Computing Power, AI Services, and Welfare in the Generative-AI Economy*

Qing Chang† Danxia Xie‡ Longtian Zhang§

May 31, 2026

Abstract

As the stock of high-quality human-generated text approaches exhaustion, frontier AI models increasingly train on data produced by AI models themselves. This is self-limiting: successive model generations trained on their predecessors' output drift from the real-world distribution, a phenomenon known as model collapse. Computing power is the underappreciated input through which fidelity to that distribution can be sustained. We develop a dynamic general-equilibrium model in which generative AI firms produce differentiated AI services, consumer data and computing power protect data quality, and the firms' accumulated AI output erodes it. Our central finding is that computing power is systematically undersupplied in every decentralized regime through the compounding of three monopoly wedges, and that this underprovision, rather than the privacy or variety margin, is the first-order source of welfare loss in the AI economy. Two corollaries follow: private entry into the AI sector falls short of the social optimum because entrants capture only a fraction of the love-of-variety gain they create downstream; and a binding cap on consumer-data supply can nonetheless raise welfare above the unregulated level by indirectly relaxing the free-entry margin and drawing in additional firms. Calibration shows the optimal cap closes only a small share of the welfare gap, whereas an optimal computing-investment subsidy closes most of it, with gains accruing gradually as compute accumulates. The results make a welfare-based case for combining data regulation with direct support for computing infrastructure.

Keywords: model collapse, AI services, computing power, data quality, generative AI, regulation

JEL Classification: O33, O41, D62, and L51

*All the authors contribute equally to this work. Xie is grateful for the financial supports from the National Science and Technology Major Project of China (Grant No. 2022ZD0120301) and the National Natural Science Foundation of China (Grant No. 72373079), and Zhang is grateful for the financial supports from the National Natural Science Foundation of China (Grant No. 72550001, 72303261). The contents of this publication are solely the responsibility of the authors.

†Institute of Economics, School of Social Sciences, Tsinghua University, chang-qing@tsinghua.edu.cn.

‡Institute of Economics, School of Social Sciences, Tsinghua University, xiedanxia@tsinghua.edu.cn.

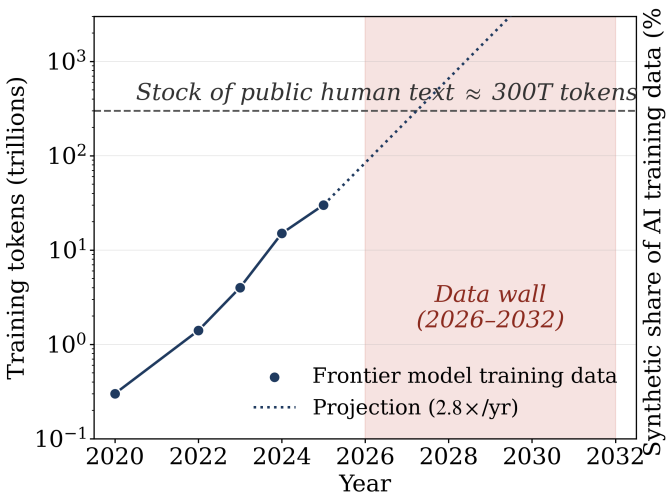
§(Corresponding author) School of International Trade and Economics, Central University of Finance and Economics, zhanglongtian@cufe.edu.cn.

1. Introduction

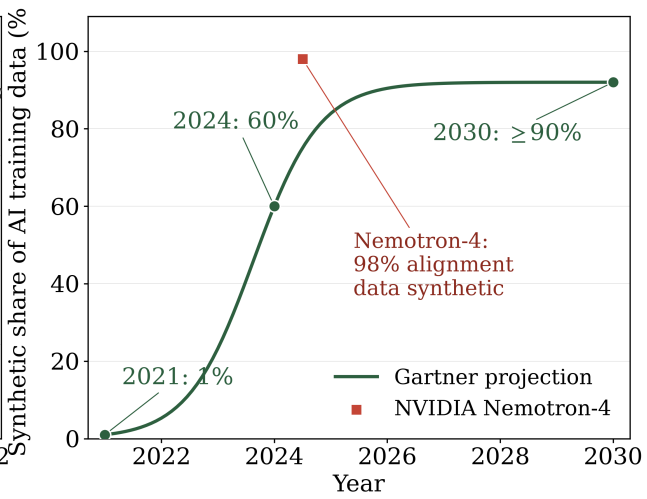
Artificial intelligence has become a central engine of contemporary economic activity. Private investment in AI exceeded USD 250 billion globally in 2024, the amortized compute cost of training a frontier generative model has grown at roughly 2.4 times per year since 2016 and is projected to exceed one billion dollars per run by 2027 (Cottier et al., 2024), and generative-AI tools are diffusing rapidly through business workflows, scientific research, and the public sector. This expansion rests on two inputs that are themselves at the center of contemporary debates in economics: data, which supply the informational content from which models learn, and computing power, which translates that content into trained systems at scale. The technical relationship between the two has been formalized through neural scaling laws, which tie model performance to training data and training compute through stable power-law exponents (Kaplan et al., 2020; Hoffmann et al., 2022), and which in turn have shaped the fixed-cost structure of the frontier AI industry.

Both inputs are now pressing against important limits. On the data side, Villalobos et al. (2022) project that the effective stock of high-quality human-generated text usable for training will be exhausted between 2026 and 2032, with comparable ceilings binding in vision and scientific corpora. Industry has responded by relying increasingly on synthetic data, that is, samples generated by AI models themselves and fed back into training. Such data already account for the overwhelming majority of alignment and instruction-tuning corpora at frontier laboratories, and industry analysts forecast that they will supply most total AI training data by 2030. Synthetic data is not, however, a neutral substitute for the real-world information from which it is distilled. When successive model generations are trained on predominantly self-generated inputs, they exhibit model collapse: a progressive loss of distributional tails, a narrowing of output diversity, and a deterioration in downstream task performance (Shumailov et al., 2024; Wenger, 2024). Because synthetic data is already in widespread use, the policy-relevant question is no longer whether it will be used but under what economic conditions its quality can be sustained as the underlying stock of real-world information grows only slowly. Figure 1 documents these three trends.

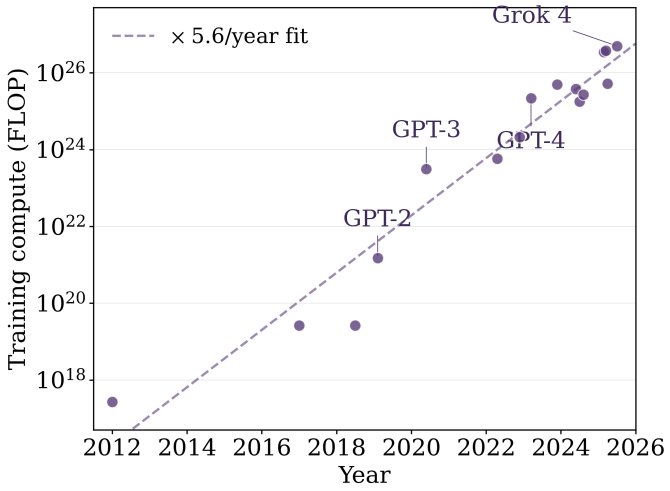
Computing power enters this question through a mechanism that the economics literature has yet to formalize. In addition to its standard role as a scale input to training, compute also governs the fidelity with which a synthetic-data pipeline preserves the statistical structure embedded in its source information: larger models, longer training, and more inference-time sampling produce synthetic samples that lie closer to the real-world distribution and that mitigate the tail-erosion driving collapse. Compute thus acts as a quality-preserving input, complementing the real-world information supplied by consumers rather than substituting for it. The supply of AI-grade compute is highly concentrated, with a narrow base of GPU producers and foundries accounting for the bulk of global frontier capacity (Cottier et al., 2024). Since training and inference also bear large up-front fixed costs that shape the market structure of the AI industry, the decentralized provision of quality-preserving compute is unlikely to coincide with the efficient level. Whether such deviations exist, how they interact with entry into the generative-AI sector, and which policy instruments can correct them, are the questions this paper addresses.



(a) Real-world data is running out



(b) Synthetic data is replacing real data



(c) Training compute grows exponentially

Figure 1. Three stylized facts motivating the analysis.

Notes. Panel (a): Frontier model training datasets grow at roughly 2.8× per year and are projected to meet the ≈300-trillion-token stock of public human text between 2026 and 2032 (Villalobos et al., 2022; dataset sizes from Epoch AI). Panel (b): Synthetic data's share of AI training data rose from about 1% in 2021 to 60% in 2024 and is projected to exceed 90% by 2030 (Gartner); NVIDIA reports that more than 98% of the data used to align Nemotron-4 is synthetic. Panel (c): Training compute for frontier AI models has grown at roughly 5× per year since 2012 (Epoch AI frontier-models database; log scale).

We study these questions in a dynamic general-equilibrium framework populated by a representative household, a finite set of monopolistically competitive final-goods producers, a finite set of monopolistically competitive generative-AI firms, a single computing-power provider, and monopolistic data intermediaries. Households supply labor inelastically and, following Cong et al. (2021), generate consumer data as a byproduct of consumption activity, bearing a convex privacy cost in the tradition of Jones and Tonetti (2020) and Farboodi and Veldkamp (2026). AI firms combine consumer

data with their accumulated AI output and with AI labor to produce differentiated AI services, which final-goods producers aggregate through a constant-elasticity-of-substitution technology featuring love of variety. The computing-power provider accumulates compute through investment and labor, facing a high depreciation rate calibrated to the rapid hardware turnover in AI accelerators. The core of the framework is a reduced-form data-quality function that specifies the quality of each AI firm's services as a Cobb-Douglas combination of consumer data and compute, normalized by the firm's accumulated AI output. A larger accumulated stock therefore erodes quality, capturing model collapse in reduced form, while consumer data and compute both protect it, with their relative weight set by a single parameter. The multiplicative pairing of quality and the accumulated stock in the aggregator ensures that AI services remain net-productive on the equilibrium path while generating the economic tensions that drive our results.

We solve the model under four regimes: a social-planner allocation; a restricted-entry equilibrium featuring a single AI firm, which proxies for the current concentration of the generative-AI market; a free-entry equilibrium; and a data-regulation regime in which the government imposes an upper bound on per-capita consumer-data supply. Our central analytical result is that in every decentralized regime computing power is strictly underprovided relative to the social optimum, and the magnitude of this underprovision admits a transparent decomposition into three stacked monopoly wedges: the computing-power provider internalizes only the residual value of compute after the AI-firm markup, AI firms internalize only the residual value of data after the final-goods markup, and the entry externality is absent entirely from private entry decisions. Because the three wedges compound multiplicatively, the distortion is quantitatively first-order rather than a minor correction, and the computing-infrastructure margin emerges as the dominant source of welfare loss in the AI economy. This finding shifts the policy focus away from the consumer-data margin that has dominated previous debates.

Two corollaries flow from this central mechanism. The first concerns AI-firm entry. Under free entry, the equilibrium number of AI firms is strictly below the optimal number because private entrants capture only a fraction of the love-of-variety gain they generate for downstream producers; the insight generalizes the entry inefficiency of monopolistic-competition growth models (Romer, 1990; Jones, 1995) to the generative-AI setting. The second, and more novel, concerns data regulation. A binding consumer-data cap can raise welfare strictly above the unregulated free-entry level, even though the cap mechanically reduces data quality: the direct quality loss from restricting consumer data is outweighed by an indirect entry-promoting effect, as a lower consumer-data bill relaxes the free-entry condition for AI firms and draws in additional varieties. We characterize the welfare-maximizing cap as the solution to an implicit first-order condition that log-linearizes transparently into direct-quality, entry, data-stock, and labor-reallocation components.

A calibrated quantitative exercise, grounded in macroeconomic evidence on computing-equipment depreciation, in machine-learning estimates of data-scaling exponents (Kaplan et al., 2020), and in the privacy-valuation benchmarks of Jones and Tonetti (2020) and Farboodi and Veldkamp (2026), confirms the dominant role of the compute margin numerically and quantifies the welfare consequences of the two policy instruments that emerge from the analysis. An optimal consumer-

data cap closes only a small share of the welfare gap and does so not by directly addressing the compute distortion but by relaxing the free-entry condition of AI firms, thereby promoting variety; a computing-investment subsidy, by contrast, acts on the distortion at its source and delivers substantially larger welfare gains. The two instruments are therefore complements rather than substitutes in the policy mix. Transition-dynamics experiments under myopic expectations further reveal that the compute subsidy, unlike the data cap, involves a genuine intertemporal tradeoff between current consumption and future productivity, since the bulk of its gains accrues only as the computing and AI-services stocks accumulate over time.

Our paper contributes to three strands of the literature. First, on the economics of data, we build on the foundational treatment of data as a nonrival input by Jones and Tonetti (2020), Farboodi and Veldkamp (2026), and Cong et al. (2021), which extends the endogenous-growth tradition of Romer (1990) and Jones (1995). Most closely related is Mihet et al. (2026), who distinguish raw from processed data and show that compute and data are complements for raw-data learning but substitutes when processed data is already available; our framework shares this compute-data interaction but formalizes compute as a quality-preserving input against model collapse rather than as a raw-data processor, and distinguishes the stock of real-world information from the firm's accumulated AI output as conceptually distinct inputs whose ratio determines data quality. Second, on the economics of generative AI, a rapidly expanding literature assesses the aggregate productivity, growth, and regulatory consequences of AI adoption (Acemoglu, 2025; Acemoglu and Lensman, 2024; Jones, 2024; Korinek and Vipra, 2024). Closest to our analysis is Korinek and Vipra (2024), who study scaling and market structure in AI; we complement their industrial-organization focus with a dynamic general-equilibrium framework that links market structure to the underprovision of a quality-preserving input. Third, our treatment of synthetic data and model collapse connects to a nascent machine-learning literature on training-pipeline stability under iterative self-consumption (Shumailov et al., 2024; Wenger, 2024), which we embed in an economic environment where it generates tractable equilibrium allocations and explicit policy implications. Our transition analysis adopts the myopic-expectations approach from behavioral macroeconomics (Gabaix, 2020; Angeletos and Huo, 2021).

The remainder of the paper is organized as follows. Section 2 sets out the economic environment, introducing the household, AI firms, the computing-power provider, final-goods producers, and data intermediaries. Section 3 solves for the social planner's allocation and the three competitive equilibria, establishes the monopoly-distortion decomposition of compute underprovision, and characterizes the welfare-maximizing consumer-data cap. Section 4 calibrates the model, reports benchmark results and comparative statics, and quantifies the two policy instruments in consumption-equivalent units. Section 5 studies the transition paths associated with entry deregulation, the optimal data cap, and the optimal computing-investment subsidy. Section 6 concludes.

2. Economic Environment

To study the coordination between computing power and AI services in an AI-driven economy, this section develops a benchmark model featuring five types of agents: a representative household; M

generative AI firms that produce differentiated AI services; a computing power provider that supplies data-processing capacity; N firms producing final consumption goods; and data intermediaries that facilitate data transactions. A government is introduced in Section 3 in the context of data regulation.

The model is set in discrete time with an infinite horizon. A central feature of the framework is the introduction of data quality, which serves as the key link between computing resources and AI services. Different AI firms produce different types of AI services, and final-goods producers combine these differentiated services via a CES aggregator. This structure provides the foundation for the welfare analysis and policy experiments conducted in the subsequent sections.

Representative Household. The economy is populated by a continuum of homogeneous consumers with total measure L. Each consumer supplies one unit of labor inelastically each period. In addition, following the setup in Cong et al. (2021), consumers generate data through their own consumption activities, hereafter referred to as "consumer data," and provide d_t units of data per capita in each period. The household's lifetime utility is given by

$$U = \sum_{t=0}^{\infty} \beta^t u(c_t, d_t) = \sum_{t=0}^{\infty} \beta^t \left[\ln c_t - \phi(d_t)\right], \tag{1}$$

where c_t denotes per capita aggregate consumption, and $\beta \in (0,1)$ represents the consumer's subjective discount factor. The function $\phi(d_t)$ captures the privacy disutility associated with data usage, reflecting potential losses arising from the exposure of personal information.

We assume $\phi(\cdot)$ is twice continuously differentiable, strictly increasing ($\phi'(d) > 0$), and strictly convex ($\phi''(d) > 0$). Strict monotonicity captures the basic premise that providing more data entails a higher privacy cost. Strict convexity is less mechanical and deserves comment. The first unit of data a household reveals is typically of low sensitivity, such as a coarse location stamp or a page-view record, and its marginal disclosure imposes only limited harm. As the household reveals additional information, however, the marginal unit becomes progressively more diagnostic of identity and behavior: small cross-tabulations of otherwise innocuous fields can identify a unique individual, and attackers or data brokers can combine new information with previously released information to construct profiles whose informational content grows faster than linearly in raw volume.[1] We do not impose a specific functional form at this stage; the functional form used for calibration is specified in Section 4.1.[2]

[1]The convex specification is standard in the economics-of-data literature. Jones and Tonetti (2020) adopt a quadratic privacy cost and interpret it as reflecting the idea that households do not donate their data freely; if the marginal privacy cost were constant or declining, the household's data-supply schedule would typically be bang-bang at an interior price, contrary to the smooth adjustment observed empirically. Farboodi and Veldkamp (2020) discuss the case where privacy is already nearly fully compromised, in which a concave segment may arise; we follow the bulk of the literature in focusing on the interior region, where convexity prevails. Economically, convexity captures re-identification risk (small combinations of attributes can uniquely identify an individual), downstream exploitation risk (profiled consumers face price discrimination, targeted persuasion, and algorithmic sorting in ways that compound with the amount of data revealed), and exposure risk to data breaches whose severity rises super-linearly in dataset size. These channels make the marginal privacy loss an increasing function of cumulative disclosure.

[2]Since consumer data is generated through consumption activities, the amount of data supplied by consumers is naturally bounded above by some multiple of consumption. We assume throughout that the convexity of $\phi(\cdot)$ is strong enough that this upper bound is slack in equilibrium, so that consumer data supply is governed entirely by the consumer's

On the consumption side, the economy features N differentiated varieties of consumption goods, denoted by $c_{i,t}$ for $i \in \{1, 2, \ldots, N\}$. Aggregate per capita consumption c_t is defined as a constant-elasticity-of-substitution (CES) aggregator of individual varieties:

$$c_t = \left(\sum_{i=1}^{N} c_{i,t}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

where σ denotes the elasticity of substitution across consumption varieties. We assume $\sigma > 1$, implying that different varieties are imperfect but substitutable. Throughout the analysis, the price of the aggregate consumption good is normalized to 1.

Generative AI Firms. In addition to consumer data, the economy features the output of generative AI firms, namely AI services, which the rest of the economy uses as a productive input. The economy is populated by M generative AI firms, indexed by $j \in \{1, 2, \ldots, M\}$. Each firm specializes in producing a distinct type of AI service. These differentiated types capture the idea that different AI firms develop models with different capabilities, for instance specialization in natural language processing, image generation, scientific reasoning, or code synthesis, so that their outputs are imperfect substitutes for one another.

To operate, each AI firm must pay a one-time fixed cost $F > 0$, denominated in units of the final good. This entry cost captures the substantial upfront investment required to develop a frontier generative AI model, including the computational resources needed for large-scale pre-training, the research effort to achieve emergent capabilities, and the costs of assembling and curating initial training datasets. Empirical evidence documents that these fixed costs have grown rapidly and constitute a significant barrier to entry: Cottier et al. (2024) estimate that the amortized compute cost of training state-of-the-art AI models has grown at roughly 2.4 times per year since 2016, with the largest training runs projected to exceed one billion dollars by 2027. The scaling laws documented by Kaplan et al. (2020) and refined by Hoffmann et al. (2022) provide the technological foundation for these patterns, showing that model performance follows a predictable power-law relationship with model size, dataset size, and training compute, so that substantial fixed investment in all three dimensions is required to reach the frontier of capabilities. In the AI literature, emergent capabilities refer to qualitative leaps in model performance that arise only when the scale of training data and computation exceeds certain thresholds, making such capabilities inaccessible to firms that cannot afford the necessary fixed investment. Whether M is determined exogenously by institutional or technological barriers or endogenously through free entry will depend on the market regime under consideration, as discussed in Section 3.[3]

Consider a representative generative AI firm j. Given its accumulated AI output at the beginning

privacy preferences.

[3] Although F is paid once at entry, in a stationary environment this is equivalent to a per-period flow cost of $(1-\beta)F$, which is the form that appears in the free-entry conditions of Section 3.

of period t, denoted by $D_{j,t}$, the flow of newly produced AI services $\tilde{D}_{j,t}$ within the period is given by

$$\tilde{D}_{j,t} = \eta \left[\lambda D_{j,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} L_{A,j,t}^{1-\alpha}, \tag{3}$$

where $L_{A,j,t}$ denotes labor employed by AI firm j, and $d_t L$ represents the aggregate consumer data in the economy (per-capita supply d_t times total population L). The AI firm uses a share of the economy-wide AI labor $L_{A,t}$, with $\sum_{j=1}^{M} L_{A,j,t} = L_{A,t}$; under the symmetric equilibria studied in Section 3, $L_{A,j,t} = L_{A,t}/M$ for all j.[4] Because consumer data is nonrival, the same $d_t L$ units are simultaneously accessible to all M generative AI firms without being diminished. The parameter $\eta > 0$ captures the efficiency of the data generation process, $\lambda \in (0,1)$ measures the relative importance of the firm's existing AI output compared to newly introduced consumer data, ε denotes the elasticity of substitution between these two inputs, and $\alpha \in (0,1)$ governs the contribution of data relative to labor in the Cobb–Douglas nest. The stock of type-j AI services evolves according to

$$D_{j,t+1} = \tilde{D}_{j,t} + (1-\delta_D) D_{j,t}, \tag{4}$$

where $\delta_D \in (0,1)$ denotes the depreciation rate of AI services.

Equation (3) highlights several key features of this technology. The production of new AI services relies on the firm's existing AI output, complemented by consumer data drawn from real-world activities. Over successive iterations, the information content embedded in the accumulated AI output may gradually drift away from real-world conditions, whereas consumer data contains fresh information directly drawn from reality. The continuous inflow of consumer data therefore helps refresh and update the information content delivered by the firm's services. In addition, because AI services are nonrival, the use of the firm's existing AI output in producing new services does not preclude it from providing the accumulated stock each period to final-goods producers.

The two inputs in Equation (3) are assumed to be substitutes, with an elasticity of substitution $\varepsilon > 1$. While producing new AI services from the firm's existing AI output is less costly than relying on consumer data, excessive reliance on accumulated AI output can compromise informational fidelity. Generative AI firms therefore face a trade-off and must optimally balance the use of consumer data and their existing AI output in producing new services, a trade-off that we formalize through the notion of data quality introduced next.

Building on the above analysis, we introduce the concept of data quality to better characterize the properties of these AI services. Beyond the mechanism captured in Equation (3), improvements in data generation efficiency and accuracy can also be achieved through sustained investment in computing power. Let $K_{j,t}$ denote the amount of computing power used by AI firm j, where $K_{j,t}$ is a share of the total computing stock K_t produced by the computing power provider (introduced below); because

[4] Symmetry across AI firms is a tractability assumption. Because the core mechanism operates through downstream markup compounding and the CES love-of-variety structure, the qualitative results of Section 3, namely compute under-provision in every decentralized regime and the existence of a welfare-improving consumer-data cap, extend to asymmetric settings in which firms differ in their fixed costs or productivity. Heterogeneity would, however, alter the quantitative magnitude of the entry inefficiency and the within-sector distribution of compute, dimensions that speak to frontier-versus-fringe dynamics in the AI industry and that we leave for future work.

computing power is rival, these shares satisfy $\sum_{j=1}^{M} K_{j,t} = K_t$, and under the symmetric equilibria studied in Section 3, $K_{j,t} = K_t/M$ for all j. Combining these two channels, namely consumer data and computing power, we define the quality of type-j AI services as

$$Q_{j,t} = \left[\frac{(d_t L)^b K_{j,t}^{1-b}}{D_{j,t}} \right]^{\tau}, \tag{5}$$

where $b \in (0,1)$ captures the relative importance of consumer data, as opposed to computing power, in determining data quality, and $\tau \in (0,1)$ governs the sensitivity of quality to the composition of inputs. The restriction $\tau < 1$ ensures that the accumulated AI output continues to contribute positively to the effective data input $Q_{j,t}D_{j,t}$, whose elasticity with respect to $D_{j,t}$ is $1 - \tau > 0$. A higher ratio of consumer data and computing power relative to the accumulated AI output implies higher quality, reflecting the idea that each unit of AI output is supported by more real-world information and computational resources.

This specification echoes a growing body of evidence from the machine learning literature on model collapse: when synthetic data dominates the training mix relative to fresh real-world information, successive generations of models exhibit degraded fidelity to the underlying data distribution (Shumailov et al., 2024), which in turn lowers the quality of the AI services they deliver. Placing $D_{j,t}$, the firm's accumulated AI output, in the denominator of $Q_{j,t}$ captures this mechanism in reduced form, with τ governing the severity of collapse and b governing the relative protective role of consumer data versus computing power. AI labor $L_{A,j,t}$, by contrast, does not appear in $Q_{j,t}$; in our formulation labor affects only the flow of newly produced AI services through Equation (3), while quality is determined by the balance between real-world information, computing resources, and the accumulated AI output.

Computing Power Provider. We assume that the economy features a single computing power provider. Modeling a single provider is adopted primarily for tractability, but also reflects two basic features of the AI compute supply chain. First, computing power is the principal input sustaining the quality of AI services: Kaplan et al. (2020) and Hoffmann et al. (2022) show that language-model performance scales as a power law in training compute, and modern frontier models generate the bulk of their alignment data through large-scale model inference,[5] which itself requires substantial compute. Second, the supply of AI-grade compute is highly concentrated: NVIDIA holds an estimated 70%–95% share of the market for AI accelerators, essentially all of which are fabricated by a single foundry, Taiwan Semiconductor Manufacturing Company (TSMC), and delivered to AI firms largely through three cloud hyperscalers (Amazon Web Services, Microsoft Azure, and Google Cloud Platform).[6] Representing this narrow supply base as a single provider captures the scarcity of AI-ready compute without modeling the internal organization of the supply chain. The provider continuously

[5]For example, over 98% of the alignment data for NVIDIA's Nemotron-4 340B was synthetically generated by large models (NVIDIA Nemotron-4 340B Technical Report, arXiv:2406.11704).

[6]NVIDIA market share figures are from industry estimates by Mizuho Securities and TrendForce, as discussed in Cottier et al. (2024). TSMC's specialized chip-on-wafer-on-substrate packaging, a process required to integrate multiple dies with high-bandwidth memory, constrains industry-wide output.

accumulates computing resources through investment and makes these resources available to the M AI firms. Unlike AI services, computing power is a rival input: a unit of computing power used by one AI firm cannot simultaneously be used by another.

Specifically, the flow of newly produced computing power in period t, denoted by $\tilde{K}_t$, is given by

$$\tilde{K}_t = \xi I_t^{\gamma} L_{K,t}^{1-\gamma},$$

where $\xi > 0$ is a scale parameter capturing the efficiency of computing power production, I_t represents investment by the computing power provider in period t, denominated in units of the final good, and $L_{K,t}$ denotes labor employed by the computing power provider.

Accordingly, the law of motion for the stock of computing power is given by

$$K_{t+1} = \tilde{K}_t + (1 - \delta_K) K_t, \tag{6}$$

where $\delta_K \in (0, 1)$ denotes the depreciation rate of computing power, and $\gamma \in (0, 1)$ captures the importance of investment in the production of new computing power. The parameter δ_K will be calibrated to a relatively high value in Section 4.1, reflecting the rapid obsolescence of AI-grade hardware.

Producers. For each consumption variety, the economy features a corresponding producer that specializes in its production. Consider a representative producer of variety i. Production of this variety requires labor input $L_{i,t}$ as well as AI services from the M generative AI firms. Because AI services are nonrival across producers, each producer i has access to the entire stock $D_{j,t}$ of every AI firm j. Producers aggregate the M differentiated types of AI services via a CES technology, yielding a common effective data input

$$\Phi_t = \left(\sum_{j=1}^{M} \left(Q_{j,t} D_{j,t} \right)^{\frac{\rho-1}{\rho}} \right)^{\frac{\rho}{\rho-1}}, \tag{7}$$

where $\rho > 1$ denotes the elasticity of substitution across different types of AI services, and $Q_{j,t} D_{j,t}$ is the quality-adjusted AI service supplied by AI firm j. Since the CES aggregator depends only on the economy-wide quantities $\{Q_{j,t}, D_{j,t}\}_{j=1}^{M}$, and these are the same for all final-goods producers, Φ_t is common to all N producers. A larger M raises Φ_t through a love-of-variety effect: holding the per-firm quality-adjusted service $Q_{j,t} D_{j,t}$ fixed, greater diversity in service types improves the effective data input available to producers.

The production function of firm i is then given by

$$Y_{i,t} = A \, \Phi_t^{\zeta} \, L_{i,t}, \tag{8}$$

where A denotes total factor productivity and $\zeta \in (0, 1)$ is the elasticity of output with respect to the effective data input Φ_t.

Data Intermediaries. A distinguishing feature of data is its nonrivalry: the same data can be used by multiple agents simultaneously without being diminished. In this economy, nonrivalry arises in two distinct markets. First, consumer data supplied by households can be used simultaneously by all M generative AI firms. Second, each type of AI service produced by an AI firm can be used simultaneously by all N final-goods producers. We introduce monopolistic data intermediaries to facilitate transactions in both markets. In the spirit of Jones and Tonetti (2020)'s treatment of data markets, each intermediary serves as the single marketplace for its own data type and is subject to a zero-profit condition, which pins down the relation between the supply-side and demand-side prices. We emphasize that these intermediaries and their associated prices only enter the analysis in the competitive equilibria characterized in Section 3; the social planner's problem, by contrast, allocates data directly without relying on market prices.

In the consumer data market, the intermediary purchases consumer data from households at the supply-side price $p^s_{d,t}$ and resells it to each of the M generative AI firms at the demand-side price $p^d_{d,t}$. Because consumer data is nonrival across AI firms, the intermediary can sell the same data to all M firms. The zero-profit condition therefore yields

$$M\, p^d_{d,t} = p^s_{d,t}, \tag{9}$$

so that the price paid by each individual AI firm equals $1/M$ of the total price received by the consumer. A larger number of AI firms raises the total demand for consumer data, which in turn increases the supply-side price $p^s_{d,t}$ that consumers receive, strengthening their incentive to supply data.

In the market for AI services, the intermediary purchases type-j AI services from AI firm j at the supply-side price $p^s_{D,j,t}$ and resells them to each of the N final-goods producers at the demand-side price $p^d_{D,j,t}$. Because each type of AI service is nonrival across producers, the zero-profit condition yields

$$N\, p^d_{D,j,t} = p^s_{D,j,t}, \quad \forall\, j \in \{1, 2, \ldots, M\}. \tag{10}$$

Each producer can use the entire stock of type-j AI services $D_{j,t}$, and the effective price it faces equals $1/N$ of the supply-side price set by the corresponding AI firm. Each AI firm j is a monopolistic supplier of its own differentiated AI service and therefore faces a downward-sloping residual demand curve with elasticity ρ (inherited from the CES aggregator (7)); it internalizes this market power when choosing $p^s_{D,j,t}$. Analogously, each final-goods producer of variety i is a monopolistic supplier of its own variety, facing a residual demand curve with elasticity σ from the consumption aggregator (2), and sets its output price as the standard CES markup over marginal cost. The computing power provider, being the sole supplier of computing resources to all M AI firms, is also a monopolist and sets the rental price $p_{K,t}$ taking into account the aggregate compute demand from AI firms.

Market Clearing and Resource Constraints. To close the model, we impose the standard market-clearing and resource-constraint conditions. Total labor is allocated across final-goods producers,

generative AI firms, and the computing power provider:

$$\sum_{i=1}^{N} L_{i,t} + L_{A,t} + L_{K,t} = L, \tag{11}$$

where $L_{A,t} = \sum_{j=1}^{M} L_{A,j,t}$ as defined below Equation (3). The final good is used for consumption and for investment in computing power, yielding the aggregate resource constraint

$$Y_t = c_t L + I_t, \tag{12}$$

where Y_t denotes aggregate output. These two conditions, together with the data-market clearing relations in (9) and (10), close the model and will be imposed throughout the equilibrium analysis in Section 3.

Table 1 provides a summary of the key equations that define the economic environment.

Table 1. Economic Environment

Component	Equation
Lifetime utility	$U = \sum_{t=0}^{\infty} \beta^t \left[\ln c_t - \phi(d_t)\right]$
CES consumption aggregator	$c_t = \left(\sum_{i=1}^{N} c_{i,t}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}$
AI-service flow (firm j)	$\tilde{D}_{j,t} = \eta\left[\lambda D_{j,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} L_{A,j,t}^{1-\alpha}$
AI-service accumulation	$D_{j,t+1} = \tilde{D}_{j,t} + (1-\delta_D)\, D_{j,t}$
Data quality (type j)	$Q_{j,t} = \left[(d_t L)^b\, K_{j,t}^{1-b} \,/\, D_{j,t}\right]^{\tau}$
CES data aggregator	$\Phi_t = \left(\sum_{j=1}^{M} (Q_{j,t} D_{j,t})^{\frac{\rho-1}{\rho}}\right)^{\frac{\rho}{\rho-1}}$
Computing power flow	$\tilde{K}_t = \xi\, I_t^{\gamma}\, L_{K,t}^{1-\gamma}$
Computing power accumulation	$K_{t+1} = \tilde{K}_t + (1-\delta_K)\, K_t$
Production function	$Y_{i,t} = A\, \Phi_t^{\zeta}\, L_{i,t}$
Data intermediary (consumer data)	$M\, p_{d,t}^{d} = p_{d,t}^{s}$
Data intermediary (AI services)	$N\, p_{D,j,t}^{d} = p_{D,j,t}^{s}$
Labor market clearing	$\sum_{i=1}^{N} L_{i,t} + L_{A,t} + L_{K,t} = L$
Aggregate resource constraint	$Y_t = c_t L + I_t$
Entry cost (AI firm)	$F > 0$ (one-time, in units of final good)

3. Solving the Model

In this section, we solve the model and characterize both the optimal allocation and the three competitive market equilibria under different market structures. We focus on steady-state allocations here and defer the analysis of transition dynamics to Section 5.

3.1 Social Planner's Problem

We first characterize the optimal allocation. Because all M generative AI firms have identical technologies, the social planner treats them symmetrically. We therefore impose the following symmetric conditions: $D_{j,t} = D_t$, $K_{j,t} = K_t/M$, $L_{A,j,t} = L_{A,t}/M$, and $Q_{j,t} = Q_t$ for all j, where D_t denotes the common per-firm stock of AI services, K_t is the total stock of computing power, and $L_{A,t}$ is the total labor employed in the AI sector.

Under symmetry, the effective data input defined in Equation (7) simplifies to

$$\Phi_t = M^{\frac{\rho}{\rho-1}} Q_t D_t, \tag{13}$$

where $Q_t = [(d_t L)^b (K_t/M)^{1-b}/D_t]^\tau$ is the common data quality across all AI firms. The factor $M^{\rho/(\rho-1)}$ captures the love-of-variety effect: holding per-firm data quality and quantity fixed, a greater number of differentiated AI-service types raises the effective data input available to producers. Combining Equations (13) and (8) with the symmetry across consumption-good producers and the labor market clearing condition yields the aggregate resource constraint

$$c_t L + I_t = N^{\frac{1}{\sigma-1}} A M^{\frac{\rho\zeta}{\rho-1}} \left\{ \left[(d_t L)^b \left(\frac{K_t}{M} \right)^{1-b} \right]^\tau D_t^{1-\tau} \right\}^\zeta (L - L_{K,t} - L_{A,t}). \tag{14}$$

The per-firm AI-service accumulation under symmetry is

$$D_{t+1} = \eta \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M} \right)^{1-\alpha} + (1-\delta_D) D_t, \tag{15}$$

where each of the M symmetric AI firms employs $L_{A,t}/M$ units of labor and uses the full stock of consumer data $d_t L$ (which is nonrival across firms). The computing power accumulation remains as in Equation (6).

Based on the structure above, the social planner's problem can be written as

$$\max_{M, c_t, d_t, D_t, L_{A,t}, L_{K,t}, I_t, K_t} \sum_{t=0}^{\infty} \beta^t \left[\ln c_t - \phi(d_t) \right] \tag{16}$$

subject to the resource constraint

$$c_0 L + I_0 + MF = N^{\frac{1}{\sigma-1}} A M^{\frac{\rho\zeta}{\rho-1}} \left\{ \left[(d_0 L)^b \left(\frac{K_0}{M} \right)^{1-b} \right]^\tau D_0^{1-\tau} \right\}^\zeta (L - L_{K,0} - L_{A,0})$$

in period $t = 0$ (which includes the one-time entry cost MF for creating M AI firms), the standard resource constraint (14) for all $t \geq 1$, and the accumulation equations (15) and (6). We use a superscript "sp" to denote the steady-state value of each variable under the optimal allocation. The complete derivation is provided in Appendix A.1. Solving this problem yields the following proposition.

Proposition 1 (Optimal Allocation)**.** *Under the optimal allocation, all variables are constant in the steady state. The four endogenous variables $(L_A^{sp}, L_K^{sp}, D^{sp}, M^{sp})$, together with consumer-data supply d^{sp}, are jointly determined by the following system:*

$$\frac{L_K^{sp}}{L - L_A^{sp} - L_K^{sp}} = \frac{1-\gamma}{\gamma} B^{sp}, \tag{17}$$

$$\frac{[1-\beta(1-\delta_D)]\, D^{sp}\, M^{sp}}{(1-\alpha)\beta\eta(\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{sp}/M^{sp})^{-\alpha}(L - L_A^{sp} - L_K^{sp})} = (1-\tau)\zeta + \frac{\alpha}{1-\alpha} \cdot \frac{L_A^{sp}}{L - L_A^{sp} - L_K^{sp}} \cdot \frac{\lambda (D^{sp})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{sp}}, \tag{18}$$

$$\phi'(d^{sp}) d^{sp} = \frac{1}{1 - B^{sp}} \left[b\tau\zeta + \frac{\alpha L_A^{sp} (1-\lambda)(d^{sp} L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L - L_A^{sp} - L_K^{sp})\Omega^{sp}} \right], \tag{19}$$

$$(1-\beta)\, F = \frac{Y^{sp}}{M^{sp}} \left[\omega - \frac{L_A^{sp}}{L - L_A^{sp} - L_K^{sp}} \right], \tag{20}$$

where $Y^{sp} \equiv c^{sp} L + I^{sp}$ is aggregate output, the auxiliary Ω^{sp} is pinned down by the steady-state AI-service-accumulation identity

$$\Omega^{sp} = \left[\frac{\delta_D D^{sp}}{\eta (L_A^{sp}/M^{sp})^{1-\alpha}} \right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} = \lambda (D^{sp})^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d^{sp} L)^{\frac{\varepsilon-1}{\varepsilon}},$$

and the parameter bundles B^{sp} and ω are

$$B^{sp} = \frac{(1-b)\tau\zeta\gamma\beta\delta_K}{1 - \beta(1-\delta_K)}, \tag{21}$$

$$\omega \equiv \frac{\rho\zeta}{\rho - 1} - (1-b)\tau\zeta. \tag{22}$$

Given the solution, per-variety output and the aggregate compute stock are

$$Y_i^{sp} = \left\{ A\, (M^{sp})^{\omega} (d^{sp} L)^{b\tau\zeta} (D^{sp})^{\zeta(1-\tau)} \frac{L - L_A^{sp} - L_K^{sp}}{N} \left[\xi^{\frac{1}{\gamma}} \delta_K^{-\frac{1}{\gamma}} (L_K^{sp})^{\frac{1}{\gamma}-1} N^{\frac{\sigma}{\sigma-1}} B^{sp} \right]^{(1-b)\gamma\tau\zeta} \right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}, \tag{23}$$

$$K^{sp} = \left[\xi^{\frac{1}{\gamma}} \delta_K^{-\frac{1}{\gamma}} (L_K^{sp})^{\frac{1}{\gamma}-1} N^{\frac{\sigma}{\sigma-1}} B^{sp} \right]^{\gamma} (Y_i^{sp})^{\gamma}.$$

Several features of the solution are worth noting. First, the labor allocation between computing power production and final-goods production, characterized by Equation (17), is independent of M^{sp}. This is because the constant B^{sp}, which summarizes the optimal share of output devoted to computing investment, depends only on production and depreciation parameters.

Second, the optimal number of AI firms M^{sp}, characterized by Equation (20), balances the marginal love-of-variety gain $\omega Y^{sp}/M^{sp}$ against the dilution cost $L_A^{sp}/(L-L_A^{sp}-L_K^{sp})\cdot Y^{sp}/M^{sp}$, amortized at rate $(1-\beta)$ and set equal to the fixed entry cost F. A lower entry cost F, a smaller elasticity of substitution ρ, or a higher importance of data in production ζ all raise M^{sp}.

Third, M^{sp} enters the AI-service Euler equation (18) and the output equation (23). In Equation (18), the $D^{sp}M^{sp}$ term in the numerator of the left-hand side reflects the fact that the per-firm stock of AI services must be evaluated in the context of M^{sp} symmetric firms. In the output equation, the love-of-variety term $(M^{sp})^{\omega}$ captures the net productivity gains from combining diverse types of AI services in production; ω is the gross Dixit–Stiglitz variety exponent $\rho\zeta/(\rho-1)$ net of the compute-dilution exponent $(1-b)\tau\zeta$ arising because each additional firm receives a smaller share K_t/M of the aggregate computing stock. Notably, M^{sp} does not appear in the consumer data supply equation (19): the M from the M identical AI-service accumulation constraints cancels with the $1/M$ from the per-firm labor allocation in the shadow price of AI services.

3.2 Competitive Market Equilibrium: Restricted Entry

We now consider a competitive equilibrium in which entry into the generative AI sector is restricted, so that the economy features a single monopolistic AI firm ($M = 1$). This regime captures a market structure in which prohibitively high entry costs, such as the massive computational and research investments required to train frontier AI models, effectively prevent new firms from entering. When $M = 1$, the effective data input simplifies to $\Phi_t = Q_t D_t$, where $Q_t = [(d_t L)^b K_t^{1-b}/D_t]^{\tau}$, and the data aggregator (7) is trivially satisfied. Moreover, the consumer data intermediary condition (9) reduces to $p_{d,t}^d = p_{d,t}^s$, meaning that the single AI firm pays the full market price for consumer data.

Under this regime, the aggregate resource constraint becomes

$$c_t L + I_t = N^{\frac{1}{\sigma-1}} A \left\{ \left[(d_t L)^b K_t^{1-b} \right]^{\tau} D_t^{1-\tau} \right\}^{\zeta} (L - L_{K,t} - L_{A,t}), \tag{24}$$

which coincides with Equation (14) evaluated at $M = 1$. The key difference from the social planner's problem lies not in the technology but in the market structure: the AI firm, the computing power provider, and each final-goods producer all exercise monopoly power in their respective markets. We now state each agent's optimization problem, then formally define the equilibrium, and finally present the steady-state solution in Proposition 2. The complete derivation is provided in Appendix A.2.

Representative Household. Given the preferences in Equation (1), the household faces a trade-off when deciding how much consumer data to supply: selling more data generates additional income $p_{d,t}^s d_t$ but also entails a higher privacy cost $\phi(d_t)$. The household maximizes lifetime utility subject to the flow budget constraint

$$c_t + a_{t+1} = w_t + p_{d,t}^s d_t + (1 + r_t) a_t + \pi_t, \tag{25}$$

where a_t denotes assets held by consumers (which finance investment in computing power through the financial market) and π_t denotes per capita dividends from the profits of monopolistic firms. The

first-order conditions with respect to consumer data d_t and asset holdings a_t yield

$$\frac{p^s_{d,t}}{c_t} = \phi'(d_t), \tag{26}$$

$$\beta c_{t-1}(1 + r_t) = c_t. \tag{27}$$

Equation (26) equates the marginal benefit of selling data (in utility terms) to the marginal privacy cost, while Equation (27) is the standard consumption Euler equation.

Final-goods producers. Each monopolistic producer of variety i chooses the quantity of AI services and labor to employ, taking data quality Q_t and the demand-side price $p^d_{D,t}$ as given. The profit maximization problem is

$$\max_{\{D_{i,t},\, L_{i,t},\, p_{i,t}\}} \; p_{i,t} Y_{i,t} - p^d_{D,t}\, Q_t D_{i,t} - w_t L_{i,t},$$

where $p^d_{D,t}$ denotes the demand-side purchase price of AI services. By symmetry across varieties, the equilibrium price simplifies to $p_{i,t} = N^{1/(\sigma-1)}$ for all i, and the standard monopolistic markup $\sigma/(\sigma-1)$ applies.

Generative AI firm. The single monopolistic AI firm purchases consumer data d_t and rents computing power K_t, hires labor $L_{A,t}$ to produce new AI services, and sells the accumulated stock D_t to the data intermediary. Because the firm has monopoly power over its AI services, the supply-side price $p^s_{D,t}(D_t)$ is a decreasing function of the quantity supplied. The firm solves

$$\max_{\{d_t, D_t, L_{A,t}, p^s_{D,t}, K_t\}} \Pi_A = \sum_{t=0}^{\infty} \frac{1}{\prod_{s=0}^{t}(1 + r_s)} \left[p^s_{D,t}(D_t)\, Q_t D_t - w_t L_{A,t} - p_{K,t} K_t - p^s_{d,t} d_t L \right], \tag{28}$$

subject to the AI-service accumulation equation (4). The effective revenue is $p^s_{D,t} Q_t D_t$: data quality enters multiplicatively, so the firm has an incentive to improve quality by increasing its use of computing power or consumer data. The problem is inherently dynamic because AI services accumulate over time. The monopolistic position of the AI firm means that it will restrict the supply of AI services below the efficient level, a distortion compounded by the monopoly power of the computing power provider upstream.

Computing power provider. The computing power provider is the sole supplier of computing resources. It accumulates computing power through investment I_t and labor $L_{K,t}$, and rents these resources to the AI firm. The provider solves

$$\max_{\{I_t, L_{K,t}, K_t, p_{K,t}\}} \Pi_K = \sum_{t=0}^{\infty} \frac{1}{\prod_{s=0}^{t}(1 + r_s)} \Big[p_{K,t}(K_t)\, K_t - w_t L_{K,t} - I_t \Big], \tag{29}$$

subject to the computing power accumulation equation (6). Its monopolistic position allows it to set a rental price $p_{K,t}(K_t)$ above the competitive level, raising the cost of computing power for the AI firm and reducing the equilibrium level of computing resources.

Equilibrium. A competitive equilibrium under restricted entry is a path of quantities $\{c_t, d_t, c_{i,t}, a_t, D_t, L_{A,t}, K_t, I_t, L_{K,t}, L_{i,t}, Y_{i,t}\}_{t=0}^{\infty}$ and prices $\{w_t, r_t, p^s_{d,t}, p^s_{D,t}, p^d_{D,t}, p_{K,t}\}_{t=0}^{\infty}$ such that: (i) the household maximizes lifetime utility (1) subject to its budget constraint (25); (ii) the generative AI firm maximizes (28) subject to (4); (iii) the computing power provider maximizes (29) subject to (6); (iv) each final-goods producer maximizes per-period profit; (v) the labor market, goods market, consumer-data market, AI-services market, and computing-capital market all clear, as specified in Section 2; and (vi) the asset market clears, $I_t = [a_{t+1} - (1 + r_t)a_t]L$.

Because the model abstracts from long-run growth, all variables are constant in the steady state. We use a superscript "re" to denote the steady-state value of each variable under the restricted-entry competitive equilibrium. Equation (27) then implies $r^{re} = 1/\beta - 1$. Substituting this result into the remaining equations yields the following proposition.

Proposition 2 (Competitive Equilibrium with Restricted Entry)**.** *In competitive equilibrium with $M = 1$, all variables are constant in the steady state. The three endogenous variables $(L_A^{re}, L_K^{re}, D^{re})$, together with consumer-data supply d^{re}, are jointly determined by the following system:*

$$\frac{L_K^{re}}{L - L_A^{re} - L_K^{re}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma} B^{re}, \tag{30}$$

$$\frac{[1-\beta(1-\delta_D)]D^{re}}{(1-\alpha)\beta\eta(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{re})^{-\alpha}(L - L_A^{re} - L_K^{re})} = (1-\tau)\zeta^2 + \frac{\alpha}{1-\alpha}\cdot\frac{L_A^{re}}{L - L_A^{re} - L_K^{re}}\cdot\frac{\lambda(D^{re})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{re}}, \tag{31}$$

$$\phi'(d^{re})d^{re} = \frac{\sigma-1}{(1-B^{re})\sigma}\left[b\tau\zeta^2 + \frac{\alpha L_A^{re}(1-\lambda)(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L - L_A^{re} - L_K^{re})\Omega^{re}}\right], \tag{32}$$

where the auxiliary Ω^{re} is pinned down by the steady-state AI-service-accumulation identity

$$\Omega^{re} = \left[\frac{\delta_D D^{re}}{\eta(L_A^{re})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} = \lambda(D^{re})^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}},$$

and the parameter bundle B^{re} is

$$B^{re} = \frac{(\sigma-1)\zeta^3(1-b)^2\tau^2\beta\gamma\delta_K}{\left[1-\beta(1-\delta_K)\right]\sigma}. \tag{33}$$

Given the solution, per-variety output and the aggregate compute stock are

$$Y_i^{re} = \left\{A(d^{re}L)^{b\tau\zeta}(D^{re})^{\zeta(1-\tau)}\frac{L - L_A^{re} - L_K^{re}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{re})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{re}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}},$$

$$K^{re} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{re})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{re}\right]^{\gamma}(Y_i^{re})^{\gamma}.$$

Two notable differences emerge relative to the optimal allocation in Proposition 1. First, comparing Equations (30) and (17), the ratio $B^{re}/B^{sp} = \zeta^2(1-b)\tau(\sigma-1)/\sigma < 1$ shows that the share of output devoted to computing investment is strictly lower under competitive equilibrium than under the

social optimum, reflecting the combined monopoly power of the AI firm and the computing power provider. Second, the data-related Equations (31) and (32) differ from their counterparts (18) and (19) through the additional powers of ζ and the markup factor $(\sigma-1)/\sigma$, reflecting the multiple layers of monopolistic distortion.

3.3 Competitive Market Equilibrium: Free Entry

We now consider an alternative market regime in which entry into the generative AI sector is unrestricted. When the entry cost F is finite, new AI firms will enter as long as the present value of expected profits exceeds F. In a symmetric steady-state equilibrium, entry occurs until per-firm profits are exactly driven to the amortized entry cost $(1-\beta)F$.

This regime captures a more competitive market for AI services. Unlike the restricted-entry case studied in Section 3.2, where a single monopolistic AI firm controls the entire supply of AI services, here M^{fe} firms each produce a differentiated type. Each firm exercises monopolistic competition power in the sense that it faces a downward-sloping demand curve for its own variety of service, with perceived elasticity ρ from the CES data aggregator (7), while at the same time competing with other firms for consumer data and computing power. Because consumer data is nonrival, it can be simultaneously used by all M^{fe} firms: each firm pays $p^s_{d,t}/M^{fe}$ through the consumer data intermediary, and the consumer receives the full price $p^s_{d,t}$. A larger number of AI firms therefore raises the total demand for consumer data and increases the price that consumers receive, strengthening their incentive to supply data. Computing power, by contrast, is rival and must be divided across firms, with each firm renting K_t/M^{fe}.

The effective data input under symmetry is $\Phi_t = (M^{fe})^{\rho/(\rho-1)} Q_t D_t$, and the aggregate resource constraint takes the same form as Equation (14) with $M = M^{fe}$. Each AI firm j takes the aggregate data composite Φ_t as given and maximizes the present discounted value of its profits, facing a CES demand curve with elasticity ρ for its variety of AI service. The computing power provider remains a monopolist and rents computing resources to all M^{fe} AI firms.

We use a superscript "fe" to denote the steady-state value of each variable under the free-entry competitive equilibrium. The complete derivation is provided in Appendix A.3. Solving the optimization problems of all agents and imposing the free entry condition yields the following proposition.

Proposition 3 (Competitive Equilibrium with Free Entry)**.** *In competitive equilibrium with free entry, all variables are constant in the steady state. The four endogenous variables* $(L_A^{fe}, L_K^{fe}, D^{fe}, M^{fe})$*, together with consumer-data supply* d^{fe}*, are jointly determined by the following system:*

$$\frac{L_K^{fe}}{L - L_A^{fe} - L_K^{fe}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma} B^{fe}, \tag{34}$$

$$\frac{[1-\beta(1-\delta_D)]\, D^{fe}\, M^{fe}}{(1-\alpha)\beta\eta(\Omega^{fe})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{fe}/M^{fe})^{-\alpha}(L - L_A^{fe} - L_K^{fe})} = \left(1-\frac{1}{\rho}\right)(1-\tau)\zeta + \frac{\alpha}{1-\alpha}\cdot\frac{L_A^{fe}}{L - L_A^{fe} - L_K^{fe}}\cdot\frac{\lambda(D^{fe})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{fe}}, \tag{35}$$

$$\phi'(d^{fe})d^{fe} = \frac{\sigma-1}{(1-B^{fe})\sigma}\left[\left(1-\frac{1}{\rho}\right)b\tau\zeta + \frac{\alpha L_A^{fe}(1-\lambda)(d^{fe}L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L-L_A^{fe}-L_K^{fe})\Omega^{fe}}\right], \tag{36}$$

$$(1-\beta)F = \frac{1}{\rho}\left(1-\frac{1}{\sigma}\right)\zeta\frac{Y^{fe}}{M^{fe}}, \tag{37}$$

where $Y^{fe} \equiv c^{fe}L + I^{fe}$ is aggregate output. The auxiliary Ω^{fe} is pinned down by the steady-state AI-service-accumulation identity

$$\Omega^{fe} = \left[\frac{\delta_D D^{fe}}{\eta(L_A^{fe}/M^{fe})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} = \lambda(D^{fe})^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d^{fe}L)^{\frac{\varepsilon-1}{\varepsilon}},$$

and the parameter bundle B^{fe} is

$$B^{fe} = \frac{(\sigma-1)(1-1/\rho)\zeta^2(1-b)^2\tau^2\beta\gamma\delta_K}{\left[1-\beta(1-\delta_K)\right]\sigma}. \tag{38}$$

Given the solution, per-variety output and the aggregate compute stock are

$$Y_i^{fe} = \left\{A\,(M^{fe})^{\omega}(d^{fe}L)^{b\tau\zeta}(D^{fe})^{\zeta(1-\tau)}\frac{L-L_A^{fe}-L_K^{fe}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{fe})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{fe}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}},$$

$$K^{fe} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{fe})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{fe}\right]^{\gamma}(Y_i^{fe})^{\gamma},$$

with aggregate output $Y^{fe} = N^{\sigma/(\sigma-1)}Y_i^{fe}$ and per-capita consumption $c^{fe} = (1-B^{fe})Y^{fe}/L$.

Relative to Proposition 2, the labor allocation (34) and the data-related Equations (35) and (36) replace one power of ζ with the CES markup factor $(1-1/\rho)$, reflecting the shift from pure monopoly to monopolistic competition in the market for AI services. The compact form of the free-entry condition (37) has a transparent interpretation: because the AI firm's technology is constant-returns-to-scale in its inputs and faces a CES demand curve with elasticity ρ, Euler's theorem implies that total input payments, namely the labor bill $w^{fe}L_A^{fe}$, the compute bill, and the consumer-data bill $p_d^{fe}d^{fe}L$, equal $(1-1/\rho)$ of gross revenue, so free entry drives the residual $(1/\rho)$-share, the CES markup rent, down to the amortized entry cost $(1-\beta)F$.[7] A detailed comparison across regimes is deferred to Section 3.5.

3.4 Competitive Market Equilibrium: Data Regulation

The previous two regimes assumed that consumers freely choose how much personal data to supply, subject only to their own privacy preferences. In practice, however, governments increasingly restrict

[7] We adopt the convention that regime-superscripted prices such as p_d^{fe} (and later p_d^{dr}) refer to the supply-side steady-state price in the corresponding regime; the supply-side superscript s is suppressed whenever a regime superscript is present, in the interest of notational parsimony.

the commercial use of consumer data. Motivated by regulations such as the European Union's General Data Protection Regulation (GDPR) and similar legislative initiatives worldwide, we now consider a regime in which the government imposes a binding cap $\bar{d}$ on the per-capita amount of consumer data that can be used in the economy.

We maintain the free-entry assumption so that M AI firms compete in the market for AI services, exactly as in Section 3.3. The regime in this subsection refers specifically to the case in which the cap is strictly binding, that is, $\bar{d} < d^{fe}$, where d^{fe} is the unconstrained per-capita data supply characterized in Proposition 3. When $\bar{d} \geq d^{fe}$ the consumer's privately optimal choice already satisfies the cap, so the cap is slack, the household's first-order condition (26) continues to hold, and the equilibrium coincides with the free-entry regime. We therefore restrict attention to $\bar{d} < d^{fe}$ throughout this subsection. Under the binding cap, $d_t = \bar{d}$ holds in every period, and the market price of consumer data adjusts to clear the market at this regulated quantity. Consumers still receive income $p^s_{d,t}\bar{d}$ from data sales, but the price is now determined by the AI firms' aggregate willingness to pay for consumer data rather than by the consumer's marginal privacy cost.

This regime highlights a fundamental trade-off between privacy protection and productive efficiency. On the one hand, restricting data usage reduces the privacy disutility $\phi(\bar{d})$ borne by consumers. On the other hand, less consumer data lowers data quality Q_t through Equation (5) and reduces the flow of new AI services through Equation (3), both of which depress output. In the extreme case $\bar{d} \to 0$, data quality falls to zero and production collapses entirely, since consumer data is an essential input to data quality. The welfare implications of data regulation therefore depend critically on the level of the cap $\bar{d}$ relative to the unconstrained free-entry equilibrium d^{fe}.

We use a superscript "dr" to denote the steady-state value of each variable under this regime. The complete derivation is provided in Appendix A.4. Because the market structure is identical to the free-entry regime, the derivation closely parallels Appendix A.3, with the consumer data supply equation replaced by the binding cap.

Proposition 4 (Competitive Equilibrium with Data Regulation)**.** *Suppose the data cap is strictly binding, $\bar{d} < d^{fe}$, where d^{fe} is the unregulated per-capita data supply characterized in Proposition 3. In competitive equilibrium with data regulation, all variables are constant in the steady state. Consumer-data supply is fixed at $d^{dr} = \bar{d}$; the four endogenous variables $(L_A^{dr}, L_K^{dr}, D^{dr}, M^{dr})$ are jointly determined by the following system:*

$$\frac{L_K^{dr}}{L - L_A^{dr} - L_K^{dr}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma} B^{dr}, \tag{39}$$

$$\frac{[1-\beta(1-\delta_D)]\, D^{dr}\, M^{dr}}{(1-\alpha)\beta\eta(\Omega^{dr})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{dr}/M^{dr})^{-\alpha}(L - L_A^{dr} - L_K^{dr})} = \left(1 - \frac{1}{\rho}\right)(1-\tau)\zeta + \frac{\alpha}{1-\alpha} \cdot \frac{L_A^{dr}}{L - L_A^{dr} - L_K^{dr}} \cdot \frac{\lambda (D^{dr})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{dr}}, \tag{40}$$

$$(1-\beta)F = \frac{(1-1/\sigma)\zeta\, Y^{dr}}{M^{dr}} \left[1 - \left(1 - \frac{1}{\rho}\right)\tau\right] - \frac{(1-1/\sigma)\, Y^{dr}\, L_A^{dr}}{(1-\alpha)M^{dr}(L - L_A^{dr} - L_K^{dr})} \left[1 - \alpha\, \frac{\lambda (D^{dr})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{dr}}\right], \tag{41}$$

where $Y^{dr} \equiv c^{dr} L + I^{dr}$ is aggregate output. The auxiliary Ω^{dr} is pinned down by the steady-state AI-service-

accumulation identity

$$\Omega^{dr} = \left[\frac{\delta_D D^{dr}}{\eta(L_A^{dr}/M^{dr})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} = \lambda(D^{dr})^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(\bar{d}L)^{\frac{\varepsilon-1}{\varepsilon}},$$

and the parameter bundle B^{dr} coincides with B^{fe} in (38), reflecting that the computing investment share is determined by market structure rather than by data regulation. Given the solution, per-variety output and the aggregate compute stock are

$$Y_i^{dr} = \left\{A\,(M^{dr})^{\omega}(\bar{d}L)^{b\tau\zeta}(D^{dr})^{\zeta(1-\tau)}\frac{L - L_A^{dr} - L_K^{dr}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{dr})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{dr}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}, \tag{42}$$

$$K^{dr} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{dr})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{dr}\right]^{\gamma}(Y_i^{dr})^{\gamma},$$

with aggregate output $Y^{dr} = N^{\sigma/(\sigma-1)}Y_i^{dr}$ and per-capita consumption $c^{dr} = (1-B^{dr})Y^{dr}/L$. Steady-state social welfare is $W^{dr} = (1-\beta)^{-1}\left[\ln c^{dr} - \phi(\bar{d})\right]$.

Two economic forces shape the regulated equilibrium. First, by forcing each consumer to sell less data at $\bar{d}$ than she would privately choose, the cap mechanically reduces the quality of the data input that AI firms use in production, which drags down per-firm productivity. Second, and less obviously, the cap also reduces each AI firm's total spending on consumer data: although the scarcer data commands a higher unit price p_d^{dr}, the product $p_d^{dr}\bar{d}$ falls because price does not rise enough to compensate for the quantity cut. Each firm therefore becomes cheaper to operate, the free-entry margin loosens, and more AI firms enter in equilibrium, so $M^{dr} > M^{fe}$. These two forces pull per-firm output and variety in opposite directions, and which one dominates for aggregate output is a quantitative question that we take up in Section 4.2. Consumers, meanwhile, collect a higher price on every unit of data they are allowed to sell, so income from data sales need not fall one-for-one with $\bar{d}$; this rent, together with the lower privacy disutility $\phi(\bar{d})$, is what the household gains in exchange for the productivity distortion.

3.5 Comparison across Regimes

Propositions 1–4 deliver closed-form steady-state characterizations for all four regimes. This subsection draws out the key economic insights that emerge from comparing these expressions, summarized in Table 2.

Underprovision of computing power. The constant B, which pins down the share of aggregate output allocated to computing investment, provides the sharpest cross-regime comparison. Computing the ratios directly from Equations (21), (33), and (38) yields

$$\frac{B^{re}}{B^{sp}} = \frac{(\sigma-1)\zeta^2(1-b)\tau}{\sigma}, \qquad \frac{B^{fe}}{B^{re}} = \frac{\rho-1}{\rho\zeta}, \qquad \frac{B^{fe}}{B^{sp}} = \frac{(\sigma-1)(1-1/\rho)\zeta(1-b)\tau}{\sigma}.$$

Table 2. Comparison across Regimes

	Social Planner	Restricted Entry	Free Entry	Data Regulation
Number of AI firms	M^{sp} (optimal)	1 (exogenous)	M^{fe} (free entry)	M^{dr} (free entry)
Consumer data	d^{sp} (optimal)	d^{re} (market)	d^{fe} (market)	$\bar{d}$ (regulated)
Data market	No distortion	Monopoly	Monopolistic competition	Monopolistic competition
Distortion factor	ζ	ζ^2	$(1-1/\rho)\zeta$	$(1-1/\rho)\zeta$
Computing share B	Highest	Lowest	Intermediate	$= B^{fe}$

Since ζ, $(1-b)\tau$, and $(\sigma-1)/\sigma$ are all strictly less than one, the first ratio satisfies $B^{re}/B^{sp} \ll 1$: the restricted-entry equilibrium devotes far less output to computing investment than the social optimum. The root cause is the compounding of three monopolistic distortions, arising in the goods market, the AI-services market, and the computing market, each of which depresses the perceived return to computing and discourages investment.

Introducing competition in the AI-services market through free entry eliminates one of these three distortions. The second ratio, $(\rho-1)/(\rho\zeta)$, is much greater than one for any $\rho > 1$, implying a dramatic increase in computing investment when multiple AI firms compete. The economic logic is straightforward: under monopoly ($M = 1$), the AI firm restricts the supply of AI services to extract rents, which in turn suppresses the derived demand for computing power. Under monopolistic competition, each firm's perceived demand elasticity rises from its low monopoly level to ρ, greatly expanding data supply and, consequently, computing demand. Nevertheless, the third ratio shows that even with free entry, computing investment remains below the social optimum. The residual gap reflects the goods-market markup $(\sigma-1)/\sigma$ and the computing-market monopoly, neither of which is addressed by entry into the AI-services market. Data regulation does not alter this conclusion, since $B^{dr} = B^{fe}$: the computing investment share depends on market structure, not on the level of data usage. In summary,

$$B^{re} \ll B^{fe} = B^{dr} < B^{sp}.$$

Anatomy of distortions. The AI-services-market distortion identified above takes a particularly transparent form. In each regime, a single "distortion factor" appears repeatedly in the AI-service Euler equation, the computing demand, and the B constant (see Table 2). In the social planner's solution, the marginal product of data enters as a single factor of ζ, reflecting the true elasticity of output with respect to effective data. Under restricted entry, this becomes ζ^2: the AI firm, facing the full demand curve for its AI services, perceives a marginal revenue that involves an additional factor of ζ from the production function's data elasticity. Under free entry, the CES demand curve replaces this additional ζ with $(1-1/\rho)$, the firm's perceived marginal revenue share, so that the distortion factor takes the values ζ (SP), ζ^2 (RE), and $(1-1/\rho)\zeta$ (FE and DR). Since $\zeta < 1$ while $(1-1/\rho)$ is close

to one for $\rho > 1$, the distortion under free entry is much smaller than under monopoly. This pattern, in which ζ is replaced by ζ^2 under monopoly and by $(1-1/\rho)\zeta$ under monopolistic competition, is the structural signature of AI-services-market power in the model, and it propagates to every equilibrium condition that involves the AI firm's optimization.

Endogenous number of AI firms. In all regimes except restricted entry, the number of AI firms is endogenous. Comparing the social planner's optimality condition (20) with the free entry conditions (37) and (41) reveals a fundamental difference in how M is determined. The social planner equates the amortized entry cost $(1-\beta)F$ to the net social benefit of an additional AI-service variety: the love-of-variety productivity gain $\omega Y/M$ minus the labor dilution cost from spreading AI workers across more firms. Private firms, by contrast, enter until per-firm profit equals the amortized entry cost. These two conditions generally yield different values of M, for two reasons.

First, the business-stealing effect tends to make free entry excessive: a new entrant captures market share from incumbents, and the private incentive to enter exceeds the social benefit because part of the entrant's revenue comes at the expense of existing firms. Second, the nonrivalry externality of consumer data works in the opposite direction: each additional firm benefits from the existing stock of consumer data without diminishing it, but this social benefit is not fully captured in private profits because the consumer data intermediary distributes the cost equally across firms. Whether M^{fe} exceeds or falls short of M^{sp} depends on the relative strength of these two forces, which is in general analytically ambiguous.

Under data regulation, the number of AI firms M^{dr} is also endogenous but is affected by the data cap $\bar{d}$. The free entry condition (41) shows that a lower $\bar{d}$ reduces the per-firm consumer data cost $p_d^{dr}\bar{d}L/M^{dr}$, relaxing the entry constraint and permitting more firms to enter, so that M^{dr} is typically decreasing in $\bar{d}$ in the neighborhood of d^{fe}.[8] This introduces an indirect channel through which data regulation affects productivity: beyond the direct effect of less consumer data on data quality, a binding cap also alters the diversity of AI services available to producers by reshaping AI firm entry, operating on the effective data input Φ_t through the love-of-variety channel.

Welfare-improving data regulation. Can a binding cap $\bar{d} < d^{fe}$ strictly improve welfare? The question is subtler than it may first appear, because the consumer at d^{fe} already solves her own optimization problem, so welfare gains can only arise from distortions she does not internalize. AI firms' monopoly markups create one such distortion: they pay only the fraction $(1-1/\rho)$ of data's marginal revenue product, so p_d^{fe} sits below data's social value and consumers under-supply data from a productivity-maximizing perspective. This markup externality alone would call for looser regulation, not tighter. Cutting the other way, the consumer's private choice overlooks general-equilibrium feedback through AI-service accumulation, labor reallocation, and the AI-firm entry channel, which may push toward tighter regulation. Proposition 5 formalizes the trade-off.

[8] Although the equilibrium price p_d^{dr} rises as $\bar{d}$ tightens (since AI firms bid more aggressively for a scarcer input), the product $p_d^{dr}\bar{d}$ declines in our calibration, lowering per-firm data expenditures. This is confirmed in Section 4.2: tightening the cap from $d^{fe} \approx 0.066$ to $\bar{d} = 0.05$ raises M^{dr} from approximately 6.6 to 7.7.

Proposition 5 (Welfare-improving data regulation). *Let $W^{dr}(\bar{d})$ denote steady-state social welfare under data regulation with cap $\bar{d}$ and let d^{fe} denote the unregulated free-entry level. Define $\Theta \equiv 1-(1-b)\gamma\tau\zeta$ and $\xi_L(\bar{d}) \equiv -L_A^{dr}(\bar{d})/[L-L_A^{dr}(\bar{d})]$. Then:*

(i) For $\bar{d} \geq d^{fe}$ the cap is non-binding and $W^{dr}(\bar{d}) = W^{fe}$. On $(0, d^{fe})$ the cap binds and $W^{dr}(\bar{d}) = (1-\beta)^{-1}[\ln c^{dr}(\bar{d}) - \phi(\bar{d})]$, which rises from $-\infty$ at $\bar{d}=0$ to W^{fe} as $\bar{d} \to d^{fe-}$.

(ii) Assume W^{dr} is continuously differentiable and strictly quasi-concave on $(0, d^{fe})$. An interior optimum $\bar{d}^* \in (0, d^{fe})$ with $W^{dr}(\bar{d}^*) > W^{fe}$ exists if and only if

$$\left.\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}}\right|_{\bar{d}=d^{fe-}} < \phi'(d^{fe}) = \frac{p_d^{fe}}{c^{fe}}. \tag{43}$$

(iii) When it exists, $\bar{d}^*$ satisfies the implicit equation

$$\phi'(\bar{d}^*)\,\Theta\,\bar{d}^* = \underbrace{b\tau\zeta}_{\text{direct}} + \underbrace{\omega\,\eta_M(\bar{d}^*)}_{\text{entry}} + \underbrace{\zeta(1-\tau)\eta_D(\bar{d}^*)}_{\text{AI-service stock}} + \underbrace{[1+(1-b)\tau\zeta(1-\gamma)]\,\xi_L(\bar{d}^*)\,\eta_{L_A}(\bar{d}^*)}_{\text{labor}}, \tag{44}$$

where $\omega \equiv \rho\zeta/(\rho-1) - (1-b)\tau\zeta$ (defined in (22)) is the net elasticity of output with respect to the number of AI firms, and $\eta_X(\bar{d}) \equiv \mathrm{d}\ln X^{dr}/\mathrm{d}\ln\bar{d}$ for $X \in \{M, D, L_A\}$ are the equilibrium elasticities in the DR system.

The detailed derivation is provided in Appendix A.5. The welfare function is piecewise: it equals W^{fe} for any $\bar{d} \geq d^{fe}$ and rises from $-\infty$ toward W^{fe} as $\bar{d}$ increases from 0 to d^{fe}. Condition (43) states that the slope of W^{dr} immediately below d^{fe} is negative, so W^{dr} overshoots W^{fe} from below as $\bar{d} \to d^{fe-}$, implying the existence of an interior maximum $\bar{d}^* < d^{fe}$ where regulation strictly improves welfare. The right-hand side of (43) is the consumer's private marginal benefit from data sales, namely the data price p_d^{fe}/c^{fe} that satisfies her FOC in free entry. The left-hand side is the full general-equilibrium marginal effect of the cap on consumption, which the consumer does not observe. Regulation improves welfare when this general-equilibrium effect on consumption is smaller than the private price the consumer receives in FE.

Equation (44) decomposes the marginal consumption response into four economically distinct channels, each of which has a clear interpretation. The direct effect $b\tau\zeta > 0$ captures the productivity gain from having more consumer data in the quality function. A higher $\bar{d}$ raises data quality Q^{dr} through the $(\bar{d}L)^{b\tau\zeta}$ factor, directly increasing output. This channel is positive and is controlled by the product $b\tau\zeta$: a higher weight of consumer data in quality (b), a stronger quality channel (τ), or a larger data share in production (ζ) all amplify it. This is the only channel that operates absent general-equilibrium feedback, and it is the largest single term pushing in favor of more data. Notably, the corresponding private price p_d^{fe} in FE reflects only a markup-scaled fraction of this direct productivity gain, so the consumer systematically under-supplies data from the direct-productivity perspective.

The entry effect $\omega\,\eta_M$ captures how the data cap shapes the equilibrium number of AI firms through the free-entry condition. Because $\eta_M < 0$ near d^{fe} (tightening the cap lowers per-firm

consumer data costs, relaxing the entry constraint and attracting more firms), this channel is negative. Its magnitude depends on two forces: the net love-of-variety elasticity $\omega = \rho\zeta/(\rho-1)-(1-b)\tau\zeta$, which nets the positive love-of-variety effect against the compute-dilution effect (spreading total compute across more firms reduces per-firm data quality), and the entry elasticity η_M itself. Smaller ρ (greater product differentiation among AI firms) amplifies ω because each additional entrant delivers larger love-of-variety gains, making this the dominant channel when AI products are highly differentiated.

The AI-service stock effect $\zeta(1-\tau)\,\eta_D$ captures the response of the stock of AI services to the cap. The coefficient $\zeta(1-\tau) > 0$ is the elasticity of output with respect to D^{dr}. The sign of η_D is ambiguous in general, but in our calibration $\eta_D > 0$: a tighter cap reduces the flow of new AI services generated, lowering D^{dr}. This channel therefore reinforces the direct effect, both pushing toward looser regulation.

The labor reallocation effect $[1+(1-b)\tau\zeta(1-\gamma)]\,\xi_L\,\eta_{L_A}$ captures the general-equilibrium reallocation of labor between data-generation and production sectors. The prefactor $1+(1-b)\tau\zeta(1-\gamma) > 0$ combines the output elasticities with respect to production labor and computing labor; $\xi_L = -L_A^{dr}/(L-L_A^{dr}) < 0$ is the labor-constraint multiplier, reflecting that a one-percent increase in L_A reduces production labor by a fraction $L_A/(L-L_A)$ of one percent. The sign of η_{L_A} depends on whether tightening the cap increases or decreases labor demand in the data-generation sector.

The four-channel decomposition above admits a more economical interpretation. Two opposing general-equilibrium externalities, neither of which the consumer internalizes at the unregulated margin, determine whether a binding cap improves welfare. The markup externality pushes for a looser cap: because AI firms pay only the fraction $(1-1/\rho)$ of data's marginal revenue product, the private price p_d^{fe} falls below data's full productivity value, and the consumer's unregulated choice already undersupplies data from a productivity-maximizing perspective. This externality is captured by the direct channel $b\tau\zeta$ and reinforced by the data-stock channel $\zeta(1-\tau)\eta_D$ under our calibration. The entry externality pushes in the opposite direction: the consumer does not internalize that her data supply at the unregulated margin raises the per-firm consumer-data bill, tightens the free-entry condition of AI firms, and crowds out additional varieties. This externality is captured by the entry channel $\omega\,\eta_M$, whose magnitude rises with the net love-of-variety elasticity ω and the entry responsiveness $|\eta_M|$. The labor-reallocation channel is a third-order refinement whose sign depends on how labor flows adjust in equilibrium; at our calibration it is quantitatively smaller than the first three.

A binding cap therefore improves welfare whenever the entry externality dominates the markup externality. The condition is easiest to satisfy when ρ is small (AI products are highly differentiated, so each additional entrant delivers large love-of-variety gains) and when entry is responsive to the per-firm data bill (so $|\eta_M|$ is large). Under our benchmark calibration in Section 4.2, the entry externality dominates by only a modest margin: the numerical results show $c^{dr}(0.05) > c^{fe}$ by roughly 0.6 percent, which implies $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d}|_{d^{fe-}} < 0$, so condition (43) holds.

4. Numerical Simulation and Policy Experiments

After deriving and analyzing the model preliminarily, this section further illustrates its main implications through numerical simulations. We first calibrate the relevant parameters in the model by drawing on the existing literature. We then discuss how several key parameters affect the values of the main variables when the economy reaches the steady state. Finally, we conduct a set of representative policy experiments to examine their potential effects.

4.1 Parameter Calibration

The parameter values used in our numerical exercises are reported in Table 3. We discuss the calibration strategy below.

Table 3. Parameter Calibration

Parameter	Definition	Value	Source
β	Subjective discount factor	0.99	Standard
γ	Investment share in compute production	0.33	Standard
δ_K	Depreciation rate of compute	0.08	BEA Fixed Assets
σ	Elasticity of substitution across varieties	4	Jones and Tonetti (2020)
ζ	Importance of data in production	0.06	Jones and Tonetti (2020)
ε	Elasticity of substitution between existing AI output and consumer data	50	Jones and Tonetti (2020)
ϕ	Privacy disutility coefficient, $\phi(d) = \phi d^2$	1	Jones and Tonetti (2020)
λ	Weight of existing AI output in service generation	0.3	Shumailov et al. (2024)
δ_D	Depreciation rate of AI services	0.2	Corrado et al. (2009)
α	Data share in AI-service generation	0.33	Kaplan et al. (2020)
η	Efficiency of AI-service generation	1	Normalized
ξ	Efficiency of compute production	1	Normalized
A	Total factor productivity	1	Normalized
b	Weight of consumer data in data quality	0.5	Normalized
L	Total population	100	Normalized
N	Number of varieties	100	Normalized
τ	Data quality adjustment parameter	0.5	Comparative statics
ρ	Elasticity of substitution across data types	2	Comparative statics
F	Entry cost for AI firms	200	Comparative statics

Several parameters take standard values from the macroeconomics literature. We set the subjective discount factor to $\beta = 0.99$ and the elasticity of substitution across consumption varieties to $\sigma = 4$, following Jones and Tonetti (2020). The investment share in the production of computing power is set to $\gamma = 0.33$, consistent with the standard capital share. The efficiency terms η, ξ, and A are each

normalized to 1, and the weight of consumer data in data quality is set to $b = 0.5$ so that computing power and consumer data play equally important roles. The steady-state propositions of Section 3 imply that the equilibrium allocation depends mainly on the distribution of labor across sectors rather than on absolute levels of population or variety, so we normalize $L = 100$ and $N = 100$ without loss of generality.

The remaining parameters require more careful calibration. Unlike conventional physical capital such as structures or general equipment, computing hardware, particularly GPUs and other AI-specific processors, depreciates rapidly due to continual advances in semiconductor technologies and the short replacement cycles of data-center infrastructure. According to the Bureau of Economic Analysis (BEA) Fixed Asset Tables, the annual depreciation rate of computers and peripheral equipment is approximately 31.5%, substantially higher than the rates applied to buildings or industrial equipment. We therefore set $\delta_K = 0.08$ on a quarterly basis, corresponding to an annual rate of roughly 30%. For the parameter ζ, which governs the importance of data in production, we follow Jones and Tonetti (2020), who estimate the share of GDP attributable to data at 5–10% and obtain $\zeta \approx 0.06$; in our model ζ plays an analogous role to their parameter η, so we adopt the same value. We set $\varepsilon = 50$, again following Jones and Tonetti (2020), who choose a large elasticity of substitution so that the CES aggregator is approximately linear, i.e., $(\varepsilon - 1)/\varepsilon \approx 1$. In our context, this implies that the firm's existing AI output and consumer data are close substitutes in the generation process, allowing the key distinction between the two inputs to operate primarily through the data quality channel.

The parameter λ governs the relative importance of the firm's existing AI output versus consumer data in the generation of new AI services. Shumailov et al. (2024) study the phenomenon of model collapse, in which iterative training on AI-generated data leads to progressive degradation of model performance. Their findings suggest that when the proportion of synthetic data in the training set exceeds approximately 50–60%, model quality deteriorates significantly, implying that the effective contribution of existing AI output is bounded well below one half. We accordingly set $\lambda = 0.3$. For the depreciation rate of AI services, data is commonly treated in the literature as a form of intangible capital. Following the estimates in Corrado et al. (2009) for the depreciation of intangible assets, and consistent with the observation that training data typically needs to be refreshed every three to five years to remain relevant, we set $\delta_D = 0.2$.

Two additional parameters, namely α and ϕ, can be calibrated using indirect empirical evidence. The parameter α governs the share of data inputs relative to labor in the AI-service generation process. Kaplan et al. (2020) document that the performance of large language models improves as a power law in training dataset size, with scaling exponents in the range 0.3–0.5. Since α controls the elasticity of AI-service output with respect to data inputs in our model, it plays an analogous role to the data scaling exponent in neural scaling laws. We set $\alpha = 0.33$, consistent with the lower end of the empirically observed range. For the privacy disutility, we assume the quadratic form $\phi(d) = \phi d^2$, following Jones and Tonetti (2020) and Farboodi and Veldkamp (2026).[9] To calibrate ϕ, we follow the

[9] Farboodi and Veldkamp (2026) note that when consumer data is already used extensively, privacy may be nearly fully exposed, so additional data use imposes only limited marginal privacy costs. In that case, a concave form may be more appropriate. However, in the more typical case in which consumer data is only partially used, the convex (quadratic) specification is well justified.

approach of Jones and Tonetti (2020), who set their privacy parameter $\kappa = 0.20$ so that full disclosure of consumer data is equivalent to a 10% reduction in consumption. Applying a similar logic to our quadratic specification, we set $\phi = 1$, which implies that at the equilibrium level of data provision the privacy cost amounts to a few percent of consumption, a plausible magnitude given the existing survey evidence on willingness to pay for privacy protection.

The remaining three parameters, namely the data quality adjustment parameter τ, the elasticity of substitution across different types of AI services ρ, and the entry cost for AI firms F, lack reliable empirical estimates. Importantly, these three parameters correspond to the three core mechanisms of the model: τ governs the strength of the data quality channel through which computing power and consumer data affect production; ρ controls the degree of differentiation among AI firms and hence the intensity of competition in the market for AI services; and F determines the entry barrier and hence the equilibrium number of AI firms. Because the values of all three parameters are uncertain and may substantially affect the model's implications, we examine their effects in detail through comparative statics in the next subsection.

For the benchmark, we set $\tau = 0.5$ and $\rho = 2$. The choice of ρ is constrained by the model's equilibrium structure: when ρ is large, different types of AI services become close substitutes, AI firms' markup $\rho/(\rho - 1)$ shrinks toward unity, narrowing the rent available to cover the entry cost. Under our calibration, however, the free-entry equilibrium remains well-defined throughout the range $\rho \in [1.1, 4]$ that we consider in the comparative statics. The upper end $\rho = 4$ effectively approximates the perfect-substitutes limit: the markup $\rho/(\rho - 1) = 4/3$ is already close to one, and the per-firm rent shrinks to the point where the number of AI firms that the economy can support becomes very small. We therefore set $\rho = 2$ as the benchmark, corresponding to a markup of 2 and substantial product differentiation among AI firms. The entry cost is set at $F = 200$, chosen so that the free-entry equilibrium features a moderate number of AI firms ($M^{fe} \approx 6.6$). For the data regulation regime, we set the data cap at $\bar{d} = 0.05$, which is strictly below the unconstrained free-entry level $d^{fe} \approx 0.066$, ensuring that the cap is binding and has real economic effects.

4.2 Numerical Results and Comparative Statics

This section presents the quantitative properties of the four regimes. Section 4.2.1 reports the benchmark results at the calibrated parameter values, establishing the welfare ranking across regimes and the role of data regulation as a second-best policy. Section 4.2.2 then examines how these results respond to variation in the key parameters τ, ρ, and F.

4.2.1 Benchmark Results

Table 4 reports the steady-state values of the key variables under all four regimes at the benchmark parameterization: $\tau = 0.5$, $\rho = 2$, $F = 200$, and $\bar{d} = 0.05$.

Several patterns stand out. First, computing power is drastically underprovided in all competitive equilibria, confirming the analytical prediction of Section 3. The ratio $K^{sp}/K^{re} \approx 1{,}400$ is striking: the restricted-entry equilibrium invests roughly three orders of magnitude less in computing than the

Table 4. Benchmark Steady-State Results across Regimes

Variable	Social Planner	Restricted Entry	Free Entry	Data Regulation
Consumer data, d	0.110	0.023	0.066	0.050
AI services, D	1.561	1.562	2.442	2.014
Compute, K	16.526	0.012	0.106	0.107
Data quality, Q	1.253	0.324	0.366	0.361
Output, Y	707.4	445.0	572.5	576.3
Consumption, c	7.043	4.450	5.725	5.762
AI firms, M	30.2	1	6.6	7.7
Social welfare, W	194.0	149.2	174.0	174.9

social optimum. Free entry narrows this gap by an order of magnitude ($K^{fe}/K^{re} \approx 9$), but even under free entry computing remains far below the optimal level ($K^{sp}/K^{fe} \approx 156$). Data regulation does not appreciably change computing investment ($K^{dr} \approx K^{fe}$), consistent with $B^{dr} = B^{fe}$.

Second, consumer data usage follows a clear ranking $d^{sp} > d^{fe} > d^{dr} > d^{re}$. The social planner supplies roughly five times as much consumer data as the restricted-entry equilibrium, reflecting the compounded monopolistic distortions that depress the private return to data. Notably, data regulation achieves a slightly higher welfare ($W^{dr} = 174.9$) than the unconstrained free-entry equilibrium ($W^{fe} = 174.0$) despite using less consumer data ($\bar{d} = 0.050 < d^{fe} = 0.066$). This occurs because the binding data cap indirectly encourages more AI firm entry ($M^{dr} = 7.7 > M^{fe} = 6.6$), and the resulting love-of-variety gain more than offsets the direct output loss from less consumer data; reduced privacy costs contribute an additional, smaller piece to the welfare improvement.

Third, the social planner operates far more AI firms ($M^{sp} \approx 30$) than the free-entry equilibrium ($M^{fe} \approx 7$). This large gap reflects the fact that private firms do not internalize the full social benefit of AI-service variety in production. Data regulation partly mitigates this gap by increasing M^{dr} to roughly 8, since the binding data cap redistributes rents and lowers per-firm data costs, making entry more attractive.

Fourth, despite higher data quality ($Q^{sp} = 1.25$ versus $Q^{fe} = 0.37$) and more AI firms, the output difference across regimes is moderate in proportional terms ($Y^{sp}/Y^{re} \approx 1.6$), reflecting the small data elasticity $\zeta = 0.06$. However, because welfare is defined as the present value of log consumption net of privacy costs, even moderate output differences translate into substantial welfare gaps: in consumption-equivalent terms, developed in Section 4.3.3, the welfare loss from restricted entry relative to the social optimum amounts to a permanent reduction of about one-third in consumption.

4.2.2 Comparative Statics on Welfare

We now examine how the three key parameters τ, ρ, and F affect welfare across the four regimes. Figure 2 presents social welfare W as a function of each parameter, alongside the most informative secondary variable that clarifies the underlying mechanism: computing power K for τ, and the

number of AI firms M for ρ and F. The remaining variables are reported in Appendix B.

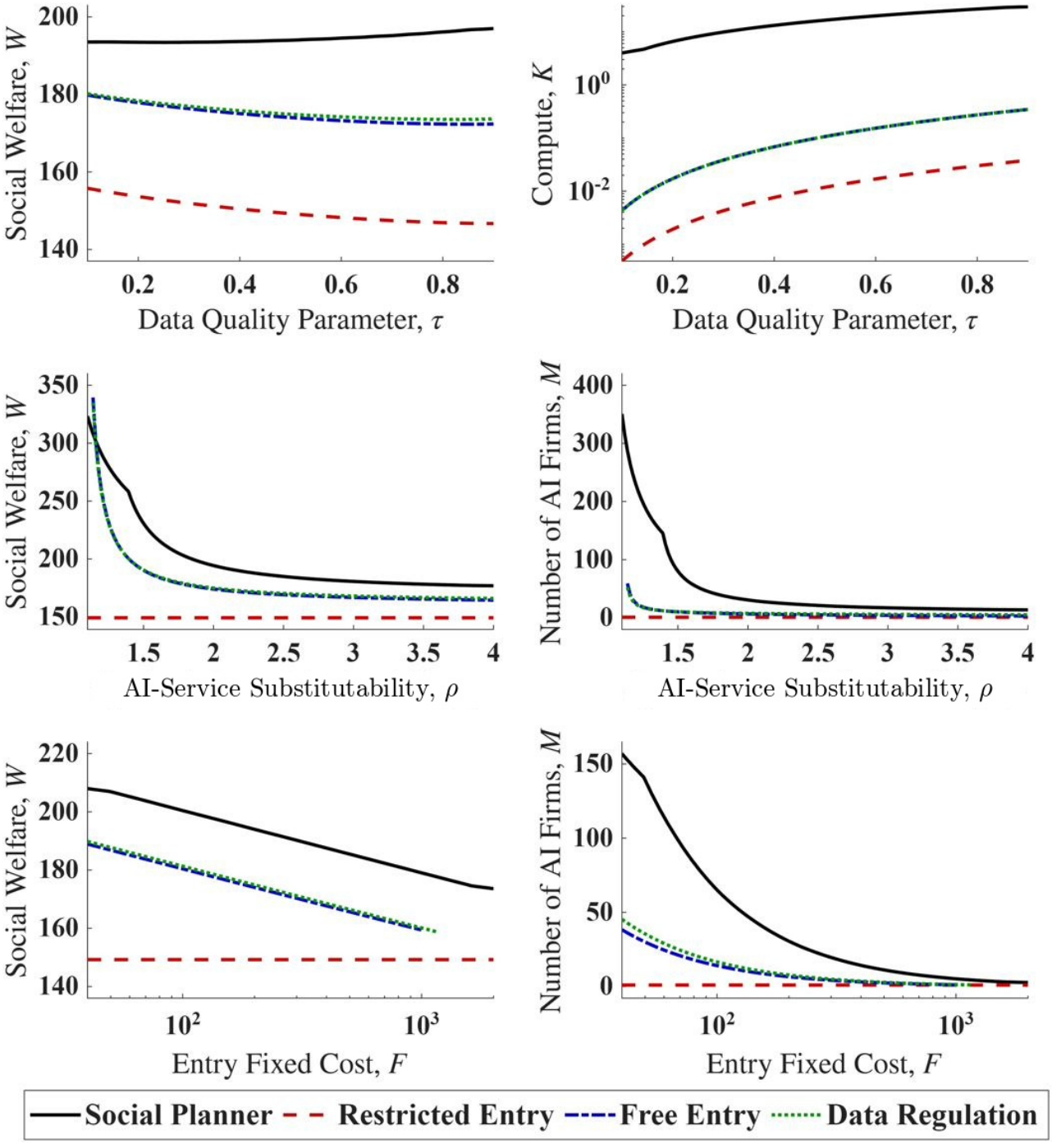


Figure 2. Comparative statics on welfare across regimes.

Notes. Each row shows a single parameter: τ (top), ρ (middle), F (bottom). Left column reports social welfare W; right column reports the mechanism variable (computing power K on log scale for τ; number of AI firms M for ρ and F). Solid black: Social Planner; dashed red: Restricted Entry; dash-dot blue: Free Entry; dotted green: Data Regulation.

Before discussing each parameter in turn, we note that the free-entry and data regulation curves (FE and DR) are truncated in the F panel of Figure 2, where they terminate near $F \approx 1{,}000$; they

remain well-defined throughout the ρ range we consider. The truncation under large F reflects a genuine feature of the model rather than a numerical limitation: when F becomes sufficiently large, the free-entry condition (37) has no interior solution with $M^{fe} \geq 1$, as a high fixed cost drives per-firm profit below the amortized entry cost $(1-\beta)F$, and the free-entry equilibrium ceases to exist. The same mechanism applies to the data regulation regime, but with a slightly higher tolerance because the binding data cap $\bar{d} < d^{fe}$ lowers per-firm data costs and redistributes rents toward AI firms, sustaining entry for marginally higher F.

Data quality parameter τ. The parameter τ governs the sensitivity of data quality to the ratio of real-world inputs (consumer data and computing power) to the accumulated AI output. As τ increases, the quality channel becomes more important, amplifying the productive role of computing power. The top-left panel of Figure 2 shows that welfare under the social optimum rises gently with τ, as the planner fully exploits the quality channel by scaling up computing investment. In contrast, welfare under restricted entry, free entry, and data regulation all decline modestly as τ increases: in these regimes the underinvestment in computing becomes more costly precisely when computing matters more for quality, so that the monopoly distortion is amplified. The decline is most pronounced under restricted entry, where the compounded monopoly distortion is most severe. Consequently, the welfare gap between SP and the competitive equilibria widens as τ rises. The right panel confirms the mechanism: the social planner's computing investment K^{sp} rises sharply with τ (note the log scale), while K^{re}, K^{fe}, and K^{dr} remain two to three orders of magnitude below the social optimum throughout, consistent with the persistent distortion $B^{re} \ll B^{fe} = B^{dr} < B^{sp}$.

Data substitutability ρ. The parameter ρ controls the substitutability across different types of AI services and hence the competitive structure of the AI industry. Smaller ρ implies greater product differentiation, stronger market power for each AI firm, larger love-of-variety gains from additional firms, and, through the larger markup, more resources to cover the entry cost. The middle-left panel of Figure 2 shows that welfare in all regimes that feature endogenous entry (SP, FE, DR) rises steeply as ρ falls toward 1, driven by the explosion of love-of-variety gains. The free-entry and data regulation equilibria remain well-defined throughout the range $\rho \in [1.1, 4]$ under our calibration. The ranking $W^{sp} > W^{dr} > W^{fe} > W^{re}$ holds across the entire range, with the gaps shrinking as ρ rises and competitive outcomes approach the restricted-entry benchmark. The right panel shows the corresponding number of AI firms: M^{sp} grows from roughly 15 at $\rho = 4$ to over 300 as $\rho \to 1$, while M^{fe} and M^{dr} remain much smaller but never collapse to 1; they rise from around 4–6 at $\rho = 4$ to roughly 70 at $\rho \to 1$. These results suggest that the welfare case for encouraging AI firm entry is strongest when different AI firms' services are highly differentiated.

Entry cost F. The bottom row of Figure 2 shows welfare and the number of AI firms as functions of F (log scale). Welfare under all regimes that feature endogenous entry declines in F: a higher entry cost reduces the equilibrium number of firms and shrinks the aggregate love-of-variety gain. The SP curve declines smoothly across the entire range, while the FE and DR curves are truncated near $F \approx 1{,}000$, where per-firm profits fall below the amortized entry cost. Welfare under restricted entry is independent of F (by construction, since $M^{re} = 1$), and appears as a horizontal line. Notably,

$W^{dr} > W^{fe}$ for all F in the feasible range, with the gap largest at low F: when entry is cheap, many AI firms enter under both regimes, and the data cap's indirect entry-encouragement effect (through reduced per-firm data costs) is most potent, allowing M^{dr} to exceed M^{fe} by a larger absolute margin. The right panel confirms this pattern: at $F \approx 30$, M^{sp} exceeds 150 while M^{fe} and M^{dr} are around 40–50, highlighting the substantial entry gap between the social optimum and decentralized equilibrium.

4.3 Policy Experiments

Having characterized the benchmark allocations and comparative statics, we now turn to policy. The analysis in Section 3 identifies two distinct sources of welfare loss in the decentralized equilibria: the underprovision of computing power and the suboptimal supply of consumer data. This subsection investigates two corresponding policy instruments, namely an optimal data cap and a computing-investment subsidy, and then quantifies their welfare impact in consumption-equivalent units. Throughout we retain the benchmark parameterization of Section 4.1.

4.3.1 Optimal data cap

Section 3.4 establishes that a binding cap $\bar{d} < d^{fe}$ can raise welfare above the free-entry level, and Proposition 5 characterizes the welfare-maximizing cap $\bar{d}^*$ as the solution to an implicit first-order condition. Figure 3 plots $W^{dr}(\bar{d})$ over the binding interval; numerically, the optimum is $\bar{d}^* \approx 0.042$, roughly 37% below the unregulated level d^{fe}, and delivers $W^{dr}(\bar{d}^*) = 175.0$, about one welfare unit above $W^{fe} = 174.0$. The welfare function is piecewise: on the binding region it rises from $-\infty$ at $\bar{d} = 0$ (where data quality collapses through the $(\bar{d}L)^{b\tau}$ channel) to W^{fe} at $\bar{d} = d^{fe}$, peaks at an interior point strictly below d^{fe}, and is flat at the FE level beyond. Its interior maximum confirms that the sufficient condition identified in Proposition 5 holds at the benchmark calibration.

Economically, Proposition 5 identifies four channels through which a binding cap affects welfare, two of which dominate numerically at our benchmark calibration, and with very different signs. The first is a direct privacy benefit: because the consumer chooses d^{fe} by balancing only the private market price p_d^{fe} against her marginal privacy cost, without internalizing that her data supply depresses AI-firm entry and the associated love-of-variety gains, the privately-chosen d^{fe} exceeds the welfare-maximizing level. Tightening the cap corrects this entry-externality-induced over-supply at the cost of somewhat lower data quality. The second channel operates directly through AI-firm entry: a tighter cap reduces per-firm consumer-data expenditures, relaxing the free-entry condition (41) and drawing in additional AI firms, so that the equilibrium number rises from $M^{fe} = 6.6$ to $M^{dr,*} = 8.2$. The resulting love-of-variety gain more than offsets the direct data-quality loss, so aggregate output at $\bar{d}^*$ sits just above its unregulated level. The decomposition in Section 4.3.3 shows that the entry-induced consumption gain is in fact the larger of the two dominant channels, with the privacy reduction contributing the smaller share of the total welfare improvement.

Two features of Figure 3 merit comment. The welfare function is remarkably flat over a wide interior region, reflecting a near-cancellation of competing forces: tightening the cap trades output for privacy, and over a broad range the two effects nearly offset. From a policy perspective, this

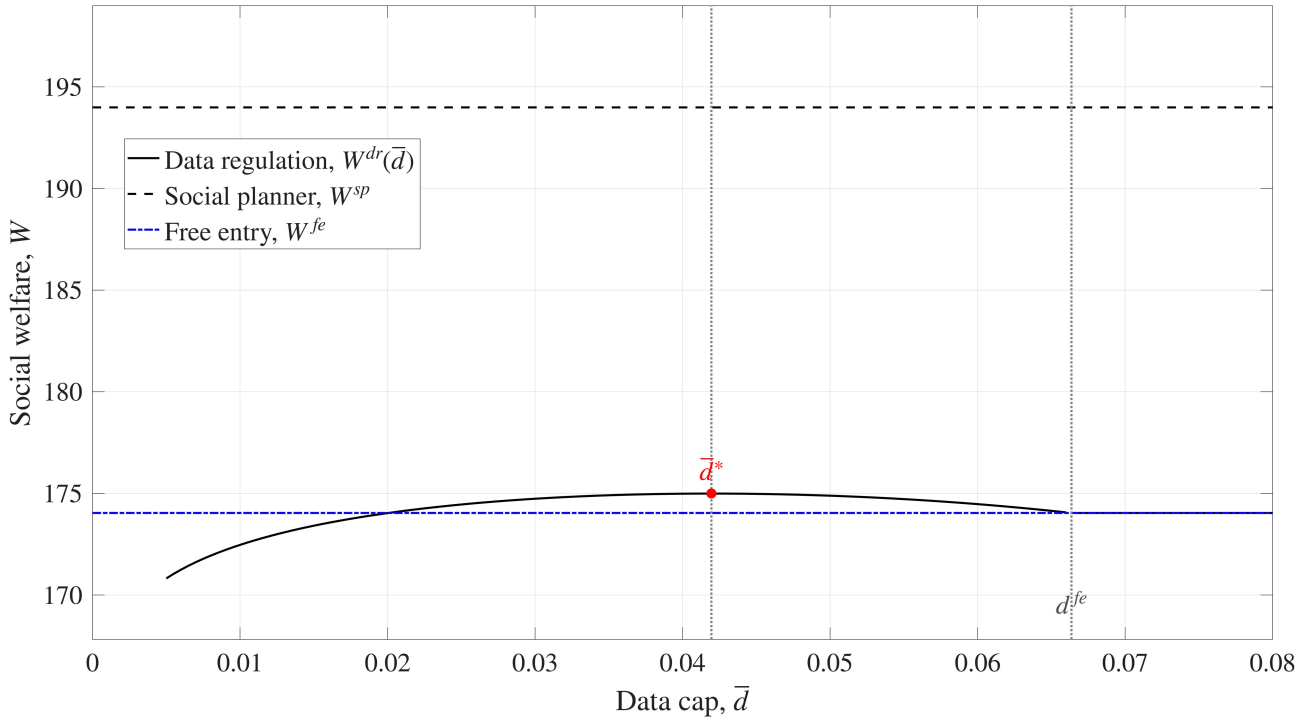


Figure 3. Social welfare under data regulation as a function of the cap $\bar{d}$.

Notes. The solid black curve shows $W^{dr}(\bar{d})$; the horizontal dashed lines mark W^{sp} and W^{fe}; the vertical dotted line at $\bar{d} = d^{fe}$ marks the point where the cap ceases to bind. The red dot marks the welfare-maximizing cap $\bar{d}^*$.

flatness is reassuring, since the precise level of the cap matters far less than the decision to impose one at all, but it also cautions against reading too much into the specific value of $\bar{d}^*$, which is sensitive to the calibration of the privacy cost function. At the same time, the absolute welfare gain of data regulation is modest: it closes less than 5% of the gap between the free-entry equilibrium and the social optimum ($W^{sp} - W^{fe} \approx 20$). Most of the remaining welfare loss stems from sources that a data cap does not address, chiefly the underprovision of computing power documented in Table 4, since $B^{dr} = B^{fe}$ by construction. We take up this distortion directly in the next subsection through a computing-investment subsidy.

4.3.2 Compute subsidy

The analysis in Section 3 shows that the three compounded monopoly distortions drive the computing-investment share B^{fe} far below the social optimum B^{sp}. A natural policy response is a subsidy paid directly to the computing-power provider, financed by a lump-sum tax on consumers. We parameterize the subsidy as a rate $\tau_s \in [0, 1)$ that scales down the provider's effective marginal cost of producing K; in steady state this raises the computing-investment share to

$$B^{fe}(\tau_s) = \frac{B^{fe}}{1 - \tau_s},$$

and likewise for the data-regulation regime (with $B^{dr}(\tau_s) = B^{fe}(\tau_s)$). Because the tax is lump-sum, the resource constraint $cL = [1 - B(\tau_s)]N^{\sigma/(\sigma-1)}Y_i$ continues to hold unchanged in form; the financing only affects how consumption is split between households and the government's investment contribution.[10] We consider two scenarios: free entry combined with an optimal subsidy, and data regulation at $\bar{d}^*$ combined with an optimal subsidy, searching in each case for the welfare-maximizing τ_s^*.

Figure 4 plots $W^{fe}(\tau_s)$ and $W^{dr}(\tau_s; \bar{d}^*)$ against the subsidy rate, and Table 5 reports the steady state at each regime's optimum. The subsidy's direct effect is mechanical: a lower effective cost of computing raises K, and with it the data quality $Q = [(dL)^b(K/M)^{1-b}/D]^\tau$ and output. Two general-equilibrium feedback channels amplify this direct effect. First, higher data quality increases AI firms' revenue, loosening the free-entry condition and drawing in additional firms, so the subsidy indirectly stimulates entry and delivers a love-of-variety gain. Second, higher output raises wages and consumer-data prices, supporting a more productive allocation throughout the economy. Working against these gains is the resource cost: the subsidy must be financed, and a subsidy large enough to push $B(\tau_s)$ substantially above the optimal share B^{sp} begins to over-invest the economy at the expense of consumption. The optimum τ_s^* balances these forces, and in our calibration sits close to the rate that aligns $B(\tau_s)$ with B^{sp}.

Table 5. Steady State with Optimal Compute Subsidy

Variable	Free Entry	FE + τ_s^*	DR($\bar{d}^*$) + τ_s^*	Social Planner
Subsidy rate, τ_s	0	0.994	0.994	n/a
Consumer data, d	0.066	0.067	0.042	0.110
Compute, K	0.106	18.02	18.08	16.526
AI firms, M	6.6	7.1	8.8	30.2
Consumption, c	5.725	6.126	6.169	7.043
Social welfare, W	174.0	180.8	181.8	194.0

At the benchmark calibration the optimum subsidy is extremely large, with $\tau_s^* \approx 0.994$ in both scenarios, and Table 5 should be read as a statement about the severity of the underlying distortion rather than as an operational policy recommendation. Because the three compounded markups drive B^{fe} to less than one percent of B^{sp}, restoring anything close to the efficient level of computing investment requires the subsidy to shoulder almost the full cost of compute production. Even then, the optimum falls well short of the social planner: the subsidy pushes K above K^{sp} and restores part of the quality gap, but the entry gap identified in Section 4.2.1 remains unaddressed, and aggregate entry ($M^{fe+s} \approx 7$) stays far below $M^{sp} \approx 30$. The combined policy of data regulation at $\bar{d}^*$ plus an optimal subsidy delivers a slightly higher welfare than the subsidy alone, closing roughly 39% versus

[10] The lump-sum financing assumption is a benchmark choice that isolates the marginal incentive effect of the subsidy from the deadweight loss of financing. Under this benchmark, the optimal τ_s^* reported in Table 5 is best read as an upper bound quantifying the severity of the compounded monopoly distortion, rather than as an operational policy rate. Introducing distortionary taxation would scale the optimum substantially downward, because the shadow cost of public funds would enter the planner's first-order condition for τ_s; the qualitative ranking that the compute subsidy closes the bulk of the FE-to-SP welfare gap while the data cap closes only a small share is invariant to this refinement.

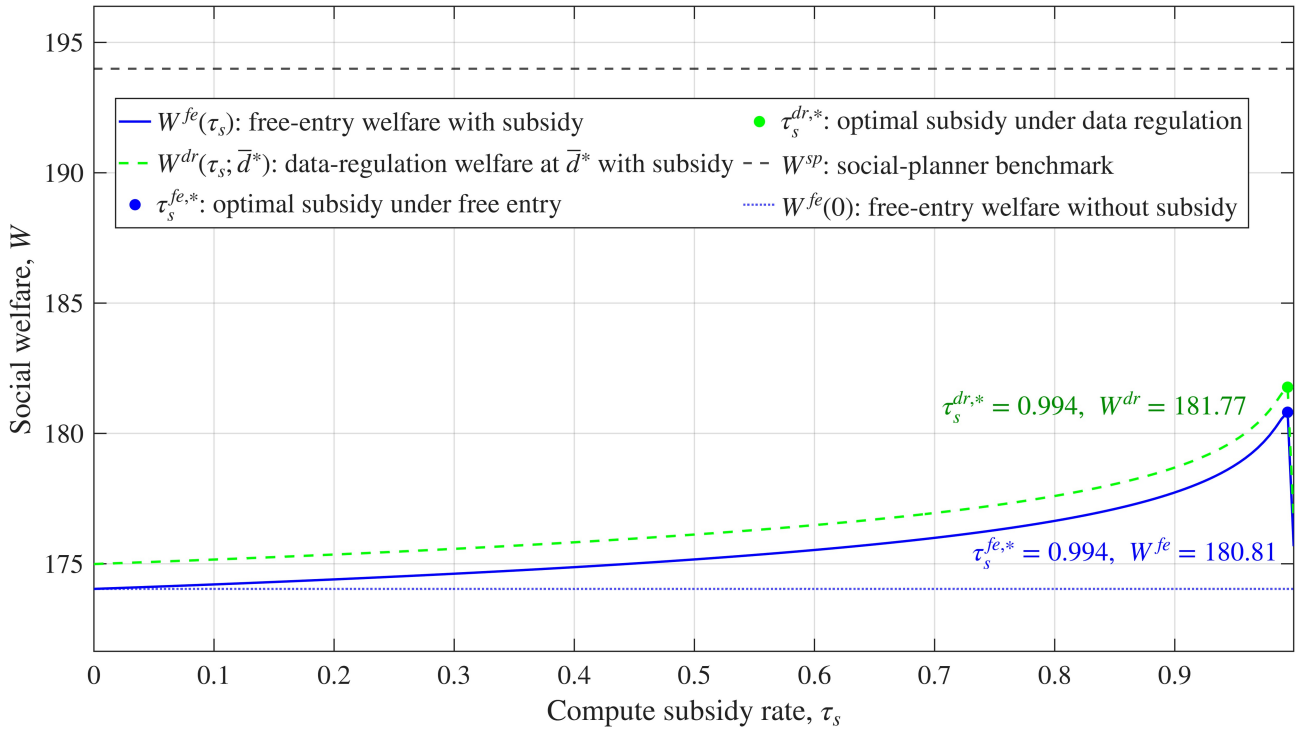


Figure 4. Social welfare as a function of the compute subsidy rate τ_s.

Notes. The solid blue curve shows $W^{fe}(\tau_s)$; the dashed green curve shows $W^{dr}(\tau_s;\bar{d}^*)$. Horizontal reference lines mark W^{sp}, $W^{fe}(0)$, and W^{re}. Filled markers identify the welfare-maximizing subsidies $\tau_s^{fe,*}$ and $\tau_s^{dr,*}$.

34% of the FE-to-SP gap; the fact that the optimal subsidy rate is essentially identical under both scenarios is an early indication that the two instruments act on different margins, a point we take up more carefully in Section 4.3.3.

The consumption column in Table 5 is worth reading alongside the K column. At the optimum, consumption rises by about seven percent above its unregulated level ($c = 6.13$ vs. $c^{fe} = 5.72$), and output rises by a comparable amount (from $Y = 572$ to $Y \approx 615$ under FE+τ_s^*). Because utility is logarithmic in consumption and periods are discounted at $\beta = 0.99$, each percentage-point increase in steady-state consumption translates into about one welfare unit, so the modest-looking consumption gain accounts for essentially the entire welfare improvement of roughly seven utility units. As noted earlier, the extreme value of τ_s^* should be read as a measure of how severe the underlying compute distortion is under our benchmark calibration, not as a recommendation for the size of an actual subsidy. In Section 4.3.3 we translate the welfare gaps reported here into consumption-equivalent units to make comparisons across policies more interpretable.

4.3.3 Consumption-equivalent welfare

The welfare levels reported in the preceding subsections are measured in utility units and are difficult to interpret directly. To facilitate comparisons across regimes and policies, we follow the standard practice in the macroeconomics literature and express all welfare differences in consumption-equivalent terms. Given two steady states a and b, the consumption-equivalent variation $\lambda_{a\to b}$ is

defined as the permanent percentage increase in consumption that, applied to every period of allocation a, would equate its lifetime welfare to that of b. Because utility is separable with logarithmic consumption, the definition admits the closed form

$$\lambda_{a\to b} = \exp\big[(1-\beta)(W^b - W^a)\big] - 1. \tag{45}$$

A positive $\lambda_{a\to b}$ indicates that allocation b is welfare-superior, with the magnitude measuring how much extra consumption the consumer in allocation a would need to be compensated into indifference.

Equation (45) can be further decomposed to isolate the two sources of welfare differences in our model, namely consumption and privacy. Writing $W^b - W^a$ as the sum of a consumption component $\Delta W_c = (1-\beta)^{-1}(\ln c^b - \ln c^a)$ and a privacy component $\Delta W_\phi = -(1-\beta)^{-1}\phi[(d^b)^2 - (d^a)^2]$, we define

$$\lambda_c = \frac{c^b}{c^a} - 1, \qquad \lambda_\phi = \exp\big[-\phi\,((d^b)^2 - (d^a)^2)\big] - 1,$$

which satisfy the exact multiplicative identity $(1+\lambda_{a\to b}) = (1+\lambda_c)(1+\lambda_\phi)$. To a close approximation for the values we encounter, this reduces to the additive decomposition $\lambda \approx \lambda_c + \lambda_\phi$. The term λ_c captures the welfare change due to different levels of steady-state consumption, while λ_ϕ captures the welfare change due to different privacy costs. Table 6 reports the three quantities for all regimes and policy experiments considered so far, taking the free-entry equilibrium as the reference allocation.

Table 6. Consumption-Equivalent Welfare (Benchmark: Free Entry)

Scenario	W	λ (%)	λ_c (%)	λ_ϕ (%)
Restricted entry (monopoly AI firm)	149.24	−21.96	−22.27	+0.39
Free entry, no intervention (benchmark)	174.04	0.00	0.00	0.00
Data regulation, fixed cap $\bar{d} = 0.05$	174.89	+0.85	+0.66	+0.19
Data regulation, welfare-maximizing $\bar{d}^*$	174.99	+0.96	+0.69	+0.26
Free entry with optimal compute subsidy	180.81	+7.01	+7.02	−0.003
Optimal data cap and optimal compute subsidy	181.77	+8.05	+7.76	+0.26
Social planner (first-best)	193.99	+22.09	+23.03	−0.76

Notes. W denotes the utility-level social welfare; λ, λ_c, and λ_ϕ are the total, consumption-component, and privacy-component consumption-equivalent welfare changes relative to the free-entry benchmark, expressed in percent.

The consumption-equivalent metric reveals the magnitude of the welfare stakes far more clearly than the raw utility levels. The welfare loss from restricted entry relative to free entry corresponds to a permanent 22% reduction in consumption, a cost of roughly the same order as the welfare gap between free entry and the social optimum, which also corresponds to 22%. Put differently, moving from $M = 1$ to $M = M^{fe}$ yields a welfare improvement roughly equal in magnitude to that of moving all the way from M^{fe} to the first-best allocation. Data regulation alone delivers less than one percent of consumption-equivalent welfare, while an optimal compute subsidy, even financed by lump-sum taxation, delivers about seven percent. Combining the two instruments produces a welfare gain of

about eight percent, still leaving approximately two-thirds of the full distortion unresolved.

The decomposition into consumption and privacy components clarifies how the two policy instruments operate. The compute subsidy's welfare gain is essentially a consumption effect: because the subsidy does not enter the consumer's first-order condition for data supply directly, equilibrium d is virtually unchanged (the minor general-equilibrium feedback through wages and consumer-data prices moves d by only about a percentage point relative to free entry), and λ_ϕ is negligible. The data cap, by contrast, generates welfare gains through both channels, with the consumption component (+0.69%) being roughly three times the privacy component (+0.26%); the consumption component arises because tightening the cap relaxes the free-entry condition and stimulates entry, boosting the love-of-variety gain and thereby output, even though the rationale for regulation is fundamentally privacy-related. Most strikingly, the total welfare effect of combining both policies (+8.05%) is almost exactly the sum of their individual effects (+7.01% and +0.96%). The two instruments therefore operate on largely orthogonal distortions: the subsidy corrects the computing-share wedge, the cap addresses the entry-externality-induced over-supply of consumer data, and the two policies can be stacked without substantial interaction.

A final observation concerns the social planner. The $\lambda_\phi = -0.76\%$ in the last row is negative: the planner's allocation imposes higher privacy costs on consumers than the free-entry equilibrium does, because the optimal level of consumer data supply ($d^{sp} = 0.11$) exceeds the free-entry level ($d^{fe} = 0.066$). This reflects the fact that the planner internalizes the full social return to data in production, including the love-of-variety spillovers across AI firms, which the private consumer does not. Seen through the consumption-equivalent lens, the planner is willing to accept a small increase in privacy cost in exchange for a much larger boost in consumption. A data cap binds consumer data below d^{fe} and therefore moves the economy in the opposite direction of the first-best on the privacy dimension; its welfare gain comes entirely from the indirect entry channel it triggers. This contrast highlights a basic tension in the policy space: the data cap corrects the welfare-relevant over-supply of consumer data that arises from the entry externality, but it does so by tightening privacy rather than by raising the private return to data, and in this narrow sense the cap moves in the opposite direction from the first-best.

5. Transition Dynamics

Sections 4.2 and 4.3 compare steady-state allocations across regimes and evaluate policy interventions by contrasting the pre- and post-policy long-run equilibria. Such steady-state comparisons implicitly assume that the economy can transit instantaneously between equilibria, an abstraction that is convenient for characterizing the direction and magnitude of welfare changes but may obscure economically important short-run dynamics. In particular, because computing capital and AI services are stock variables that evolve through capital- and AI-service-accumulation equations, any policy intervention affecting their long-run levels must operate through a gradual transition path, during which consumption, output, and welfare can differ substantially from either the initial or the final steady state.

This section analyzes the transition paths of the economy under three policy experiments. Section 5.1 sets out the equilibrium concept used throughout. Section 5.2 presents the three experiments: the relaxation of entry restrictions from the restricted-entry regime to free entry, which serves as the benchmark because the associated welfare gap is the largest in magnitude and the transition involves substantial adjustments in every state variable; the imposition of the optimal data cap from Section 4.3.1; and the introduction of the optimal compute subsidy from Section 4.3.2. Section 5.3 distills the common lessons.

5.1 Methodology

The full-information rational-expectations version of the model would require agents to form forward-looking expectations of the entire future path of prices and state variables, yielding a boundary-value problem in which shadow prices, stock variables, and flow variables for every period are jointly determined as a single large nonlinear system. For a horizon of 200 quarters with eight unknowns per period (two stocks, four flow variables, and two Lagrange multipliers on the AI-service- and compute-accumulation constraints), the resulting system has roughly sixteen hundred equations to be solved simultaneously. Such a computation is technically demanding, sensitive to the quality of the initial guess, and difficult to reconcile transparently with the steady-state analysis of Section 3. We therefore adopt a semi-dynamic equilibrium concept that retains the core intertemporal structure, namely stock accumulation, while closing the model in each period with static decision rules for flow variables.

Formally, the two stock variables D_t and K_t evolve according to their accumulation equations, while flow variables that lack accumulation dynamics are assumed to jump instantaneously to their new-regime steady-state values at $t = 0$. This treatment applies to labor allocations $(L_{A,t}, L_{K,t})$ and to consumer data supply d_t: these are static decisions in the model and, absent explicit adjustment costs or habit formation in the privacy cost, agents have no reason to transition gradually. In particular, the flat trajectory of d_t visible in Figures 6 and 7 after the policy shock reflects this modeling choice rather than a numerical artifact; introducing a privacy habit would generate smoother d_t transitions but is beyond the scope of the present analysis. Given period-t stocks and the jumped-in flow values, the number of AI firms M_t is determined by the static free-entry condition $\pi_t = (1 - \beta)F$. Because the output technology is a power function of M_t, this condition admits the closed-form solution $M_t = [K_1 \cdot N^{\sigma/(\sigma-1)} Y_0(D_t, K_t)/((1 - \beta)F)]^{1/(1-\omega)}$, derived in Appendix A.6. This avoids the need to invoke a nonlinear solver each period and thereby ensures the numerical reliability and reproducibility of the transition paths. The static form of the free-entry condition is the operational content of the myopic-expectations assumption: under perfect foresight, entering firms would instead require the present discounted value of their future profits to equal the entry cost, a forward-looking condition that couples period-t entry to the entire future path of the economy. This is consistent with the behavioral macroeconomics literature on myopic or cognitive-discounting expectations (Gabaix, 2020; Angeletos and Huo, 2021), which shows that such simplified expectational structures can be viewed as the rational response of agents facing complex or unfamiliar policy environments, and it has been used in dynamic macroeconomic models with structural change (Chen et al., 2008). Under

rational expectations, forward-looking investment and entry would shift K_t and M_t closer to their new steady-state values earlier in the transition, deepening short-run consumption adjustments but leaving the qualitative ranking of policies intact.

Given initial stocks (D_0, K_0) inherited from the old steady state, the two stock variables evolve according to the AI-service and computing-power accumulation Equations (4) and (6), with the period's flows $\tilde{D}_t$ and I_t determined by the static equilibrium conditions. By construction, the path converges to the new-regime steady state as $t \to \infty$. The simulation horizon is set at $T = 200$ quarters (50 years), long enough for all three experiments to converge essentially to their new steady states and for cumulative discounted welfare to be computed with negligible truncation error. For visual clarity, the transition figures below display only the first 60 quarters (15 years): convergence is economically complete by that horizon in all three experiments, with every variable within a small fraction of its new steady-state level, so the subsequent 140 quarters contribute no additional dynamic content worth plotting.

5.2 Transition Paths

We now apply the semi-dynamic equilibrium concept of Section 5.1 to the three policy experiments of Section 4.3. Each experiment starts from the relevant initial steady state, imposes the new policy at $t = 0$, and propagates the economy forward to the new steady state. The figures plot six variables: c_t, $\bar{d}_t$, the stocks D_t and K_t, the number of AI firms M_t, and cumulative welfare ΔW_t relative to staying in the initial regime. The qualitative features, namely impact jumps, overshoots, and consumption dips, are what carry the economic content.

5.2.1 Entry Deregulation

Consider an economy initially in the restricted-entry steady state ($M = 1$), in which a single monopolistic AI firm serves the entire market for AI services. At $t = 0$ the government removes the barrier to entry, and additional AI firms begin operating until the free-entry condition is satisfied. This experiment captures the dynamic counterpart of the welfare comparison W^{re} vs W^{fe}, the single largest welfare gap in our analysis. Figure 5 plots the transition paths of six key variables.

Three features of the transition deserve comment. The number of AI firms M_t jumps discretely at $t = 0$ from 1 to a value close to its FE steady-state level. This reflects our semi-dynamic assumption that the static free-entry condition must hold each period: at $t = 0$, although the stocks (D_0, K_0) remain at their RE levels, nothing prevents new firms from entering immediately, so M_t pins itself to whatever level equates current-period profit to the amortized entry cost. M is therefore effectively a jump variable in our analysis. A richer model with entry frictions such as learning-by-doing, bureaucratic lags, or limited investor attention would smooth this adjustment, and the quick initial rise in M_t should be read as the dynamic counterpart of the RE-to-FE steady-state comparison under an assumption that AI firms can enter as soon as the policy is announced.

Consumption c_t jumps up at $t = 0$ and then rises more gradually toward its new steady state. The immediate jump reflects the impact effect of moving from monopoly to monopolistic competition: the

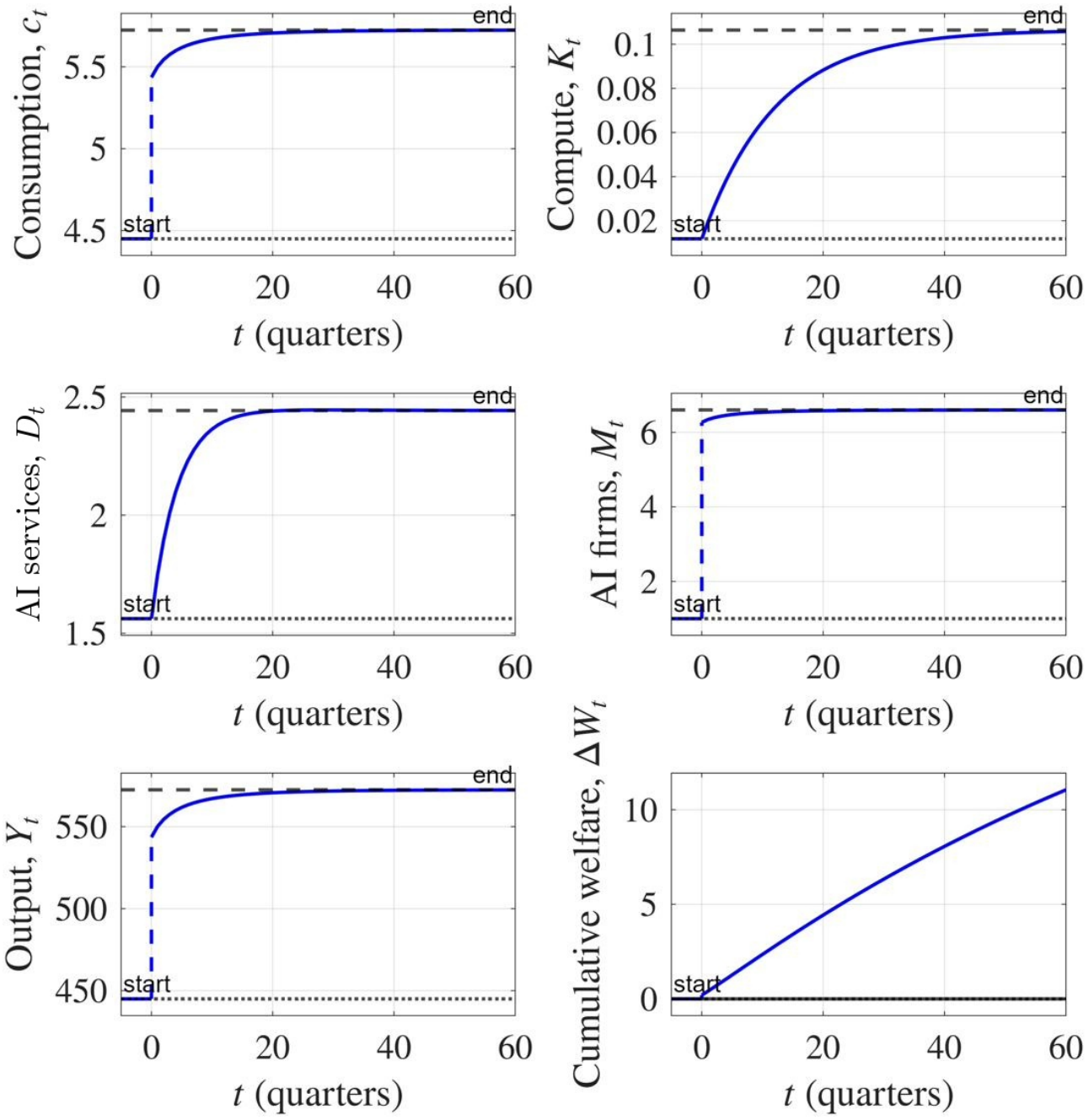


Figure 5. Transition from restricted entry to free entry.

Notes. Initial stocks (D_0, K_0) are inherited from the RE steady state at $t = 0$; the free-entry regime takes effect immediately. The horizontal dotted line in each panel marks the starting (RE) steady-state level; the dashed line marks the ending (FE) steady-state level. The final panel reports cumulative discounted welfare relative to a counterfactual in which the economy remains in RE indefinitely.

sharp rise in M delivers love-of-variety gains in production that feed through to consumption within the period. The subsequent gradual rise traces the slow accumulation of K and D. Since computing depreciates much more slowly than AI services, with a quarterly depreciation rate of about a twelfth for compute versus a fifth for AI services, the two stocks converge at very different speeds, with D

approaching its new steady state several times faster than K. Since both stocks enter the data-quality function, but with computing contributing through a much larger accumulation channel, the overall speed of the transition is governed primarily by computing-capital accumulation.

Cumulative welfare ΔW_t, defined as the difference between actual discounted utility along the transition and the counterfactual of remaining in RE forever, is positive from the very first period and rises monotonically. There is no J-curve in this experiment, which reflects a key economic feature of entry deregulation: the policy does not require any aggregate resource sacrifice. New AI firms enter by bearing the fixed cost F out of their own expected profits; the household sector is not taxed and consumption need not fall to finance the adjustment. Deregulation is therefore a rare case in which immediate gains coincide with ongoing accumulation gains, and the welfare improvement reported by the steady-state comparison in Table 6 is essentially realized within the first decade and a half.

5.2.2 Imposing the Optimal Data Cap

The second experiment begins with the economy in the free-entry steady state and imposes the welfare-maximizing data cap $\bar{d}^*$ derived in Section 4.3.1 at $t = 0$. Figure 6 displays the resulting transition paths.

The dynamics are driven by a single force: the exogenous tightening of consumer data at $t = 0$. The stock of AI services D_t then begins a monotonic decline toward its new lower steady state, reflecting the reduction in raw data input into the AI-service generation function. Computing capital barely moves, since the policy does not touch the computing-investment margin, so K_t changes only through the indirect general-equilibrium effect of M_t and Y_t. The number of AI firms jumps upward on impact and then drifts slightly as D_t stabilizes, the dynamic counterpart of the entry-stimulating comparative-statics effect analyzed in Section 4.3.1.

Consumption shows a modest but economically interesting overshoot. It jumps up at $t = 0$ above its new steady-state level, then gradually declines toward the new long-run value. The mechanism is straightforward: on impact, the economy is still operating with the full legacy stock of AI services D^{fe} but is freed from the high privacy cost of d^{fe}; as the data cap causes D_t to depreciate toward its lower steady-state level, the output-side loss from fewer AI services erodes the initial consumption gain. This overshoot reflects the asymmetric timing of two effects: privacy cost adjusts instantly when the cap is imposed, while the output consequences of less data accumulate only gradually. Cumulative welfare ΔW_t is nevertheless positive from the first period and rises monotonically, and the full steady-state welfare gain is realized well within the simulation horizon. The transition imposes no welfare loss relative to remaining in FE, and the policy is therefore attractive from a timing perspective as well as a long-run one.

5.2.3 Introducing the Optimal Compute Subsidy

The third experiment introduces the welfare-maximizing compute subsidy from an initial FE steady state. Unlike data regulation, the compute subsidy asks the economy to undertake a fundamental change in its resource-allocation composition: a much larger share of output must be devoted to

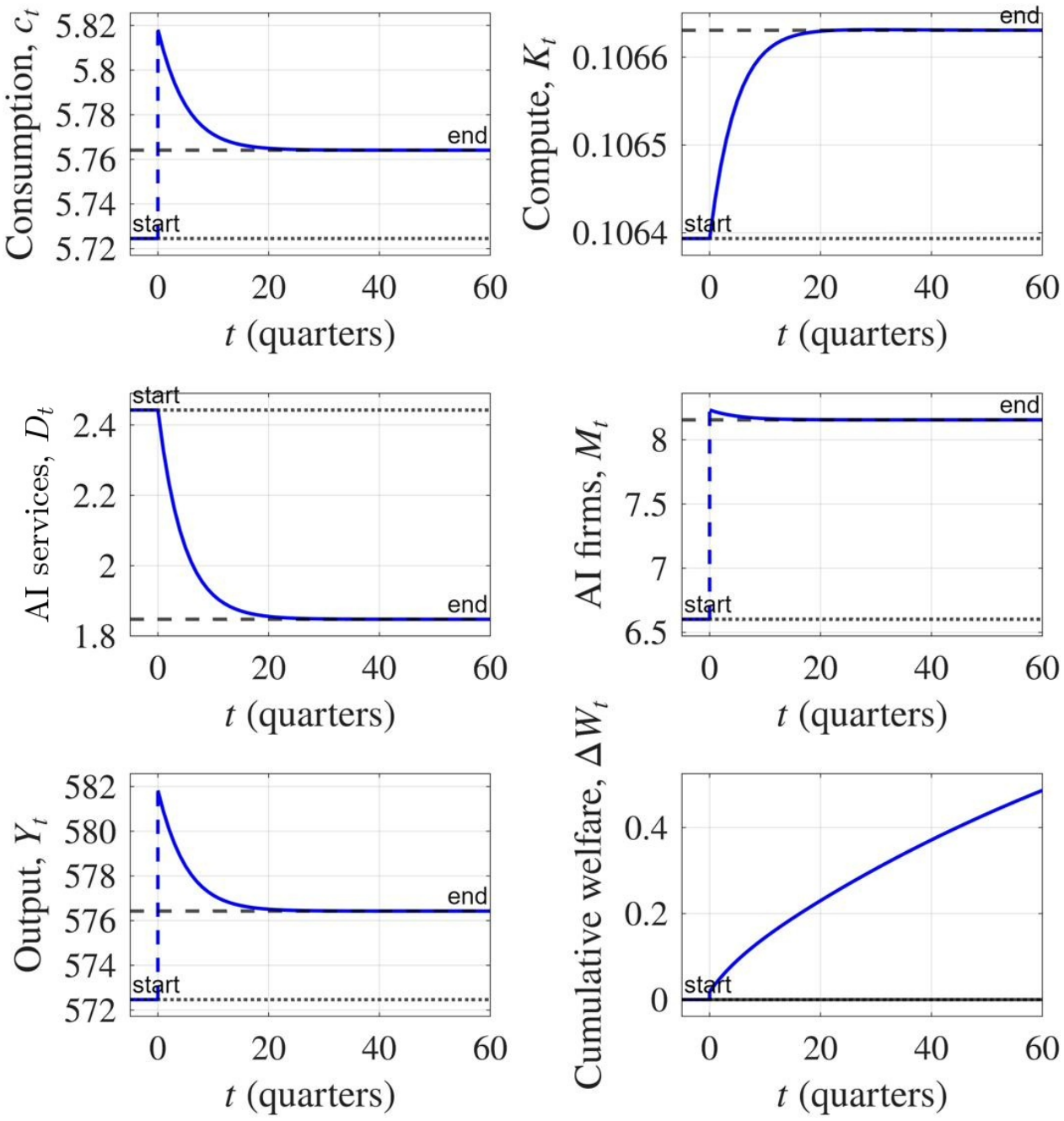


Figure 6. Transition from the free-entry equilibrium to the data-regulation equilibrium at the optimal cap $\bar{d}^*$.

Notes. Starting stocks (D_0, K_0) are inherited from the FE steady state. The horizontal dotted line in each panel marks the starting (FE) steady-state level; the dashed line marks the ending (DR at $\bar{d}^*$) steady-state level.

investment in order to build computing capital from the low private-equilibrium level to the much higher subsidized-equilibrium level. Figure 7 displays the transition paths.

The distinctive feature of the compute-subsidy transition is the magnitude of the required adjustment in computing capital. The new steady state calls for a far higher level of K than the initial FE equilibrium, and the gap can be closed only through gradual accumulation bounded by physical de-

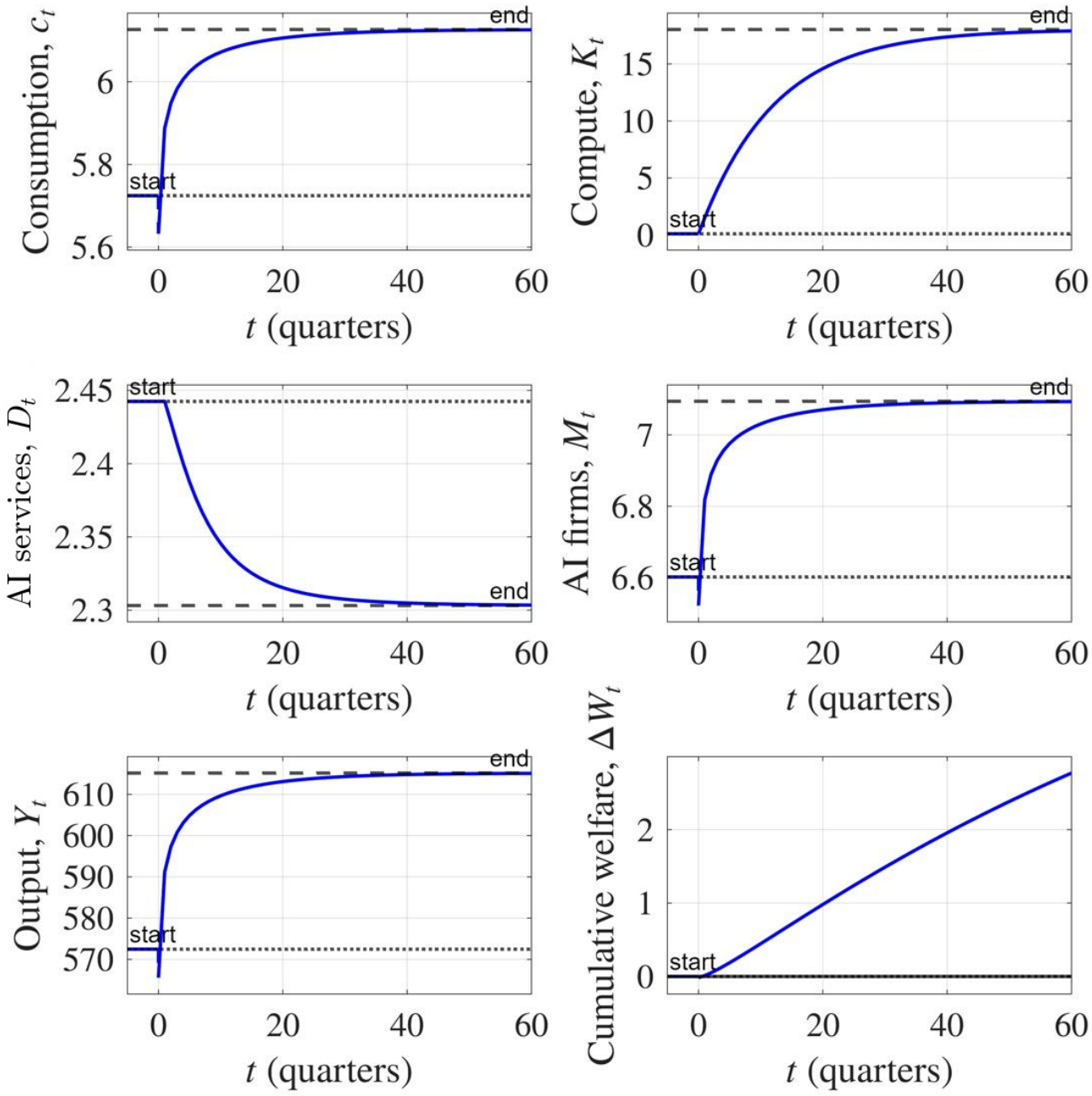


Figure 7. Transition from the free-entry equilibrium to the free-entry equilibrium with the welfare-maximizing compute subsidy.

Notes. Starting stocks (D_0, K_0) are inherited from the FE steady state. The horizontal dotted line in each panel marks the starting (FE) steady-state level; the dashed line marks the ending (FE+τ_s^*) steady-state level.

preciation. The stock of AI services D_t adjusts only modestly, since the subsidy does not directly target the AI-service margin and the general-equilibrium feedback through wages and consumer-data prices is small; the dynamics of D_t are a second-order effect relative to the computing accumulation. The number of AI firms rises gradually as computing capital grows and per-firm productivity improves, so M_t traces a smooth upward path rather than the discrete jumps seen in the other experiments.

The consumption response is the most economically revealing part of the figure. On impact, consumption drops slightly below its starting level: the subsidy immediately increases the share of output devoted to investment, and this reallocation necessarily comes at the expense of consumption in the same period. This is the J-curve that one expects from any policy that induces a large capital-accumulation response. In our simulation the dip is modest and cumulative welfare turns positive within the first year, but the quantitative depth of the dip should not be taken literally. Our semi-dynamic equilibrium concept has investment each period tracking current output rather than anticipating future productivity; a fully forward-looking computing-investment decision would bring forward investment more aggressively, producing a larger initial consumption decline and a slower recovery. What Figure 7 establishes robustly is the qualitative structure of the transition: unlike the entry-deregulation and data-regulation experiments, the compute subsidy involves a genuine intertemporal trade-off between current consumption and future productivity. The short-run cost is present even if its precise magnitude depends on expectational assumptions.

5.3 Policy Implications

The three policy experiments yield a sharp sequencing logic for AI-sector policy, grounded in two distinct rankings of the instruments. The first is a ranking on welfare gain: the compute subsidy, by addressing the compounded monopoly distortion at its source, delivers the largest consumption-equivalent improvement; entry deregulation, by shifting the economy from $M = 1$ to monopolistic competition, delivers the second largest; the optimal data cap, by relaxing the free-entry margin indirectly, delivers the smallest. The second is a ranking on timing: entry deregulation and the data cap reallocate resources that already exist and therefore improve welfare from the first period onward, while the compute subsidy requires households to consume less today in exchange for productivity gains that accrue only as the computing stock accumulates over roughly a decade. Crucially, these two rankings are not aligned: the instrument with the largest long-run gain also carries the longest adjustment horizon.

A corollary of this ranking is that the familiar privacy-versus-innovation framing of data regulation is misleading. Debates around the EU's GDPR and California's CCPA (Capasso, 2025; Bartholdy, 2026) typically assume that tighter data rules must cost the AI sector, but in our calibration the opposite can hold in a relevant range: a moderate binding cap reduces each firm's total data bill, relaxes the free-entry margin, and draws in additional firms, so that the love-of-variety gain from a more crowded AI sector offsets most of the direct loss from lower data quality. Privacy regulation and a vibrant AI sector can be complements rather than substitutes, and the policy question is therefore where to set the cap, not whether to have one.

The implementation order follows directly from combining the two rankings. Allocative instruments should be deployed first because they cost nothing on impact, stimulate AI-firm entry, and raise the private return to subsequent computing investment; a compute subsidy should then be calibrated to the residual gap rather than to the full FE-to-SP distance, since part of that distance is absorbed once the allocative margins have been corrected. The actual policy packages now being deployed are broadly consistent with this logic. China's "Data Twenty Articles," which build a data-factor

market alongside targeted public compute investment, and the UK's 2025 AI Opportunities Action Plan, which pairs regulatory streamlining with public compute capacity, combine the allocative and compute margins explicitly. Other large packages, such as the U.S. CHIPS and Science Act and the EU Chips Act, concentrate fiscal weight on the compute margin; our framework implies that their realized welfare gains will depend partly on how quickly complementary allocative conditions evolve alongside them.

6. Conclusion

This paper has developed a dynamic general-equilibrium framework for an AI-driven economy in which real-world data is increasingly scarce, synthetic data has become a primary training input, and computing power acts as the principal mechanism preserving the statistical fidelity of the synthetic-data pipeline. Households, monopolistically competitive generative-AI firms, a compute provider, final-goods producers, and monopolistic data intermediaries interact through a reduced-form data-quality function that formalizes the model-collapse phenomenon documented in the machine-learning literature. Our central result is that computing power is systematically underprovided in every decentralized regime through a three-stacked-wedge monopoly distortion, making the computing-infrastructure margin the first-order source of welfare loss in the AI economy. Two corollaries follow: private entry into the AI sector falls short of the social optimum because of an uninternalized love-of-variety externality, and a binding consumer-data cap can raise welfare above the unregulated free-entry level by indirectly relaxing the free-entry margin. A calibrated analysis confirms this ranking quantitatively: the data cap closes only a small share of the welfare gap, while an optimal computing-investment subsidy closes most of the remaining gap, though its gains accrue gradually through capital accumulation rather than as a one-time reallocation.

Several simplifications in the framework point to productive directions for future research. The data-quality function is written in reduced form, and providing it with explicit micro-foundations tied to measurable features of frontier models would sharpen both the positive and normative content of the analysis. The symmetric-firms and myopic-expectations assumptions similarly abstract from dimensions that matter for market structure and for transition dynamics under major policy changes. Three extensions appear particularly fruitful: opening the economy to cross-border flows of data and compute, so that the framework can speak to compute export controls and data-localization debates; introducing household and labor-market heterogeneity, so that the distributional consequences of AI-industry policies can be quantified; and empirically validating the central mechanism, namely whether additional computing investment is reflected in measurably higher AI-service quality and downstream productivity. We view the present framework as a first step toward a tractable macroeconomic account of the interaction between computing power and data quality in the age of generative AI.

References

Acemoglu, D. (2025). The Simple Macroeconomics of AI. *Economic Policy 40*(121), 13–58.

Acemoglu, D. and T. Lensman (2024). Regulating Transformative Technologies. *American Economic Review: Insights 6*(3), 359–376.

Angeletos, G.-M. and Z. Huo (2021). Myopia and Anchoring. *American Economic Review 111*(4), 1166–1200.

Bartholdy, M. (2026). Positioning Synthetic Data under EU Data Protection Law. *Computer Law & Security Review 61*, 106310.

Capasso, M. (2025). Synthetic Data as Meaningful Data: On Responsibility in Data Ecosystems. *Big Data & Society 12*(4), 1–12.

Chen, H.-J., M.-C. Li, and Y.-J. Lin (2008). Chaotic Dynamics in an Overlapping Generations Model with Myopic and Adaptive Expectations. *Journal of Economic Behavior & Organization 67*(1), 48–56.

Cong, L. W., D. Xie, and L. Zhang (2021). Knowledge Accumulation, Privacy, and Growth in a Data Economy. *Management Science 67*(10), 6480–6492.

Corrado, C., C. Hulten, and D. Sichel (2009). Intangible Capital and U.S. Economic Growth. *Review of Income and Wealth 55*(3), 661–685.

Cottier, B., R. Rahman, L. Fattorini, N. Maslej, and D. Owen (2024). The Rising Costs of Training Frontier AI Models. arXiv:2405.21015.

Farboodi, M. and L. Veldkamp (2020). Long-Run Growth of Financial Data Technology. *American Economic Review 110*(8), 2485–2523.

Farboodi, M. and L. Veldkamp (2026). A Model of the Data Economy. *Review of Economic Studies, forthcoming*.

Gabaix, X. (2020). A Behavioral New Keynesian Model. *American Economic Review 110*(8), 2271–2327.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. (2022). Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Volume 35, pp. 30016–30030.

Jones, C. I. (1995). R&D-Based Models of Economic Growth. *Journal of Political Economy 103*(4), 759–784.

Jones, C. I. (2024). The AI Dilemma: Growth versus Existential Risk. *American Economic Review: Insights 6*(4), 575–590.

Jones, C. I. and C. Tonetti (2020). Nonrivalry and the Economics of Data. *American Economic Review 110*(9), 2819–2858.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.

Korinek, A. and J. Vipra (2024). Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence. CEPR Discussion Paper DP19319.

Mihet, R., K. Rishabh, and O. Gomes (2026). Is it AI or Data that Drives Firm Market Power? *Journal of Monetary Economics 157*, 103878.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy 98*(5), S71–S102.

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature 631*, 755–759.

Villalobos, P., J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho (2022). Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. arXiv:2211.04325.

Wenger, E. (2024). AI Produces Gibberish When Trained on Too Much AI-Generated Data. *Nature 631*, 742–743.

Online Appendix

A. Derivation of Equilibrium Allocations

This appendix provides the detailed derivation of the equilibrium allocations under all four regimes and analyzes the optimal consumer data cap under data regulation. Section A.1 derives the optimal allocation, Sections A.2–A.4 derive the three competitive equilibria, and Section A.5 analyzes the optimal consumer data cap.

A.1 Social Planner

This appendix provides the detailed derivation of Proposition 1. Under the symmetry conditions $D_{j,t} = D_t$, $K_{j,t} = K_t/M$, $L_{A,j,t} = L_{A,t}/M$, and $Q_{j,t} = Q_t$ for all j, the social planner's Lagrangian associated with problem (16) can be written as

$$\begin{aligned}\mathcal{L} = &\sum_{t=0}^{\infty} \beta^t \left[\ln c_t - \phi(d_t)\right] \\ &+ \sum_{t=0}^{\infty} \lambda_t \left(N^{\frac{1}{\sigma-1}} A\, M^{\frac{\rho\zeta}{\rho-1}} \left\{ \left[(d_t L)^b \left(\frac{K_t}{M}\right)^{1-b} \right]^{\tau} D_t^{1-\tau} \right\}^{\zeta} (L - L_{K,t} - L_{A,t}) - c_t L - I_t \right) \\ &+ \sum_{t=0}^{\infty} M\mu_t \left\{ \eta \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} + (1-\delta_D) D_t - D_{t+1} \right\} \\ &+ \sum_{t=0}^{\infty} \nu_t \left[\xi I_t^{\gamma} L_{K,t}^{1-\gamma} + (1-\delta_K) K_t - K_{t+1} \right]\end{aligned}$$

where λ_t, μ_t, and ν_t are the Lagrange multipliers associated with the resource constraint, the per-firm AI-service accumulation equation, and the compute accumulation equation, respectively. The factor M multiplying μ_t accounts for the fact that there are M identical AI-service accumulation constraints. To simplify notation, denote

$$\Phi_t^{\zeta} \equiv M^{\frac{\rho\zeta}{\rho-1}} \left\{ \left[(d_t L)^b \left(\frac{K_t}{M}\right)^{1-b} \right]^{\tau} D_t^{1-\tau} \right\}^{\zeta}, \qquad \Omega_t \equiv \lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}}.$$

Taking first-order conditions with respect to c_t, d_t, K_t, D_t, $L_{K,t}$, $L_{A,t}$, and I_t, respectively, yields:

$$\frac{\partial \mathcal{L}}{\partial c_t} = \frac{\beta^t}{c_t} - \lambda_t L = 0 \tag{A.1}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial d_t} = & -\beta^t \phi'(d_t) + \lambda_t \frac{b\tau\zeta}{d_t} N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} (L - L_{K,t} - L_{A,t}) \\ & + M\mu_t \alpha(1-\lambda)\eta \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} (d_t L)^{-\frac{1}{\varepsilon}} L \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} = 0\end{aligned} \tag{A.2}$$

In the FOC for d_t, the first term is the marginal privacy cost, the second captures consumer data's contribution to output through data quality, and the third reflects its contribution to AI-service generation across all M firms (hence the factor M).

$$\frac{\partial \mathcal{L}}{\partial K_t} = \lambda_t \frac{(1-b)\tau\zeta}{K_t} N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} (L - L_{K,t} - L_{A,t}) + \nu_t (1 - \delta_K) - \nu_{t-1} = 0 \tag{A.3}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial D_t} = & \lambda_t \frac{(1-\tau)\zeta}{D_t} N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} (L - L_{K,t} - L_{A,t}) \\ & + M\mu_t \left\{ \alpha\lambda\eta \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} D_t^{-\frac{1}{\varepsilon}} \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} + 1 - \delta_D \right\} - M\mu_{t-1} = 0\end{aligned} \tag{A.4}$$

In the FOC for D_t, the first term captures the contribution of AI services to output through the effective data input Φ_t. Because the Lagrangian is written under symmetry with D_t as the common per-firm stock, varying D_t changes all M firms' AI services simultaneously, and the derivative $\partial \Phi_t^{\zeta} / \partial D_t = (1-\tau)\zeta \Phi_t^{\zeta} / D_t$ applies directly. The second and third terms carry factors of M from the M identical accumulation constraints.

$$\frac{\partial \mathcal{L}}{\partial L_{K,t}} = -\lambda_t N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} + \nu_t (1-\gamma)\xi I_t^{\gamma} L_{K,t}^{-\gamma} = 0 \tag{A.5}$$

$$\frac{\partial \mathcal{L}}{\partial L_{A,t}} = -\lambda_t N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} + M\mu_t (1-\alpha)\eta \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{-\alpha} \frac{1}{M} = 0$$

Note that the FOC for $L_{A,t}$ equates the marginal output loss from reallocating one unit of labor away from production to the marginal gain in data generation. The gain equals M firms times the per-firm marginal product $(1/M)$, which simplifies so that:

$$\lambda_t N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta} = \mu_t (1-\alpha)\eta \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{-\alpha}$$

$$\frac{\partial \mathcal{L}}{\partial I_t} = -\lambda_t + \nu_t \gamma \xi I_t^{\gamma-1} L_{K,t}^{1-\gamma} = 0 \tag{A.6}$$

We now combine these first-order conditions to derive the steady-state equations in Proposition 1. From Equations (A.5) and (A.6), we obtain

$$N^{\frac{\sigma}{\sigma-1}} \frac{Y_{i,t}}{L - L_{A,t} - L_{K,t}} = \frac{(1-\gamma) I_t}{\gamma L_{K,t}} \tag{A.7}$$

From Equations (A.1) and (A.6), we obtain the shadow price of computing power:

$$\nu_t = \frac{\beta^t}{L\gamma\xi c_t I_t^{\gamma-1} L_{K,t}^{1-\gamma}} \tag{A.8}$$

From the simplified FOC for $L_{A,t}$ and Equation (A.1), we obtain the shadow price of AI services:

$$\mu_t = \frac{\beta^t N^{\frac{1}{\sigma-1}} A \, \Phi_t^{\zeta}}{(1-\alpha)\eta L c_t \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}} (L_{A,t}/M)^{-\alpha}} \tag{A.9}$$

Substituting Equations (A.1) and (A.9) into Equation (A.2) yields

$$\phi'(d_t) = \frac{N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{c_t d_t L} \left[b\tau\zeta + \frac{\alpha L_{A,t}}{(1-\alpha)(L - L_{A,t} - L_{K,t})} \cdot \frac{(1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega_t} \right] \tag{A.10}$$

Note that the factor M from the M identical AI-service accumulation constraints cancels with the $1/M$ from $(L_{A,t}/M)$ in the shadow price μ_t, so that M does not appear in the final expression.

In the steady state, all variables remain constant, and the Lagrange multipliers satisfy $\mu_t = \beta\mu_{t-1}$ and $\nu_t = \beta\nu_{t-1}$. Simplifying Equation (A.3) and substituting Equations (A.1) and (A.8), we obtain

$$(1-b)\tau\zeta N^{\frac{\sigma}{\sigma-1}} \frac{Y_i^{sp}}{K^{sp}} = \frac{1-\beta(1-\delta_K)}{\beta\gamma\xi (I^{sp})^{\gamma-1} (L_K^{sp})^{1-\gamma}} \tag{A.11}$$

This equation has the same form as in the single-firm case because the elasticity of Φ_t^{ζ} with respect to K_t is $(1-b)\tau\zeta$ regardless of M. Substituting the steady-state compute accumulation $\delta_K K^{sp} = \xi (I^{sp})^{\gamma} (L_K^{sp})^{1-\gamma}$ into Equation (A.11) and combining with the resource constraint yields

$$\frac{c^{sp} L}{N^{\frac{\sigma}{\sigma-1}} Y_i^{sp}} = 1 - \frac{(1-b)\tau\zeta\gamma\beta\delta_K}{1-\beta(1-\delta_K)} \equiv 1 - B^{sp} \tag{A.12}$$

Combining Equations (A.7) and (A.11), and then substituting the steady-state compute accumulation, yields

$$\frac{L_K^{sp}}{L - L_A^{sp} - L_K^{sp}} = \frac{1-\gamma}{\gamma} B^{sp} \tag{A.13}$$

We now derive the data-related equations. Simplifying Equation (A.4) in the steady state and substituting Equations (A.1) and (A.9), we obtain

$$\frac{[1-\beta(1-\delta_D)]\, D^{sp}\, M}{(1-\alpha)\beta\eta (\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}} (L_A^{sp}/M)^{-\alpha} (L - L_A^{sp} - L_K^{sp})} = (1-\tau)\zeta + \frac{\alpha}{1-\alpha} \cdot \frac{L_A^{sp}}{L - L_A^{sp} - L_K^{sp}} \cdot \frac{\lambda (D^{sp})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{sp}} \tag{A.14}$$

Substituting Equation (A.12) into Equation (A.10) yields the steady-state consumer data supply

condition:

$$\phi'(d^{sp})d^{sp} = \frac{1}{1-B^{sp}}\left[b\tau\zeta + \frac{\alpha L_A^{sp}(1-\lambda)(d^{sp}L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L-L_A^{sp}-L_K^{sp})\Omega^{sp}}\right] \tag{A.15}$$

In the steady state, combining Equations (3) and (15) yields $\tilde{D}^{sp} = \delta_D D^{sp}$. Substituting into the definition of $\tilde{D}_{j,t}$ and solving for Ω^{sp}, we obtain

$$\Omega^{sp} = \left[\frac{\delta_D D^{sp}}{\eta(L_A^{sp}/M)^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}}.$$

From the definition of Ω_t, the steady-state consumer data supply is

$$d^{sp} = \left\{\Omega^{sp} - \lambda(D^{sp})^{\frac{\varepsilon-1}{\varepsilon}}\right\}^{\frac{\varepsilon}{\varepsilon-1}} \frac{1}{(1-\lambda)^{\frac{\varepsilon}{\varepsilon-1}}L}.$$

Therefore, jointly considering Equations (A.13), (A.14), and (A.15), together with the steady-state relationships above, we can solve for the allocation under the social optimum.

Based on the results obtained above, we can further derive the other key variables. The steady-state output per variety is

$$Y_i^{sp} = \left\{A\,M^{\omega}(d^{sp}L)^{b\tau\zeta}(D^{sp})^{\zeta(1-\tau)}\frac{L-L_A^{sp}-L_K^{sp}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{sp})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{sp}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}$$

The total stock of computing power is

$$K^{sp} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{sp})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{sp}\right]^{\gamma}(Y_i^{sp})^{\gamma}$$

The common data quality across AI firms is

$$Q^{sp} = \left[\frac{(d^{sp}L)^b(K^{sp}/M^{sp})^{1-b}}{D^{sp}}\right]^{\tau}$$

Optimal number of AI firms. We now derive the optimal M^{sp} condition stated in Equation (20). The social planner chooses M at $t = 0$, paying a total entry cost MF. In the steady state ($t \geq 1$), no further entry occurs and the per-period welfare is $W(M) = \ln c^{sp}(M) - \phi(d^{sp}(M))$. The planner maximizes the present discounted value of welfare net of entry costs. By the envelope theorem, the marginal effect of M on the per-period steady-state Lagrangian (evaluated at the optimal allocation conditional on M) is given by the direct partial derivatives only, since the indirect effects through the optimally chosen variables are zero.

The direct effect of M on the resource constraint (14) is

$$\frac{\partial}{\partial M}\left[N^{\frac{1}{\sigma-1}}A\,(\Phi^{sp})^{\zeta}(L-L_K^{sp}-L_A^{sp})\right] = \frac{\omega}{M^{sp}}\cdot Y^{sp},$$

where $\omega \equiv \rho\zeta/(\rho-1) - (1-b)\tau\zeta$ is the net elasticity of Φ^ζ with respect to M, and $Y^{sp} = c^{sp}L + I^{sp}$ is aggregate output.

The direct effect of M on the AI-service accumulation is obtained by differentiating the total AI-service contribution $M \cdot [\eta(\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{sp}/M)^{1-\alpha} - \delta_D D^{sp}]$ with respect to M. Using the steady-state condition $\delta_D D^{sp} = \eta(\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{sp}/M^{sp})^{1-\alpha}$, this simplifies to

$$\frac{\partial}{\partial M}\left[M \cdot \left(\eta(\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}}\left(\frac{L_A^{sp}}{M}\right)^{1-\alpha} - \delta_D D^{sp}\right)\right] = -(1-\alpha)\delta_D D^{sp}.$$

Multiplying this derivative by the shadow price μ^{sp} of the AI-service accumulation constraint yields its contribution to the steady-state Lagrangian. The first-order condition for M equates the present value of the two marginal effects to the entry cost F:

$$\frac{1}{1-\beta}\left[\lambda^{sp}\frac{\omega\, Y^{sp}}{M^{sp}} - \mu^{sp}(1-\alpha)\delta_D D^{sp}\right] = \lambda^{sp} F.$$

Dividing by λ^{sp} and using the relationship $\mu^{sp}/\lambda^{sp} = Y^{sp}/[(1-\alpha)\eta(\Omega^{sp})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{sp}/M^{sp})^{-\alpha}(L - L_A^{sp} - L_K^{sp})]$ derived from the FOC for $L_{A,t}$, together with the steady-state AI-service accumulation condition, we can show that

$$\frac{\mu^{sp}}{\lambda^{sp}}(1-\alpha)\delta_D D^{sp} = \frac{Y^{sp}\, L_A^{sp}}{M^{sp}(L - L_A^{sp} - L_K^{sp})}.$$

Substituting this result and rearranging yields the optimal M^{sp} condition:

$$(1-\beta)\, F = \frac{Y^{sp}}{M^{sp}}\left[\omega - \frac{L_A^{sp}}{L - L_A^{sp} - L_K^{sp}}\right].$$

The left-hand side is the amortized per-period entry cost. The right-hand side is the net per-period benefit of one additional AI firm: the love-of-variety gain $\omega\, Y^{sp}/M^{sp}$ minus the data dilution cost $Y^{sp}L_A^{sp}/[M^{sp}(L - L_A^{sp} - L_K^{sp})]$ arising from spreading the same total AI labor across more firms.

A.2 Restricted Entry

This appendix provides the detailed derivation of Proposition 2. Under the restricted-entry regime with $M = 1$, there is a single generative AI firm, and the effective data input reduces to $\Phi_t = Q_t D_t$, where $Q_t = [(d_t L)^b K_t^{1-b}/D_t]^\tau$. The aggregate resource constraint coincides with Equation (24). We derive the equilibrium conditions for each agent in turn.

Consumer. From the utility maximization problem of the representative consumer, the Lagrangian can be written as

$$\mathcal{L}_C = \sum_{t=0}^{\infty}\beta^t\left[\ln c_t - \phi(d_t)\right] + \sum_{t=0}^{\infty}\lambda_t\left[w_t + p_{d,t}^s d_t + (1+r_t)a_t + \pi_t - c_t - a_{t+1}\right]$$

where λ_t is the Lagrange multiplier associated with this problem. Taking first-order conditions with respect to c_t, d_t, and a_t, respectively, yields

$$\frac{\partial \mathcal{L}_C}{\partial c_t} = \frac{\beta^t}{c_t} - \lambda_t = 0 \tag{A.16}$$

$$\frac{\partial \mathcal{L}_C}{\partial d_t} = -\beta^t \phi'(d_t) + \lambda_t p^s_{d,t} = 0 \tag{A.17}$$

$$\frac{\partial \mathcal{L}_C}{\partial a_t} = (1 + r_t)\lambda_t - \lambda_{t-1} = 0 \tag{A.18}$$

Combining Equations (A.16) and (A.17) yields Equation (26). Combining Equations (A.16) and (A.18) yields Equation (27).

Final-goods producers. The first-order conditions of the producer's problem yield the inverse demand for AI services and the labor demand condition:

$$p^s_{D,t} = \left(1 - \frac{1}{\sigma}\right) \zeta N^{\frac{\sigma}{\sigma-1}} A \left(Q_t D_t\right)^{\zeta-1} L_{i,t}, \tag{A.19}$$

$$w_t = \left(1 - \frac{1}{\sigma}\right) N^{\frac{1}{\sigma-1}} A \left(Q_t D_t\right)^{\zeta}. \tag{A.20}$$

Generative AI firm. Based on these results, we can write the Lagrangian for the generative AI firm. Since the firm has monopoly power in the market for AI services, it internalizes the demand for AI services from the data intermediary. Substituting Equation (A.19) into the firm's profit function (28), the revenue term becomes $(1 - 1/\sigma)\zeta N^{\sigma/(\sigma-1)} A (Q_t D_t)^{\zeta} L_{i,t}$, which depends on data quality and the stock of AI services. The Lagrangian is therefore

$$\mathcal{L}_A = \sum_{t=0}^{\infty} \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left\{ \left(1 - \frac{1}{\sigma}\right) \zeta N^{\frac{\sigma}{\sigma-1}} A \left[(d_t L)^{b\tau} K_t^{(1-b)\tau} D_t^{1-\tau} \right]^{\zeta} L_{i,t} - w_t L_{A,t} - p_{K,t} K_t - p^s_{d,t} d_t L \right\}$$

$$+ \sum_{t=0}^{\infty} \mu_t \left\{ \eta \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} L_{A,t}^{1-\alpha} + (1-\delta_D) D_t - D_{t+1} \right\}$$

where μ_t is the Lagrange multiplier associated with this problem. Taking first-order conditions with respect to d_t, K_t, $L_{A,t}$, and D_t, respectively, yields:

$$\frac{\partial \mathcal{L}_A}{\partial d_t} = \frac{\left(1 - \frac{1}{\sigma}\right) \zeta^2 b\tau N^{\frac{\sigma}{\sigma-1}} Y_{i,t} d_t^{-1} - p^s_{d,t} L}{\prod_{s=0}^{t}(1+r_s)} + \mu_t \alpha\eta(1-\lambda) L^{\frac{\varepsilon-1}{\varepsilon}} \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} d_t^{-\frac{1}{\varepsilon}} L_{A,t}^{1-\alpha} = 0 \tag{A.21}$$

$$\frac{\partial \mathcal{L}_A}{\partial K_t} = \left(1 - \frac{1}{\sigma}\right) \zeta^2 (1-b)\tau N^{\frac{\sigma}{\sigma-1}} Y_{i,t} K_t^{-1} - p_{K,t} = 0 \tag{A.22}$$

$$\frac{\partial \mathcal{L}_A}{\partial L_{A,t}} = -\frac{w_t}{\prod_{s=0}^{t}(1+r_s)} + \mu_t \eta(1-\alpha) \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} L_{A,t}^{-\alpha} = 0 \tag{A.23}$$

$$\frac{\partial \mathcal{L}_A}{\partial D_t} = \frac{\left(1-\frac{1}{\sigma}\right)\zeta^2(1-\tau)N^{\frac{\sigma}{\sigma-1}}Y_{i,t}D_t^{-1}}{\prod_{s=0}^{t}(1+r_s)} + \mu_t\left\{\alpha\eta\lambda\left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_tL)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} D_t^{-\frac{1}{\varepsilon}} L_{A,t}^{1-\alpha} + 1 - \delta_D\right\} - \mu_{t-1} = 0 \tag{A.24}$$

To simplify notation, define the CES data aggregate

$$\Omega_t \equiv \lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_tL)^{\frac{\varepsilon-1}{\varepsilon}}, \tag{A.25}$$

so that $\tilde{D}_t = \eta\,\Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}} L_{A,t}^{1-\alpha}$. Combining Equations (A.21) and (A.23) yields

$$\frac{p_{d,t}^s d_t L - \left(1-\frac{1}{\sigma}\right)\zeta^2 b\tau N^{\frac{\sigma}{\sigma-1}}Y_{i,t}}{w_t} = \frac{\alpha(1-\lambda)(d_tL)^{\frac{\varepsilon-1}{\varepsilon}}L_{A,t}}{(1-\alpha)\Omega_t}. \tag{A.26}$$

Rearranging Equation (A.22) yields

$$p_{K,t}K_t = \left(1-\frac{1}{\sigma}\right)\zeta^2(1-b)\tau N^{\frac{\sigma}{\sigma-1}}Y_{i,t}. \tag{A.27}$$

Combining Equations (A.21) and (A.24) yields

$$\frac{\left(1-\frac{1}{\sigma}\right)\zeta^2(1-\tau)N^{\frac{\sigma}{\sigma-1}}Y_{i,t}}{(1+r_t)D_t} = \frac{p_{d,t-1}d_{t-1}L - \left(1-\frac{1}{\sigma}\right)\zeta^2 b\tau N^{\frac{\sigma}{\sigma-1}}Y_{i,t-1}}{\alpha\eta(1-\lambda)\Omega_{t-1}^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(d_{t-1}L)^{\frac{\varepsilon-1}{\varepsilon}}L_{A,t-1}^{1-\alpha}} - \frac{p_{d,t}^s d_t L - \left(1-\frac{1}{\sigma}\right)\zeta^2 b\tau N^{\frac{\sigma}{\sigma-1}}Y_{i,t}}{1+r_t}\cdot\frac{\alpha\eta\lambda\Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}D_t^{-\frac{1}{\varepsilon}}L_{A,t}^{1-\alpha}+1-\delta_D}{\alpha\eta(1-\lambda)\Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(d_tL)^{\frac{\varepsilon-1}{\varepsilon}}L_{A,t}^{1-\alpha}}. \tag{A.28}$$

Computing power provider. We derive the optimality conditions for the computing power provider. Since the provider has monopoly power, it internalizes the demand for computing power from the generative AI firm. Substituting Equation (A.27) into the provider's profit function (29) yields a revenue term that depends on the effective contribution of computing power to output. The Lagrangian can be written as:

$$\mathcal{L}_K = \sum_{t=0}^{\infty}\frac{1}{\prod_{s=0}^{t}(1+r_s)}\left[\left(1-\frac{1}{\sigma}\right)\zeta^2(1-b)\tau N^{\frac{\sigma}{\sigma-1}}A\left[(d_tL)^{b\tau}K_t^{(1-b)\tau}D_t^{1-\tau}\right]^{\zeta}L_{i,t} - w_tL_{K,t} - I_t\right] + \sum_{t=0}^{\infty}\nu_t\left[\xi I_t^{\gamma}L_{K,t}^{1-\gamma} + (1-\delta_K)K_t - K_{t+1}\right]$$

where ν_t is the Lagrange multiplier associated with this problem. Taking first-order conditions with respect to I_t, $L_{K,t}$, and K_t, respectively, yields:

$$\frac{\partial \mathcal{L}_K}{\partial I_t} = -\frac{1}{\prod_{s=0}^{t}(1+r_s)} + \nu_t\xi\gamma I_t^{\gamma-1}L_{K,t}^{1-\gamma} = 0 \tag{A.29}$$

$$\frac{\partial \mathcal{L}_K}{\partial L_{K,t}} = -\frac{w_t}{\prod_{s=0}^{t}(1+r_s)} + \nu_t \xi (1-\gamma) I_t^{\gamma} L_{K,t}^{-\gamma} = 0 \tag{A.30}$$

$$\frac{\partial \mathcal{L}_K}{\partial K_t} = \frac{\left(1-\frac{1}{\sigma}\right)\zeta^3 (1-b)^2 \tau^2 N^{\frac{\sigma}{\sigma-1}} Y_{i,t} K_t^{-1}}{\prod_{s=0}^{t}(1+r_s)} + \nu_t (1-\delta_K) - \nu_{t-1} = 0 \tag{A.31}$$

Combining Equations (A.29) and (A.30) yields

$$\frac{1}{w_t} = \frac{\gamma L_{K,t}}{(1-\gamma) I_t}. \tag{A.32}$$

Combining Equations (A.29) and (A.31) yields

$$\left(1-\frac{1}{\sigma}\right)\zeta^3 (1-b)^2 \tau^2 N^{\frac{\sigma}{\sigma-1}} \frac{Y_{i,t}}{K_t} + \frac{1-\delta_K}{\xi \gamma I_t^{\gamma-1} L_{K,t}^{1-\gamma}} = \frac{1+r_t}{\xi \gamma I_{t-1}^{\gamma-1} L_{K,t-1}^{1-\gamma}}. \tag{A.33}$$

Steady-state characterization. Since $r_t = r^{re} = 1/\beta - 1$ in the steady state, and all variables remain constant, Equation (A.28) can be simplified. Substituting the steady-state interest rate and setting all time-varying variables equal to their steady-state values, we obtain

$$\begin{aligned}\frac{\left(1-\frac{1}{\sigma}\right)\beta\zeta^2(1-\tau) N^{\frac{\sigma}{\sigma-1}} Y_i^{re}}{D^{re}} = {} & \frac{p_d^{re} d^{re} L - \left(1-\frac{1}{\sigma}\right)\zeta^2 b \tau N^{\frac{\sigma}{\sigma-1}} Y_i^{re}}{\alpha\eta(1-\lambda)(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}}(L_A^{re})^{1-\alpha}} \\ & \times \left[1-\beta\left(\alpha\eta\lambda(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(D^{re})^{-\frac{1}{\varepsilon}}(L_A^{re})^{1-\alpha}+1-\delta_D\right)\right].\end{aligned} \tag{A.34}$$

Similarly, Equation (A.33) can be simplified in the steady state to

$$\left(1-\frac{1}{\sigma}\right)\zeta^3(1-b)^2\tau^2 N^{\frac{\sigma}{\sigma-1}} \frac{Y_i^{re}}{K^{re}} = \frac{1-\beta(1-\delta_K)}{\beta\xi\gamma(I^{re})^{\gamma-1}(L_K^{re})^{1-\gamma}} \tag{A.35}$$

At the same time, the computing power accumulation equation (6) implies that in the steady state,

$$\delta_K K^{re} = \xi (I^{re})^{\gamma} (L_K^{re})^{1-\gamma}. \tag{A.36}$$

Substituting Equation (A.36) into Equation (A.35) and rearranging, we can express the ratio of investment to output as

$$\begin{aligned}&\left(1-\frac{1}{\sigma}\right)\zeta^3(1-b)^2\tau^2\beta\gamma\delta_K N^{\frac{\sigma}{\sigma-1}} Y_i^{re} = \left[1-\beta(1-\delta_K)\right] I^{re} \\ \Rightarrow \quad & \frac{I^{re}}{N^{\frac{\sigma}{\sigma-1}} Y_i^{re}} = \frac{(\sigma-1)\zeta^3(1-b)^2\tau^2\beta\gamma\delta_K}{\left[1-\beta(1-\delta_K)\right]\sigma} \equiv B^{re}\end{aligned} \tag{A.37}$$

The constant B^{re} summarizes the share of output devoted to computing power investment in the steady state. Substituting the above expression into the resource constraint (12) yields the

steady-state consumption-output ratio:

$$\frac{c^{re}L}{N^{\frac{\sigma}{\sigma-1}}Y_i^{re}} = 1 - B^{re} \tag{A.38}$$

Next, combining Equations (A.20) and (A.32), we obtain

$$(1-\gamma)\sigma I^{re}L_i^{re} = (\sigma-1)\gamma N^{\frac{1}{\sigma-1}}Y_i^{re}L_K^{re} \tag{A.39}$$

Substituting Equation (A.37) into Equation (A.39) and rearranging yields

$$\frac{L_K^{re}}{L - L_A^{re} - L_K^{re}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma}B^{re} \tag{A.40}$$

On the other hand, combining Equations (3) and (4) in the steady state yields $\tilde{D}^{re} = \delta_D D^{re}$. Substituting this into the definition of $\tilde{D}_t$ and solving for Ω^{re}, we obtain

$$\Omega^{re} = \left[\frac{\delta_D D^{re}}{\eta(L_A^{re})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} \tag{A.41}$$

At the same time, from the definition of Ω_t in Equation (A.25), together with Equation (A.41), we can also obtain the relationship between consumer data and AI services:

$$d^{re} = \left\{\left[\frac{\delta_D D^{re}}{\eta(L_A^{re})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} - \lambda(D^{re})^{\frac{\varepsilon-1}{\varepsilon}}\right\}^{\frac{\varepsilon}{\varepsilon-1}} \frac{1}{(1-\lambda)^{\frac{\varepsilon}{\varepsilon-1}}L} \tag{A.42}$$

Substituting Equations (26) and (A.38) into Equation (A.34), respectively, and rearranging yields

$$\frac{\beta\zeta^2(1-\tau)(\sigma-1)}{D^{re}\sigma} = \frac{(1-B^{re})\phi'(d^{re})d^{re} - \left(1-\frac{1}{\sigma}\right)\zeta^2 b\tau}{\alpha\eta(1-\lambda)(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}}(L_A^{re})^{1-\alpha}}\left[1-\beta\left(\alpha\eta\lambda(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(D^{re})^{-\frac{1}{\varepsilon}}(L_A^{re})^{1-\alpha}+1-\delta_D\right)\right] \tag{A.43}$$

Furthermore, combining Equations (A.26) and (A.32), and substituting in Equations (26) and (A.37), we obtain

$$\frac{(1-B^{re})\phi'(d^{re})d^{re} - \left(1-\frac{1}{\sigma}\right)\zeta^2 b\tau}{B^{re}} = \frac{\alpha(1-\lambda)(1-\gamma)(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}}L_A^{re}}{(1-\alpha)\gamma\Omega^{re}L_K^{re}} \tag{A.44}$$

Combining Equations (A.43) and (A.44), and further substituting Equation (A.40), we obtain the steady-state condition governing the intertemporal accumulation of AI services:

$$\frac{[1-\beta(1-\delta_D)]D^{re}}{(1-\alpha)\beta\eta(\Omega^{re})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{re})^{-\alpha}(L-L_A^{re}-L_K^{re})} = (1-\tau)\zeta^2 + \frac{\alpha}{1-\alpha}\cdot\frac{L_A^{re}}{L-L_A^{re}-L_K^{re}}\cdot\frac{\lambda(D^{re})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{re}}$$

This corresponds to Equation (31) in Proposition 2. Given the relationships summarized in Equations (A.41) and (A.42), this equation implicitly defines the relationship between AI services

D^{re} and labor employed by generative AI firms L_A^{re}. In addition, Equation (A.44) can be further rewritten, after substituting Equation (A.38), as

$$\phi'(d^{re})d^{re} = \frac{\sigma - 1}{(1 - B^{re})\sigma}\left[b\tau\zeta^2 + \frac{\alpha L_A^{re}(1-\lambda)(d^{re}L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L - L_A^{re} - L_K^{re})\Omega^{re}}\right]$$

This corresponds to Equation (32) in Proposition 2, which characterizes the steady-state supply of consumer data. Finally, substituting Equations (A.40) and (A.44) into the labor market clearing condition (11), we can solve for labor employed in each sector, namely L_A^{re}, L_K^{re}, and L_i^{re}. These results in turn allow us to recover AI services D^{re} and consumer data d^{re}.

Given the values of all the above variables, we can further solve for the remaining variables. Combining Equations (A.35), (A.36), and the production function for consumption goods in Equation (8), and rearranging, we obtain consumption-good output Y_i^{re}, computing power K^{re}, and data quality Q^{re} as follows:

$$Y_i^{re} = \left\{A(d^{re}L)^{b\tau\zeta}(D^{re})^{\zeta(1-\tau)}\frac{L - L_A^{re} - L_K^{re}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{re})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{re}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}$$

$$K^{re} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{re})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{re}\right]^{\gamma}(Y_i^{re})^{\gamma}$$

$$Q^{re} = \left[\frac{(d^{re}L)^b(K^{re})^{1-b}}{D^{re}}\right]^{\tau}$$

A.3 Free Entry

This appendix provides the detailed derivation of Proposition 3. Under the free-entry regime, there are M symmetric generative AI firms, each producing a differentiated type of AI service. In this appendix, we suppress the firm index j and work directly with per-firm symmetric quantities: D_t, K_t/M, $L_{A,t}/M$.

Producers and AI-service demand. The final-goods producer's cost minimization over the M types of quality-adjusted AI services yields the CES demand for type j:

$$p_{D,j,t}^d = p_{\Phi,t}\left(\frac{Q_{j,t}D_{j,t}}{\Phi_t}\right)^{-\frac{1}{\rho}},$$

where $p_{\Phi,t} = (1 - 1/\sigma)\zeta N^{1/(\sigma-1)}A\Phi_t^{\zeta-1}L_{i,t}$ is the shadow price of the composite data input. Under symmetry, $Q_{j,t}D_{j,t} = Q_tD_t$ for all j and $\Phi_t = M^{\rho/(\rho-1)}Q_tD_t$, so:

$$p_{D,t}^d = p_{\Phi,t}\,M^{\frac{1}{\rho-1}}, \qquad p_{D,t}^s = N\,p_{D,t}^d.$$

Per-firm AI-service revenue (supply side) is therefore:

$$p^s_{D,t}\, Q_t D_t = \frac{N\, p_{\Phi,t}\, \Phi_t}{M} = \frac{(1-1/\sigma)\zeta\, N^{\sigma/(\sigma-1)} Y_{i,t}}{M}.$$

AI firm's problem. Each AI firm j faces a CES demand curve with elasticity ρ for its variety. Taking the aggregate Φ_t as given, the firm's perceived revenue is proportional to $(Q_{j,t}D_{j,t})^{(\rho-1)/\rho}$. The Lagrangian for firm j can be written as:

$$\begin{aligned}\mathcal{L}_A = \sum_{t=0}^{\infty} \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left\{ \left(1-\frac{1}{\rho}\right) \frac{(1-1/\sigma)\zeta\, N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M} - w_t \frac{L_{A,t}}{M} - p_{K,t}\frac{K_t}{M} - \frac{p^s_{d,t}}{M} d_t L \right\} \\ + \sum_{t=0}^{\infty} \mu_t \left\{ \eta \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(d_t L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} + (1-\delta_D)D_t - D_{t+1} \right\}\end{aligned}$$

where the factor $(1-1/\rho)$ in the revenue term reflects the CES markup: the firm's marginal revenue is $(1-1/\rho)$ times its average revenue. The consumer data cost is $p^s_{d,t} d_t L/M$ because the intermediary condition (9) implies each firm pays $1/M$ of the total consumer data price.

Taking first-order conditions with respect to d_t, K_t, $L_{A,t}$, and D_t yields:

$$\begin{aligned}\frac{\partial \mathcal{L}_A}{\partial d_t} = \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left[\left(1-\frac{1}{\rho}\right)(1-1/\sigma)\zeta\, b\tau\, \frac{N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M\, d_t} - \frac{p^s_{d,t} L}{M} \right] \\ + \mu_t\, \alpha(1-\lambda)\eta\, \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} (d_t L)^{-\frac{1}{\varepsilon}} L \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} = 0\end{aligned} \tag{A.45}$$

$$\frac{\partial \mathcal{L}_A}{\partial K_t} = \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left[\left(1-\frac{1}{\rho}\right)(1-1/\sigma)\zeta\,(1-b)\tau\, \frac{N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M\, K_t/M} - p_{K,t} \right] = 0 \tag{A.46}$$

From Equation (A.46), the per-firm compute demand is:

$$p_{K,t}\frac{K_t}{M} = \left(1-\frac{1}{\rho}\right)(1-1/\sigma)\zeta\,(1-b)\tau\, \frac{N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M}. \tag{A.47}$$

Summing over M firms gives total compute demand:

$$p_{K,t} K_t = \left(1-\frac{1}{\rho}\right)\left(1-\frac{1}{\sigma}\right)\zeta\,(1-b)\tau\, N^{\frac{\sigma}{\sigma-1}} Y_{i,t}. \tag{A.48}$$

$$\frac{\partial \mathcal{L}_A}{\partial L_{A,t}} = -\frac{w_t}{M \prod_{s=0}^{t}(1+r_s)} + \mu_t(1-\alpha)\eta\, \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{-\alpha} \frac{1}{M} = 0 \tag{A.49}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_A}{\partial D_t} = \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left(1-\frac{1}{\rho}\right)(1-1/\sigma)\zeta\,(1-\tau)\, \frac{N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M\, D_t} \\ + \mu_t \left\{ \alpha\lambda\eta\, \Omega_t^{\frac{\alpha\varepsilon}{\varepsilon-1}-1} D_t^{-\frac{1}{\varepsilon}} \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} + 1 - \delta_D \right\} - \mu_{t-1} = 0\end{aligned} \tag{A.50}$$

Computing power provider. The computing power provider substitutes the total compute demand (A.48) into its profit function. Following the same steps as in Appendix A.2, but with $(1-1/\rho)(1-1/\sigma)\zeta(1-b)\tau$ replacing $(1-1/\sigma)\zeta^2(1-b)\tau$, the provider's FOC yields:

$$\left(1-\frac{1}{\rho}\right)\left(1-\frac{1}{\sigma}\right)\zeta^2(1-b)^2\tau^2 N^{\frac{\sigma}{\sigma-1}}\frac{Y_i^{fe}}{K^{fe}} = \frac{1-\beta(1-\delta_K)}{\beta\xi\gamma(I^{fe})^{\gamma-1}(L_K^{fe})^{1-\gamma}}$$

Following the same derivation as in Appendix A.2, namely substituting the steady-state compute accumulation, combining with the resource constraint, and using the labor-investment relationship, we obtain:

$$B^{fe} = \frac{(\sigma-1)(1-1/\rho)\zeta^2(1-b)^2\tau^2\beta\gamma\delta_K}{[1-\beta(1-\delta_K)]\sigma}$$

and the labor allocation

$$\frac{L_K^{fe}}{L-L_A^{fe}-L_K^{fe}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma}B^{fe}. \tag{A.51}$$

Steady-state characterization. In the steady state, using $r^{fe} = 1/\beta - 1$ and combining the AI firm's FOCs, we derive the AI-service Euler equation. From Equations (A.45), (A.49), and (A.50), following the same algebraic steps as in Appendix A.2 but with the CES markup $(1-1/\rho)$ replacing the monopoly factor ζ, we obtain:

$$\frac{[1-\beta(1-\delta_D)]\,D^{fe}\,M^{fe}}{(1-\alpha)\beta\eta(\Omega^{fe})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{fe}/M^{fe})^{-\alpha}(L-L_A^{fe}-L_K^{fe})} = \left(1-\frac{1}{\rho}\right)(1-\tau)\zeta + \frac{\alpha}{1-\alpha}\cdot\frac{L_A^{fe}}{L-L_A^{fe}-L_K^{fe}}\cdot\frac{\lambda(D^{fe})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{fe}} \tag{A.52}$$

For the consumer data supply condition, combining the consumer's FOC $p_d^{fe}/c^{fe} = \phi'(d^{fe})$ with the AI firm's FOC for d and the resource constraint yields:

$$\phi'(d^{fe})d^{fe} = \frac{\sigma-1}{(1-B^{fe})\sigma}\left[\left(1-\frac{1}{\rho}\right)b\tau\zeta + \frac{\alpha L_A^{fe}(1-\lambda)(d^{fe}L)^{\frac{\varepsilon-1}{\varepsilon}}}{(1-\alpha)(L-L_A^{fe}-L_K^{fe})\Omega^{fe}}\right] \tag{A.53}$$

As in the social planner's case, the factor M^{fe} from aggregating across M identical firms' AI-service accumulation constraints cancels with the $1/M$ from the per-firm labor allocation (L_A/M) in the shadow price of AI services, so that M does not appear in the final expression.

The steady-state AI-service accumulation gives:

$$\Omega^{fe} = \left[\frac{\delta_D D^{fe}}{\eta(L_A^{fe}/M^{fe})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}} \tag{A.54}$$

and consumer data:

$$d^{fe} = \left\{\Omega^{fe} - \lambda(D^{fe})^{\frac{\varepsilon-1}{\varepsilon}}\right\}^{\frac{\varepsilon}{\varepsilon-1}}\frac{1}{(1-\lambda)^{\frac{\varepsilon}{\varepsilon-1}}L} \tag{A.55}$$

Free entry condition. In the steady state, each AI firm earns per-period profit π^{fe}. The free entry condition requires that the present value of profits equals the entry cost:

$$\frac{\pi^{fe}}{1-\beta} = F \qquad \Longrightarrow \qquad \pi^{fe} = (1-\beta)F.$$

Per-firm profit is revenue minus costs:

$$\pi^{fe} = \underbrace{\frac{(1-1/\sigma)\zeta\, N^{\frac{\sigma}{\sigma-1}} Y_i^{fe}}{M^{fe}}}_{\text{per-firm AI-service revenue}} - \underbrace{p_K^{fe}\frac{K^{fe}}{M^{fe}}}_{\text{compute cost}} - \underbrace{w^{fe}\frac{L_A^{fe}}{M^{fe}}}_{\text{labor cost}} - \underbrace{\frac{p_d^{fe}\, d^{fe} L}{M^{fe}}}_{\text{consumer data cost}}$$

Using the compute demand condition (A.47), the compute cost can be written as $(1-1/\rho)(1-b)\tau$ times per-firm revenue. Substituting the wage equation $w^{fe} = (1-1/\sigma)N^{1/(\sigma-1)}A(\Phi^{fe})^{\zeta}$ and the consumer data price $p_d^{fe} = c^{fe}\phi'(d^{fe})$, and defining aggregate output $Y^{fe} \equiv c^{fe}L + I^{fe} = N^{\sigma/(\sigma-1)}Y_i^{fe}$, the free entry condition can be expressed as:

$$(1-\beta)F = \frac{(1-1/\sigma)\zeta\, Y^{fe}}{M^{fe}}\left[1-\left(1-\frac{1}{\rho}\right)(1-b)\tau\right] - \frac{w^{fe}L_A^{fe}}{M^{fe}} - \frac{p_d^{fe}\, d^{fe} L}{M^{fe}} \tag{A.56}$$

where the first term is per-firm revenue net of compute cost and the remaining terms are per-firm labor and consumer data costs. Equivalently, under the constant-returns-to-scale structure of the AI firm's steady-state technology, Euler's theorem implies that total input payments equal $(1-1/\rho)$ of per-firm revenue, so the free entry condition can also be written in the compact form $(1-\beta)F = (1/\rho)(1-1/\sigma)\zeta Y^{fe}/M^{fe}$, which makes transparent that entry drives the CES markup rent down to the amortized entry cost.

Remaining variables. After solving the system of Equations (A.51), (A.52), (A.53), (A.56), together with (A.54) and (A.55), for L_K^{fe}, L_A^{fe}, D^{fe}, d^{fe}, and M^{fe}, the remaining variables can be derived. Output per variety is:

$$Y_i^{fe} = \left\{ A\,(M^{fe})^{\omega}(d^{fe}L)^{b\tau\zeta}(D^{fe})^{\zeta(1-\tau)}\frac{L-L_A^{fe}-L_K^{fe}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{fe})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{fe}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}$$

Total computing power is:

$$K^{fe} = \left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{fe})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{fe}\right]^{\gamma}(Y_i^{fe})^{\gamma}$$

Data quality (common across AI firms) is:

$$Q^{fe} = \left[\frac{(d^{fe}L)^b(K^{fe}/M^{fe})^{1-b}}{D^{fe}}\right]^{\tau}$$

A.4 Data Regulation

This appendix provides the detailed derivation of Proposition 4. The data regulation regime combines the free-entry market structure of Appendix A.3 with a binding government cap $\bar{d}$ on consumer data provision. Because the market structure, consisting of M symmetric AI firms with monopolistic competition, a monopolistic computing power provider, and N monopolistic final-goods producers, is identical to the free-entry regime, we focus on the modifications arising from the data cap and refer to Appendix A.3 for the unchanged derivations.

Consumer. With the binding data cap, the consumer maximizes Equation (1) subject to the budget constraint (25) and the additional constraint $d_t \leq \bar{d}$. Since the cap binds:

$$d^{dr} = \bar{d}.$$

The consumer's first-order condition for data supply (26) no longer holds; in general, $p_d^{dr}/c^{dr} \neq \phi'(\bar{d})$. The data price p_d^{dr} is instead determined by the AI firms' demand. The consumption Euler equation (27) continues to hold, yielding $r^{dr} = 1/\beta - 1$ in the steady state.

Producers and AI firms. The final-goods producers' optimization is identical to that in Appendix A.3, yielding the same CES demand for each type of AI service and the same shadow price of the composite data input. Each AI firm j faces the same CES demand curve with elasticity ρ as in the free-entry regime and solves the same optimization problem, with the sole difference that $d_t = \bar{d}$ is exogenous rather than chosen optimally.

The AI firm's Lagrangian takes the same form as in Appendix A.3, with $\bar{d}$ replacing d_t:

$$\begin{aligned}\mathcal{L}_A = &\sum_{t=0}^{\infty} \frac{1}{\prod_{s=0}^{t}(1+r_s)} \left\{ \left(1 - \frac{1}{\rho}\right) \frac{(1-1/\sigma)\zeta N^{\frac{\sigma}{\sigma-1}} Y_{i,t}}{M} - w_t \frac{L_{A,t}}{M} - p_{K,t}\frac{K_t}{M} - \frac{p^s_{d,t}}{M}\bar{d}L \right\} \\ &+ \sum_{t=0}^{\infty} \mu_t \left\{ \eta \left[\lambda D_t^{\frac{\varepsilon-1}{\varepsilon}} + (1-\lambda)(\bar{d}L)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\alpha\varepsilon}{\varepsilon-1}} \left(\frac{L_{A,t}}{M}\right)^{1-\alpha} + (1-\delta_D)D_t - D_{t+1} \right\}\end{aligned}$$

Since $d_t = \bar{d}$ is not a choice variable, the firm does not take a first-order condition with respect to d_t. The first-order conditions with respect to K_t, $L_{A,t}$, and D_t are identical in form to Equations (A.46), (A.49), and (A.50), evaluated at $d_t = \bar{d}$.

The consumer data price is determined residually by the AI firms' aggregate willingness to pay. In the symmetric steady state, each of the M^{dr} firms demands consumer data at the marginal revenue product evaluated at $\bar{d}$, and the intermediary condition $M^{dr} p_d^{d,dr} = p_d^{dr}$ determines the market price:

$$p_d^{dr} = M^{dr}\left[\left(1-\frac{1}{\rho}\right)\left(1-\frac{1}{\sigma}\right)\zeta\, b\tau\, \frac{N^{\frac{\sigma}{\sigma-1}} Y_i^{dr}}{M^{dr}\,\bar{d}L} + \frac{\mu^{dr}}{\lambda_0^{dr}}\,\alpha(1-\lambda)\eta\,(\Omega^{dr})^{\frac{\alpha\varepsilon}{\varepsilon-1}-1}(\bar{d}L)^{-\frac{1}{\varepsilon}}\left(\frac{L_A^{dr}}{M^{dr}}\right)^{1-\alpha}\right],$$

where $\lambda_0^{dr} \equiv 1/\prod_{s=0}^{t}(1+r_s)$ evaluated in the steady state (so μ^{dr}/λ_0^{dr} is the current-value shadow price of AI services) and μ^{dr} denotes the steady-state value of the corresponding multiplier. In

general, $p_d^{dr}/c^{dr} \neq \phi'(\bar{d})$: when $\bar{d} < d^{fe}$, the AI firms' willingness to pay exceeds the consumer's marginal privacy cost, generating a surplus for consumers from data sales.

Computing power provider. The computing power provider's problem is identical to that in Appendix A.3. The total compute demand has the same functional form, with $\bar{d}$ replacing d^{fe}. The steady-state optimality condition yields:

$$B^{dr} = \frac{(\sigma-1)(1-1/\rho)\zeta^2(1-b)^2\tau^2\beta\gamma\delta_K}{[1-\beta(1-\delta_K)]\sigma} = B^{fe}.$$

Steady-state characterization. Following the same algebraic steps as in Appendix A.3, but with $\bar{d}$ replacing the endogenous d, we obtain the labor allocation condition:

$$\frac{L_K^{dr}}{L - L_A^{dr} - L_K^{dr}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma} B^{dr}. \tag{A.57}$$

The AI-service Euler equation is:

$$\frac{[1-\beta(1-\delta_D)]\, D^{dr}\, M^{dr}}{(1-\alpha)\beta\eta(\Omega^{dr})^{\frac{\alpha\varepsilon}{\varepsilon-1}}(L_A^{dr}/M^{dr})^{-\alpha}(L - L_A^{dr} - L_K^{dr})} = \left(1-\frac{1}{\rho}\right)(1-\tau)\zeta + \frac{\alpha}{1-\alpha}\cdot\frac{L_A^{dr}}{L - L_A^{dr} - L_K^{dr}}\cdot\frac{\lambda(D^{dr})^{\frac{\varepsilon-1}{\varepsilon}}}{\Omega^{dr}},$$

where

$$\Omega^{dr} = \left[\frac{\delta_D D^{dr}}{\eta(L_A^{dr}/M^{dr})^{1-\alpha}}\right]^{\frac{\varepsilon-1}{\alpha\varepsilon}}. \tag{A.58}$$

The free entry condition requires that per-firm steady-state profit equals the amortized entry cost:

$$(1-\beta)F = \frac{(1-1/\sigma)\zeta\, Y^{dr}}{M^{dr}}\left[1-\left(1-\frac{1}{\rho}\right)(1-b)\tau\right] - \frac{w^{dr}L_A^{dr}}{M^{dr}} - \frac{p_d^{dr}\,\bar{d}L}{M^{dr}}, \tag{A.59}$$

where $Y^{dr} \equiv c^{dr}L + I^{dr} = N^{\sigma/(\sigma-1)}Y_i^{dr}$ is aggregate output; the first term is per-firm revenue net of compute cost, with the factor $(1-1/\rho)(1-b)\tau$ being the compute cost share of per-firm revenue implied by Equation (A.47) evaluated at $\bar{d}$, and the remaining terms are per-firm labor and consumer data costs. Unlike the free-entry regime, the binding data cap $d^{dr} = \bar{d}$ breaks the constant-returns-to-scale structure in the firm's choice variables $(K_{j,t}, L_{A,j,t})$ (since the exogenous $\bar{d}L$ enters the quality index as a fixed input with exponent sum $(1-b)\tau + (1-\tau) = 1 - b\tau < 1$), so the compact identity $(1-\beta)F = R/\rho$ available in Appendix A.3 does not apply here.

The system (A.57)–(A.59), together with (A.58) and the exogenous constraint $d^{dr} = \bar{d}$, constitutes a closed system in L_K^{dr}, L_A^{dr}, D^{dr}, and M^{dr}.

Remaining variables. Output per variety is:

$$Y_i^{dr} = \left\{A\,(M^{dr})^{\omega}(\bar{d}L)^{b\tau\zeta}(D^{dr})^{\zeta(1-\tau)}\frac{L - L_A^{dr} - L_K^{dr}}{N}\left[\xi^{\frac{1}{\gamma}}\delta_K^{-\frac{1}{\gamma}}(L_K^{dr})^{\frac{1}{\gamma}-1}N^{\frac{\sigma}{\sigma-1}}B^{dr}\right]^{(1-b)\gamma\tau\zeta}\right\}^{\frac{1}{1-(1-b)\gamma\tau\zeta}}$$

Total computing power is:

$$K^{dr} = \left[\xi^{\frac{1}{\gamma}} \delta_K^{-\frac{1}{\gamma}} (L_K^{dr})^{\frac{1}{\gamma}-1} N^{\frac{\sigma}{\sigma-1}} B^{dr}\right]^{\gamma} (Y_i^{dr})^{\gamma}$$

Data quality (common across AI firms) is:

$$Q^{dr} = \left[\frac{(\bar{d}L)^b (K^{dr}/M^{dr})^{1-b}}{D^{dr}}\right]^{\tau}$$

Steady-state social welfare is:

$$W^{dr} = \frac{1}{1-\beta}\left[\ln c^{dr} - \phi(\bar{d})\right],$$

where $c^{dr} = (1 - B^{dr})N^{\sigma/(\sigma-1)}Y_i^{dr}/L$. The welfare effect of varying $\bar{d}$ involves two opposing channels: a lower $\bar{d}$ reduces the privacy disutility $\phi(\bar{d})$ but also reduces output Y_i^{dr} (and hence consumption c^{dr}) through lower data quality and reduced AI-service generation. In addition, a lower $\bar{d}$ affects the equilibrium number of AI firms M^{dr} through the free entry condition (A.59), introducing a further indirect channel through the love-of-variety effect.

A.5 Optimal Consumer Data and Welfare-Improving Regulation

This subsection provides the detailed derivation of Proposition 5. The derivation proceeds in five steps: (i) expressing W^{dr} as a function of $\bar{d}$; (ii) establishing boundary behavior on the binding interval; (iii) deriving the first-order condition for an interior optimum; (iv) comparing W^{dr} and W^{fe} at the boundary $\bar{d} = d^{fe}$; and (v) log-linearizing output to obtain an explicit decomposition of $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d}$ and the implicit equation for $\bar{d}^*$.

Step 1: Welfare as a function of $\bar{d}$. Consumer data supply is capped at $\bar{d}$. The effective equilibrium depends on whether this cap binds, which in turn depends on the comparison between $\bar{d}$ and the unregulated free-entry level d^{fe}.

Non-binding case. If $\bar{d} \geq d^{fe}$, the consumer's unconstrained optimum d^{fe} already satisfies the cap, so the equilibrium reverts to free entry:

$$d = d^{fe}, \qquad L_A^{dr} = L_A^{fe}, \qquad L_K^{dr} = L_K^{fe}, \qquad D^{dr} = D^{fe}, \qquad M^{dr} = M^{fe}, \qquad c^{dr} = c^{fe},$$

and the privacy cost is $\phi(d^{fe}) = \phi\,(d^{fe})^2$. Welfare therefore equals W^{fe} for every $\bar{d} \geq d^{fe}$.

Binding case. If $\bar{d} < d^{fe}$, the cap is strictly binding and $d = \bar{d}$. The DR equilibrium is characterized by the four equations in Proposition 4: the labor allocation (39) pins down L_K^{dr}/L_p^{dr} via B^{dr}; the AI-service Euler equation (40) links L_A^{dr}, D^{dr}, and M^{dr}; the free entry condition (41) determines M^{dr}; and the steady-state Ω^{dr} consistency condition links D^{dr}, L_A^{dr}, M^{dr}, and $\bar{d}$. This system jointly

determines $(L_A^{dr}, L_K^{dr}, D^{dr}, M^{dr})$ as functions of $\bar{d}$. Output follows from (42) and consumption from

$$c^{dr}(\bar{d}) = \frac{(1 - B^{dr})\, N^{\sigma/(\sigma-1)}\, Y_i^{dr}(\bar{d})}{L}, \tag{A.60}$$

with B^{dr} independent of $\bar{d}$ (it depends only on deep parameters, as shown by the formula for $B^{dr} = B^{fe}$ in Appendix A.4).

Combining both cases:

$$W^{dr}(\bar{d}) = \begin{cases} \dfrac{1}{1-\beta}\left[\ln c^{dr}(\bar{d}) - \phi(\bar{d})\right] & \text{if } \bar{d} < d^{fe}, \\ W^{fe} & \text{if } \bar{d} \geq d^{fe}. \end{cases} \tag{A.61}$$

Continuity at $\bar{d} = d^{fe}$. The upper and lower expressions coincide at $\bar{d} = d^{fe}$. To see this, substitute $d = d^{fe}$ into the DR system: the resulting equations are identical to the FE system in Proposition 3 (with $\bar{d}$ in place of the endogenous d, but the imposed value happens to equal the FE equilibrium). The unique solution is therefore the FE allocation, and (A.60) gives $c^{dr}(d^{fe}) = c^{fe}$. The upper expression of (A.61) evaluates to $(1-\beta)^{-1}[\ln c^{fe} - \phi(d^{fe})] = W^{fe}$, matching the lower expression.

Step 2: Boundary behavior on the binding interval. We examine the behavior of $W^{dr}(\bar{d})$ on the relevant interval $(0, d^{fe}]$.

Limit at the right endpoint. By continuity established in Step 1, $\lim_{\bar{d}\to d^{fe-}} W^{dr}(\bar{d}) = W^{fe}$.

Limit at the left endpoint. As $\bar{d} \to 0^+$, we show that $c^{dr}(\bar{d}) \to 0$, hence $\ln c^{dr} \to -\infty$, and consequently $W^{dr}(\bar{d}) \to -\infty$.

The argument proceeds by tracing the dependence of output on $\bar{d}$ through the production chain. Data quality is

$$Q^{dr} = \left[\frac{(\bar{d}L)^b (K^{dr}/M^{dr})^{1-b}}{D^{dr}}\right]^{\tau}.$$

The exponent $b\tau > 0$ implies that for any bounded K^{dr}/M^{dr} and any bounded D^{dr}, $Q^{dr} \to 0$ as $\bar{d} \to 0^+$. The effective data input is

$$\Phi^{dr} = (M^{dr})^{\rho/(\rho-1)}\, Q^{dr}\, D^{dr},$$

so Φ^{dr} also tends to zero at the same rate. Output satisfies $Y_i^{dr} \propto (\Phi^{dr})^{\zeta/\Theta}$ up to terms involving $L_A^{dr}, L_K^{dr}, M^{dr}$, all of which remain bounded as $\bar{d} \to 0^+$ (they are constrained by $L_A + L_K \leq L$ and the free entry condition, which prevents M^{dr} from diverging). Therefore $Y_i^{dr} \to 0$ and $c^{dr} \to 0$. Since $\phi(\bar{d}) \to \phi(0)$ (a finite constant) as $\bar{d} \to 0^+$, the dominant term is $\ln c^{dr} \to -\infty$, giving

$$\lim_{\bar{d}\to 0^+} W^{dr}(\bar{d}) = \lim_{\bar{d}\to 0^+} \frac{1}{1-\beta}\left[\ln c^{dr}(\bar{d}) - \phi(\bar{d})\right] = -\infty.$$

Combining the two limits: on $(0, d^{fe}]$, W^{dr} is continuous, takes the value $-\infty$ at the left endpoint and W^{fe} at the right endpoint. Whether W^{dr} exceeds W^{fe} at some interior point depends on local

properties of the function, to which we now turn.

Step 3: First-order condition on the binding interval. On $(0, d^{fe})$ where the cap binds, differentiating (A.61) with respect to $\bar{d}$:

$$\begin{aligned}\frac{\mathrm{d}W^{dr}}{\mathrm{d}\bar{d}} &= \frac{1}{1-\beta}\left[\frac{1}{c^{dr}}\frac{\mathrm{d}c^{dr}}{\mathrm{d}\bar{d}} - \phi'(\bar{d})\right] \\ &= \frac{1}{1-\beta}\left[\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}} - \phi'(\bar{d})\right].\end{aligned} \tag{A.62}$$

An interior critical point $\bar{d}^* \in (0, d^{fe})$ must satisfy $\mathrm{d}W^{dr}/\mathrm{d}\bar{d} = 0$, or equivalently

$$\left.\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}}\right|_{\bar{d}=\bar{d}^*} = \phi'(\bar{d}^*). \tag{A.63}$$

At the optimum, the general-equilibrium marginal effect of the cap on (log) consumption equals the marginal privacy cost.

Step 4: Comparison with FE at the boundary. To determine the existence of a welfare-improving cap, evaluate (A.62) as $\bar{d} \to d^{fe-}$. The FE consumer's first-order condition $\phi'(d^{fe}) = p_d^{fe}/c^{fe}$ gives

$$\left.\frac{\mathrm{d}W^{dr}}{\mathrm{d}\bar{d}}\right|_{\bar{d}=d^{fe-}} = \frac{1}{1-\beta}\left[\left.\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}}\right|_{\bar{d}=d^{fe-}} - \frac{p_d^{fe}}{c^{fe}}\right]. \tag{A.64}$$

Assume W^{dr} is continuously differentiable and strictly quasi-concave on $(0, d^{fe})$. We distinguish two cases.

Case (a): $\mathrm{d}W^{dr}/\mathrm{d}\bar{d}|_{d^{fe-}} \geq 0$**.** The welfare function is non-decreasing as $\bar{d}$ approaches the right endpoint. By strict quasi-concavity, $W^{dr}(\bar{d}) \leq W^{dr}(d^{fe-}) = W^{fe}$ for all $\bar{d} \in (0, d^{fe})$. No binding cap improves welfare, and the optimal choice is $\bar{d} \geq d^{fe}$ (i.e., no binding regulation).

Case (b): $\mathrm{d}W^{dr}/\mathrm{d}\bar{d}|_{d^{fe-}} < 0$**.** The welfare function is strictly decreasing as $\bar{d} \to d^{fe-}$. Hence for some $\varepsilon > 0$, $W^{dr}(d^{fe} - \varepsilon) > W^{fe}$. Combined with $W^{dr}(\bar{d}) \to -\infty$ as $\bar{d} \to 0^+$, strict quasi-concavity and the intermediate value theorem imply the existence of a unique interior maximum $\bar{d}^* \in (0, d^{fe})$ satisfying (A.63), with $W^{dr}(\bar{d}^*) > W^{fe}$.

Equation (A.64) shows that Case (b) is equivalent to

$$\left.\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}}\right|_{\bar{d}=d^{fe-}} < \frac{p_d^{fe}}{c^{fe}},$$

which is condition (43) in Proposition 5. Thus condition (43) is both necessary and sufficient for the existence of a welfare-improving binding cap.

Step 5: Log-linearization of output and decomposition of $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d}$**.** To obtain an explicit characterization of the left-hand side of (43), we log-linearize the output expression (42). Taking the

logarithm of both sides:

$$\ln Y_i^{dr} = \frac{1}{\Theta}\left\{\ln A + \omega \ln M^{dr} + b\tau\zeta \ln\left(\bar{d}L\right) + \zeta(1-\tau)\ln D^{dr} + \ln\frac{L_p^{dr}}{N} + (1-b)\gamma\tau\zeta \cdot \ln \Xi_t\right\},$$

where $\Theta \equiv 1-(1-b)\gamma\tau\zeta$ (assumed positive, which holds under plausible parameter values since $(1-b)\gamma\tau\zeta$ is typically small), $L_p^{dr} \equiv L - L_A^{dr} - L_K^{dr}$ is production labor, and

$$\Xi_t \equiv \xi^{1/\gamma}\delta_K^{-1/\gamma}(L_K^{dr})^{1/\gamma-1}N^{\sigma/(\sigma-1)}B^{dr}.$$

Since B^{dr} is independent of $\bar{d}$, all $\bar{d}$-dependence of Ξ_t flows through L_K^{dr}: $\ln \Xi_t = (1/\gamma-1)\ln L_K^{dr} + \text{const}$. Substituting and multiplying by Θ:

$$\Theta \ln Y_i^{dr} = \text{const} + \omega_M \ln M^{dr} + \omega_d \ln \bar{d} + \omega_D \ln D^{dr} + \omega_{L_p}\ln L_p^{dr} + \omega_{L_K}\ln L_K^{dr}, \tag{A.65}$$

with

$$\omega_M = \omega = \frac{\rho\zeta}{\rho-1} - (1-b)\tau\zeta, \quad \omega_d = b\tau\zeta, \quad \omega_D = \zeta(1-\tau), \quad \omega_{L_p} = 1, \quad \omega_{L_K} = (1-b)\gamma\tau\zeta\cdot\left(\frac{1}{\gamma}-1\right) = (1-b)\tau\zeta(1-\gamma).$$

Labor constraint. The labor allocation (39) gives

$$\frac{L_K^{dr}}{L_p^{dr}} = \frac{(1-\gamma)\sigma}{(\sigma-1)\gamma}B^{dr} \equiv \kappa,$$

a constant independent of $\bar{d}$. Hence $L_K^{dr} = \kappa L_p^{dr}$, or equivalently $\ln L_K^{dr} = \ln L_p^{dr} + \ln\kappa$. Together with $L = L_A^{dr} + L_K^{dr} + L_p^{dr}$, we obtain

$$L_p^{dr} = \frac{L - L_A^{dr}}{1+\kappa}, \qquad L_K^{dr} = \frac{\kappa(L - L_A^{dr})}{1+\kappa}.$$

Differentiating $\ln L_p^{dr} = \ln\left(L - L_A^{dr}\right) - \ln(1+\kappa)$ with respect to $\bar{d}$:

$$\frac{\mathrm{d}\ln L_p^{dr}}{\mathrm{d}\bar{d}} = -\frac{1}{L - L_A^{dr}}\cdot\frac{\mathrm{d}L_A^{dr}}{\mathrm{d}\bar{d}} = -\frac{L_A^{dr}}{L - L_A^{dr}}\cdot\frac{\mathrm{d}\ln L_A^{dr}}{\mathrm{d}\bar{d}} = \frac{\xi_L}{\bar{d}}\,\eta_{L_A},$$

where $\xi_L \equiv -L_A^{dr}/(L - L_A^{dr}) < 0$ and $\eta_{L_A} \equiv \mathrm{d}\ln L_A^{dr}/\mathrm{d}\ln\bar{d}$. Since $\ln L_K^{dr} = \ln L_p^{dr} + \text{const}$, the same relation holds for L_K^{dr}.

Decomposition. Differentiating (A.65) with respect to $\bar{d}$ and using $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d} = \mathrm{d}\ln Y_i^{dr}/\mathrm{d}\bar{d}$ (since $c^{dr} \propto Y_i^{dr}$ by (A.60)):

$$\Theta\cdot\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}} = \omega_M\cdot\frac{\mathrm{d}\ln M^{dr}}{\mathrm{d}\bar{d}} + \frac{\omega_d}{\bar{d}} + \omega_D\cdot\frac{\mathrm{d}\ln D^{dr}}{\mathrm{d}\bar{d}} + (\omega_{L_p}+\omega_{L_K})\cdot\frac{\mathrm{d}\ln L_p^{dr}}{\mathrm{d}\bar{d}}.$$

Substituting the labor constraint and multiplying through by $\bar{d}$:

$$\Theta\bar{d}\cdot\frac{\mathrm{d}\ln c^{dr}}{\mathrm{d}\bar{d}} = \underbrace{\omega_d}_{\text{direct}} + \underbrace{\omega_M\,\eta_M}_{\text{entry}} + \underbrace{\omega_D\,\eta_D}_{\text{AI-service stock}} + \underbrace{(\omega_{L_p}+\omega_{L_K})\,\xi_L\,\eta_{L_A}}_{\text{labor reallocation}}, \tag{A.66}$$

where $\eta_X \equiv \mathrm{d}\ln X^{dr}/\mathrm{d}\ln\bar{d}$ for $X \in \{M, D, L_A\}$.

Sign analysis of each channel.

- The direct effect $\omega_d = b\tau\zeta > 0$: a higher $\bar{d}$ raises data quality through the $(\bar{d}L)^{b\tau\zeta}$ factor.
- The entry effect $\omega_M\eta_M$: the net love-of-variety elasticity $\omega_M = \omega = \rho\zeta/(\rho-1) - (1-b)\tau\zeta > 0$ for $\rho > 1$, since $\rho/(\rho-1) > 1 > (1-b)\tau$. This elasticity nets the positive love-of-variety gain $\rho\zeta/(\rho-1)$ against the compute-dilution loss $(1-b)\tau\zeta$ (spreading total compute across more firms reduces per-firm data quality). The entry elasticity $\eta_M < 0$ in a neighborhood of d^{fe}, as can be seen from the free entry condition (41): the per-firm consumer data cost $p_d^{dr}\bar{d}L/M^{dr}$ is a major $\bar{d}$-sensitive component, and reducing $\bar{d}$ relaxes the entry constraint, permitting more firms to enter. Hence this term is negative.
- The AI-service stock effect $\omega_D\eta_D$: $\omega_D = \zeta(1-\tau) > 0$; the sign of η_D is ambiguous in general, as it is determined jointly by the AI-service Euler equation and the Ω^{dr} consistency relation. In our numerical calibration, $\eta_D > 0$ in a neighborhood of d^{fe} (comparing $D^{fe} = 2.44$ at $\bar{d} = 0.066$ with $D^{dr} = 2.01$ at $\bar{d} = 0.05$: tightening the cap lowers D^{dr}), so this term is positive and reinforces the direct effect.
- The labor reallocation term $(\omega_{L_p}+\omega_{L_K})\xi_L\eta_{L_A}$: $(\omega_{L_p}+\omega_{L_K}) = 1+(1-b)\tau\zeta(1-\gamma) > 0$ and $\xi_L < 0$; the sign of η_{L_A} follows from how labor is reallocated between data generation and production as the cap varies and is generally ambiguous without further calculation.

The direct effect and the AI-service stock effect are positive; the entry effect is unambiguously negative. Whether the total can become negative, which is required for condition (43), depends on whether the entry effect is large enough in magnitude to offset the other terms. The key parameter is ρ: small ρ raises ω_M and amplifies the entry effect. In our numerical calibration with $\rho = 2$, the entry effect dominates by a modest margin: the numerical results show $c^{dr}(0.05) > c^{fe}$, which implies $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d} < 0$ near d^{fe}, so condition (43) is satisfied.

Explicit FOC. Substituting (A.66) evaluated at $\bar{d}^*$ into the first-order condition $\mathrm{d}\ln c^{dr}/\mathrm{d}\bar{d}|_{\bar{d}^*} = \phi'(\bar{d}^*)$ and multiplying by $\Theta\bar{d}^*$:

$$\phi'(\bar{d}^*)\,\Theta\,\bar{d}^* = b\tau\zeta + \omega\,\eta_M(\bar{d}^*) + \zeta(1-\tau)\eta_D(\bar{d}^*) + [1+(1-b)\tau\zeta(1-\gamma)]\,\xi_L(\bar{d}^*)\,\eta_{L_A}(\bar{d}^*).$$

This is an implicit equation in $\bar{d}^*$ because the elasticities $\eta_X(\bar{d}^*)$ themselves depend on $\bar{d}^*$ through the four-equation DR system. In principle, one could compute the η_X by log-linearizing the four DR equations (labor allocation, AI-service Euler equation, free entry condition, and the Ω^{dr} consistency

relation) around the equilibrium, yielding a linear system in the elasticities. In the quantitative analysis of Section 4.2, we instead solve the full DR system directly.

A.6 Closed-Form Solution for M_t along Transition Paths

We derive the closed-form expression for the number of AI firms M_t used in the transition analysis of Section 5, which determines M_t each period from the static free-entry condition given the current state variables and the flow values that have jumped to their new-regime steady state.

Along the transition, the computing stock K_t and the stock of AI services D_t are predetermined state variables (inherited from the previous period through their accumulation equations), and the flow variables $L_{A,t}$, $L_{K,t}$, d_t are assumed to take their new-regime steady-state values from $t = 0$ onward. Given $(D_t, K_t, L_{A,t}, L_{K,t}, d_t)$, the data quality and effective data input are

$$Q_t = \left[\frac{(d_t L)^b (K_t/M_t)^{1-b}}{D_t}\right]^\tau, \qquad \Phi_t = M_t^{\rho/(\rho-1)}\, Q_t\, D_t.$$

Substituting the expression for Q_t into Φ_t:

$$\Phi_t^\zeta = M_t^\omega\, (d_t L)^{b\tau\zeta}\, K_t^{(1-b)\tau\zeta}\, D_t^{(1-\tau)\zeta},$$

where $\omega \equiv \rho\zeta/(\rho-1) - (1-b)\tau\zeta$ is the net love-of-variety elasticity defined in (22). Per-variety output is $Y_{i,t} = A\Phi_t^\zeta \cdot L_{p,t}/N$ with $L_{p,t} \equiv L - L_{A,t} - L_{K,t}$, and aggregate output is $Y_t = N^{\sigma/(\sigma-1)} Y_{i,t}$. Writing

$$Y_0(D_t, K_t, d_t, L_{p,t}) \equiv A\,(d_t L)^{b\tau\zeta}\, K_t^{(1-b)\tau\zeta}\, D_t^{(1-\tau)\zeta}\, \frac{L_{p,t}}{N},$$

we have $Y_{i,t} = Y_0 \cdot M_t^\omega$ and $Y_t = N^{\sigma/(\sigma-1)}\, Y_0\, M_t^\omega$. From the wage equation $w_t = (1-1/\sigma) N^{1/(\sigma-1)} A\Phi_t^\zeta$ and $A\Phi_t^\zeta = Y_{i,t} N/L_{p,t}$, we obtain $w_t = (1-1/\sigma)\, Y_t/L_{p,t}$.

The per-period free-entry condition $(1-\beta)F = \pi_t$ with

$$\pi_t = \frac{(1-1/\sigma)\zeta Y_t}{M_t}\left[1 - (1-1/\rho)(1-b)\tau\right] - \frac{w_t L_{A,t}}{M_t} - \frac{p_{d,t}^s\, d_t L}{M_t}$$

can be multiplied through by M_t and, after substituting the expression for w_t and the consumer data price $p_{d,t}^s = 2\phi d_t c_t$ with $c_t = (1-B) N^{\sigma/(\sigma-1)} Y_{i,t}/L$, rearranged to give

$$(1-\beta) F\, M_t = K_1(L_{A,t}, L_{p,t}, d_t) \cdot N^{\sigma/(\sigma-1)}\, Y_0 \cdot M_t^\omega,$$

where

$$K_1(L_{A,t}, L_{p,t}, d_t) \equiv \left(1 - \tfrac{1}{\sigma}\right)\left[\zeta\Theta' - \frac{L_{A,t}}{L_{p,t}}\right] - 2\phi\, d_t^2\,(1-B),$$

and $\Theta' \equiv 1 - (1-1/\rho)(1-b)\tau$ is the computing-cost correction factor familiar from (37). Since K_1 depends only on the flow variables (which have jumped to their new-regime values at $t = 0$) and

not on M_t, the free-entry condition reduces to

$$M_t = \left[\frac{K_1 \cdot N^{\sigma/(\sigma-1)} Y_0(D_t, K_t)}{(1-\beta)F} \right]^{1/(1-\omega)}. \tag{A.67}$$

Thus M_t is a simple power function of the current state (D_t, K_t), available without invoking a nonlinear solver each period. Given M_t, the period's output $Y_{i,t}$, consumption c_t, investment I_t, and AI-service flow $\tilde{D}_t$ follow from their respective definitions, and the stock variables advance via the accumulation equations.

Under entry deregulation (RE to FE), the same formula applies because the free-entry condition takes effect at $t = 0$; the initial jump in M_t is simply (A.67) evaluated at the inherited RE stocks. Under data regulation, $d_t = \bar{d}^*$ replaces d^{fe} in Y_0 and K_1; under the compute subsidy, B is replaced by $B(\tau_s)$ and Y_0 is unchanged in functional form. The exponent $1/(1-\omega)$ governs the sensitivity of entry to the state: because $1-\omega$ is a stable positive number for plausible parameter values, M_t is well-defined and varies smoothly along the transition.

B. Supplementary Comparative Statics Figures

This appendix presents the complete comparative statics results for all variables across the four regimes. Each figure displays the steady-state values of consumer data d, AI services D, data quality Q, output Y, computing power K (or number of AI firms M), and consumption c as functions of one parameter, holding the other parameters at their benchmark values. We discuss the key patterns below.

Figure B.1 reveals several interesting patterns in the response to the data quality parameter τ. First, consumer data usage d rises with τ in all regimes except data regulation (where $\bar{d}$ is fixed at 0.05 by construction), reflecting the increased marginal value of real-world data through the quality channel. The ranking $d^{sp} > d^{fe} > d^{re}$ is preserved across the entire range. Second, the stock of AI services D declines sharply with τ in all regimes, because as the quality channel strengthens, agents substitute away from accumulated AI output (which enters output through the $(1-\tau)$ weight) toward higher-quality alternatives generated from consumer data and computing. Third, output Y rises with τ only under SP, while it is approximately flat or slightly declining under FE and DR and clearly declining under RE, illustrating how the decentralized regimes fail to benefit from the strengthening quality channel because their computing investment remains suppressed. Fourth, the number of AI firms M^{sp} rises modestly from around 28 to 33 as τ increases, while M^{fe} and M^{dr} remain modest (in the range 6–8); the planner's response reflects the increased return to AI-service variety when data quality matters more for production. Finally, consumption c mirrors the output pattern, with SP dominant throughout and the RE curve clearly declining.

Figure B.2 shows how the allocation responds to the substitutability parameter ρ. Consumer data d is approximately flat under SP and RE (since ρ does not enter their consumer data supply equations directly) and rises modestly under FE as ρ increases; DR is pinned at $\bar{d} = 0.05$ by construction. The stock of AI services D rises monotonically with ρ across all regimes that feature endogenous entry,

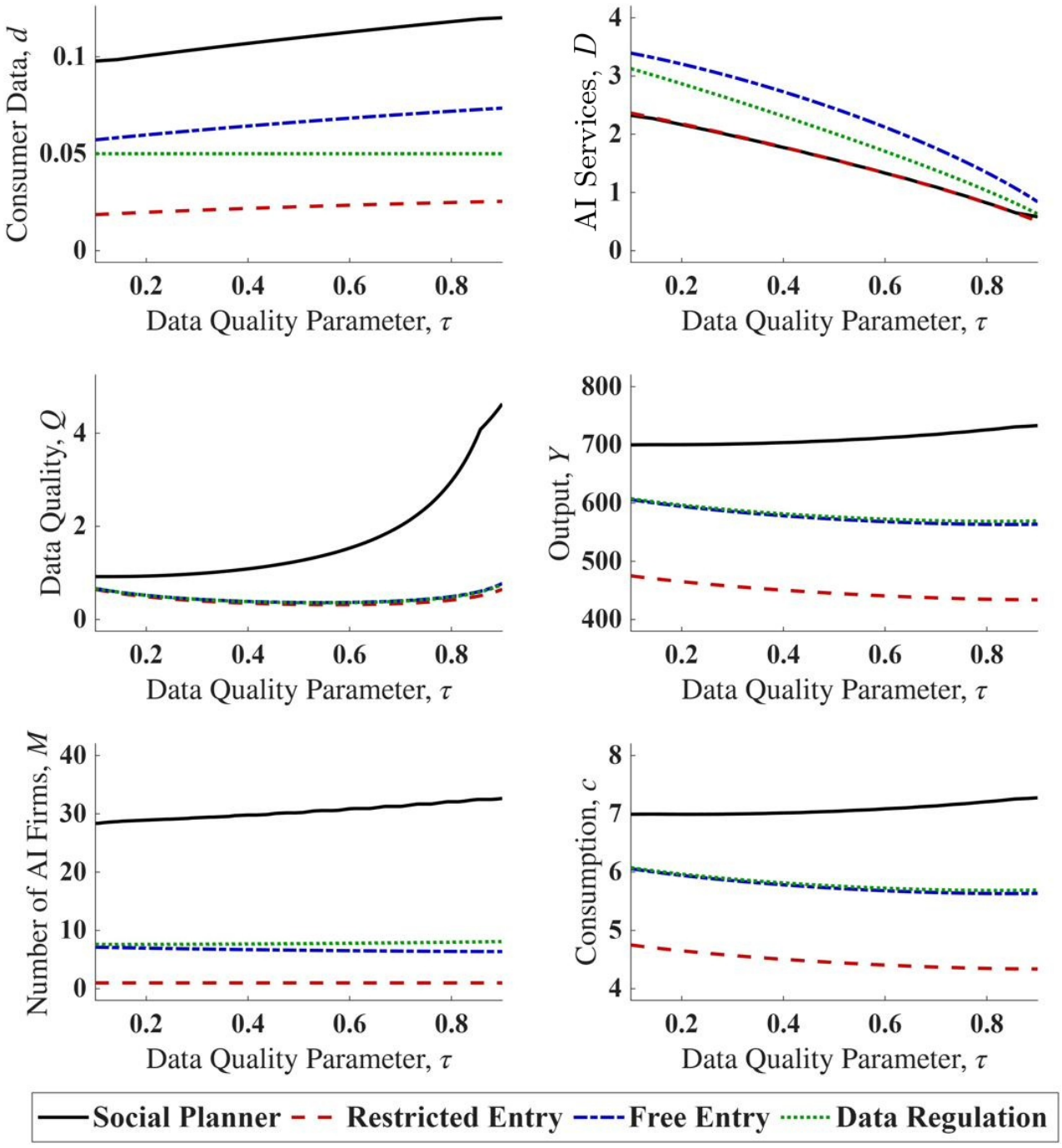


Figure B.1. Comparative statics with respect to τ (data quality parameter).

Notes. The six panels plot steady-state values of consumer data d, AI services D, data quality Q, output Y, computing power K, and consumption c as τ varies over $[0.1, 0.9]$, with the remaining parameters fixed at $\rho = 2$ and $F = 200$. Solid black: Social Planner (SP); dashed red: Restricted Entry (RE); dash-dot blue: Free Entry (FE); dotted green: Data Regulation (DR) at $\bar{d} = 0.05$. DR's consumer data is pinned at $\bar{d}$ by construction; the other three regimes reflect the endogenous response of data supply, the stock of AI services, and compute investment to a stronger quality channel.

with D^{fe} exhibiting the steepest increase because higher substitutability weakens AI firms' markup

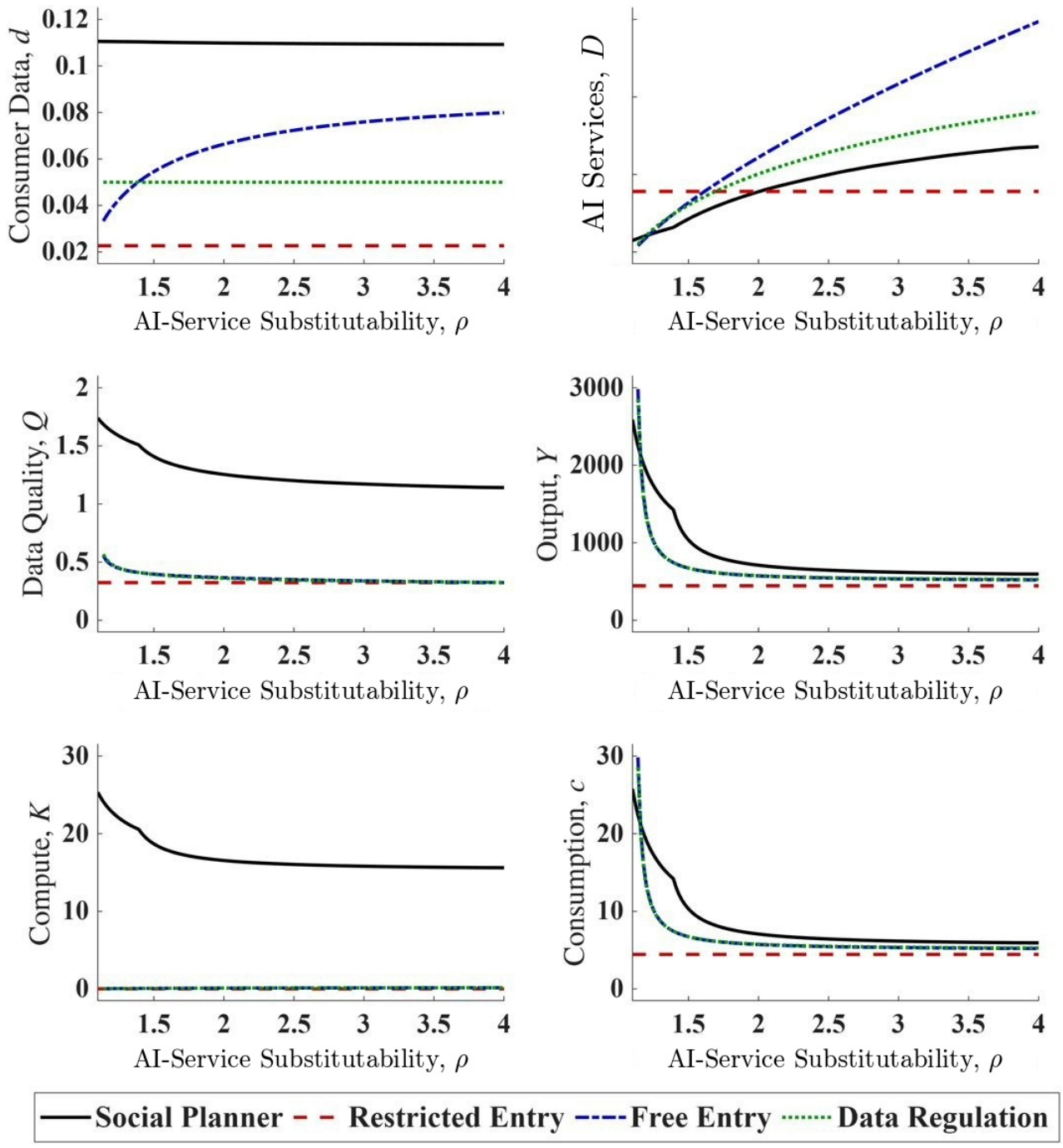


Figure B.2. Comparative statics with respect to ρ (AI-service substitutability).

Notes. The six panels plot steady-state values of consumer data d, AI services D, data quality Q, output Y, computing power K, and consumption c as ρ varies over $[1.1, 4]$, with the remaining parameters fixed at $\tau = 0.5$ and $F = 200$. Solid black: Social Planner (SP); dashed red: Restricted Entry (RE); dash-dot blue: Free Entry (FE); dotted green: Data Regulation (DR) at $\bar{d} = 0.05$. Smaller ρ corresponds to greater product differentiation among AI firms and hence larger love-of-variety gains from additional entrants; as $\rho \to 1$, the markup $\rho/(\rho-1)$ explodes and per-firm rents support very large numbers of AI firms in the endogenous-entry regimes.

and hence shifts production toward larger per-firm stocks. Data quality Q declines under SP as ρ increases (since the planner substitutes away from per-firm quality toward greater variety at low ρ) and is approximately constant across competitive equilibria. Output Y and consumption c decline steeply with ρ in all regimes with endogenous entry, with the steepest decline in FE and DR at very low ρ, where markups explode, entry surges, and love-of-variety gains are enormous, pushing Y above its social-planner counterpart. Despite this, SP maintains the highest welfare throughout because the planner balances the love-of-variety gain against the privacy cost of high data provision, whereas competitive equilibria over-provide output relative to the welfare-optimal allocation. Computing power K^{sp} declines accordingly with ρ, while the competitive K values remain negligibly small throughout.

Figure B.3 shows the response to the entry cost F. Consumer data usage is approximately constant within each regime: $d^{sp} \approx 0.11$, $d^{fe} \approx 0.065$, $\bar{d} = 0.05$, and $d^{re} \approx 0.02$, with mild declines under SP and FE as F rises. This near-invariance of d to F reflects the fact that the consumer data supply equation does not depend directly on F; F affects d only indirectly through equilibrium feedback. The stock of AI services D rises with F under SP, FE, and DR, because as fewer AI firms enter (lower M), each surviving firm accumulates more AI output to sustain aggregate production; under RE, D is flat since $M = 1$ regardless of F. Data quality Q declines gently under SP (driven by lower K per firm) and is approximately constant in competitive equilibria. Output Y and consumption c decline with F in all regimes with endogenous entry, with the FE and DR curves truncating near $F \approx 1{,}000$, where the free-entry condition can no longer be satisfied. The number of AI firms M^{sp} falls sharply from around 155 at $F = 30$ to below 5 at $F = 1{,}000$, while M^{fe} and M^{dr} start near 40–45 at $F = 30$ and collapse to 1 as the truncation points are approached. Throughout, the persistent gap $M^{sp} \gg M^{fe}$ confirms that private entry is inefficiently low across the entire range of plausible entry costs.

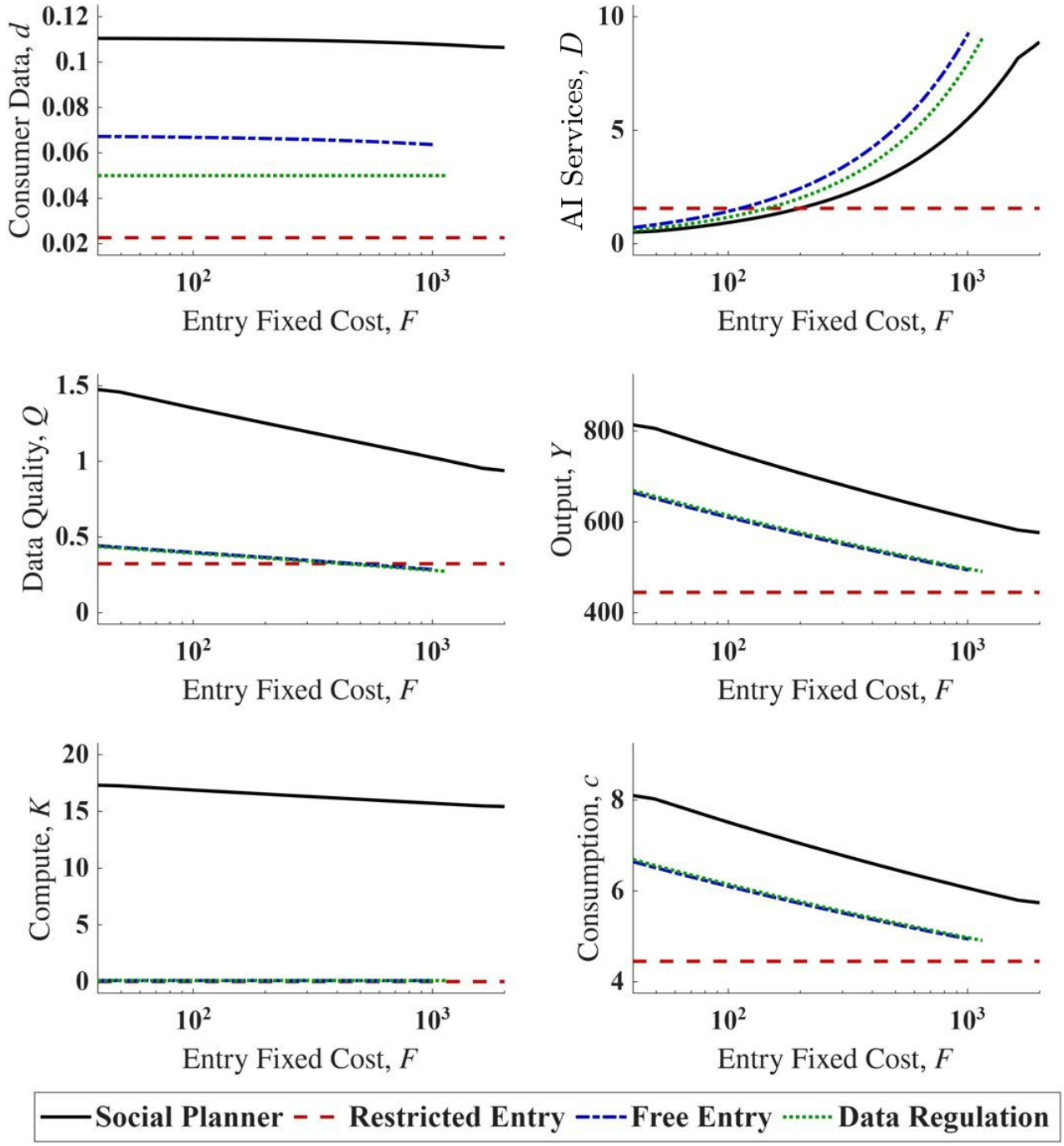


Figure B.3. Comparative statics with respect to F (entry cost).

Notes. The six panels plot steady-state values of consumer data d, AI services D, data quality Q, output Y, computing power K, and consumption c as F varies over $[30, 1{,}000]$ on a log scale, with the remaining parameters fixed at $\tau = 0.5$ and $\rho = 2$. Solid black: Social Planner (SP); dashed red: Restricted Entry (RE); dash-dot blue: Free Entry (FE); dotted green: Data Regulation (DR) at $\bar{d} = 0.05$. RE is independent of F by construction since $M^{re} = 1$; the FE and DR curves truncate near $F \approx 1{,}000$, where per-firm profits fall below the amortized entry cost $(1 - \beta)F$ and the free-entry equilibrium ceases to exist.