is evaluated.

After training, the transformer is evaluated at every $(\mu^s, Z)$. For numerical stability, the resulting perceived distributional forecasts and marginal-utility predictions are damped across outer iterations:

$$\widehat{m}'^{,(h)}_{sZ} = (1-\lambda_\theta)\widehat{m}'^{,(h-1)}_{sZ} + \lambda_\theta \widehat{m}'^{,\mathrm{raw}}_{sZ},$$
$$\widehat{\ell}^{(h)}_{sZ} = (1-\lambda_\theta)\widehat{\ell}^{(h-1)}_{sZ} + \lambda_\theta \widehat{\ell}^{\mathrm{raw}}_{sZ},$$

where $h$ indexes the outer equilibrium iteration. The firm problem is then resolved using the associated interpolation weights, debt prices, and default probabilities.

The updated capital and debt policies are also damped,

$$g^{(h)} = (1-\lambda_g)g^{(h-1)} + \lambda_g g^{\mathrm{raw},(h)}, \qquad g = (g_k, g_b).$$

The resulting policies are simulated using $\mathcal{Y}_F$, imposing (41) at each date. The new distributional transitions update the transformer training sample, and the procedure is repeated until the policy and perceived-law convergence criteria in Section 3 are satisfied.

### A.3.7 Numerical Implementation

The economic and machine-learning components are implemented separately. NumPy is used for dense array operations, Numba for JIT-compiled dynamic programming and distribution transitions, and PyTorch for transformer training, inference, and attention extraction. Bellman updates, expected-continuation arrays, debt-price schedules, and distributional transitions are parallelized across CPU cores where possible. Transformer operations are run on CUDA, Apple `mps`, or CPU depending on available hardware.

Attention matrices are extracted under no-gradient evaluation from the same transformer

that supplies the perceived distributional law used in (38) and (39).

# B Mathematical Appendix

This appendix collects the proofs of Remark 1, Remark 2, and Proposition 1, describes the treatment of the default/exit decision and the associated choice-specific shocks, and derives the policy-sensitivity decomposition for the quantitative transformer.

## B.1 Proof of Remark 1

*Proof.* Since $K = \sum_{j=1}^{J} s_j$, an additional unit of capital raises $K$ one-for-one regardless of the bin in which it is located. Hence

$$\frac{\partial K'}{\partial s_j} = \frac{\partial K'}{\partial K}\frac{\partial K}{\partial s_j} = \kappa\alpha^2 Z K^{\alpha-1}$$

for every $j$, and therefore

$$\omega_j^E = \frac{\kappa\alpha^2 Z K^{\alpha-1}}{J\,\kappa\alpha^2 Z K^{\alpha-1}} = \frac{1}{J}.$$

On the transformer side, state-independent attention implies

$$\frac{\partial r^W}{\partial s_j} = \omega_j^T.$$

The chain rule then gives

$$\frac{\partial \widehat{K}'}{\partial s_j} = g_r(r^W, Z)\omega_j^T.$$

Using $\sum_{\ell=1}^{J} \omega_\ell^T = 1$,

$$\frac{\partial \widehat{K}'/\partial s_j}{\sum_{\ell=1}^{J} \partial \widehat{K}'/\partial s_\ell} = \frac{g_r(r^W, Z)\omega_j^T}{g_r(r^W, Z)\sum_{\ell=1}^{J}\omega_\ell^T} = \omega_j^T.$$

Finally, because $\widehat{K}' = K'$ on an open set, their partial derivatives coincide. It follows that

$$\omega_j^T = \omega_j^E = \frac{1}{J}.$$

□

## B.2 Proof of Remark 2

*Proof.* From $s_j = \mu_j k_j$,

$$K' = \kappa_\alpha Z \sum_{j=1}^{J} \mu_j^{1-\alpha} s_j^\alpha,$$

and therefore

$$\frac{\partial K'}{\partial s_j} = \kappa_\alpha \alpha Z \mu_j^{1-\alpha} s_j^{\alpha-1} = \kappa_\alpha \alpha Z k_j^{\alpha-1}.$$

Hence

$$\frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \frac{k_j^{\alpha-1}}{\sum_{\ell=1}^{J} k_\ell^{\alpha-1}}.$$

Under the frozen-attention representation,

$$r^W(\widetilde{s}; s) = \sum_{j=1}^{J} \omega_j^T(s) \widetilde{s}_j,$$

so that

$$\left.\frac{\partial \widehat{K}'}{\partial \widetilde{s}_j}\right|_{\widetilde{s}=s} = g_r(r^W, Z)\omega_j^T(s).$$

First-order equality between $\widehat{K}'$ and $K'$ therefore implies

$$g_r(r^W, Z)\omega_j^T(s) = \frac{\partial K'}{\partial s_j}.$$

Summing over $j$ and using $\sum_j \omega_j^T(s) = 1$ yields

$$g_r(r^W, Z) = \sum_{\ell=1}^{J} \frac{\partial K'}{\partial s_\ell}.$$

Substitution gives

$$\omega_j^T(s) = \frac{\partial K'/\partial s_j}{\sum_{\ell=1}^{J} \partial K'/\partial s_\ell} = \omega_j^E(s).$$

Since $0 < \alpha < 1$, $k^{\alpha-1}$ is decreasing in $k$, completing the result. □

## B.3 Proof of Proposition 1

*Proof.* Fix an aggregate state $z$ and a distributional state $s = (s_1, \ldots, s_J)$ in the open set on which the transformer represents the equilibrium mapping exactly. By assumption,

$$F(s, z) = \widehat{y} = g(r(s, z), z),$$

where the attention representation is

$$r(s, z) = \sum_{m=1}^{J} \omega_m^T(s, z) v_m(s, z), \qquad r(s, z) \in \mathbb{R}^d.$$

Here, $\omega_m^T(s, z)$ is scalar, while $v_m(s, z) \in \mathbb{R}^d$.

Consider a local perturbation of a single distributional component $s_j$, holding the aggregate state $z$ fixed. Because the equality

$$F(s, z) = g(r(s, z), z)$$

holds on an open set, rather than only at an isolated point, the partial derivatives of the two mappings with respect to $s_j$ must also coincide. Applying the chain rule therefore gives

$$\frac{\partial F(s, z)}{\partial s_j} = \nabla_r g(r(s, z), z)^\top \frac{\partial r(s, z)}{\partial s_j}.$$

There is no additional derivative through $z$, since $z$ is held fixed when taking the partial derivative with respect to $s_j$.

It remains to characterize the response of the attention representation $r(s, z)$. Differentiating

$$r(s, z) = \sum_{m=1}^{J} \omega_m^T(s, z) v_m(s, z)$$

with respect to $s_j$, and applying the product rule to each term in the sum, yields

$$\begin{aligned} \frac{\partial r(s, z)}{\partial s_j} &= \sum_{m=1}^{J} \frac{\partial}{\partial s_j} \left[ \omega_m^T(s, z) v_m(s, z) \right] \\ &= \sum_{m=1}^{J} \frac{\partial \omega_m^T(s, z)}{\partial s_j} v_m(s, z) + \sum_{m=1}^{J} \omega_m^T(s, z) \frac{\partial v_m(s, z)}{\partial s_j}. \end{aligned}$$

The first term captures changes in the aggregate representation that arise because a perturbation to $s_j$ changes how attention is allocated across tokens. The second captures changes

in the information being aggregated, holding the attention weights themselves fixed.

Substituting this expression into the chain rule gives

$$\frac{\partial F(s,z)}{\partial s_j} = \nabla_r g(r,z)^\top \left[ \sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} v_m(s,z) + \sum_{m=1}^{J} \omega_m^T(s,z) \frac{\partial v_m(s,z)}{\partial s_j} \right],$$

which is the decomposition stated in Proposition 1.

The normalization of the attention weights provides an equivalent representation of the first term. Since

$$\sum_{m=1}^{J} \omega_m^T(s,z) = 1$$

for every admissible $(s,z)$, differentiating this identity with respect to $s_j$ gives

$$\sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} = 0.$$

Hence,

$$\sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} v_m(s,z) = \sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} \left[ v_m(s,z) - r(s,z) \right].$$

The policy sensitivity can therefore equivalently be written as

$$\frac{\partial F(s,z)}{\partial s_j} = \nabla_r g(r,z)^\top \left[ \sum_{m=1}^{J} \frac{\partial \omega_m^T(s,z)}{\partial s_j} \left( v_m(s,z) - r(s,z) \right) + \sum_{m=1}^{J} \omega_m^T(s,z) \frac{\partial v_m(s,z)}{\partial s_j} \right].$$

This form makes clear that the attention-reallocation channel operates by shifting weight across value representations relative to the current attention aggregate, while the value channel captures how the representations themselves respond to the distributional state. □

## B.4 Smooth Approximation to the Default Decision

In the model presented in the main text, equity holders compare the value of continuing, $V_t^c(\varepsilon, k, b; \Omega_t)$, with the value of default, $V^d$, which is normalized to zero. Under the deterministic default rule, the firm value is

$$V_t^{\text{hard}}(\varepsilon, k, b; \Omega_t) = \max \left\{ V_t^c(\varepsilon, k, b; \Omega_t), V^d \right\}, \tag{42}$$

and the default decision is

$$d_t^{\text{hard}}(\varepsilon, k, b; \Omega_t) = \mathbf{1}\left\{V_t^c(\varepsilon, k, b; \Omega_t) < V^d\right\}.$$

This rule introduces a kink in the firm value at $V_t^c(\varepsilon, k, b; \Omega_t) = V^d$ and a discontinuity in the default policy. For the numerical implementation, I replace the hard maximum with a smooth log-sum-exp approximation.

Specifically, consider transitory choice-specific shocks to the continuation and default options. Equity holders compare

$$V_t^c(\varepsilon, k, b; \Omega_t) + \xi_t^c \qquad \text{and} \qquad V^d + \xi_t^d,$$

where $\xi_t^c$ and $\xi_t^d$ are independent type-I extreme-value random variables with common scale parameter $\sigma_d > 0$. The shocks are independent across firms and over time and affect only the current default decision. They are integrated out before the continuation value is evaluated and are not included as additional state variables.

Conditional on the economic state

$$s_t \equiv (\varepsilon, k, b; \Omega_t),$$

the firm defaults whenever

$$V^d + \xi_t^d > V_t^c(s_t) + \xi_t^c.$$

Because the difference of two independent type-I extreme-value variables follows a logistic distribution, the conditional default probability is

$$\begin{aligned} p_t^d(s_t) &\equiv \Pr\left[V^d + \xi_t^d > V_t^c(s_t) + \xi_t^c \mid s_t\right] \\ &= \frac{\exp\left(V^d/\sigma_d\right)}{\exp\left(V_t^c(s_t)/\sigma_d\right) + \exp\left(V^d/\sigma_d\right)} \\ &= \left[1 + \exp\left(\frac{V_t^c(s_t) - V^d}{\sigma_d}\right)\right]^{-1}. \end{aligned} \tag{43}$$

The corresponding continuation probability is

$$\begin{aligned} p_t^c(s_t) &\equiv 1 - p_t^d(s_t) \\ &= \frac{\exp\left(V_t^c(s_t)/\sigma_d\right)}{\exp\left(V_t^c(s_t)/\sigma_d\right) + \exp\left(V^d/\sigma_d\right)} \\ &= \left[1 + \exp\left(\frac{V^d - V_t^c(s_t)}{\sigma_d}\right)\right]^{-1}. \end{aligned}$$

After integrating over the choice-specific shocks, the expected firm value is, up to an additive constant that is common across choices,

$$V_t^{\sigma_d}(s_t) = \sigma_d \log \left[ \exp\left( \frac{V_t^c(s_t)}{\sigma_d} \right) + \exp\left( \frac{V^d}{\sigma_d} \right) \right]. \tag{44}$$

This is the smooth counterpart of the hard maximum in (42). An equivalent expression is

$$\begin{aligned} V_t^{\sigma_d}(s_t) = & \max\left\{ V_t^c(s_t), V^d \right\} \\ & + \sigma_d \log \left[ 1 + \exp\left( - \frac{\left| V_t^c(s_t) - V^d \right|}{\sigma_d} \right) \right]. \end{aligned} \tag{45}$$

Equation (45) is also the numerically stable expression used in the implementation, since it avoids exponentiating large value-function levels.

The derivative of the smoothed value with respect to the continuation value is

$$\frac{\partial V_t^{\sigma_d}(s_t)}{\partial V_t^c(s_t)} = \frac{\exp\left( V_t^c(s_t)/\sigma_d \right)}{\exp\left( V_t^c(s_t)/\sigma_d \right) + \exp\left( V^d/\sigma_d \right)} = p_t^c(s_t).$$

Thus, the marginal effect of a change in the continuation value on the ex-ante firm value equals the probability that the firm continues. Similarly,

$$\frac{\partial V_t^{\sigma_d}(s_t)}{\partial V^d} = p_t^d(s_t).$$

The second derivative is

$$\frac{\partial^2 V_t^{\sigma_d}(s_t)}{\partial V_t^c(s_t)^2} = \frac{1}{\sigma_d} p_t^c(s_t) p_t^d(s_t),$$

which is largest near the default boundary and converges to zero away from it.

**Relation to the hard default rule.** The smooth formulation nests the deterministic model as a limiting case. For any state such that $V_t^c(s_t) \neq V^d$,

$$\lim_{\sigma_d \to 0^+} p_t^d(s_t) = \mathbf{1}\left\{ V_t^c(s_t) < V^d \right\},$$

and

$$\lim_{\sigma_d \to 0^+} p_t^c(s_t) = \mathbf{1}\left\{ V_t^c(s_t) > V^d \right\}.$$

At the indifference point $V_t^c(s_t) = V^d$, both choices receive probability one half. Moreover,

$$\lim_{\sigma_d \to 0^+} V_t^{\sigma_d}(s_t) = \max\left\{ V_t^c(s_t), V^d \right\}.$$

The approximation error satisfies

$$0 \leq V_t^{\sigma_d}(s_t) - \max\left\{V_t^c(s_t), V^d\right\} \leq \sigma_d \log 2.$$

Hence, the smoothed value converges uniformly to the hard maximum as $\sigma_d \to 0^+$.

**Use in the quantitative implementation.** The baseline implementation sets $\sigma_d = 10^{-2}$. This value is small relative to the variation in firm continuation values and therefore makes the smooth policy close to the deterministic default rule. Its purpose is numerical: it replaces the discontinuous default indicator with a continuous probability, stabilizes the distributional transition, and avoids abrupt changes in debt prices and policy functions when a firm moves across the default boundary.

In particular, the debt-pricing equation uses the smoothed default probability together with the recovery rate defined in equation (24),

$$q_t(\varepsilon, k', b') = \mathbb{E}_t\left[M_{t,t+1}\left(p_{t+1}^c(\varepsilon_{t+1}, k', b'; \Omega_{t+1}) + p_{t+1}^d(\varepsilon_{t+1}, k', b'; \Omega_{t+1})\mathcal{R}(k', b')\right)\right],$$

where $p_{t+1}^c = 1 - p_{t+1}^d$.

Likewise, the Young distributional transition assigns the mass $p_t^c(\varepsilon, k, b; \Omega_t)$ to the continuing-firm transition and the mass $p_t^d(\varepsilon, k, b; \Omega_t)$ to the default-and-entry transition. Thus, the smoothing does not alter the set of economic choices or introduce an additional persistent state. It only replaces the hard allocation of firm mass at the default boundary with its smooth probabilistic counterpart. The quantitative implementation uses (43)–(44) as a numerical approximation. As $\sigma_d$ becomes small, the smoothed economy converges to the deterministic-default economy.

## B.5 Policy-Sensitivity Decomposition

This appendix extends the decomposition in Proposition 1 to the multi-layer transformer used in the quantitative application. The purpose is to characterize the local response of aggregate outcomes to economically feasible redistributions of firm mass and to separate the two channels generated within self-attention from the remaining paths in the quantitative architecture and from the direct current-state term in the distributional forecasting law.

**Compensated mass-reallocation directions.** The quantitative exercise considers local, mass-preserving reallocations across capital–debt–productivity cells. Recall that the

distributional forecast coordinates for cell $j$ are

$$m_{j,t} = \mu_{j,t}\left(\bar{k}_{j,t}, 1, \bar{b}_{j,t}, \bar{\varepsilon}_{j,t}\right).$$

For a cell $j$ with $0 < \mu_{j,t} < 1$, introduce a scalar perturbation $\eta$. A positive perturbation reallocates mass toward cell $j$,

$$m_{j,t}(\eta) = (\mu_{j,t} + \eta)\left(\bar{k}_{j,t}, 1, \bar{b}_{j,t}, \bar{\varepsilon}_{j,t}\right),$$

while all other cells lose mass proportionally,

$$m_{i,t}(\eta) = \left(1 - \frac{\eta}{1 - \mu_{j,t}}\right) m_{i,t}, \qquad i \neq j.$$

A negative value of $\eta$ implements the reverse reallocation. Total probability mass is therefore preserved:

$$\mu_{j,t} + \eta + \sum_{i \neq j} \left(1 - \frac{\eta}{1 - \mu_{j,t}}\right) \mu_{i,t} = 1.$$

Let

$$\partial_j \equiv \left.\frac{d}{d\eta}\right|_{\eta=0}$$

denote the directional derivative associated with this compensated reallocation. The corresponding derivatives of the distributional coordinates are

$$\partial_j m_{j,t} = \left(\bar{k}_{j,t}, 1, \bar{b}_{j,t}, \bar{\varepsilon}_{j,t}\right), \qquad \partial_j m_{i,t} = -\frac{m_{i,t}}{1 - \mu_{j,t}}, \quad i \neq j.$$

In particular,

$$\partial_j \mu_{j,t} = 1, \qquad \partial_j \mu_{i,t} = -\frac{\mu_{i,t}}{1 - \mu_{j,t}}, \quad i \neq j.$$

The transformer does not take the mass-weighted moment vectors $m_{i,t}$ directly as inputs. Let $\phi_i$ denote the map from the distributional forecast coordinates to the economic coordinates supplied to firm token $i$. For a positive-mass cell,

$$\phi_i(m_{i,t}) = \left(\frac{m_{i,k,t}}{m_{i,\mu,t}}, m_{i,\mu,t}, \frac{m_{i,b,t}}{m_{i,\mu,t}}, \frac{m_{i,\varepsilon,t}}{m_{i,\mu,t}}\right) = \left(\bar{k}_{i,t}, \mu_{i,t}, \bar{b}_{i,t}, \bar{\varepsilon}_{i,t}\right).$$

The perturbed moment vectors are first mapped back into these economic coordinates and then used to construct the transformer input $X_t(\eta)$. Because the perturbation scales all coordinates of $m_{i,t}$ proportionally within each cell, away from admissibility clipping it leaves $\bar{k}_{i,t}$, $\bar{b}_{i,t}$, and $\bar{\varepsilon}_{i,t}$ unchanged. The local experiment therefore changes the allocation of firm

mass across the cross section while holding within-cell characteristics fixed.

**Direct and transformer-mediated sensitivity.** The transformer forecasts changes in the mass-weighted distributional coordinates,

$$\widehat{m}_{t+1}(\eta) = m_t(\eta) + \widehat{\Delta m}_{t+1}(\eta), \qquad \widehat{\Delta m}_{t+1}(\eta) = \mathcal{T}_{\theta,m}(X_t(\eta)),$$

where the network output has been transformed back into economic units.

Let

$$\widehat{Y}_{t+1}(\eta) = A(\widehat{m}_{t+1}(\eta))$$

denote a scalar aggregate outcome implied by the predicted distribution. The two outcomes used in the quantitative decomposition are aggregate capital,

$$A_K(m) = \sum_i m_{i,k},$$

and aggregate leverage,

$$A_{\text{lev}}(m) = \frac{\sum_i m_{i,b}}{\sum_i m_{i,k}}.$$

Neither object is a separate transformer prediction target.

The directional sensitivity of the implied aggregate outcome is

$$\partial_j \widehat{Y}_{t+1} = \nabla_m A(\widehat{m}_{t+1})^\top \left[\partial_j m_t + \partial_j \widehat{\Delta m}_{t+1}\right]. \tag{46}$$

I refer to

$$\mathcal{D}_j \equiv \nabla_m A(\widehat{m}_{t+1})^\top \partial_j m_t$$

as the *direct current-state component*, and to

$$\mathcal{T}_j \equiv \nabla_m A(\widehat{m}_{t+1})^\top \partial_j \widehat{\Delta m}_{t+1}$$

as the *transformer-mediated component*. This distinction arises from the explicit residual forecasting law $\widehat{m}_{t+1} = m_t + \mathcal{T}_{\theta,m}(X_t)$ and is separate from the internal residual connections of the transformer architecture.

**Self-attention channels.** Consider transformer block $\ell$ and attention head $h$. Suppressing the time subscript, the pre-normalized hidden states generate

$$Q_h^{(\ell)}, \qquad K_h^{(\ell)}, \qquad V_h^{(\ell)},$$

and the attention matrix and head output are

$$\Omega_h^{(\ell)} = \mathrm{softmax}\left(\frac{Q_h^{(\ell)} K_h^{(\ell)\top}}{\sqrt{d_h}}\right), \qquad R_h^{(\ell)} = \Omega_h^{(\ell)} V_h^{(\ell)}.$$

Taking the directional derivative along the compensated mass-reallocation direction gives

$$\partial_j R_h^{(\ell)} = \underbrace{(\partial_j \Omega_h^{(\ell)}) V_h^{(\ell)}}_{\text{attention-reallocation channel}} + \underbrace{\Omega_h^{(\ell)} (\partial_j V_h^{(\ell)})}_{\text{value channel}}. \tag{47}$$

Equation (47) is the multi-token, multi-head counterpart of equation (7). The first term measures the response generated because the perturbation changes the attention weights, while the second measures the response through the values being aggregated holding those weights locally fixed.

After concatenating heads and applying the multi-head output projection, define

$$\partial_j A_R^{(\ell)} = \left[(\partial_j \Omega_1^{(\ell)}) V_1^{(\ell)}, \ldots, (\partial_j \Omega_H^{(\ell)}) V_H^{(\ell)}\right] W_O^{(\ell)},$$

and

$$\partial_j A_V^{(\ell)} = \left[\Omega_1^{(\ell)} (\partial_j V_1^{(\ell)}), \ldots, \Omega_H^{(\ell)} (\partial_j V_H^{(\ell)})\right] W_O^{(\ell)}.$$

Hence,

$$\partial_j A^{(\ell)} = \partial_j A_R^{(\ell)} + \partial_j A_V^{(\ell)}.$$

**Propagation through the full transformer.** The quantitative architecture contains internal residual connections and tokenwise feed-forward blocks in addition to self-attention. Let

$$\widetilde{H}^{(\ell)} = H^{(\ell-1)} + A^{(\ell)}$$

denote the representation after the attention residual connection, and write the feed-forward block as

$$M^{(\ell)} = \mathcal{M}^{(\ell)}\left(\widetilde{H}^{(\ell)}\right), \qquad H^{(\ell)} = \widetilde{H}^{(\ell)} + M^{(\ell)}.$$

The full local differential therefore satisfies

$$\partial_j H^{(\ell)} = \partial_j H^{(\ell-1)} + \partial_j A_R^{(\ell)} + \partial_j A_V^{(\ell)} + \partial_j M^{(\ell)},$$

where

$$\partial_j M^{(\ell)} = J_{\mathcal{M}^{(\ell)}}\left[\partial_j H^{(\ell-1)} + \partial_j A_R^{(\ell)} + \partial_j A_V^{(\ell)}\right].$$

For accounting purposes, I separate three sources of transformer-mediated sensitivity. Let $R$ and $V$ denote the contributions generated by the two terms in the self-attention product rule, and let $O$ collect the remaining network paths. At the embedded input,

$$\partial_j H_O^{(0)} = \partial_j H^{(0)}, \qquad \partial_j H_R^{(0)} = 0, \qquad \partial_j H_V^{(0)} = 0.$$

The learned CLS token and positional embeddings are fixed with respect to the distributional perturbation, so they do not generate an additional source of local sensitivity.

At transformer block $\ell$, first define

$$\partial_j \widetilde{H}_O^{(\ell)} = \partial_j H_O^{(\ell-1)},$$

$$\partial_j \widetilde{H}_R^{(\ell)} = \partial_j H_R^{(\ell-1)} + \partial_j A_R^{(\ell)},$$

and

$$\partial_j \widetilde{H}_V^{(\ell)} = \partial_j H_V^{(\ell-1)} + \partial_j A_V^{(\ell)}.$$

The feed-forward response is

$$\partial_j M^{(\ell)} = J_{\mathcal{M}^{(\ell)}} \left[ \partial_j \widetilde{H}_O^{(\ell)} + \partial_j \widetilde{H}_R^{(\ell)} + \partial_j \widetilde{H}_V^{(\ell)} \right].$$

Consistent with the quantitative accounting, changes generated through the feed-forward residual path are assigned to the remaining-network component:

$$\partial_j H_O^{(\ell)} = \partial_j \widetilde{H}_O^{(\ell)} + \partial_j M^{(\ell)},$$

$$\partial_j H_R^{(\ell)} = \partial_j \widetilde{H}_R^{(\ell)}, \qquad \partial_j H_V^{(\ell)} = \partial_j \widetilde{H}_V^{(\ell)}.$$

By construction,

$$\partial_j H^{(\ell)} = \partial_j H_O^{(\ell)} + \partial_j H_R^{(\ell)} + \partial_j H_V^{(\ell)}$$

at every layer.

Let $\mathcal{G}$ denote the final layer normalization, the tokenwise distributional readout, and the transformation of the predicted changes back into economic units. Linearizing this final mapping gives

$$\partial_j \widehat{\Delta m}_{t+1,c} = J_{\mathcal{G}}\, \partial_j H_c^{(L)}, \qquad c \in \{O, R, V\},$$

and therefore

$$\partial_j \widehat{\Delta m}_{t+1} = \partial_j \widehat{\Delta m}_{t+1,O} + \partial_j \widehat{\Delta m}_{t+1,R} + \partial_j \widehat{\Delta m}_{t+1,V}.$$

The corresponding contributions to the implied aggregate outcome are

$$\mathcal{T}_{j,c} = \nabla_m A(\widehat{m}_{t+1})^\top \partial_j \widehat{\Delta m}_{t+1,c}, \qquad c \in \{O, R, V\}.$$

Hence the exact signed local accounting is

$$\partial_j \widehat{Y}_{t+1} = \underbrace{\mathcal{D}_j}_{\text{direct current state}} + \underbrace{\mathcal{T}_{j,O}}_{\text{remaining network paths}} + \underbrace{\mathcal{T}_{j,R}}_{\text{attention reallocation}} + \underbrace{\mathcal{T}_{j,V}}_{\text{value}}. \tag{48}$$

The attention-reallocation and value terms in equation (48) are the quantitative counterparts of the two channels in Proposition 1. The remaining-network term collects paths that are present in the implemented multi-layer architecture but are outside the simple self-attention aggregation in the proposition, including the internal residual and feed-forward paths. Subsequent transformer layers, final normalization, and the output readout propagate these components to the distributional forecast.

**Finite-difference implementation and reported magnitudes.** For the numerical central difference, the perturbation size for cell $j$ is

$$\delta_{j,t} = \min\left\{\varepsilon,\, 0.45\mu_{j,t},\, 0.45(1-\mu_{j,t})\right\},$$

which ensures that both the positive and negative perturbations remain inside the probability simplex. The finite-difference sensitivity is

$$S_{j,t}^{Y,\mathrm{FD}} = \frac{\widehat{Y}_{t+1}(\delta_{j,t}) - \widehat{Y}_{t+1}(-\delta_{j,t})}{2\delta_{j,t}}.$$

These finite differences are used as a numerical check on the derivatives obtained by automatic differentiation.

The quantitative results report ergodic averages of absolute local directional sensitivities. Let $\varpi_{j,t}$ denote the cell-mass weight used when averaging across perturbation directions. For component $c$, the reported magnitude is

$$\overline{S}_c = \frac{\sum_{t,j} \varpi_{j,t} \left|\mathcal{S}_{j,t,c}\right|}{\sum_{t,j} \varpi_{j,t}},$$

with the sample restricted to the endogenous default region when reporting default-margin statistics. Because absolute values are taken before averaging, these reported magnitudes

are not additive shares:

$$\overline{S}_{\text{total}} \neq \overline{S}_{\text{direct}} + \overline{S}_O + \overline{S}_R + \overline{S}_V$$

in general. The corresponding signed local decomposition in equation (48), however, closes up to numerical precision.

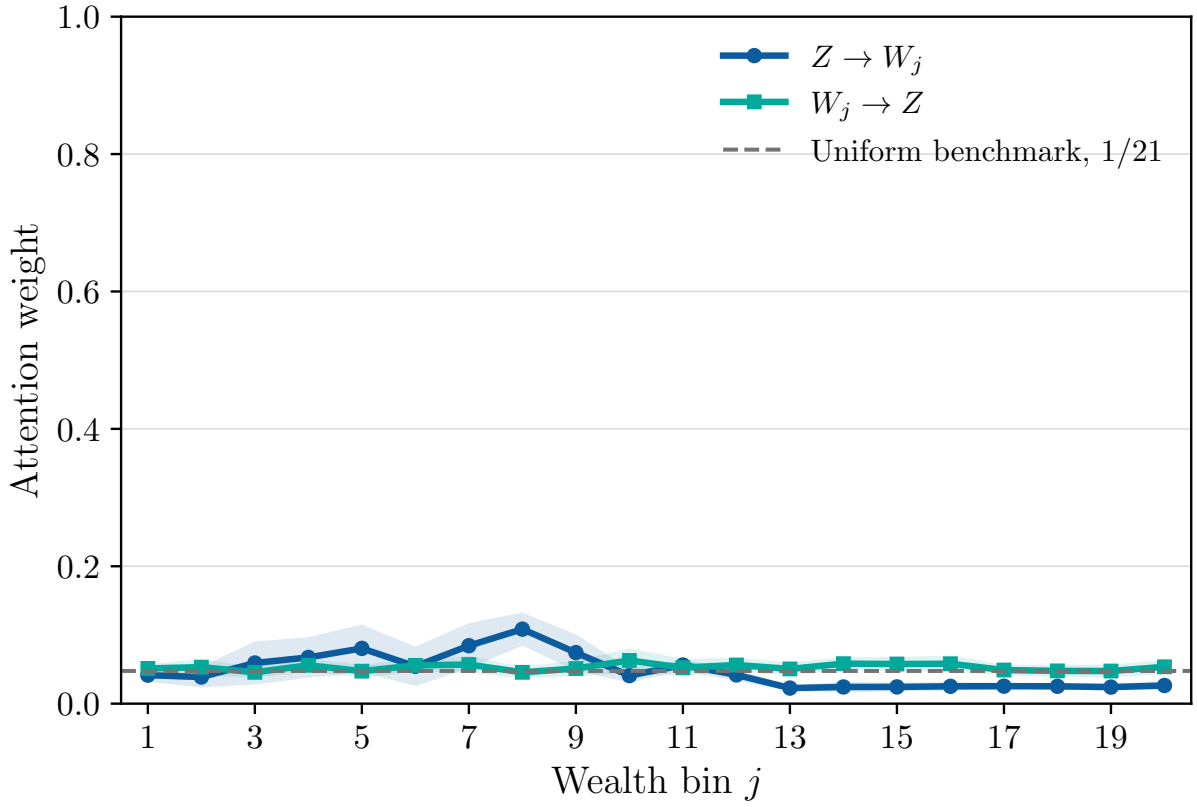
**Relation to the default-exposure diagnostic.** The default-sensitivity exercise in Appendix C is conceptually distinct from the decomposition above. There, the local default-probability map is held fixed and aggregate default exposure under the perturbed distribution is

$$D_t(\eta) = \sum_i \mu_{i,t}(\eta) p^d_{i,t}.$$

That exercise therefore measures the direct effect of reallocating firm mass across regions with different default probabilities. It is not a transformer-mediated policy-sensitivity decomposition and is not included among the aggregate outcomes reported in Table 2.

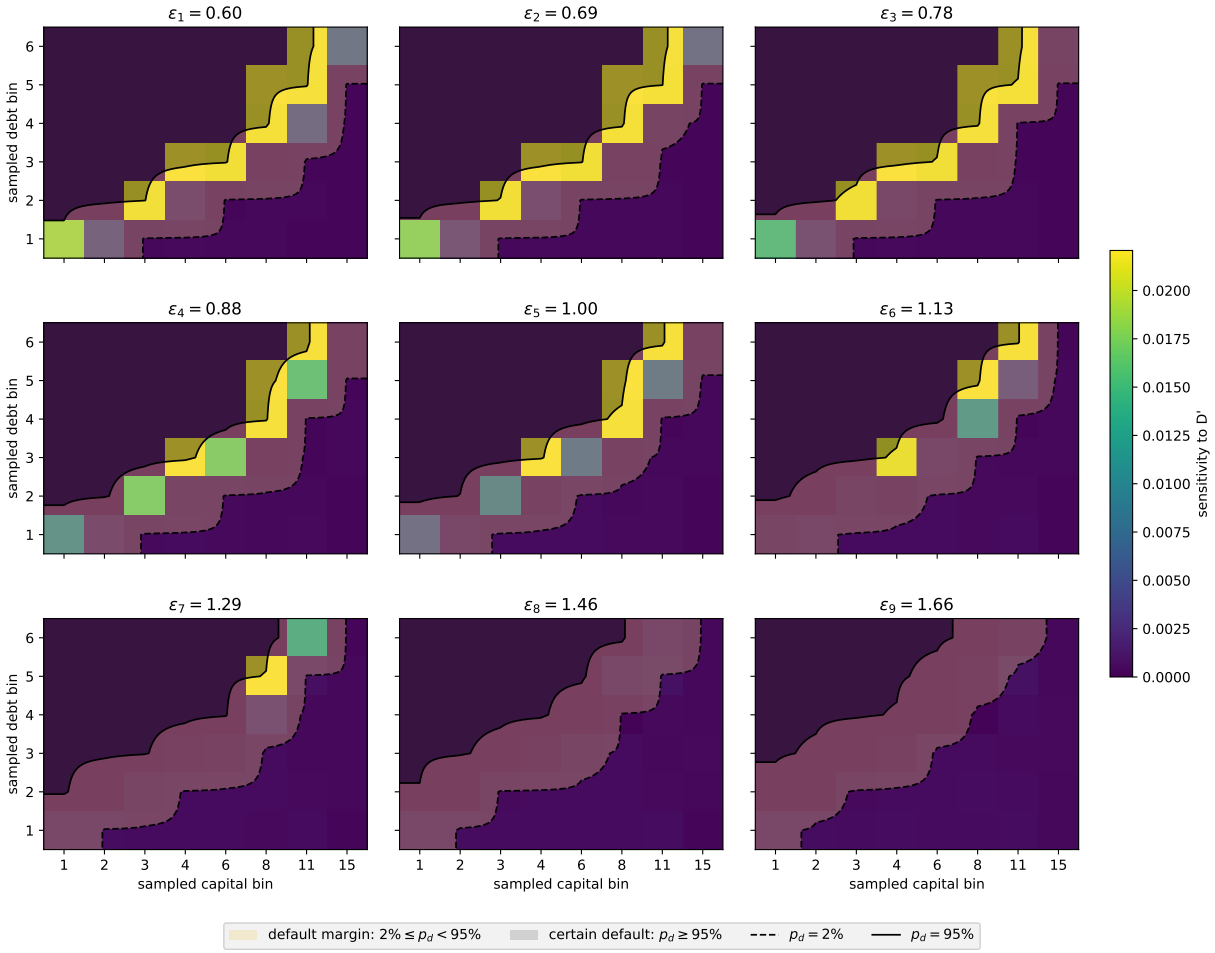
# C Additional Exhibits

Figure 7: Attention Between Wealth and Aggregate Productivity




*Notes:* The figure reports token-level directional attention between each wealth-bin token $W_j$ and aggregate productivity $Z$. The $Z \to W_j$ series measures the attention assigned to wealth bin $j$ when updating the aggregate-productivity token, while $W_j \to Z$ measures the attention assigned to aggregate productivity when updating wealth-bin token $j$. Lines report cross-run means across ten independent model solutions, each based on a distinct realization of aggregate productivity shocks and a 50,000-period simulation; shaded regions denote 95% confidence intervals across runs. The dashed horizontal line is the uniform-attention benchmark, $1/(J+1) = 1/21$.

Figure 8: Local Distributional Sensitivity of Aggregate Default



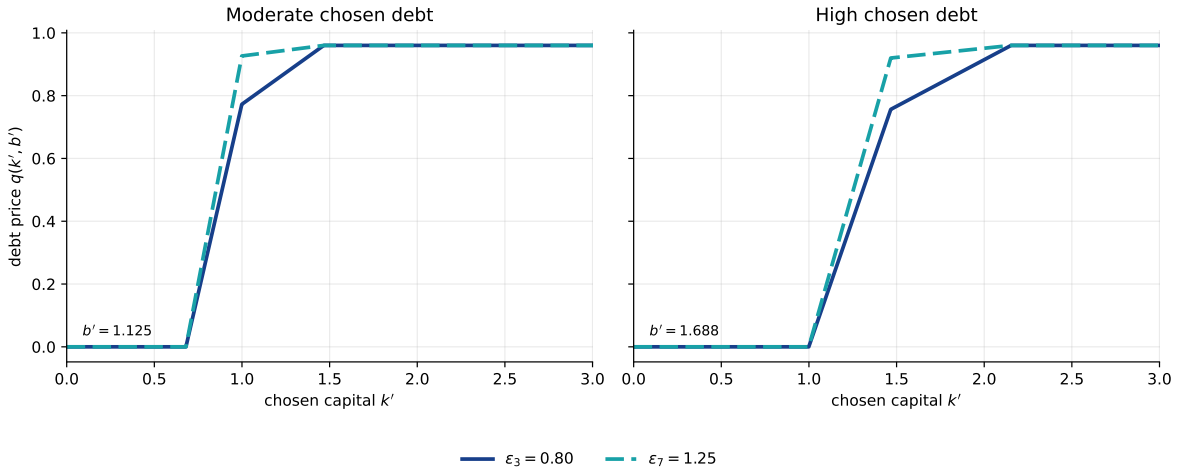

*Notes.* The figure reports the normalized magnitude of the local sensitivity of aggregate default exposure to mass-preserving perturbations of each capital–debt–productivity token, holding local default probabilities fixed. Sensitivities are averaged over the ergodic simulation. The solid and dashed contours denote the same endogenous default-risk regions used in Figure 5.

Table 2: Local Policy Sensitivity of Implied Aggregate Outcomes

| | | | | Transformer-mediated component | | | |
|---|---|---|---|---|---|---|---|
| Outcome | Region | Total | Direct | Transformer | Reallocation | Value | Other paths |
| Capital | All states | 0.320 | 0.521 | 0.209 | 0.077 | 0.043 | 0.254 |
| Capital | Default margin | 0.500 | 1.082 | 0.585 | 0.094 | 0.098 | 0.596 |
| Leverage | All states | 0.138 | 0.189 | 0.067 | 0.032 | 0.030 | 0.093 |
| Leverage | Default margin | 0.090 | 0.047 | 0.068 | 0.059 | 0.032 | 0.057 |

*Notes:* Entries report mass-weighted ergodic averages of absolute local directional sensitivities. "Default margin" refers to regions with $2\% \leq p^d < 95\%$. Capital and leverage are implied by the transformer-predicted distribution and are not separate prediction targets. "Total" is the sensitivity of the implied aggregate outcome to the distributional perturbation. "Direct" is the component operating through the current-state term $m_t$ in the residual forecasting law, while "Transformer" is the component operating through $\mathcal{T}_{\theta,m}(X_t)$. "Reallocation" and "Value" are the two channels in equation (7), propagated through the trained network. "Other paths" collects the remaining transformer-mediated paths outside the simple attention aggregation. Because the table reports averages of absolute magnitudes, the columns are not additive shares: the corresponding signed components can offset one another.

Figure 9: Stationary Debt Prices




*Notes:* The figure reports equilibrium debt prices $q(k', b')$ as functions of chosen next-period capital $k'$ for selected idiosyncratic-productivity states $\varepsilon$. The left and right panels fix moderate and high chosen debt, respectively.