Generative AI Meets Data Quality: Innovation or Risk?*

Qing Chang† Danxia Xie‡ Longtian Zhang§

December 27, 2025

Abstract

The widespread adoption of Generative AI raises concerns about potential risks, particularly those arising from excessive reliance on AI. This paper examines both the benefits and drawbacks of this emerging technology through the lens of data quality. We develop a semi-endogenous growth model in which production depends on two types of data: AI-generated data and producer data, the latter representing real-world information. Although AI-generated data are substantially cheaper to produce, their use involves a trade-off in the form of lower data quality, which leads to higher error rates in production. Our analysis shows that firms, operating under competitive equilibrium, tend to underutilize both types of data relative to the optimal allocation. We further demonstrate that, while multiple Generative AI firms exist in the market, the optimal number is one. These findings support the case for government intervention in the AI industry.

Keywords: Data factor, data quality, Generative AI, economic growth

JEL Classification: H43, I31, O31, and O41

---

*All the authors contribute equally to this work. The authors are especially grateful to Charles I. Jones for his detailed feedback and invaluable guidance. They also thank the conference participants at the 9th China Center for Economic Research Summer Institute at Peking University, Chinese Economists Society 2025 Annual Conference at Sun Yat-sen University, and 2026 AEA Annual Meeting. 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. 72303261, 72550001). 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

In recent years, the significance of data has been widely acknowledged by both academia and industry. As newly emerging industries that leverage data as a key factor of production and innovation—such as robot manufacturing, drones, and self-driving vehicles—continue to develop, another transformative technology, represented by Generative AI, primarily in the form of large language models (LLMs), is reshaping various aspects of our lives by producing an increasing volume of content. Trained on human-written texts and other real-world information, Generative AIs derive their understanding of language structure and meaning. They appear to possess their own judgment and can respond to (seemingly) any input provided. Since the introduction of ChatGPT, Generative AI have been regarded as one of the most significant drivers of economic growth in the coming decades. Meanwhile, as documented by Bail (2024), this technology is reshaping the social sciences by streamlining routine research tasks such as writing, data cleaning, and software programming.

The integration of AI into economic systems has transformed industries by enabling data-driven decision-making and predictive analytics, while the rapid adoption of Generative AI has led to an unprecedented surge in AI-generated content. According to projections by the European Union Law Enforcement Agency, AI-generated content could eventually account for up to 90% of online material.[1] del Rio-Chanona et al. (2023) show that the introduction of LLMs such as ChatGPT has reduced human-generated contributions on platforms like Stack Overflow, and Brooks et al. (2024) document the growing presence of AI-generated content on Wikipedia. However, the performance and reliability of AI systems critically depend on the quality of the data they employ. Poor data quality—characterized by inaccuracies, biases, or incompleteness—poses significant risks to the integrity of AI applications, potentially leading to flawed economic outcomes and systemic vulnerabilities, commonly referred to as "AI hallucinations." Li et al. (2023) introduce an index called "HaluEval" within a ChatGPT-based framework to assess the accuracy of various prevalent LLMs. In addition, the company Vectara proposes the "Hughes Hallucination Evaluation Model" to evaluate how frequently an LLM introduces hallucinations when summarizing a document, and periodically releases a leaderboard, as shown in Figure 1.[2] Meanwhile, serious security incidents arising from the misuse of Generative AI have also emerged.[3] Given the prevailing trends documented

[1]For more information, see the report titled *Facing Reality? Law Enforcement and the Challenge of Deepfakes: An Observatory Report from the Europol Innovation Lab*.

[2]For more information, please refer to https://github.com/vectara/hallucination-leaderboard. Note that this evaluation is relatively conservative, as it focuses solely on one of the most fundamental tasks typically assigned to LLMs. For more complex tasks, such as generating original content, LLMs may be even more prone to producing inaccurate information.

[3]For example, Samsung Electronics experienced three incidents of confidential information leaks within just 20 days due to employees' improper use of ChatGPT; the U.S. tech news site CNET faced widespread factual

above, the implications of data quality deficiencies remain underexplored from an academic perspective—particularly with regard to their broader economic and societal impacts.

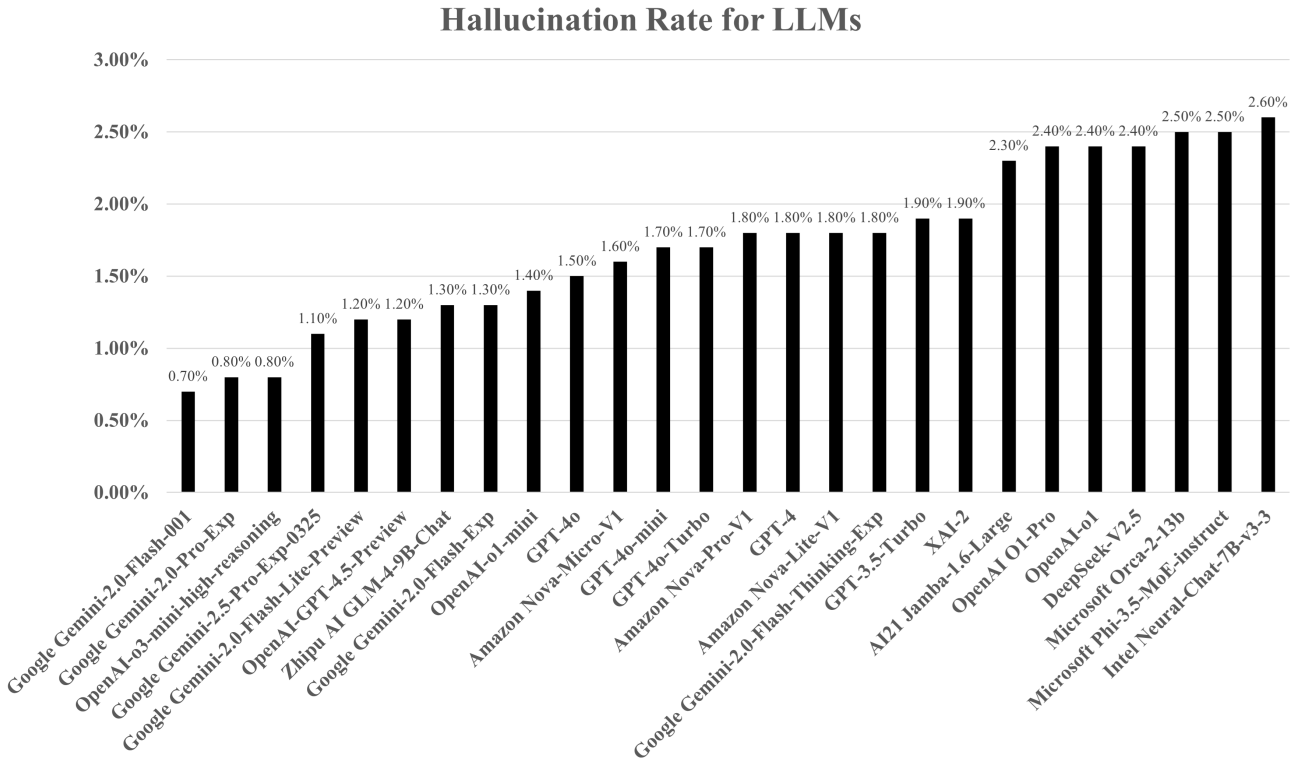


Figure 1: Hallucination Rate for Selected LLMs

Note. This figure presents the hallucination rates for the top 26 LLMs, updated as of March 25, 2025. The rates are computed using the HHEM-2.1 hallucination evaluation model developed by Vectara. This metric measures how frequently an LLM introduces hallucinations when summarizing a document.

In light of this, we propose a semi-endogenous growth model incorporating two types of data, each representing different levels of real-world information, and analyze the model's behavior with an emphasis on data quality. Specifically, we begin with a simple framework featuring only exogenous growth and AI-generated data. This new type of data is derived from existing datasets, with labor introduced as the sole marginal input in the generation process. We model AI-generated data as significantly cheaper to produce than conventional data types and as a key input in production. However, its usage increases error rates in production, thereby reducing total economic output through an additional channel. Our model thus captures a trade-off between enhancing production efficiency and elevating the likelihood of production errors.

errors after using AI software to mass-produce news articles; and iFLYTEK's AI learning device was found to have trained on inappropriate content due to a failure in its review mechanism, ultimately triggering a public backlash and causing a significant loss in the company's market value.

Furthermore, we extend our framework to a semi-endogenous growth model in which AI-generated data and producer data jointly contribute to the production process. The latter type of data originates from real-world production activities and enhances overall dataset quality, although its associated costs are higher due to the absence of the multiplier effects present in AI-generated data. We emphasize a key concept, data quality, which is determined by the ratio of the two types of data used and directly influences the error rate in production. We compare the outcomes under two settings—competitive equilibrium and optimal allocation—and find that production firms tend to underutilize producer data, even at the risk of losing their monopoly power due to excessively high error rates. Consequently, we conclude that governments should implement policies to mitigate potential negative effects, such as imposing an upper limit on AI-generated data usage or mandating its combination with real-world data by adjusting labor allocations through taxes and subsidies.

More specifically, our results can be categorized into three regimes: AI-generated data dominance, producer data dominance, and balanced growth of both data types. In the first regime, beyond the main finding that firms consistently underutilize producer data, we identify a key distinction: under the optimal allocation, the social planner assigns a large fraction of labor to the AI-generated data sector, whereas under the competitive equilibrium, the labor share in this sector remains low. We also find that the growth rate of data quality is significantly higher under the optimal allocation than under the competitive equilibrium. These results underscore the importance of regulating the use of different data types to mitigate the declining trend in data quality caused by the reliance on AI-generated data. In the second regime, although differences in labor allocation persist, AI-generated data become negligible across both settings, leading to labor concentration in the producer data sector. Finally, in the third regime, while only numerical solutions can be obtained, the results are consistent with our main conclusions, demonstrating that the model exhibits stable behavior across a wide range of parameter values.

Additionally, we examine how many Generative AI firms exist in the market and what their optimal number should be. This is a central question raised when modeling the newly introduced Generative AI sector in this paper. As the number of firms increases, fewer resources (e.g., labor) are allocated to each firm. Thus, the less concentrated the industry is, the more difficult it becomes for AI-generated data to accumulate in the economy. We find that the competitive equilibrium tends to permit multiple firms to enter the market, while the optimal allocation always involves a single firm. This result calls for further discussion on how the government should address the monopolistic tendencies in the AI industry.

The potential risks of Generative AI have emerged as a prominent topic in recent years. Inspired by Acemoglu and Lensman (2024) and Jones (2024), which provide only concise models to illustrate key features in this field, we extend their approach by developing a

more integrated framework. These studies highlight the potential risks associated with the widespread application of AI technologies, arguing that such technologies may negatively impact production or even pose existential threats to human lives. Other research in computer science, such as Shumailov et al. (2024) and Wenger (2024), also emphasizes that AI models can collapse and start generating meaningless content if real-world data are not continuously incorporated. On a broader scale, the economics of Generative AI has gained increasing attention in recent years. For example, Acemoglu (2024) estimate the potential effect of AI on total factor productivity growth, and Korinek and Vipra (2024) analyze the concentration of AI firms from the perspective of cost structures, and Brynjolfsson et al. (2025) examine the potential impact of AI on employment and find that AI can assist new workers.

We link the economics of AI with recent studies on the data economy and offer a new perspective for evaluating the advantages and disadvantages of these emerging technologies. Our framework on the data economy builds upon the foundational work of Jones and Tonetti (2020), Farboodi and Veldkamp (2023), and Cong et al. (2021). These studies all trace their origins to the seminal endogenous growth models introduced by Romer (1990) and Jones (1995), and our paper further contributes to the extensive literature on economic growth. Beyond these studies on the long-run effects of data, research on the data economy also explores micro-level aspects, such as information and privacy. For example, recent studies analyze the nonrival nature of data and highlight the competitive dynamics among digital platforms and intermediaries (Ichihashi, 2021; Yang, 2022; Bergemann and Bonatti, 2024). Our research approaches this topic from the perspective of data quality and calls for further investigation into the micro-foundations of this emerging concept.

The remainder of the paper is organized as follows. Section 2 introduces a simple model that considers only the effect of AI-generated data and derives preliminary results on their economic impact. Section 3 extends this model into a general framework by incorporating both AI-generated data and producer data. Section 4 examines the optimal allocation and provides an initial discussion of the model's implications. Section 5 then analyzes the competitive equilibrium and presents the corresponding solutions. Building on the findings from the previous two sections, Section 6 conducts comparative statics across various dimensions and explores the policy implications of this study. Finally, Section 7 concludes.

2. A Simple Model with AI-Generated Data

We begin with a simple model that features only the key type of data analyzed in this paper—AI-generated data. Suppose the economy consists of N varieties of intermediate goods, which are combined to produce final goods for consumption. The population consists

of homogeneous consumers, denoted as L_t at time t, growing at a constant rate n, with an initial population of L_0. The production of intermediate goods relies not only on labor, as is conventional in related models, but also on data. In this simplified setting, firms exclusively use AI-generated data rather than other types of data (e.g., consumer data or producer data), with these AI-generated data acting as a productivity-enhancing factor in production.

There is a single Generative AI firm in the economy that supplies AI-generated data. For simplicity, we assume that the firm must incur a substantial upfront investment before it can begin providing AI-generated data to other firms, denoted by a fixed cost $\mathcal{G}$. Let $D_{A,t}$ denote the stock of AI-generated data at time t. The generation process of this type of data is given by:

$$\boxed{\dot{D}_{A,t} = \psi D_{A,t}^{\zeta} L_{A,t} - \delta_A D_{A,t},} \tag{1}$$

where $L_{A,t}$ denotes the labor employed in the Generative AI firm at time t. The parameters ψ, ζ, and $0 < \delta_A < 1$ represent the efficiency parameter, the contribution of existing AI-generated data to the creation of new data (i.e., the spillover effect), and the depreciation rate of AI-generated data, respectively. We typically assume $0 < \zeta < 1$, as AI-generated data are expected to exhibit diminishing returns in the generation process. In other words, the production of new AI-generated data becomes increasingly difficult or inefficient as the existing stock accumulates. This accumulation formulation captures the self-generating nature of AI-generated data, which is its most distinguishing characteristic and fundamentally differentiates it from traditional capital.

The fixed cost $\mathcal{G}$ can be interpreted as capturing the emergent capabilities of AI—referring to the phenomenon whereby a large-scale AI system unexpectedly exhibits new abilities or behaviors once it surpasses a certain threshold in terms of parameters, training data, or computational resources. These capabilities are not explicitly programmed into the model and may not be observable at smaller scales. In this context, we interpret the constant $\mathcal{G}$ as the threshold at which AI models become operational and begin generating new data. Cottier et al. (2025) estimate the rising cost of training frontier AI models and report that the most expensive publicly announced training runs to date are OpenAI's GPT-4 at $40 million and Google's Gemini Ultra at $30 million.

If $D_{A,t}$ grows at a constant rate g_{D_A} in the long run—corresponding to the balanced growth path analyzed in subsequent sections—then from equation (1), we obtain:

$$\frac{\dot{D}_{A,t}}{D_{A,t}} = \psi D_{A,t}^{\zeta-1} L_{A,t} - \delta_A = g_{D_A}.$$

Thus, along the balanced growth path, the term $D_{A,t}^{\zeta-1}L_{A,t}$ must remain constant, implying:

$$(\zeta-1)g_{D_A}+n=0 \quad \Rightarrow \quad g_{D_A}=\frac{1}{1-\zeta}n. \tag{2}$$

We then derive the steady-state level of AI-generated data as:

$$D_{A,t}^{*}=\left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{1}{1-\zeta}}(L_{A,t}^{*})^{\frac{1}{1-\zeta}}. \tag{3}$$

This result shows that $D_{A,t}$ is directly related to $L_{A,t}$ with a proportionality factor in the long run. Moreover, the elasticity of $D_{A,t}$ with respect to $L_{A,t}$ is given by $1/(1-\zeta)$, which exceeds 1 under standard parameter values of ζ. This reflects the low cost of AI-generated data, a key feature in the subsequent analysis.

Next, we examine the usage of AI-generated data. Following the framework of Jones and Tonetti (2020), the production function for variety i of the intermediate good at time t, denoted as $Y_{i,t}$, is given by:

$$Y_{i,t}=D_{A,t}^{\eta}L_{i,t}, \tag{4}$$

where $L_{i,t}$ represents the labor employed in production, and the parameter η captures the importance of data in this process. Due to the nonrival nature of data, $D_{A,t}$ can be utilized not only in its own generation process but also in the production of intermediate goods. However, unlike other types of data studied in previous literature, AI-generated data are inherently prone to errors, as highlighted by Shumailov et al. (2024) and Wenger (2024). That is, firms relying on these data in production may encounter errors, which, rather than enhancing productivity, may lead to efficiency losses. For simplicity, we assume that the probability of errors occurring is an exogenous parameter, denoted as $0<e<1$, and focus on the variation of this parameter in subsequent sections. When an error occurs, the production of the corresponding variety of intermediate goods falls to zero for that period. Thus, the parameter e can also be interpreted as the error rate associated with AI-generated data. Aggregating over all varieties of intermediate goods affected by AI errors, i.e., within the range $[0, eN]$, the total consumption of the final good Y_t is given by the combination of the remaining $(N-eN)$ varieties of intermediate goods:

$$Y_t=\left(\int_{eN}^{N}Y_{i,t}^{\frac{\sigma-1}{\sigma}}\mathrm{d}i\right)^{\frac{\sigma}{\sigma-1}}, \tag{5}$$

where σ represents the elasticity of substitution among different varieties of intermediate goods. Since we do not assume heterogeneity among these varieties, substituting (4) into (5)

and considering the balanced growth path, we obtain:

$$\begin{aligned} Y_t^* &= [(1-e)N]^{\frac{\sigma}{\sigma-1}} Y_{i,t}^* \\ &= (1-e)^{\frac{\sigma}{\sigma-1}} N^{\frac{1}{\sigma-1}} \left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_A^*)^{\frac{\eta}{1-\zeta}} (1-l_A^*) L_t^{1+\frac{\eta}{1-\zeta}}, \end{aligned} \tag{6}$$

where l_A^* is the labor share allocated in the Generative AI firm. The second line follows from the labor market clearing condition, given by $L_{A,t} + \int_0^N L_{i,t} \mathrm{d}i \leq L_t$.

From equation (6), we derive the growth path of the economy's output level, Y_t^*, which depends on three main driving forces: population growth L_t, the nonrival nature of data $N^{\frac{1}{\sigma-1}}$, and the error rate of AI-generated data e. The first force, population growth, is ubiquitous in growth models and represents the scale of the economy. The second force, the nonrival nature of data, is also present in models of the data economy, such as Jones and Tonetti (2020), Cong et al. (2021), and Cong et al. (2022). This property introduces a multiplier effect of N^η in our simplified framework.[4] The third force, the error rate of AI-generated data, is the primary focus of our study, as it introduces a negative relationship between AI errors and economic growth. This leads to a fundamental insight of this paper: since AI-generated data are prone to errors, their application should be carefully managed, and measures should be implemented to mitigate their negative effects.

The richer model developed in the remainder of this paper builds upon this simple framework. We endogenize the variety of intermediate goods by introducing an innovation sector, leading to a semi-endogenous growth model. Additionally, we endogenize the error rate by incorporating the substitution effect arising from the use of an alternative data type—producer data—as well as other potential costs associated with mitigating the negative effects of AI-generated data.

3. The General Model

In this section, we introduce a general model with data generated from multiple sources. In addition to the AI-generated data discussed in the simple model, we also consider data derived from the production process, referred to as "producer data." The properties of producer data have been explored in emerging literature such as Farboodi and Veldkamp

[4]Note that if data were rivalrous, then equation (4) would take the form $Y_{i,t} = N^{-\eta} D_{A,t}^{\eta} L_{i,t}$. Following similar derivations, the output level along a balanced growth path would be:

$$(Y_t^*)_{\text{rival}} = (1-e)^{\frac{\sigma}{\sigma-1}} N^{\frac{1}{\sigma-1}-\eta} \left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_A^*)^{\frac{\eta}{1-\zeta}} (1-l_A^*) L_t^{1+\frac{\eta}{1-\zeta}}.$$

Thus, the multiplier effect arising from the nonrival nature of data is given by $Y_t^*/(Y_t^*)_{\text{rival}} = N^\eta$.

(2023) and Xie and Zhang (2023), and the combination of these two data types provides a more comprehensive framework for understanding the data economy.[5] We contribute to the literature by introducing the concept of data quality, which influences the overall error rate in production. Data quality is endogenously determined by firms' choices regarding the composition of different data types. We first outline the economic environment and discuss in detail the distinctive properties of AI-generated data. We then derive results under both the optimal allocation and the competitive equilibrium.

3.1 Economic Environment

Similar to the simple model, the economy consists of representative consumers, a final goods producer, intermediate goods producers, and several Generative AI firms. In addition, we introduce an innovation sector to ensure endogenous growth and incorporate various data intermediaries to facilitate data transactions in the competitive equilibrium. Some notations are reused or redefined in this section to enhance the clarity of the model.

Representative consumer. The population of the representative consumer, denoted as L_t, grows at a constant rate n, and utility is derived from the consumption of the final good. Additionally, consumers care about the quality of the goods they consume. To capture this concern, we introduce an additional term related to data quality.[6] Moreover, each consumer supplies one unit of labor inelastically and allocates it across four sectors: intermediate goods production, innovation, and data generation (which includes both producer data and AI-generated data). The utility function of consumers is specified as follows:

$$U = \int_0^\infty e^{-(\rho-n)t} \ln c_t \mathrm{d}t, \tag{7}$$

where c_t represents per capita consumption and N_t is the number of varieties. The parameters ρ corresponds to the consumers' discount rate.

Final good producer. There is a single final goods producer that assembles intermediate goods to produce the final good for consumption in a competitive environment. The

[5]Some previous studies, such as Jones and Tonetti (2020) and Cong et al. (2021), focus on data generated from consumers, referred to as "consumer data," which may introduce privacy costs. In this paper, we do not examine consumer data and instead focus on the interaction between producer data and AI-generated data for the following reason: producer data and AI-generated data share greater similarities, as neither contains personal information and thus does not pose privacy risks. For tractability, this similarity allows us to better highlight the unique characteristics of AI-generated data and the trade-offs involved in its application.

[6]If firms excessively rely on AI-generated data, leading to a high probability of errors, consumers may become dissatisfied with the goods produced and even lose confidence in AI-related industries as a whole.

production function is given by:

$$Y_t = \left(\int_0^{N_t} Y_{i,t}^{\frac{\sigma-1}{\sigma}} \mathrm{d}i \right)^{\frac{\sigma}{\sigma-1}}. \tag{8}$$

Other variables are defined analogously to those in the simple model.

In addition to production, the final goods producer also generates data as a byproduct, referred to as producer data. Following the framework of Farboodi and Veldkamp (2023) and Xie and Zhang (2023), the producer employs labor for data collection and cleaning, and the volume of producer data generated increases with the scale of production. Consequently, we specify the generation function of producer data as:

$$D_{P,t} = y_t^{\theta} L_{P,t}, \tag{9}$$

where $y_t = Y_t / L_t$ represents per capita output, $L_{P,t}$ denotes the labor allocated to this process, and the parameter θ captures the importance of output in data generation. Notably, we use per capita output y_t instead of total output Y_t in this function, as the former eliminates the scale effect of the economy, making it a more suitable reference variable rather than a direct input in the data generation process.

Generative AI firms. Unlike the simple model, we assume that there are $M > 0$ homogeneous Generative AI firms in the economy. Each firm generates AI-generated data independently, and for simplicity, we assume that data from these firms are perfect substitutes—that is, their sum constitutes the total stock of AI-generated data in the economy. As a result, the production and accumulation function of AI-generated data is given by:

$$\boxed{\dot{d}_{A,t} = \frac{\psi}{M} d_{A,t}^{\zeta} L_{A,t} - \delta_A d_{A,t},} \tag{10}$$

where $d_{A,t}$ denotes the AI-generated data produced by a single firm. Thus, the aggregate AI-generated data is given by $D_{A,t} = M d_{A,t}$. At any time t, intermediate goods producers rent these AI-generated data in a nonrival manner. As shown in the calibration analysis in Section 6.1, the depreciation rate of AI-generated data, δ_A, is typically very high (approaching 1). Consequently, the primary driver of AI-generated data growth stems from labor input and the recursive use of existing data to generate new data.

Intermediate good producers. There are N_t intermediate good producers, each specializing in a distinct variety of intermediate goods at any time t. In addition to employing labor in the production process, these producers utilize a combination of data sourced from both the final goods producer and the Generative AI firm. Formally, the overall dataset available

to an intermediate goods producer is given by:

$$\boxed{D_{i,t} = \left[\beta D_{P,i,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,i,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}},} \tag{11}$$

where $D_{i,t}$ represents the overall dataset, $D_{P,i,t}$ denotes producer data, and $D_{A,i,t}$ corresponds to AI-generated data for variety i. The parameter ε captures the elasticity of substitution between the two types of data.

From equation (10), we observe that AI-generated data accumulate by using themselves as one of the inputs, which results in a multiplier effect and lower costs compared to producer data in the data generation process. However, as discussed in the simple model, concerns regarding the quality of AI-generated data remain significant. Formally, there exists a probability $e_{i,t}$ that the output of an intermediate goods producer specializing in variety i falls to zero due to errors arising from the use of AI-generated data. This probability is given by:

$$\boxed{e_{i,t} = e_0 \cdot \exp(-\xi Q_{i,t}),} \tag{12}$$

where $Q_{i,t} \geq 0$ denotes the quality of the overall dataset for variety i at time t and ξ represents the sensitivity of quality to the error rate.[7] Furthermore, data quality $Q_{i,t}$ is determined by the ratio of producer data to AI-generated data used in forming the overall dataset, as specified by:

$$\boxed{Q_{i,t} = \left(\frac{D_{P,i,t}}{D_{A,i,t}}\right)^{\tau},} \tag{13}$$

where $\tau > 0$ denotes the elasticity of quality with respect to this ratio. When the proportion of AI-generated data in the overall dataset increases, quality $Q_{i,t}$ declines, whereas a higher share of producer data leads to an improvement in quality. This is the key variable we focus on in this paper, and we will provide further discussion in Section 3.2.

Based on the evolution of data quality discussed above, the allocations of producer data and AI-generated data are determined to optimize intermediate goods production, which inherently involves a certain risk of zero output. Formally, the production function for an intermediate goods producer specializing in variety i is given by:

$$\boxed{Y_{i,t} = (1-e_{i,t}) D_{i,t}^{\eta} L_{i,t},} \tag{14}$$

which is analogous to equation (4) in the simple model, except that the overall dataset $D_{i,t}$ is

[7] Alternatively, we could assume that the output of an intermediate goods producer falls to a level below that of a scenario without data usage. However, for simplicity and tractability, we assume zero output in the event of an error.

used instead of relying solely on AI-generated data.

Innovation sector. We follow the framework of Romer (1990) to model the innovation sector, distinguishing between data and ideas. In this paper, an idea is defined as the blueprint for creating a new variety of intermediate goods, denoted as N_t, and each blueprint is developed by employing χ units of labor. Formally, the innovation process is specified as follows:

$$\dot{N}_t = \frac{1}{\chi} L_{R,t}, \tag{15}$$

where $L_{R,t}$ represents the labor allocated to innovation, and $\chi > 0$ is a constant reflecting entry costs. As the number of intermediate goods varieties increases, the output level grows in a sustained manner, ensuring that the model follows a balanced growth path in the long run.

Loss of business. Finally, we consider the possibility that when an incumbent firm excessively relies on AI-generated data in production, leading to a high probability of errors, potential intermediate goods producers may have an opportunity to displace it. Formally, we assume that ownership of variety i changes according to a Poisson process with an arrival rate $\delta(e_{i,t})$, which is specified in quadratic form for simplicity:

$$\delta(e_{i,t}) = \delta_0 e_{i,t}^2, \tag{16}$$

where δ_0 parameterizes the arrival rate when $e_{i,t}$ reaches its maximum value, e_0. Importantly, equation (16) is not an explicit component of the economic environment. That is, while the social planner does not account for this process in decision-making, it plays a crucial role when analyzing the competitive equilibrium.

3.2 Discussions

The economic environment of the general model is summarized in Table 1. Several important issues should be clarified before we proceed to solve the model. We summarize them in the following three points. First, we clarify the selection of producer data. In our framework, in addition to AI-generated data, which serve as the primary focus of our discussion, we introduce another type of data that represent real-world information. Although producer data emerge from the production process and enhance productivity, this type of data possesses distinct properties compared to the "learning by doing" concept developed in Arrow (1962). In our setting, the generation of producer data requires labor as an input and can be traded among firms, unlike the "learning by doing" effect, which remains an internalized knowledge accumulation process within a firm. Given that producer data are directly linked

to the production process, we select this type of data to interact with AI-generated data in our analysis.

Table 1: Economic Environment

Meanings	Equations
Utility	$U = \int_0^\infty e^{-(\rho-n)t} \ln c_t \mathrm{d}t$
Final good production	$Y_t = \left(\int_0^{N_t} Y_{i,t}^{\frac{\sigma-1}{\sigma}} \mathrm{d}i \right)^{\frac{\sigma}{\sigma-1}}$, with $\sigma > 1$
Producer data generation	$D_{P,t} = y_t^\theta L_{P,t}$, with $\theta \in (0,1)$
AI-generated data accumulation	$\dot{d}_{A,t} = \frac{\psi}{M} d_{A,t}^\zeta L_{A,t} - \delta_A d_{A,t}$, with $\zeta \in (0,1)$ and $\psi > 0$
AI-generated data summation	$D_{A,t} = M d_{A,t}$, with $M > 0$
Formation of overall dataset	$D_{i,t} = \left[\beta D_{P,i,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,i,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon}{\varepsilon-1}}$, with $\varepsilon > 1$
Intermediate good production	$Y_{i,t} = (1 - e_{i,t}) D_{i,t}^\eta L_{i,t}$, with $\eta \in (0,1)$
Error rate	$e_{i,t} = e_0 \cdot \exp(-\xi Q_{i,t})$, with $e_0 > 0$ and $\xi > 0$
Quality of data	$Q_{i,t} = \left(\frac{D_{P,i,t}}{D_{A,i,t}} \right)^\tau$, with $\tau \in (0,1)$
Innovation (new varieties)	$\dot{N}_t = \frac{1}{\chi} L_{R,t}$
Labor resource constraint	$L_{P,t} + L_{A,t} + \int_0^{N_t} L_{i,t} \mathrm{d}i + L_{R,t} = L_t$
Nonrivalry of data	$D_{P,i,t} \leq D_{P,t}$ and $D_{A,i,t} \leq D_{A,t}$
Population growth (exogenous)	$L_t = L_0 e^{nt}$
Aggregate output	$Y_t = c_t L_t$
Per capita output	$y_t = Y_t / L_t$
Loss of business due to errors	$\delta(e_{i,t}) = \delta_0 e_{i,t}^2$

Secondly, the assumptions regarding Generative AI firms need to be specified. This new type of firm leverages data and computing power to develop automated solutions such as text generation, image synthesis, and code creation, significantly reducing production costs and enhancing productivity. These firms operate with high fixed costs and low marginal costs, benefiting from economies of scale and network effects. Since AI models require substantial investment in training but can be deployed at near-zero marginal cost, market dynamics often lead to oligopolistic competition, where leading firms establish dominance through data access and advanced algorithms. As a result, although we introduce a fixed cost $\mathcal{G}$ in this section, it has no effect in the long run. We also consider the general case when there

are multiple Generative AI firms in the economy for further implications on the scale of AI industry. Furthermore, we do not use real-world data (represented as producer data) as an input for generating new AI-generated data but instead consider labor as the sole input. This assumption is made because the overall dataset applied in production already consists of a combination of different types of data.

Moreover, we should further elaborate on the microfoundation of data quality. As documented in Acemoglu and Lensman (2024) and Jones (2024), the use of transformative technologies such as generative AI may increase risk, despite their potential to enhance productivity. We argue that the key underlying reason is that AI cannot generate genuinely new information about reality, nor can it make decisions based on new events occurring in the real world. As a result, AI-generated data merely replicate and reformulate past information, gradually deviating from optimal production decisions, which ultimately leads to a decline in overall data quality. To mitigate this negative effect, new information must be continuously updated by incorporating real-world data—represented as producer data in our model. This motivates our definition of data quality, which is simply expressed as the ratio of the two types of data.

4. Optimal Allocation

We first define and characterize the optimal allocation in our environment. Since we do not introduce heterogeneity among different varieties of intermediate goods and given the nonrival nature of data (including both producer data and AI-generated data), the structure of the economy can be significantly simplified. Specifically, all subscripts i in equations (11), (12), and (13) can be omitted, and equation (14) is then rewritten as:

$$Y_{i,t} = (1 - e_t) D_t^{\eta} L_{i,t},$$

indicating that each intermediate goods producer utilizes the entire dataset available in the economy. Moreover, by combining the formation of the overall dataset (11), the definition of the error rate (12), the intermediate goods production function (14), and the labor market clearing condition, the final goods production function is derived as:

$$\begin{aligned} Y_t &= N_t^{\frac{\sigma}{\sigma-1}} Y_{i,t} \\ &= \left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right). \end{aligned} \tag{17}$$

Based on the above analysis, we now formally state the social planner's problem. The key allocations to be determined are the labor shares across the four different sectors, given the

parameter values. The optimal allocation solves:

$$\max_{L_{P,t},L_{A,t},L_{R,t}} \int_0^\infty e^{-(\rho-n)t} \ln c_t \mathrm{d}t,$$

subject to:

$$c_t = y_t = \frac{Y_t}{L_t}, \tag{18}$$
$$L_t = L_0 e^{nt},$$

and the constraints given by equations (10), (9), (13), (15), and (17). To simplify the presentation of results, we introduce the following parameter:

$$\mathcal{A} \equiv \frac{(1-\zeta)\left[\rho + (1-\zeta)\delta_A\right]}{n + \delta_A(1-\zeta)}, \tag{19}$$

which is always positive given the definitions of the related parameters.

The planner seeks to allocate labor across the four sectors, with particular emphasis on the labor employed in the generation processes of the two types of data, as this allocation directly influences the error rate. The planner prefers to utilize more AI-generated data rather than producer data if the former grows at a higher rate. However, given the additional utility derived from higher data quality—which, in turn, results from a greater reliance on producer data—the planner must balance the benefits and costs associated with the use of AI-generated data. Conversely, if producer data grows at a higher rate, AI-generated data becomes negligible. As a result, the comparison of the growth rates between the two types of data plays a central role in our analysis. Accordingly, the optimal allocations under different regimes are presented in the following two propositions: Proposition 1 provides the results for the case where the two types of data grow at different rates, while we leave the results for the case where they grow at the same rate in Appendix A. In the following propositions, we use the superscript "sp" to denote variables derived in the optimal allocation.

Proposition 1. *(The Optimal Allocation When Producer Data and AI-Generated Data Grow at Different Rates) Along a balanced growth path, as the population L_t grows large, the optimal allocation converges to the following results.*

The growth rates of producer data and AI-generated data are given by:

$$g_{D_P}^{sp} = \begin{cases} \left[1 + \theta\left(\dfrac{1}{\sigma-1} + \dfrac{\eta}{1-\zeta}\right)\right] n, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 \leq \dfrac{1-\theta\eta}{1-\zeta}, \\ \left[1 + \dfrac{\theta}{1-\theta\eta}\left(\dfrac{1}{\sigma-1} + \eta\right)\right] n, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 > \dfrac{1-\theta\eta}{1-\zeta}. \end{cases}$$

and

$$g_{D_A}^{sp} = \begin{cases} \dfrac{\tau-\eta}{\tau}\left[1+\dfrac{\theta}{1-\theta\eta}\left(\dfrac{1}{\sigma-1}+\eta\right)\right]n, & \text{if} \quad \dfrac{\tau-\eta}{\tau}\left(\dfrac{\theta}{\sigma-1}+1\right) < \dfrac{1-\theta\eta}{1-\zeta} < \dfrac{\theta}{\sigma-1}+1, \\ \dfrac{1}{1-\zeta}n, & \text{otherwise}. \end{cases}$$

The growth rates of labor shares allocated to producer data and AI-generated data are given by:

$$g_{l_P}^{sp} = 0,$$

and

$$g_{l_A}^{sp} = \begin{cases} \left[\dfrac{(\tau-\eta)(1-\zeta)}{\tau(1-\theta\eta)}\left(\dfrac{\theta}{\sigma-1}+1\right)-1\right]n, & \text{if} \quad \dfrac{\tau-\eta}{\tau}\left(\dfrac{\theta}{\sigma-1}+1\right) < \dfrac{1-\theta\eta}{1-\zeta} < \dfrac{\theta}{\sigma-1}+1, \\ 0, & \text{otherwise}. \end{cases}$$

The labor shares allocated to different sectors are given by:

$$l_P^{sp} \to \begin{cases} 0, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{\eta\rho(\sigma-1)}{n+\rho(1+\eta)(\sigma-1)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 > \dfrac{1-\theta\eta}{1-\zeta}, \end{cases}$$

$$l_A^{sp} \to \begin{cases} \dfrac{\eta\rho(\sigma-1)}{\mathcal{A}n+\rho(\sigma-1)(\mathcal{A}+\eta)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ 0, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 > \dfrac{1-\theta\eta}{1-\zeta}, \end{cases} \tag{20}$$

and

$$l_R^{sp} \to \begin{cases} \dfrac{\mathcal{A}n}{\mathcal{A}n+\rho(\sigma-1)(\mathcal{A}+\eta)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{n}{n+\rho(1+\eta)(\sigma-1)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 > \dfrac{1-\theta\eta}{1-\zeta}, \end{cases}$$

where $\mathcal{A}$ is a constant defined in equation (19).

Given the labor shares derived above (l_P^{sp}, l_A^{sp}, and l_R^{sp}), other variables are determined as follows:

$$c_t^{sp} = y_t^{sp}$$

$$= \begin{cases} \dfrac{(1-\beta)^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(1-e_0)\left[\dfrac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}\left(l_A^{sp}\right)^{\frac{\eta}{1-\zeta}}\cdots \\ \left(l_R^{sp}\right)^{\frac{1}{\sigma-1}}\left(1-l_P^{sp}-l_A^{sp}-l_R^{sp}\right)L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1<\dfrac{1-\theta\eta}{1-\zeta}, \\ \left[\dfrac{\beta^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}\left(l_P^{sp}\right)^{\eta}\left(l_R^{sp}\right)^{\frac{1}{\sigma-1}}\left(1-l_P^{sp}-l_R^{sp}\right)\right]^{\frac{1}{1-\theta\eta}}L_t^{\frac{1}{1-\theta\eta}\left(\eta+\frac{1}{\sigma-1}\right)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1>\dfrac{1-\theta\eta}{1-\zeta}. \end{cases} \tag{21}$$

$$D_{A,t}^{sp} = \left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{1}{1-\zeta}}\left(l_A^{sp}\right)^{\frac{1}{1-\zeta}}L_t^{\frac{1}{1-\zeta}},$$

$$D_{P,t}^{sp} = \left(c_t^{sp}\right)^{\theta} l_P^{sp} L_t,$$

$$N_t^{sp} = \frac{1}{n\chi} l_R^{sp} L_t, \tag{22}$$

and

$$\boxed{g_c^{sp} = g_y^{sp} = \begin{cases} \left(\dfrac{1}{\sigma-1}+\dfrac{\eta}{1-\zeta}\right)n, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1<\dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{1}{1-\theta\eta}\left(\dfrac{1}{\sigma-1}+\eta\right)n, & \text{otherwise.} \end{cases}} \tag{23}$$

Proof. See online Appendix A. □

The most important result in the proposition is the solution for per capita output when the two types of data grow at different rates, as shown in equation (21). Since the labor shares allocated across different sectors converge to constant values along a balanced growth path, these equations indicate that per capita output is proportional to the scale of the economy (or, equivalently, the population) raised to some power. From equation (23), we observe that the growth rates of per capita output and consumption are divided into two regimes, with the threshold determining which type of data grows at a higher rate. In both regimes, the growth rates reflect the degree of increasing returns to population growth and consist of two additive components. The first term, $1/(\sigma-1)$, corresponds to the well-documented "love of variety" effect, which diminishes as the elasticity of substitution among different varieties of intermediate goods increases. The second term is novel and captures the differential effects of the two types of data, which are linked to their respective generation processes. Beyond the common parameter η appearing in both regimes, which signifies the importance of data to the economy, each regime incorporates an additional distinct parameter representing the characteristics of its respective data type. In the first regime, ζ enters the solution, denoting the spillover effect of AI-generated data in generating new data of the same type. Conversely,

in the second regime, θ appears instead of ζ, reflecting the reference effect of per capita output in the generation of producer data. Notably, in the second regime, the "love of variety" effect is further amplified by the producer data generation process through the term $1/(1-\theta\eta)$, highlighting the stronger link between the production process and this type of data.

The remaining results under the optimal allocation yield several important implications. First, the labor shares allocated across the two data-generation sectors, l_P^{sp} and l_A^{sp}, concentrate entirely in one sector when the economy is in the corresponding regime. Specifically, l_P^{sp} shrinks to zero when AI-generated data dominate the economy, and similarly, l_A^{sp} becomes negligible when producer data are the dominant type. Although the error rate in intermediate goods production converges to one when AI-generated data accumulate much faster than producer data—resulting in a near-zero survival rate for the corresponding intermediate goods—the economy still grows at a positive rate due to the low cost and large volume of AI-generated data used in production. In contrast, in the regime where producer data dominate, AI-generated data become ineffective, as they can no longer accumulate efficiently and thus lose their competitive advantage relative to other data types. Moreover, it is notable that while the spillover effect of AI-generated data, ζ, influences labor allocation through the constant $\mathcal{A}$, the counterpart parameter for producer data generation—the reference effect of output, θ—has no impact on this allocation. A more influential parameter is the importance of data in intermediate goods production, η, which appears throughout the expressions for labor shares across regimes. This is the key parameter in our model and has been carefully calibrated in Jones and Tonetti (2020).

Next, equation (22) indicates that the optimal level of variety, N_t^{sp}, is proportional to the population in the economy. This proportionality factor depends on the population growth rate n, the entry cost χ, and the labor share allocated to the innovation sector, l_R^{sp}. Although a larger population leads to a higher level of variety, a higher population growth rate slows down the growth of variety. This outcome reflects the trade-off between economic scale and population growth in determining variety: while a larger economy allows more labor to be allocated to the innovation sector, thereby fostering greater variety creation, a higher population growth rate introduces a dilution effect that crowds out resources allocated to the innovation process.

We will revisit these results after analyzing the allocations in the competitive equilibrium within this environment. The labor shares allocated across different sectors and the resulting data quality under various allocations will play a crucial role in the comparison.

5. Competitive Equilibrium

We now consider an allocation in which consumers own the generative AI firm. From equation (3), it is evident that the generative AI firm exhibits increasing returns to scale, as the exponent on its sole input, $L_{A,t}$, is greater than one. In other words, the firm's profit will be positive if it possesses some degree of monopolistic power that prevents potential competitors from entering the market. In this section, we assume that the M Generative AI firm operating are all operated in a monopolistic environment and that their profits are used to pay for the investments before entering the economy. Throughout the paper, both sellers and buyers are assumed to be price takers in the data market.

5.1 Decision Problems

Household Problem. Households supply one unit of labor inelastically at a wage rate w_t. They hold assets a_t, which earn a return at rate r_t. These assets can also be interpreted as the profits redistributed to households due to the monopolistic power of firms. The representative household then solves the following optimization problem:

$$U_0 = \max_{\{c_t\}} \int_0^\infty e^{-(\rho-n)t} \ln c_t \mathrm{d}t \tag{24}$$

subject to

$$\dot{a}_t = (r_t - n)a_t + w_t - c_t. \tag{25}$$

Here, we normalize the price of the final consumption good to 1.

Final Good Producer Problem. The final good producer must determine two inputs: the intermediate goods used in final good production and the labor employed in generating producer data. Since producer data are typically considered a byproduct of production, these two decisions should be treated jointly. Thus, the final good producer solves the following optimization problem:

$$\max_{\{Y_{i,t}, L_{P,t}\}} Y_t + p_{D_P,t} \left(\frac{Y_t}{L_t}\right)^\theta L_{P,t} - \int_0^{N_t} p_{i,t} Y_{i,t} \mathrm{d}i - w_t L_{P,t}, \tag{26}$$

subject to equation (8). In the above equation, $p_{D_P,t}$ and $p_{i,t}$ represent the prices of producer data and intermediate goods of variety i, respectively. Taking the first-order condition with

respect to $Y_{i,t}$, we obtain the demand function for intermediate goods:

$$\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)=p_{i,t}. \tag{27}$$

Generative AI Firm Problem. Each Generative AI firm operates in a monopolistic environment, where they can determine the price of AI-generated data from the demand behavior of intermediate good producers. Due to the nonrival nature of data, the Generative AI firm can continue using AI-generated data to produce new data even after selling it to other firms. Meanwhile, it must also determine the amount of labor employed in each period. Since the firm engages in intertemporal decision-making, its optimization problem is formulated as follows:

$$\max_{\{d_{A,t},L_{A,t}\}}\Pi=\int_0^{\infty}\exp\left(-\int_0^t r_\tau \mathrm{d}\tau\right)\left[p_{D_A,t}(d_{A,t})\cdot d_{A,t}-w_t\frac{L_{A,t}}{M}\right]\mathrm{d}t, \tag{28}$$

subject to the evolution equation of $d_{A,t}$ given in equation (10). Here, $p_{D_A,t}$ represents the price of AI-generated data, which can be derived from the intermediate good producer problem. Considering the emergent capabilities $\mathcal{G}$, the free-entry condition of Generative AI firms is written as $\Pi=\mathcal{G}$.

Intermediate Good Producer Problem. Given the demand function for intermediate goods in equation (27), each intermediate good producer across different varieties must determine the optimal usage of the two types of data and the employment of labor. Letting $V_{i,t}$ denote the market value of variety i at time t, the firm's optimization problem is formulated as follows:

$$\begin{aligned} r_tV_{i,t}=\max_{\{L_{i,t},D_{P,i,t},D_{A,i,t}\}} & Y_t^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)Y_{i,t}^{1-\frac{1}{\sigma}}-w_tL_{i,t}-p^d_{D_P,t}D_{P,i,t}-p^d_{D_A,t}D_{A,i,t}+ \\ & \dot{V}_{i,t}-\delta(e_{i,t})V_{i,t}, \end{aligned} \tag{29}$$

where $p^d_{D_P,t}$ and $p^d_{D_A,t}$ represent the demand-side prices of producer data and AI-generated data, respectively. These prices differ from those faced by the supply side of data, an issue that will be further examined after introducing the role of data intermediaries. Similar to the optimal allocation, while the joint usage of both types of data enhances firm production, their relative proportions may have counteracting effects. Specifically, a higher reliance on AI-generated data increases the error rate in production, raising the likelihood of output disruptions. Additionally, firms must account for the "loss of business" effect induced by an increasing error rate— a consideration absent in the optimal allocation framework.

Data Intermediary Problem. We follow the framework established in Jones and Tonetti (2020) to model data intermediaries. In this setting, we introduce two types of data intermediaries: one that handles producer data and another that handles AI-generated data. We assume that both intermediaries operate as monopolists in their respective data markets but are constrained by free entry into data intermediation. As a result, the optimization problem for the producer data intermediary is formulated as follows:

$$\max_{p^d_{D_P,t}, D_{P,t}} \; p^d_{D_P,t} \int_0^{N_t} D_{P,i,t} \mathrm{d}i - p_{D_P,t} D_{P,t}, \tag{30}$$

subject to

$$D_{P,i,t} \leq D_{P,t}. \tag{31}$$

Due to the nonrival nature of data, the data intermediary earns zero profit in equilibrium. Similarly, the optimization problem for the AI-generated data intermediary follows the same structure and is omitted here for brevity.

Innovation Problem. Potential firms must employ labor $L_{R,t}$ before entering the economy as an intermediate good producer. Additionally, incumbent firms are subject to the "loss of business" effect if their error rates are excessively high. The free entry condition is given by:

$$\chi w_t = V_{i,t} + \frac{\int_0^{N_t} \delta(e_{i,t}) V_{i,t} \mathrm{d}i}{\dot{N}_t}. \tag{32}$$

The left-hand side represents the entry cost. The right-hand side consists of two components: the first term corresponds to the value of a new variety, while the second term captures the additional value arising from the "loss of business" effect, which benefits each potential entrant upon entering the economy. This condition follows a similar structure to that in Jones and Tonetti (2020), but the interpretation of the second term on the right-hand side differs, leading to distinct analytical implications.

5.2 Equilibrium Definition

The equilibrium in which the generative AI firm operates as a monopolist is an allocation where all households choose $\{c_t, a_t\}$ to maximize their discounted utility. The final good producer optimizes over $\{Y_{i,t}, L_{P,t}\}$, the generative AI firm selects $\{L_{A,t}\}$, and intermediate goods producers determine $\{L_{i,t}, D_{P,i,t}, D_{A,i,t}\}$ to maximize their respective profits. The two types of data intermediaries set $\{p^d_{D_P,t}, D_{P,t}\}$ and $\{p^d_{D_A,t}, D_{A,t}\}$, respectively, but both operate under zero-profit conditions. The evolution of $\{w_t, r_t, p_{i,t}, p_{D_P,t}, p_{D_A,t}\}$ is governed

by market-clearing conditions for labor, assets, intermediate goods, producer data, and AI-generated data. The number of varieties $\{N_t\}$ evolves according to the innovation possibility frontier given in equation (15). Finally, the resource constraint $Y_t = c_t L_t$ holds at all times, and population L_t grows exogenously at rate n.

5.3 Solving the Model

Unlike the optimal allocation, intermediate good producers do not account for the additional utility derived from higher data quality, while households cannot directly influence the usage of different types of data. Consequently, we expect labor to be concentrated in one of the data-generating sectors when one type of data exhibits a higher growth rate. Specifically, at least the labor share allocated to the dominated data sector shrinks to zero along the balanced growth path. We use the superscript "dc" to denote the results derived in the competitive equilibrium and define the following constant:

$$C \equiv \begin{cases} \dfrac{n\chi(\sigma-1)(\rho+\delta_0 e_0^2)}{\left(n+\delta_0 e_0^2\right)\left(1+\eta-\sigma\eta\right)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1<\dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{\chi\rho(\sigma-1)}{1+\eta-\sigma\eta}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1>\dfrac{1-\theta\eta}{1-\zeta}. \end{cases} \tag{33}$$

We do not provide the definition of C for the case $1+\theta/(\sigma-1) = (1-\theta\eta)/(1-\zeta)$, as this parameter does not influence the economy in that regime. We then state the following proposition.

Proposition 2. *(The Competitive Equilibrium) When $(1-\tau)\varepsilon > 1$, which is usually the case given the standard values of parameters, along a balanced growth path, as the population L_t grows large, the competitive equilibrium converges to the following results.*

The growth rates of producer data and AI-generated data are given by:

$$g_{D_P}^{dc} = \begin{cases} \leq \dfrac{1}{1-\tau}\left(\dfrac{\theta\eta-\tau}{1-\zeta}+\dfrac{\theta}{\sigma-1}+1\right)n, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1<\dfrac{1-\theta\eta}{1-\zeta}, \\ = \left[1+\dfrac{\theta}{1-\theta\eta}\left(\dfrac{1}{\sigma-1}+\eta\right)\right]n, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1\geq\dfrac{1-\theta\eta}{1-\zeta} \end{cases}$$

and

$$g_{D_A}^{dc} = \begin{cases} \dfrac{1}{\zeta\varepsilon-1}\left[\dfrac{\varepsilon-1}{1-\theta\eta}\left(\dfrac{\theta}{\sigma-1}+1\right)-\varepsilon\right]n, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1>\dfrac{1-\theta\eta}{1-\zeta} \quad \text{and} \quad \zeta\varepsilon<1, \\ \dfrac{1}{1-\zeta}n, & \text{otherwise}. \end{cases} \tag{34}$$

The growth rates of the labor shares allocated to generating producer data and AI-generated data

are:

$$g_{l_P}^{dc} = \begin{cases} \dfrac{\tau}{1-\tau}\left(\dfrac{\theta}{\sigma-1} - \dfrac{1-\theta}{1-\zeta} + 1\right) n < 0, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ 0, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 \geq \dfrac{1-\theta\eta}{1-\zeta} \end{cases}$$

and

$$g_{l_A}^{dc} = \begin{cases} \dfrac{\varepsilon-1}{\zeta\varepsilon-1}\left[\dfrac{1-\zeta}{1-\theta\eta}\left(\dfrac{\theta}{\sigma-1}+1\right)-1\right] n < 0, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 > \dfrac{1-\theta\eta}{1-\zeta} \quad \text{and} \quad \zeta\varepsilon < 1, \\ 0, & \text{otherwise}. \end{cases}$$

The labor shares allocated to different sectors converge to:

$$l_P^{dc} \to \begin{cases} 0, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{C\eta}{C(1+\eta)+n\chi}, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 > \dfrac{1-\theta\eta}{1-\zeta}, \end{cases}$$

$$l_A^{dc} \to \begin{cases} \dfrac{C\eta^2}{C(\mathcal{A}+\eta^2)+n\chi\mathcal{A}}, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ 0, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 > \dfrac{1-\theta\eta}{1-\zeta}, \end{cases}$$

and

$$l_R^{dc} \to \begin{cases} \dfrac{n\chi\mathcal{A}}{C(\mathcal{A}+\eta^2)+n\chi\mathcal{A}}, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ \dfrac{n\chi}{C(1+\eta)+n\chi}, & \text{if} \quad \dfrac{\theta}{\sigma-1} + 1 > \dfrac{1-\theta\eta}{1-\zeta}. \end{cases}$$

Here, C is a constant defined in equation (33). The labor employed by a single intermediate goods producer converges to a constant:

$$L_i^{dc} = n\chi \frac{1 - l_P^{dc} - l_A^{dc} - l_R^{dc}}{l_R^{dc}} = C.$$

For the labor shares when $g_{D_A}^{dc} = g_{D_P}^{dc}$, we can only obtain numerical solutions, which are provided in Appendix B.2.

Finally, given the labor shares derived in this proposition, other key variables such as $D_{A,t}^{dc}$, $D_{P,t}^{dc}$, c_t^{dc}, y_t^{dc}, and N_t^{dc} can be determined analogously to Proposition 1, with the growth rate of the economy remaining identical to that in equation (23).

Proof. See online Appendix Section B. □

In addition to the labor shares and the growth rates of key variables presented in Proposition 2, the competitive equilibrium also determines the prices of various types of goods. We now state the following proposition. For brevity, we present only the growth rates of prices in this proposition, as they play a crucial role in the subsequent analysis. Since these variables exist solely in the competitive equilibrium, we omit the superscript "dc" for clarity.

Proposition 3. *(Prices in the Competitive Equilibrium) Along a balanced growth path, as the population L_t grows large and given the labor shares presented in Proposition 2, the prices of various types of goods in the competitive equilibrium converge to the following results.*

The price of intermediate goods is given by:

$$p_{i,t} = \frac{(l_R^{dc})^{\frac{1}{\sigma-1}}}{(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]} L_t^{\frac{1}{\sigma-1}}.$$

Given $(1-\tau)\varepsilon > 1$, the growth rates of wages, the market value of intermediate goods, and the prices of producer data and AI-generated data are as follows:

$$g_w = g_V = \begin{cases} \left(\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)n, & \text{if} \quad \frac{\theta}{\sigma-1}+1 \le \frac{1-\theta\eta}{1-\zeta}, \\ \frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)n, & \text{if} \quad \frac{\theta}{\sigma-1}+1 > \frac{1-\theta\eta}{1-\zeta}, \end{cases}$$

$$g_{p_{D_P}} = \begin{cases} (1-\theta)\left(\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)n, & \text{if} \quad \frac{\theta}{\sigma-1}+1 < \frac{1-\theta\eta}{1-\zeta}, \\ \frac{1-\theta}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)n, & \text{if} \quad \frac{\theta}{\sigma-1}+1 \ge \frac{1-\theta\eta}{1-\zeta}, \end{cases}$$

and

$$g_{p_{D_A}} = \begin{cases} \left\{\left[1-\frac{\zeta\theta(\varepsilon-1)}{\zeta\varepsilon-1}\right]\frac{1}{1-\theta\eta}\frac{1}{\sigma-1}+\ldots \right. \\ \left. \frac{1}{\zeta\varepsilon-1}+\frac{1}{1-\theta\eta}\left(\eta-\theta\eta-\frac{1-\zeta}{\zeta\varepsilon-1}\right)\right\}n, & \text{if} \quad \frac{\theta}{\sigma-1}+1 > \frac{1-\theta\eta}{1-\zeta} \text{ and } \zeta\varepsilon < 1, \\ \left(\frac{1}{\sigma-1}+\frac{\eta-1}{1-\zeta}+1\right)n, & \text{otherwise}. \end{cases}$$

Proof. See online Appendix Section B. □

Similar to the optimal allocation, the balanced growth path also becomes asymptotic with $l_P^{dc} \to 0$ when AI-generated data dominate, while the results become reversed as producer

data become dominant. The underlying reason is that intermediate goods producers do not internalize the externality arising from the dynamics of data quality, specifically the ratio of producer data to AI-generated data usage. Although we introduce the "loss of business" effect, $\delta(e_{i,t})V_{i,t}$, to constrain excessive reliance on AI-generated data in the decision-making process of intermediate good producers, this effect proves to be relatively insignificant in the long run.

To illustrate this effect, we derive the first-order conditions for the problem defined by equation (29) with respect to $D_{P,i,t}$ and $D_{A,i,t}$ as follows:

$$\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\frac{\partial Y_{i,t}}{\partial D_{P,i,t}}=\underbrace{\frac{p_{D_P,t}}{N_t}-2\xi\tau\delta_0 e_{i,t}^2 D_{P,i,t}^{\tau-1}D_{A,i,t}^{-\tau}V_{i,t}}_{g_{p_{D_P}}-g_N\leq g_V+(\tau-1)g_{D_P}^{dc}-\tau g_{D_A}^{dc}}, \tag{35}$$

and

$$\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\frac{\partial Y_{i,t}}{\partial D_{A,i,t}}=\underbrace{\frac{p_{D_A,t}}{N_t}+2\xi\tau\delta_0 e_{i,t}^2 D_{P,i,t}^{\tau}D_{A,i,t}^{-\tau-1}V_{i,t}}_{g_{p_{D_A}}-g_N>g_V+\tau g_{D_P}^{dc}-(1+\tau)g_{D_A}^{dc}}. \tag{36}$$

The right-hand sides of equations (35) and (36) can be interpreted as the effective prices of producer data and AI-generated data, respectively. The second terms in these expressions capture the marginal effects of $D_{A,i,t}$ and $D_{P,i,t}$ embedded in the "loss of business" effect, which reduces the effective price of producer data while increasing that of AI-generated data. This mechanism also reflects the incentive for intermediate goods producers to enhance data quality.

When $g_{D_A}^{dc} < g_{D_P}^{dc}$, indicating that producer data dominate the economy, the "loss of business" effect becomes negligible since $e_{i,t} \to 0$ in this regime. When $g_{D_A}^{dc} = g_{D_P}^{dc}$, numerical results in Section C.2 show that the "loss of business" effect become trivial compared with other externalities, which is consistent with the findings presented here. However, when $g_{D_A}^{dc} > g_{D_P}^{dc}$, implying that AI-generated data dominate the economy, the analysis becomes more complex. From equation (36), we know that the "loss of business" effect term can be omitted since the growth rate of the AI-generated data price term is higher. Meanwhile, Proposition 3 states that the condition $(1-\tau)\varepsilon > 1$ implies that the sensitivity of data quality should not be excessively large and that the elasticity of substitution between the two types of data should not be too low. This also suggests that intermediate good producers will not face a rapidly increasing risk of replacement as they adopt more AI-generated data, while this type of data can substitute for producer data at a relatively high level. Under this

assumption, considering equation (35), we determine that the growth rate of $l_{P,t}$ is negative in this regime, leading to $l_{P,t} \to 0$.[8] Consequently, although the "loss of business" effect plays a significant role in determining the effective price of producer data, the declining labor share in generating producer data, $l_{P,t}$, remains irreversible, which does not change the conclusion of the model.

5.4 Further Comparisons Between Optimal Allocation and Competitive Equilibrium

Data quality. A key insight from the competitive equilibrium is that the growth rates of the two types of data differ from those derived in the optimal allocation. This result is notable, as semi-endogenous growth models typically exhibit equal growth rates along the balanced growth path in both settings. This deviation also leads to differing growth rates of data quality between the two settings, which are given by:

$$\bar{Q}_t^{sp} \propto \begin{cases} L_t^{\tau\left(1+\frac{\theta}{\sigma-1}-\frac{1-\theta\eta}{1-\zeta}\right)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 \leq \dfrac{1-\theta\eta}{1-\zeta}, \\ L_t^{\eta\left[1+\frac{\theta}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)\right]}, & \text{if} \quad \dfrac{\tau-\eta}{\tau}\left(\dfrac{\theta}{\sigma-1}+1\right) < \dfrac{1-\theta\eta}{1-\zeta} < \dfrac{\theta}{\sigma-1}+1, \\ L_t^{\tau\left[\frac{1}{1-\theta\eta}\left(\frac{\theta}{\sigma-1}+1\right)-\frac{1}{1-\zeta}\right]}, & \text{if} \quad \dfrac{\tau-\eta}{\tau}\left(\dfrac{\theta}{\sigma-1}+1\right) \geq \dfrac{1-\theta\eta}{1-\zeta} \end{cases}$$

and

$$\bar{Q}^{dc} \propto \begin{cases} L_t^{\frac{\tau}{1-\tau}\left(\frac{\theta}{\sigma-1}-\frac{1-\theta\eta}{1-\zeta}+1\right)}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 < \dfrac{1-\theta\eta}{1-\zeta}, \\ L_t^{\frac{\tau}{\zeta\varepsilon-1}\left[\frac{\varepsilon(\zeta-1)}{1-\theta\eta}\left(\frac{\theta}{\sigma-1}+1\right)+\varepsilon\right]}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 \geq \dfrac{1-\theta\eta}{1-\zeta} \quad \text{and} \quad \zeta\varepsilon < 1, \\ L_t^{\tau\left[\frac{1}{1-\theta\eta}\left(\frac{\theta}{\sigma-1}+1\right)-\frac{1}{1-\zeta}\right]}, & \text{if} \quad \dfrac{\theta}{\sigma-1}+1 \geq \dfrac{1-\theta\eta}{1-\zeta} \quad \text{and} \quad \zeta\varepsilon \geq 1. \end{cases}$$

It can be shown that the growth rate of $\bar{Q}^{sp}$ is always greater than that of $\bar{Q}^{dc}$. Further numerical results are provided in Section 6. This discrepancy arises due to the asymptotic balanced growth path, where the growth rates of the two types of data diverge. As labor shares in the dominated data sector shrink to zero, the negative growth rates of these labor shares reduce the growth rates of the corresponding type of data. For example, when $D_{A,t}$ is dominated by $D_{P,t}$, the growth rate $g_{D_A}^{dc}$ is lower than when $D_{A,t}$ is the dominant data type. However, since the usage of the dominated data type is significantly lower than that of the dominant type (due to their differing growth rates), we consider only the effect of the

[8]The condition $(1-\tau)\varepsilon > 1$ is derived through substituting the growth rate of variables involved in equation (36). For details on the derivation, please refer to Appendix B.1.

dominant type in the long run.

Number of Generative AI firms. We further examine the number of Generative AI firms under both the optimal allocation and the competitive equilibrium. In the former case, the number is determined by maximizing social welfare, while in the latter, it is determined by a free-entry condition. We restrict our analysis to the case in which AI-generated data dominate the economy ($g_{D_A} \geq g_{D_P}$), as the number of Generative AI firms is economically relevant only in this scenario. Proposition 4 presents the solution.

Proposition 4. *(Number of Generative AI Firms) When $g_{D_A} \geq g_{D_P}$—indicating that AI-generated data dominate the economy—along a balanced growth path, the optimal number of Generative AI firms is ONE, regardless of the magnitude of emergent capabilities $\mathcal{G}$. In contrast, under the competitive equilibrium, this number converges to:*

$$M^{dc} = \left\{ \frac{(\mathcal{A}-\eta)(\sigma-1)(1-e_0)}{\sigma\eta(\rho-n)(n\chi)^{\frac{1}{\sigma-1}}\mathcal{G}} \left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)} \right]^{\frac{\eta}{1-\zeta}} (l_R^{dc})^{\frac{1}{\sigma-1}} (l_A^{dc})^{1+\frac{\eta}{1-\zeta}} L_0^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}} \right\}^{\frac{1-\zeta}{1-\zeta(1-\eta)}}, \quad (37)$$

where l_P^{dc}, l_A^{dc}, and l_R^{dc} are derived from Proposition 2.

Proof. See online Appendix Sections A.3 and B.4. □

As the number of Generative AI firms increases, the economy faces repeated investment in emergent capabilities. At the same time, labor allocated to each Generative AI firm decreases, which in turn reduces the production of AI-generated data. As a result, the optimal number of such firms is one, allowing all available resources to be concentrated in generating AI-generated data for production. In contrast, under the competitive equilibrium, Generative AI firms can earn positive profits when operating in a monopolistic environment, which induces potential entrants. As more firms enter, profits decline and eventually converge to zero. Clearly, from equation (37), the equilibrium number of firms exceeds one and depends on various parameters such as η, ζ, and $\mathcal{G}$. Numerical results are presented in Section 6.3 for further discussion.

6. Numerical Examples and Further Discussions

In this section, we present several numerical examples to support our further discussion of the model, including comparative statics on key parameters and comparisons between the competitive equilibrium and the optimal allocation. Due to the lack of precise estimates for

some of the relevant parameters, our results should not be interpreted as a formal simulation of real-world dynamics. Instead, they serve as a useful exercise to illustrate how the various forces in the model interact.

6.1 Calibration

Table 2 presents the estimated parameter values, selected to align the model more closely with real-world conditions. Most parameters either adopt standard values from the literature or are estimated based on reasonable projections. For parameters that are novel and lack reliable estimates, we conduct extensive robustness checks over a wide range of values. In all subsequent analyses, we set the initial population as $L_0 = 100$, where one unit of labor corresponds to one million people.

Table 2: Parameters Values

Parameters	Meaning	Value	Comment
ρ	Subjective discount rate	0.03	Standard
n	Population growth rate	0.02	Standard
σ	Elasticity of substitution (goods)	4	Standard
χ	Labor cost of entry	0.01	Standard
η	Importance of data in production	0.06	Jones and Tonetti (2020)
δ_A	Depreciation rate of AI-generated data	0.2	Estimated
ζ	Contribution of existing AI-generated data	0.25	Estimated
θ	Importance of output in producer data generation	0.81	Calculated from model
ψ	Efficiency term in AI-generated data generation	1	Normalized
ε	Elasticity of substitution (data)	50	Discretionary
e_0	Basic error rate	0.95	Discretionary
β	Share of producer data in overall dataset	0.5	Normalized
ξ	Sensitivity of data quality to error rate	1	Normalized
τ	Elasticity of data ratio to data quality	0.5	To be discussed
κ	Weight on data quality versus consumption	0.10	To be discussed
δ_0	"Loss of business" effect	0.4	To be discussed

Note: Baseline parameter values for the numerical examples. When $g_{D_A} = g_{D_P}$, parameters ζ and θ are set simultaneously. We also discuss other values of these two parameters when $g_{D_A} \neq g_{D_P}$.

First, we assign standard parameter values that are widely used in related studies. For

example, the subjective discount rate ρ, also referred to as the rate of time preference, is set to 0.03, which is higher than the population growth rate $n = 0.02$. Furthermore, the elasticity of substitution σ among different varieties of intermediate goods is set to 4, ensuring that these goods are substitutes and that the degree of increasing returns in the absence of data is approximately $1/(\sigma - 1) = 0.33$. For the labor cost of entry χ, we set it to 0.01, implying that the invention of a new patent requires approximately 100 researchers.

Second, we refer to empirical evidence to determine the values of certain parameters. Following the estimation in Jones and Tonetti (2020), we set the importance of data in production, η, to 0.06. In that paper, the authors reference studies in computer science to derive a rough estimate and conduct robustness checks, concluding that 0.06 is the most suitable value for analysis. On the other hand, the depreciation rate of AI-generated data, δ_A, is estimated at 0.2, consistent with the widely used depreciation rate of intangible capital. Data are typically regarded as a special form of intangible capital for firms, which generally exhibits a much higher depreciation rate than physical capital. We choose this value under the assumption that data will be fully depreciated in approximately five years. For the contribution of existing AI-generated data to new data generation, ζ, we take an unconventional approach by directly consulting ChatGPT on how many existing data it uses when generating new content. After verifying multiple responses, we estimate this proportion to be between 25% and 40%, depending on the type of content being generated. Given that, in our model, data are used to enhance production—a process more creative than routine tasks—we adopt the most conservative estimate of 0.25 for ζ. Once the values of σ, η, and ζ are determined, we compute the importance of output in producer data generation, θ, to ensure that the growth rates of the two types of data are equal. This yields a value of 0.81. In the subsequent analysis, we also conduct robustness checks by varying ζ between 0.02 and 0.99 while keeping θ fixed at this value, allowing us to examine allocations when the growth rates of the two types of data differ.

Third, two parameters are selected to simplify our analysis. For the elasticity of substitution between the two types of data, ε, we choose a large value of 50 to ensure that the two data types can substitute each other nearly perfectly. Additionally, the basic error rate, e_0, which represents the maximum error rate when AI-generated data dominate the economy, is set to 0.95, slightly below 1. Meanwhile, three parameters are chosen to avoid introducing additional heterogeneity. We set the share of producer data in the overall dataset, β, to 0.5, and the sensitivity parameter related to data quality, ξ, is fixed at 1.

Last, there are three parameters that cannot be precisely determined: the elasticity of the data ratio with respect to data quality, τ, the weight on data quality, κ, and the "loss of business" effect, δ_0. While we can reference the estimations discussed in Jones and Tonetti (2020) for the latter two parameters, their interpretations in that paper differ from those in

our model. These three parameters represent two key channels in our framework: τ and κ influence utility, which is crucial in the optimal allocation, while δ_0 affects the decisions of intermediate good producers, playing a central role in the competitive equilibrium. Given these considerations, we set $\tau = 0.5$, $\kappa = 0.10$, and $\delta_0 = 0.4$ as baseline values to ensure well-behaved results, and we further explore the impact of these parameters on the economy through sensitivity analyses, which are shown in Appendix C.2.

6.2 Numerical Examples When $g_{D_A} \neq g_{D_P}$

The key parameter in this subsection is the contribution of existing AI-generated data to their own generation process, ζ, which determines the growth rate of $D_{A,t}$. As ζ increases from 0.02 to 0.99, the economy transitions from a regime where producer data dominate to one where AI-generated data become the primary data source.[9] We first examine the labor allocations across the four different sectors and then analyze the growth rates of consumption and data quality to illustrate the model's behavior.

Labor allocations. Given the values of other parameters, labor allocations across different sectors in both settings with respect to ζ are shown in Figure 2. From the figures, we first observe that labor allocated to the intermediate goods production sector constitutes the largest share in both settings. This result is intuitive, as most labor is expected to be devoted to production in order to enhance consumption and, consequently, welfare. In contrast, we find that under both of the regimes, labor employed in the producer data sector shifts entirely to the AI-generated data sector once ζ crosses a certain threshold. The reallocation of labor exhibits a discrete shift as ζ transitions across regimes, reflecting significant structural changes induced by variations in the growth rates of the two types of data. Finally, we observe that labor in the innovation sector declines, while labor in the AI-generated data sector increases. This suggests that AI-generated data become increasingly influential and, in some cases, even substitute for the effects of innovation as their multiplier effect strengthens. Moreover, it is noteworthy that the social planner allocates more labor to the AI-generated data sector and less to the innovation sector than in the competitive equilibrium. This indicates a failure to fully utilize AI-generated data under the latter, as multiple Generative AI firms compete for limited labor resources, as discussed in Proposition 4. At the same time, insufficient producer data are generated by the final goods producers, leading to lower data quality.

[9] Referring to equation (34), we start our analysis of ζ from 0.02 since $g_{D_A}^{dc}$ becomes negative when $\zeta < 0.02$, which is inconsistent with the assumptions of our model.

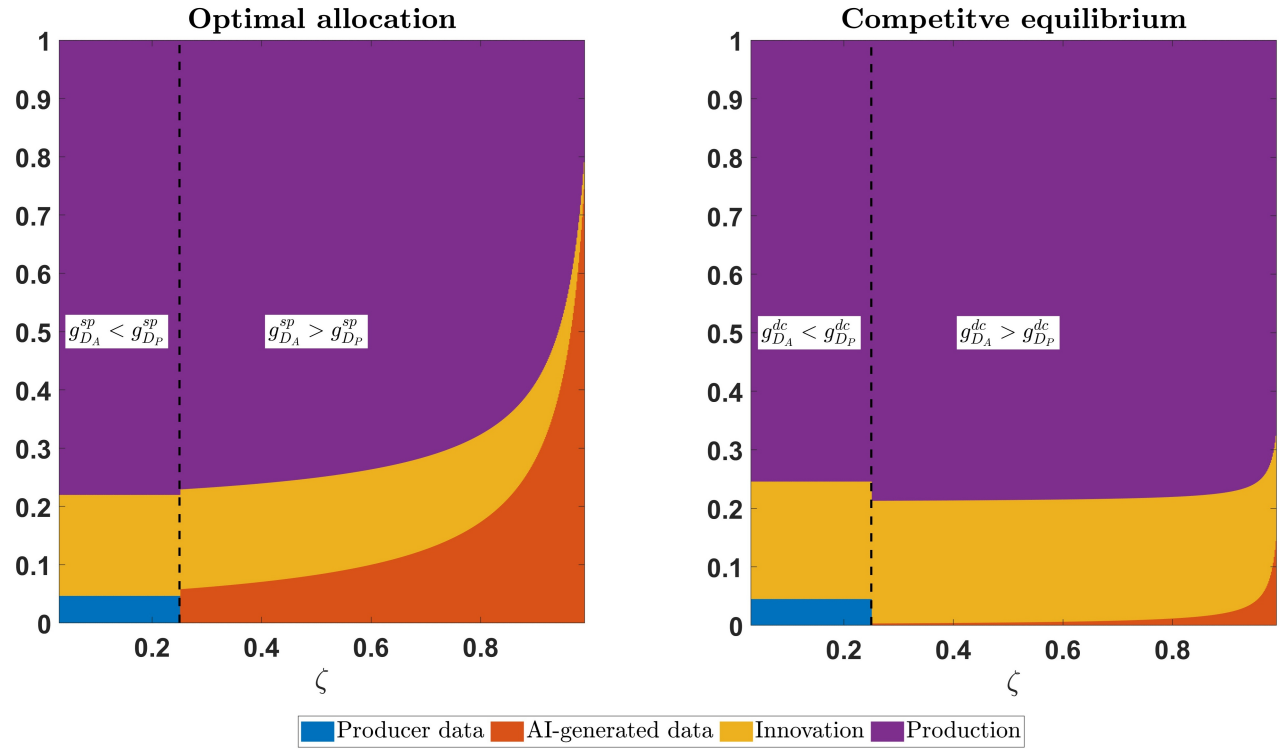


Figure 2: Labor allocations in the four different sectors

Note. These figures illustrate labor allocation in both the optimal allocation and the competitive equilibrium. The purple area represents the intermediate good production sector, the yellow area corresponds to the innovation sector, the blue area denotes the producer data generation sector, and the red area indicates labor employed in the AI-generated data sector. A black dashed line separates the regimes where $g_{D_A} < g_{D_P}$ (left region) and where $g_{D_A} > g_{D_P}$ (right region).

Growth rates. The differences between the competitive equilibrium and the optimal allocation can be further illustrated by examining the growth rates of per capita consumption and data quality, as shown in Figure 3. From the figure, we first observe that per capita consumption grows at a higher rate as ζ increases, reflecting the positive impact of AI-generated data on the economy. However, when considering data quality, the welfare analysis becomes more complex—both consumption and data quality must be accounted for in calculating welfare. Generally, data quality in the optimal allocation always grows at a rate no lower than that in the competitive equilibrium, indicating a tendency to insufficiently use producer data in the absence of encouraging policies, particularly when AI-generated data dominate the economy. As shown in the figure, the growth rate of data quality declines sharply as ζ increases, while the increase in consumption is relatively modest and fails to offset this negative effect. In this regime, the potential risks associated with data quality become a significant concern for the economy, underscoring the need for regulatory intervention in this sector.

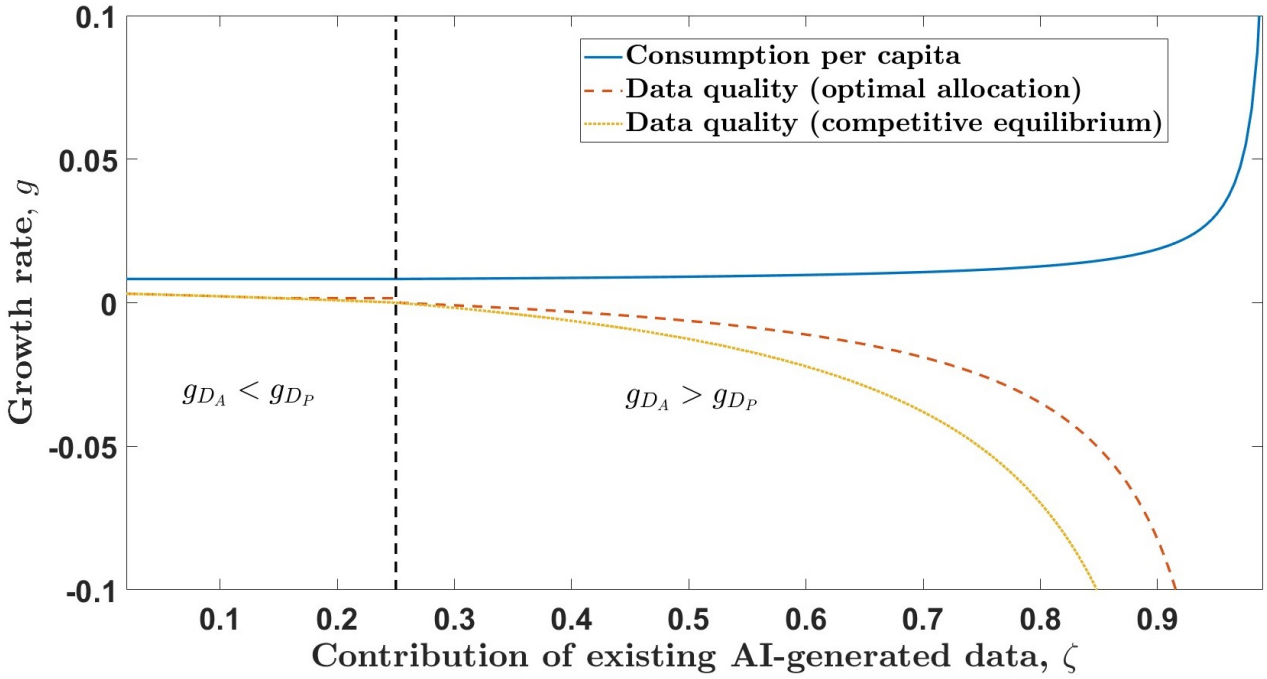


Figure 3: Growth rates of consumption per capita and data quality

Note. These figures illustrate the growth rates of consumption and data quality. While the growth rates of per capita consumption remain the same across different settings, the growth rates of data quality vary. The blue solid line represents per capita consumption, the orange dashed line indicates data quality in the optimal allocation, and the yellow dotted line denotes data quality in the competitive equilibrium. A black vertical dashed line separates the two regimes: $g_{D_A} < g_{D_P}$ (left region) and $g_{D_A} > g_{D_P}$ (right region). It is important to note that in the competitive equilibrium, when $g^{dc}_{D_A} > g^{dc}_{D_P}$, we derive only an upper bound for the growth rate of producer data. Consequently, the actual growth rate of data quality in this setting may be even lower than the dotted line depicted in the figure.

6.3 Number of Generative AI Firms

In Proposition 4, we show that multiple Generative AI firms may exist under the competitive equilibrium, whereas the optimal number of such firms is always one. Figure 4 illustrates how this number varies with respect to two key parameters, ζ and $\mathcal{G}$. We present only the case in which ζ is sufficiently large, as Generative AI firms have a negligible impact on the economy when producer data dominate. The emergent capabilities parameter $\mathcal{G}$ is shown in logarithmic form, given its relatively large absolute value. All other parameters take standard values as reported in Table 2.

From the figure, we observe that the number of Generative AI firms decreases as $\mathcal{G}$ increases, which is consistent with our analysis. Currently, a potential entrant to the Generative AI industry must make substantial upfront investments. This explains why only a few oligopolistic enterprises are currently developing LLMs. If these investment requirements

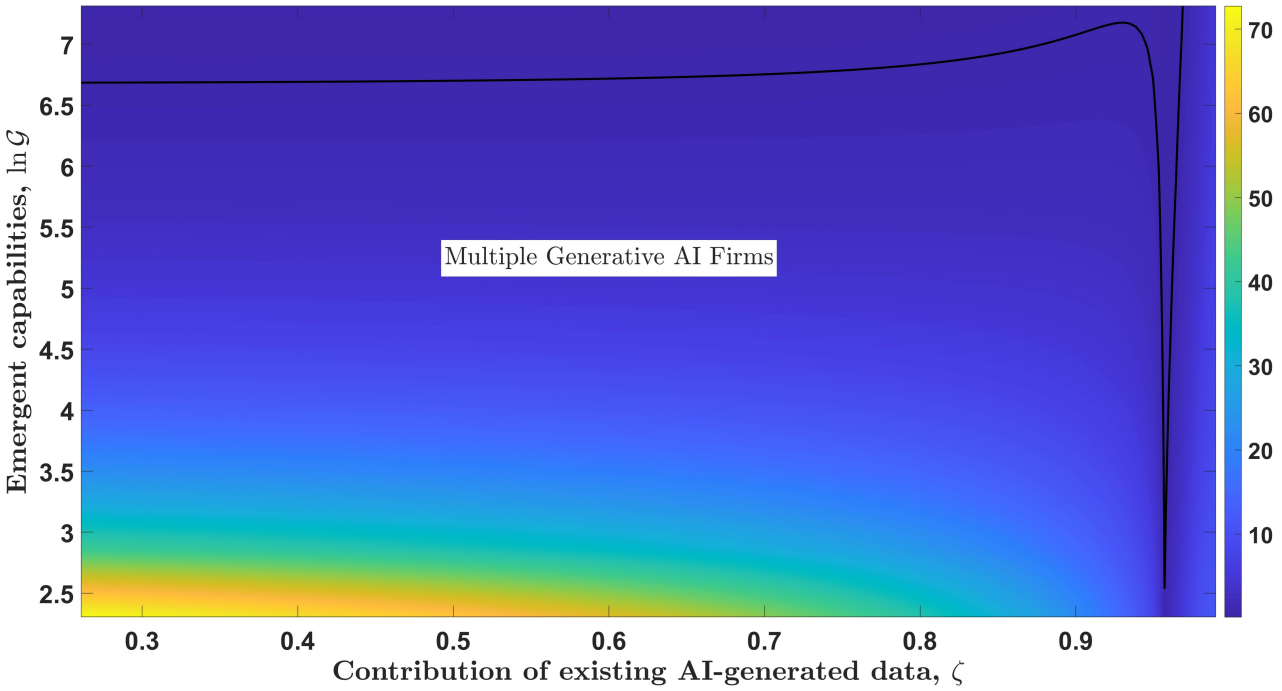


Figure 4: Number of Generative AI firms in Competitive Equilibrium

Note. This figure presents the quantitative analysis of the number of Generative AI firms in the competitive equilibrium across a wide range of values for ζ and $\mathcal{G}$. We consider only the case in which AI-generated data dominate the economy, and we use the logarithmic values of the emergent capabilities parameter $\mathcal{G}$ to more clearly illustrate changes in the number of firms. The yellow and light-colored regions correspond to parameter regimes in which the number is relatively large, while the blue and dark-colored regions indicate the opposite. To facilitate comparison with the optimal allocation, we include a contour line representing the case in which the number of Generative AI firms is one.

decline in the future, more firms will likely enter the industry. On the other hand, it is noteworthy that the number of Generative AI firms also declines as ζ increases. As the contribution of existing AI-generated data to the production of new data becomes larger, firms must accumulate a sufficient stock of data before earning profits. In this case, the presence of more firms disperses the stock of AI-generated data available to any single firm, thereby limiting the number of firms the market can sustain.

6.4 Policy implications

Several issues related to the development of generative AI and the mitigation of potential risks are highlighted by this framework. First, although AI-generated data serve as a powerful substitute for conventional data in many fields of production, we must remain cautious about the dystopian aspects of their widespread use. The potential risks associated with this new type of data may be more severe than those posed by other data types, as AI-generated data

often contain limited or even inaccurate information about the real world.[10] As a result, AI-generated data should always complement, rather than replace, other types of data collected from consumers or production processes. This is crucial for ensuring sustainable economic growth, as our model demonstrates that firms in the competitive equilibrium tend to neglect the usage of real-world data, such as producer data. Given this, while encouraging the usage of AI-generated data, governments should introduce policies to impose an lower limit on the use of producer data (or other real-world data) to mitigate potential risks, including the existential threats associated with AI overuse, as discussed in Jones (2024). Alternative policies could include providing subsidies for the combination of AI-generated data and real-world data sales.

Another important conclusion to highlight is that, in most cases, there are too many Generative AI firms in the market. The cumulative nature of AI-generated data makes such data more desirable when the resources used to produce them are concentrated within a single firm. The presence of multiple firms leads to redundant investment, and there appears to be little justification for supporting this market structure. This conclusion aligns with the scenario discussed in Begenau et al. (2018), where the authors show that large firms, due to their longer histories and greater data accumulation, are more likely to generate profits than smaller firms. As a result, the government need not be overly concerned about the concentration trend in the AI industry. Instead, regulatory efforts should focus on ensuring that such firms do not harm overall societal welfare.

7. Conclusion

We develop an endogenous growth model that incorporates AI-generated data from the perspective of data quality to highlight the potential risks of production errors arising from the widespread use of Generative AI. In addition, we emphasize the importance of integrating real-world data, such as the producer data considered in our model. The conclusions of the model are categorized into three regimes based on the relative growth rates of different types of data, with each regime exhibiting distinct properties under both the optimal allocation and the competitive equilibrium. Overall, we find that firms tend to underutilize producer data due to its relatively high cost and the absence of a multiplier effect, unlike the generation process of AI-generated data. We also find that the number of Generative AI firms in the market exceeds the socially optimal level, supporting the concentration trend of the AI indus-

[10] We acknowledge that other types of data, such as consumer data, contain personal information about individuals, but privacy risks associated with these data can be mitigated through technologies like desensitization. Moreover, producer data, which are more commonly used in production, pose significantly fewer risks compared to the concerns surrounding AI-generated data discussed in this paper.

try. These findings raise important policy considerations regarding whether governments should regulate the use of AI-generated data and oversee the broader development of the AI industry.

In addition to the two recently published papers that examine the risks of AI technologies, Jones (2024) and Acemoglu and Lensman (2024), we develop an integrated model that incorporates production, innovation, and the selection between different types of data. Our study contributes to the literature by advancing growth theory on risks associated with AI technologies and data as a factor of production. For the sake of tractability and focus, we have, by necessity, omitted several important aspects, such as discussions on property rights of AI-generated data and the integration of AI-generated data with other data types beyond the producer data analyzed in this paper. Thus, our study should be viewed as an initial attempt to comprehensively examine AI-related risks rather than as a precise representation of current reality. We hope future research will extend our framework to explore these issues further.

References

Acemoglu, D. (2024). The Simple Macroeconomics of AI. NBER Working Papers 32487, National Bureau of Economic Research.

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

Arrow, K. J. (1962). The Economic Implications of Learning by Doing. *Review of Economic Studies 29*(3), 155–173.

Bail, C. (2024). Can Generative Artificial Intelligence Improve Social Science? *Proceedings of the National Academy of Sciences 121*(21), e2314021121.

Begenau, J., M. Farboodi, and L. Veldkamp (2018). Big Data in Finance and the Growth of Large Firms. *Journal of Monetary Economics 97*, 71–87.

Bergemann, D. and A. Bonatti (2024). Data, Competition, and Digital Platforms. *American Economic Review 114*(8), 2553–2595.

Brooks, C., S. Eggert, and D. Peskoff (2024). The Rise of AI-Generated Content in Wikipedia. arXiv:2410.08044.

Brynjolfsson, E., D. Li, and L. R. Raymond (2025). Generative AI at Work. *Quarterly Journal of Economics 140*(2), 889–942.

Cong, L. W., W. Wei, D. Xie, and L. Zhang (2022). Endogenous Growth Under Multiple Uses of Data. *Journal of Economic Dynamics and Control 141*, 104395.

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

Cottier, B., R. Rahman, L. Fattorini, N. Maslej, T. Besiroglu, and D. Owen (2025). The Rising Costs of Training Frontier AI Models. arXiv:2405.21015v2.

del Rio-Chanona, M., N. Laurentsyeva, and J. Wachs (2023). Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow. arXiv:2307.07367.

Farboodi, M. and L. Veldkamp (2023). A Model of the Data Economy. NBER Working Papers 28427, National Bureau of Economic Research.

Ichihashi, S. (2021). Competing Data Intermediaries. *RAND Journal of Economics 52*(3), 515–537.

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.

Korinek, A. and J. Vipra (2024). Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence. NBER Working Papers 33139, National Bureau of Economic Research.

Li, J., X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv:2305.11747v3.

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.

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

Xie, D. and L. Zhang (2023). A Generalized Model of Growth in the Data Economy. Available at SSRN 4033576.

Yang, K. H. (2022). Selling Consumer Data for Profit: Optimal Market-Segmentation Design and Its Consequences. *American Economic Review 112*(4), 1364–1393.

Online Appendix

A. Optimal Allocation

The social planner's problem is formulated as:

$$\max_{L_{P,t},L_{A,t},L_{R,t}} \int_0^\infty e^{-(\rho-n)t} \ln c_t \mathrm{d}t,$$

subject to

$$\begin{aligned} \dot{d}_{A,t} &= \frac{\psi}{M} d_{A,t}^{\zeta} L_{A,t} - \delta_A d_{A,t}, \\ D_{A,t} &= M d_{A,t}, \\ D_{P,t} &= c_t^{\theta} L_{P,t}, \\ \dot{N}_t &= \frac{1}{\chi} L_{R,t}, \end{aligned} \tag{A.1}$$

$$c_t L_t = \left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right), \tag{A.2}$$

$$Q_t = \left(\frac{D_{P,t}}{D_{A,t}}\right)^{\tau}. \tag{A.3}$$

Here, equations (A.1) and (A.2) follow from the resource constraint (18). The first and second condition can be combined to form the law of motion of $D_{A,t}$:

$$\dot{D}_{A,t} = \frac{\psi}{M^{\zeta}} D_{A,t}^{\zeta} L_{A,t} - \delta_A D_{A,t}. \tag{A.4}$$

Next, we define the current-value Hamiltonian with state variables $D_{A,t}$ and N_t, control variables $\{L_{P,t}, L_{A,t}, L_{R,t}\}$, co-state variables λ_t and μ_t, and shadow prices $\pi_{D,t}$, $\pi_{c,t}$, and $\pi_{Q,t}$:

$$\begin{aligned} \mathcal{H} = &\ln c_t + \lambda_t \left(\frac{\psi}{M^{\zeta}} D_{A,t}^{\zeta} L_{A,t} - \delta_A D_{A,t}\right) + \frac{\mu_t}{\chi} L_{R,t} + \pi_{D,t} \left(c_t^{\theta} L_{P,t} - D_{P,t}\right) \\ &+ \pi_{c,t} \left\{ \left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right) - c_t L_t \right\} \end{aligned}$$

$$+ \pi_{Q,t}\left[\left(\frac{D_{P,t}}{D_{A,t}}\right)^{\tau} - Q_t\right].$$

The FOCs with respect to c_t, Q_t, $D_{A,t}$, $L_{A,t}$, $D_{P,t}$, $L_{P,t}$, N_t, $L_{R,t}$ are shown as follows:

$$\frac{\partial \mathcal{H}}{\partial c_t} = \frac{1}{c_t} + \theta \pi_{D,t} c_t^{\theta-1} L_{P,t} - \pi_{c,t} L_t = 0, \tag{A.5}$$

$$\frac{\partial \mathcal{H}}{\partial Q_t} = \pi_{c,t}\xi e_0 \exp\{-\xi Q_t\} N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) - \pi_{Q,t} = 0, \tag{A.6}$$

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial D_{A,t}} =& \lambda_t \left(\frac{\psi\zeta}{M^{\zeta}} D_{A,t}^{\zeta-1} L_{A,t} - \delta_A\right) + (1-\beta)\eta\pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon(\eta-1)+1}{\varepsilon-1}} \\ & D_{A,t}^{-\frac{1}{\varepsilon}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) - \pi_{Q,t}\tau D_{P,t}^{\tau} D_{A,t}^{-\tau-1} = -\dot{\lambda}_t + (\rho - n)\lambda_t,\end{aligned} \tag{A.7}$$

$$\frac{\partial \mathcal{H}}{\partial L_{A,t}} = \frac{\psi}{M^{\zeta}} \lambda_t D_{A,t}^{\zeta} - \pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} = 0, \tag{A.8}$$

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial D_{P,t}} =& -\pi_{D,t} + \beta\eta\pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon(\eta-1)+1}{\varepsilon-1}} \\ & D_{P,t}^{-\frac{1}{\varepsilon}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) + \pi_{Q,t}\tau D_{P,t}^{\tau-1} D_{A,t}^{-\tau} = 0,\end{aligned} \tag{A.9}$$

$$\frac{\partial \mathcal{H}}{\partial L_{P,t}} = \pi_{D,t} c_t^{\theta} - \pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} = 0, \tag{A.10}$$

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial N_t} =& \frac{1}{\sigma-1}\pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}-1} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \\ & (L_t - L_{P,t} - L_{A,t} - L_{R,t}) = -\dot{\mu}_t + (\rho - n)\mu_t,\end{aligned} \tag{A.11}$$

$$\frac{\partial \mathcal{H}}{\partial L_{R,t}} = \frac{\mu_t}{\chi} - \pi_{c,t}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} = 0. \tag{A.12}$$

A.1 Growth Rates Along the Balanced Growth Path

First, observe that from the innovation possibility frontier (15), we obtain

$$\frac{\dot{N}_t}{N_t} = \frac{1}{\chi} \cdot \frac{L_{R,t}}{N_t}.$$

Along a balanced growth path, the labor share $L_{R,t}$ grows at the same rate as the total population, while N_t expands at a constant rate. Hence, the growth rate of N_t is given by $g_N = n$. Consequently, we derive

$$\frac{L_{R,t}}{N_t} = n\chi. \tag{A.13}$$

Next, we determine the growth rate of the overall dataset D_t. From the definition (11), it follows that

$$g_D = \begin{cases} g_{D_A}, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ g_{D_P}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \tag{A.14}$$

In the simple model, we derive the growth rate of $D_{A,t}$ as given in equation (2). However, this rate may change when $g_{D_A} < g_{D_P}$ since $l_{A,t}$ may shrink to zero. Therefore, we define the growth rate of $D_{A,t}$ as

$$g_{D_A} = \begin{cases} \dfrac{1}{1-\zeta} n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \dfrac{n + g_{l_A}}{1-\zeta}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases}$$

Meanwhile, from equation (A.1), we obtain $g_{D_P} = \theta g_c + n$, where g_c denotes the growth rate of consumption. Substituting this into equation (A.14), we further derive

$$g_D = \begin{cases} \dfrac{1}{1-\zeta} n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \theta g_c + n, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \tag{A.15}$$

Next, from equation (13), we observe that along a balanced growth path, data quality Q_t follows one of the following patterns: $Q_t \to \infty$ if $g_{D_A} < g_{D_P}$, $Q_t \to 0$ if $g_{D_A} > g_{D_P}$, or Q_t converges to a constant if $g_{D_A} = g_{D_P}$. Since Q_t determines the long-run trajectory of the error rate e_t, we obtain

$$e_t = \begin{cases} 1, & \text{if} \quad g_{D_A} > g_{D_P}, \\ e_0 \exp\left(-\xi \bar{Q}\right), & \text{if} \quad g_{D_A} = g_{D_P}, \\ 0, & \text{if} \quad g_{D_A} < g_{D_P}, \end{cases} \tag{A.16}$$

where $\bar{Q}$ is determined by the values of $D_{A,t}$ and $D_{P,t}$ when their growth rates are equal.

From the resource constraint (A.2), the growth rate of consumption is derived as:

$$g_c = \frac{1}{\sigma-1} g_N + \eta g_D = \begin{cases} \left(\frac{1}{\sigma-1} + \frac{\eta}{1-\zeta}\right) n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1} + \eta\right) n, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases}$$

Here, the threshold is obtained by substituting the growth rates of $D_{A,t}$ and $D_{P,t}$. Specifically, we require

$$\frac{1-\theta\eta}{1-\zeta} < \frac{\theta}{\sigma-1} + 1$$

when $g_{D_A} < g_{D_P}$, with the inequality sign reversing when $g_{D_A} > g_{D_P}$. Consequently, the growth rate of producer data is given by

$$g_{D_P} = \theta g_c + n = \begin{cases} \left[1 + \theta\left(\frac{1}{\sigma-1} + \frac{\eta}{1-\zeta}\right)\right] n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \left[1 + \frac{\theta}{1-\theta\eta}\left(\frac{1}{\sigma-1} + \eta\right)\right] n, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \tag{A.17}$$

We now derive the growth rate of $l_{A,t}$ when $g_{D_A} < g_{D_P}$. From equation (A.6), we obtain

$$g_{\pi_c} + \frac{1}{\sigma-1} g_N + \eta g_D + n = g_{\pi_Q}. \tag{A.18}$$

Meanwhile, from equations (A.10) and (A.2), we have

$$\pi_{D,t} c_t^\theta = \frac{\pi_{c,t} c_t}{1 - l_{P,t} - l_{A,t} - l_{R,t}}. \tag{A.19}$$

Furthermore, from equation (A.5), it follows that

$$\begin{aligned} & 1 + \theta \pi_{D,t} c_t^\theta L_{P,t} = \pi_{c,t} c_t L_t, \\ \Rightarrow \quad & 1 + \theta \pi_{c,t} c_t L_t \frac{l_{P,t}}{1 - l_{P,t} - l_{A,t} - l_{R,t}} = \pi_{c,t} c_t L_t, \\ \Rightarrow \quad & \pi_{c,t} c_t L_t = \frac{1 - l_{P,t} - l_{A,t} - l_{R,t}}{1 - (1+\theta) l_{P,t} - l_{A,t} - l_{R,t}}. \end{aligned} \tag{A.20}$$

Thus, we obtain

$$g_{\pi_c} + g_c + n = 0. \tag{A.21}$$

Additionally, from equation (A.10), we have

$$g_{\pi_D} + \theta g_c = g_{\pi_c} + g_c. \tag{A.22}$$

Substituting equations (A.18), (A.21), and (A.22) into (A.9), we find that the growth rate of the second term exceeds that of the first term. Consequently, we obtain

$$\begin{aligned} & g_{\pi_c} + (\eta - 1)g_{D_P} + g_c + n = g_{\pi_Q} + (\tau - 1)g_{D_P} - \tau g_{D_A}, \\ \Rightarrow \quad & (\eta - 1)g_{D_P} + g_c + n = \frac{1}{\sigma - 1}n + \eta g_{D_P} + n + (\tau - 1)g_{D_P} - \tau g_{D_A}, \\ \Rightarrow \quad & g_{D_A} = \frac{\tau - \eta}{\tau} g_{D_P}. \end{aligned}$$

Then, we derive

$$\begin{aligned} & \frac{n + g_{l_A}}{1 - \zeta} = \frac{\tau - \eta}{\tau}\left[1 + \frac{\theta}{1 - \theta\eta}\left(\frac{1}{\sigma - 1} + \eta\right)\right] n, \\ \Rightarrow \quad & g_{l_A}|_{g_{D_A} < g_{D_P}} = \left[\frac{(\tau - \eta)(1 - \zeta)}{\tau(1 - \theta\eta)}\left(\frac{\theta}{\sigma - 1} + 1\right) - 1\right] n, \end{aligned}$$

and

$$g_{D_A}|_{g_{D_A} < g_{D_P}} = \frac{\tau - \eta}{\tau}\left[1 + \frac{\theta}{1 - \theta\eta}\left(\frac{1}{\sigma - 1} + \eta\right)\right] n.$$

Since the growth rate of the labor share $l_{A,t}$ must be negative, we require

$$(\tau - \eta)\left(\frac{\theta}{\sigma - 1} + 1\right) < \frac{\tau(1 - \theta\eta)}{1 - \zeta}.$$

Otherwise, we obtain the corner solution: $g_{l_A} = 0$, $l_{A,t} = 0$, and $D_{A,t} \to 0$.

By now, we derive the growth rates of all the variables.

A.2 Labor Shares Allocated in Different Sectors

First, we derive the relationship between the co-state variables and the shadow prices. From equation (A.10), the relationship between the two shadow prices, $\pi_{D,t}$ and $\pi_{c,t}$, is given by

$$\frac{\pi_{D,t}}{\pi_{c,t}} = c_t^{-\theta}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma - 1}} \left[\beta D_{P,t}^{\frac{\varepsilon - 1}{\varepsilon}} + (1 - \beta) D_{A,t}^{\frac{\varepsilon - 1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon - 1}}. \tag{A.23}$$

Similarly, from equation (A.8), we obtain the relationship between the co-state variable λ_t and the shadow price $\pi_{c,t}$:

$$\frac{\lambda_t}{\pi_{c,t}} = \frac{M^\zeta}{\psi} D_{A,t}^{-\zeta}\left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma - 1}} \left[\beta D_{P,t}^{\frac{\varepsilon - 1}{\varepsilon}} + (1 - \beta) D_{A,t}^{\frac{\varepsilon - 1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon - 1}}. \tag{A.24}$$

By combining equations (A.23) and (A.24), we derive the relationship between the co-state variable λ_t and the shadow price $\pi_{D,t}$ as

$$\frac{\lambda_t}{\pi_{D,t}} = \frac{M^{\zeta}}{\psi} D_{A,t}^{-\zeta} c_t^{\theta}. \tag{A.25}$$

Substituting equations (A.24) and (A.25) into (A.9), we obtain the relationship between the co-state variable λ_t and the shadow price $\pi_{Q,t}$:

$$\frac{\pi_{Q,t}}{\lambda_t} = \frac{\psi}{\tau M^{\zeta}} D_{A,t}^{\zeta+\tau} D_{P,t}^{1-\tau} \left\{ c_t^{-\theta} - \beta\eta \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-1} D_{P,t}^{-\frac{1}{\varepsilon}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) \right\}. \tag{A.26}$$

From equations (A.24) and (A.26), we derive the relationship between the two shadow prices, $\pi_{Q,t}$ and $\pi_{c,t}$:

$$\begin{aligned} \frac{\pi_{Q,t}}{\pi_{c,t}} = \frac{\pi_{Q,t}}{\lambda_t} \cdot \frac{\lambda_t}{\pi_{c,t}} = \frac{1}{\tau} D_{A,t}^{\tau} D_{P,t}^{1-\tau} \left[1 - e_0 \exp(-\xi Q_t) \right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \cdots \\ \left\{ c_t^{-\theta} - \beta\eta \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-1} D_{P,t}^{-\frac{1}{\varepsilon}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) \right\}. \end{aligned} \tag{A.27}$$

Finally, from equation (A.12), the relationship between the shadow price $\pi_{c,t}$ and the co-state variable μ_t is given by

$$\frac{\pi_{c,t}}{\mu_t} = \frac{1}{\chi} \left[1 - e_0 \exp(-\xi Q_t) \right]^{-1} N_t^{-\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-\frac{\eta\varepsilon}{\varepsilon-1}} = \frac{1}{\chi} \cdot \frac{1 - l_{P,t} - l_{A,t} - l_{R,t}}{c_t}. \tag{A.28}$$

Next, we proceed to derive the labor share and the levels of the variables. Substituting equations (A.24) and (A.26) into (A.7), we obtain

$$\begin{aligned} & \lambda_t \left(\psi M^{-\zeta} \zeta D_{A,t}^{\zeta-1} L_{A,t} - \delta_A \right) + \lambda_t (1-\beta)\eta\psi M^{-\zeta} D_{A,t}^{\zeta-\frac{1}{\varepsilon}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-1} \cdots \\ & (L_t - L_{P,t} - L_{A,t} - L_{R,t}) - \lambda_t \psi M^{-\zeta} D_{A,t}^{\zeta-1} D_{P,t} \left\{ c_t^{-\theta} - \beta\eta \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-1} \cdots \right. \\ & \left. D_{P,t}^{-\frac{1}{\varepsilon}} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) \right\} = -\dot{\lambda}_t + (\rho - n)\lambda_t \\ \Rightarrow \quad & \lambda_t \left(\psi M^{-\zeta} \zeta D_{A,t}^{\zeta-1} L_{A,t} - \delta_A \right) - \lambda_t \psi M^{-\zeta} D_{A,t}^{\zeta-1} c_t^{-\theta} D_{P,t} + \lambda_t \psi M^{-\zeta} \eta D_{A,t}^{\zeta-1} (L_t - L_{P,t} - L_{A,t} - L_{R,t}) \\ & = -\dot{\lambda}_t + (\rho - n)\lambda_t \\ \Rightarrow \quad & \psi M^{-\zeta} D_{A,t}^{\zeta-1} \left[\zeta L_{A,t} - c_t^{-\theta} D_{P,t} + \eta (L_t - L_{P,t} - L_{A,t} - L_{R,t}) \right] = -\frac{\dot{\lambda}_t}{\lambda_t} + \rho - n + \delta_A \end{aligned}$$

$$\Rightarrow \quad \psi M^{-\zeta} D_{A,t}^{\zeta-1} \left[\zeta L_{A,t} - L_{P,t} + \eta(L_t - L_{P,t} - L_{A,t} - L_{R,t})\right] = -\frac{\dot{\lambda}_t}{\lambda_t} + \rho - n + \delta_A$$

$$\Rightarrow \quad \psi M^{-\zeta} D_{A,t}^{\zeta-1} \left[\eta L_t - (1+\eta) L_{P,t} - (\eta - \zeta) L_{A,t} - \eta L_{R,t}\right] = -\frac{\dot{\lambda}_t}{\lambda_t} + \rho - n + \delta_A \tag{A.29}$$

$$\Rightarrow \quad \left(\frac{\dot{D}_{A,t}}{D_{A,t}} + \delta_A\right) \left[\frac{\eta - (1+\eta) l_{P,t} - (\eta - \zeta) l_{A,t} - \eta l_{R,t}}{l_{A,t}}\right] = -\frac{\dot{\lambda}_t}{\lambda_t} + \rho - n + \delta_A \tag{A.30}$$

Here, the fourth equation follows from the market clearing condition for producer data (A.1), while the final equation is derived from the law of motion of $D_{A,t}$ in equation (10). The variables $l_{P,t}$, $l_{R,t}$, and $l_{A,t}$ represent the labor shares employed in their respective sectors, all of which remain constant along a balanced growth path.

Next, substituting equations (A.24) and (A.25) into (A.5), we obtain

$$\begin{aligned}
&\frac{1}{c_t} + \lambda_t \theta \psi M^{-\zeta} D_{A,t}^{\zeta} c_t^{-1} L_{P,t} - \lambda_t \psi M^{-\zeta} D_{A,t}^{\zeta} \left[1 - e_0 \exp(-\xi Q_t)\right]^{-1} N_t^{-\frac{1}{\sigma-1}} \ldots \\
&\left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{-\frac{\eta\varepsilon}{\varepsilon-1}} L_t = 0 \\
\Rightarrow \quad &\frac{1}{c_t} + \lambda_t \psi M^{-\zeta} D_{A,t}^{\zeta} c_t^{-1} \left[\theta L_{P,t} - (L_t - L_{P,t} - L_{A,t} - L_{R,t})\right] = 0 \qquad \text{(A.31)} \\
\Rightarrow \quad &\frac{\dot{\lambda}_t}{\lambda_t} + \zeta \frac{\dot{D}_{A,t}}{D_{A,t}} + n = 0 \qquad \text{(A.32)} \\
\Rightarrow \quad &\left(\frac{\dot{D}_{A,t}}{D_{A,t}} + \delta_A\right) \left[\zeta - \frac{\eta - (1+\eta) l_{P,t} - (\eta - \zeta) l_{A,t} - \eta l_{R,t}}{l_{A,t}}\right] + \rho + (1-\zeta)\delta_A = 0 \\
\Rightarrow \quad &\left(g_{D_A} + \delta_A\right) \left[\zeta - \frac{\eta - (1+\eta) l_{P,t} - (\eta - \zeta) l_{A,t} - \eta l_{R,t}}{l_{A,t}}\right] + \rho + (1-\zeta)\delta_A = 0 \\
\Rightarrow \quad &(1+\eta) l_{P,t} + \left[\eta + \frac{\rho + (1-\zeta)\delta_A}{g_{D_A} + \delta_A}\right] l_{A,t} + \eta l_{R,t} - \eta = 0. \qquad \text{(A.33)}
\end{aligned}$$

Here, the third equation is obtained by expressing variables in terms of growth rates, noting that the labor shares employed in all four sectors grow at the same rate n, which matches the growth rate of the total population. The fourth equation follows from substituting the result derived in equation (A.30). Since we have established the growth rates of c_t and $D_{A,t}$ along a balanced growth path, equation (A.33) serves as the first key equation for the allocation of labor shares across different sectors.

Next, we derive the second key equation for the labor shares. From equation (A.32), we obtain

$$\frac{\dot{\lambda}_t}{\lambda_t} = -\zeta \frac{\dot{D}_{A,t}}{D_{A,t}} - n = -\frac{1}{1-\zeta} n.$$

Then, combining the law of motion of $D_{A,t}$ from equation (10) with the resource constraint

(A.2), and using equation (A.24), we derive

$$\begin{aligned}
&\frac{\lambda_t}{\pi_{c,t}} = \frac{L_{A,t}}{(g_{D_A}+\delta_A)D_{A,t}} \cdot \frac{c_t}{(1-l_{P,t}-l_{A,t}-l_{R,t})} \\
\Rightarrow \quad &\frac{\dot{\pi}_{c,t}}{\pi_{c,t}} = \frac{\dot{\lambda}_t}{\lambda_t} - n + \frac{\dot{D}_{A,t}}{D_{A,t}} - \frac{\dot{c}_t}{c_t} \\
\Rightarrow \quad &\frac{\dot{\pi}_{c,t}}{\pi_{c,t}} = \begin{cases} -\left[\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}+1\right]n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ -\left[\frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)+1\right]n, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases}
\end{aligned}$$

From equation (A.28), it follows that

$$\frac{\dot{\mu}_t}{\mu_t} = \frac{\dot{\pi}_{c,t}}{\pi_{c,t}} + \frac{\dot{c}_t}{c_t} = -n. \tag{A.34}$$

Substituting (A.28) into (A.11), we obtain

$$\begin{aligned}
&\frac{1}{\chi(\sigma-1)} N_t^{-1} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right) = -\frac{\dot{\mu}_t}{\mu_t} + \rho - n, \qquad \text{(A.35)} \\
&\Rightarrow \quad \frac{n}{\sigma-1}\left(\frac{1-l_{P,t}-l_{A,t}-l_{R,t}}{l_{R,t}}\right) = -\frac{\dot{\mu}_t}{\mu_t} + \rho - n, \\
&\Rightarrow \quad \frac{n}{\sigma-1}\left(\frac{1-l_{P,t}-l_{A,t}-l_{R,t}}{l_{R,t}}\right) = \rho, \\
\Rightarrow \quad &l_{P,t} + l_{A,t} + \left[1 + \frac{\rho(\sigma-1)}{n}\right] l_{R,t} - 1 = 0. \qquad \text{(A.36)}
\end{aligned}$$

The second equation follows from equation (A.13), while the third equation results from substituting the growth rate of μ_t from equation (A.34). We thus obtain the second key equation that determines the allocation of labor shares.

Finally, we derive the third key equation for the labor shares. Substituting (A.27) into (A.6), we obtain

$$\begin{aligned}
&\pi_{c,t}\xi e_0 \exp(-\xi Q_t) N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right) - \ldots \\
&\frac{1}{\tau}\pi_{c,t} D_{A,t}^{\tau} D_{P,t}^{1-\tau} \left[1 - e_0 \exp(-\xi Q_t)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} \ldots \\
&\left\{c_t^{-\theta} - \beta\eta \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{-1} D_{P,t}^{-\frac{1}{\varepsilon}} \left(L_t - L_{P,t} - L_{A,t} - L_{R,t}\right)\right\} = 0 \\
\Rightarrow \quad &\pi_{c,t}\xi \frac{e_0 \exp(-\xi Q_t)}{1 - e_0 \exp(-\xi Q_t)} c_t L_t - \frac{1}{\tau Q_t} \pi_{c,t} \frac{c_t L_t}{1 - l_{P,t} - l_{A,t} - l_{R,t}} \ldots
\end{aligned}$$

$$\left\{ l_{P,t} - \beta\eta \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{-1} D_{P,t}^{1-\frac{1}{\varepsilon}} (1 - l_{P,t} - l_{A,t} - l_{R,t}) \right\} = 0. \tag{A.37}$$

When $g_{D_A} > g_{D_P}$, equation (A.37) simplifies to

$$\begin{aligned}
& \pi_{c,t} \xi \frac{e_0 \exp(-\xi Q_t)}{1 - e_0 \exp(-\xi Q_t)} c_t L_t - \frac{1}{\tau Q_t} \pi_{c,t} \frac{c_t L_t}{1 - l_{P,t} - l_{A,t} - l_{R,t}} \cdots \\
& \left\{ l_{P,t} - \beta\eta \left[\beta \left(\frac{D_{P,t}}{D_{A,t}} \right)^{\frac{\varepsilon-1}{\varepsilon}} + 1 - \beta \right]^{-1} \left(\frac{D_{P,t}}{D_{A,t}} \right)^{1-\frac{1}{\varepsilon}} (1 - l_{P,t} - l_{A,t} - l_{R,t}) \right\} = 0 \\
\Rightarrow \quad & \xi \frac{e_0}{1 - e_0} - \frac{1}{\tau Q_t} \frac{l_{P,t}}{1 - l_{P,t} - l_{A,t} - l_{R,t}} = 0 \\
\Rightarrow \quad & Q_t = \frac{l_{P,t}(1 - e_0)}{(1 - l_{P,t} - l_{A,t} - l_{R,t}) \tau \xi e_0}.
\end{aligned} \tag{A.38}$$

The second equation follows from the fact that $Q_t \to 0$ when $g_{D_A} > g_{D_P}$, while the third equation follows from the observation that the growth rate of $\pi_{c,t} c_t L_t$ is zero, given that

$$\frac{\dot{\pi}_{c,t}}{\pi_{c,t}} + \frac{\dot{c}_t}{c_t} + n = 0.$$

Since we have $l_{P,t} \to 0$ in this scenario, equation (A.38) always holds but cannot serve as a key equation for determining the labor shares.

When $g_{D_A} < g_{D_P}$, i.e., as $Q_t \to \infty$, the first and last terms on the left-hand side of equation (A.37) converge to zero in the long run, while the second term remains constant. Consequently, equation (A.37) simplifies to

$$\pi_{c,t} \xi \frac{e_0 \exp(-\xi Q_t)}{1 - e_0 \exp(-\xi Q_t)} c_t L_t = 0,$$

which trivially holds but cannot serve as a key equation for determining the labor shares.

When $g_{D_A} = g_{D_P}$, we have $(D_{P,t}/D_{A,t})^{\tau} = \bar{Q} \in (0, +\infty)$. Consequently, equation (A.37) simplifies to

$$\begin{aligned}
& \frac{\kappa}{\bar{Q}} + \pi_{c,t} \xi \frac{e_0 \exp(-\xi \bar{Q})}{1 - e_0 \exp(-\xi \bar{Q})} c_t L_t - \pi_{c,t} \frac{1}{\tau \bar{Q}} \frac{c_t L_t}{1 - l_{P,t} - l_{A,t} - l_{R,t}} \cdots \\
& \left\{ l_{P,t} - \beta\eta \left(\beta \bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} + 1 - \beta \right)^{-1} \bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} (1 - l_{P,t} - l_{A,t} - l_{R,t}) \right\} = 0 \\
\Rightarrow \quad & \frac{\kappa}{\bar{Q}} + \xi \frac{e_0 \exp(-\xi \bar{Q})}{1 - e_0 \exp(-\xi \bar{Q})} \frac{1 - l_{P,t} - l_{A,t} - l_{R,t}}{1 - (1+\theta) l_{P,t} - l_{A,t} - l_{R,t}} - \frac{1}{\tau \bar{Q}} \frac{1}{1 - (1+\theta) l_{P,t} - l_{A,t} - l_{R,t}} \cdots
\end{aligned}$$

$$\left[l_{P,t} - \beta\eta\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} + 1 - \beta\right)^{-1}\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}(1 - l_{P,t} - l_{A,t} - l_{R,t})\right] = 0, \tag{A.39}$$

where the second equation follows from substituting equation (A.20) into the expression. Substituting equation (3), $\bar{Q}$ can be derived as

$$\bar{Q} = \left(\frac{D_{P,t}}{D_{A,t}}\right)^{\tau} = \left\{\frac{c_t^{\theta} L_{P,t}}{\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n + \delta_A(1-\zeta)}\right]^{\frac{1}{1-\zeta}} (L_{A,t})^{\frac{1}{1-\zeta}}}\right\}^{\tau}.$$

In the above equation, consumption c_t is given by

$$\begin{aligned} c_t &= \left[1 - e_0 \exp\left(-\xi\bar{Q}\right)\right] N_t^{\frac{1}{\sigma-1}} \left[\beta\left(\frac{D_{P,t}}{D_{A,t}}\right)^{\frac{\varepsilon-1}{\varepsilon}} + 1 - \beta\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} D_{A,t}^{\eta}(1 - l_{P,t} - l_{A,t} - l_{R,t}) \\ &= \left[1 - e_0 \exp\left(-\xi\bar{Q}\right)\right] \left(\frac{l_{R,t}}{n\chi}\right)^{\frac{1}{\sigma-1}} \left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} + 1 - \beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}} \left[\frac{\psi M^{-\zeta}(1-\zeta)}{n + \delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} l_{A,t}^{\frac{\eta}{1-\zeta}} \cdots \\ &\quad (1 - l_{P,t} - l_{A,t} - l_{R,t}) L_t^{\frac{1}{\sigma-1} + \frac{\eta}{1-\zeta}}. \end{aligned}$$

Thus, we obtain

$$\begin{aligned} \bar{Q}^{\frac{1}{\tau}} = &\left\{\left[1 - e_0 \exp\left(-\xi\bar{Q}\right)\right] \left(\frac{1}{n\chi}\right)^{\frac{1}{\sigma-1}} \left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} + 1 - \beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}} (1 - l_{P,t} - l_{A,t} - l_{R,t})\right\}^{\theta} \cdots \\ &\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n + \delta_A(1-\zeta)}\right]^{\frac{\eta-1}{1-\zeta}} l_{P,t} l_{R,t}^{\frac{\theta}{\sigma-1}} l_{A,t}^{\frac{\theta\eta-1}{1-\zeta}}. \end{aligned} \tag{A.40}$$

Then, $\bar{Q}$ can be determined from the above equation, given the solutions for labor shares across different sectors. Given $\bar{Q}(l_{P,t}, l_{A,t}, l_{R,t})$, equation (A.39) serves as the third key equation for determining labor shares in this regime.

Let

$$\mathcal{A} \equiv \frac{\rho + (1-\zeta)\delta_A}{g_{D_A} + \delta_A} = \frac{(1-\zeta)\left[\rho + (1-\zeta)\delta_A\right]}{n + \delta_A(1-\zeta)},$$

We now solve for the labor shares. When $g_{D_A} > g_{D_P}$, we have $l_{P,t} \to 0$, using equations (A.33) and (A.36), we obtain

$$l_A^{sp}|_{g_{D_A} > g_{D_P}} = \frac{\eta\rho(\sigma-1)}{\mathcal{A}n + \rho(\sigma-1)(\mathcal{A}+\eta)}$$

and

$$l_R^{sp}|_{g_{D_A}>g_{D_P}} = \frac{\mathcal{A}n}{\mathcal{A}n + \rho(\sigma-1)(\mathcal{A}+\eta)}.$$

When $g_{D_A} < g_{D_P}$, we have $l_A^{sp} \to 0$. Then, using equations (A.33) and (A.36), we obtain

$$l_P^{sp}|_{g_{D_A}<g_{D_P}} = \frac{\eta\rho(\sigma-1)}{n + \rho(1+\eta)(\sigma-1)}$$

and

$$l_R^{sp}|_{g_{D_A}<g_{D_P}} = \frac{n}{n + \rho(1+\eta)(\sigma-1)}.$$

When $g_{D_A} = g_{D_P}$, the labor shares can be derived from equations (A.33), (A.36), (A.39), and (A.40). However, due to the complexity of these equations, the labor shares in this regime can only be analyzed numerically.

A.3 Levels of the Key Variables

Given the labor shares l_P^{sp}, l_A^{sp}, and l_R^{sp} derived in the previous subsection, we can further determine the levels of some key variables.

First, from equation (A.13), we obtain

$$N_t^{sp} = \frac{1}{n\chi} l_R^{sp} L_t.$$

Similarly, from equation (3), we derive

$$D_{A,t}^{sp} = \left[\frac{\psi M^{-\zeta}(1-\zeta)}{n + \delta_A(1-\zeta)}\right]^{\frac{1}{1-\zeta}} (l_A^{sp})^{\frac{1}{1-\zeta}} L_t^{\frac{1}{1-\zeta}}.$$

Next, the level of consumption c_t must be analyzed under different regimes. When $g_{D_A} > g_{D_P}$, using equation (A.2), we obtain that consumption per capita can be derive from the following equation:

$$c_t^{sp}|_{g_{D_A}>g_{D_P}} = (1-e_0)(N_t^{sp})^{\frac{1}{\sigma-1}} \left[\beta \left(\frac{D_{P,t}^{sp}}{D_{A,t}^{sp}}\right)^{\frac{\varepsilon-1}{\varepsilon}} + 1 - \beta\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} D_{A,t}^{\eta}(1 - l_P^{sp} - l_A^{sp} - l_R^{sp})$$

$$= \frac{(1-\beta)^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(1-e_0)\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n + \delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_A^{sp})^{\frac{\eta}{1-\zeta}} (l_R^{sp})^{\frac{1}{\sigma-1}} (1 - l_P^{sp} - l_A^{sp} - l_R^{sp}) L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}.$$

When $g_{D_A} < g_{D_P}$, we have

$$\begin{aligned} c_t^{sp}|_{g_{D_A}<g_{D_P}} =& (N_t^{sp})^{\frac{1}{\sigma-1}} \left[\beta + (1-\beta)\left(\frac{D_{A,t}^{sp}}{D_{P,t}^{sp}}\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\eta\varepsilon}{\varepsilon-1}} (D_{P,t}^{sp})^{\eta}(1-l_P^{sp}-l_A^{sp}-l_R^{sp}) \\ =& \frac{\beta^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(c_t^{sp})^{\theta\eta}(l_P^{sp})^{\eta}(l_R^{sp})^{\frac{1}{\sigma-1}}(1-l_P^{sp}-l_A^{sp}-l_R^{sp})L_t^{\eta+\frac{1}{\sigma-1}} \\ =& \left[\frac{\beta^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(l_P^{sp})^{\eta}(l_R^{sp})^{\frac{1}{\sigma-1}}(1-l_P^{sp}-l_A^{sp}-l_R^{sp})\right]^{\frac{1}{1-\theta\eta}} L_t^{\frac{1}{1-\theta\eta}(\eta+\frac{1}{\sigma-1})}. \end{aligned}$$

When $g_{D_A} = g_{D_P}$, consumption is given by

$$\begin{aligned} c_t =& \frac{1-e_0\exp(-\xi\bar{Q})}{(n\chi)^{\frac{1}{\sigma-1}}}\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}}\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}(l_A^{sp})^{\frac{\eta}{1-\zeta}}(l_R^{sp})^{\frac{1}{\sigma-1}}\ldots \\ & (1-l_P^{sp}-l_A^{sp}-l_R^{sp})L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}. \end{aligned}$$

Finally, the usage of producer data can be determined from equation (A.1) as

$$D_{P,t}^{sp} = (c_t^{sp})^{\theta} l_P^{sp} L_t.$$

Thus, we have derived the levels of all key variables.

Finally, from $c_t^{sp}|_{g_{D_A}>g_{D_P}}$ we know that the consumption decreases as M increases along a BGP, and we can further derive that M act negatively to the social welfare in the long run. Thus, the optimal number of Generative AI firms in the optimal allocation is $M^{sp} = 1$.

B. Competitive Equilibrium

In this section, the co-state variables and shadow prices are redefined to simplify the notation. From the household problem defined by equations (24) and (25), and considering the homogeneity among different varieties of intermediate goods, we define the current-value Hamiltonian equation as:

$$\mathcal{H}(c_t, a_t, \lambda_t) = \ln c_t + \lambda_t\left[(r_t - n)a_t + w_t - c_t\right].$$

Notably, in this problem, the household does not control data generation and therefore takes data quality Q_t as given. The first-order conditions with respect to c_t and a_t are:

$$\frac{\partial \mathcal{H}}{\partial c_t} = \frac{1}{c_t} - \lambda_t = 0 \tag{B.1}$$

and

$$\frac{\partial \mathcal{H}}{\partial a_t} = \lambda_t(r_t - n) = -\dot{\lambda}_t + (\rho - n)\lambda_t. \tag{B.2}$$

From equations (B.1) and (B.2), we obtain

$$\frac{\dot{c}_t}{c_t} = r_t - \rho. \tag{B.3}$$

From the final good producer problem defined by equation (26), the first-order conditions with respect to $Y_{i,t}$ and $L_{P,t}$ are given by

$$\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}} \left(1 + \theta p_{D_P,t} Y_t^{\theta-1} \frac{L_{P,t}}{L_t^{\theta}}\right) = p_{i,t} \tag{B.4}$$

and

$$p_{D_P,t} \left(\frac{Y_t}{L_t}\right)^{\theta} = w_t. \tag{B.5}$$

From the intermediate good producer problem defined by equations (29) and (B.4), the firm's optimization problem can be rewritten as

$$r_t V_{i,t} = \max_{\{L_{i,t}, D_{P,i,t}, D_{A,i,t}\}} \left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}} \left(1 + \theta p_{D_P,t} Y_t^{\theta-1} \frac{L_{P,t}}{L_t^{\theta}}\right) Y_{i,t} - w_t L_{i,t} - p^d_{D_P,t} D_{P,i,t} - p^d_{D_A,t} D_{A,i,t} + \dot{V}_{i,t} - \delta(e_{i,t}) V_{i,t}. \tag{B.6}$$

The following partial derivatives hold:

$$\frac{\partial Y_{i,t}}{\partial L_{i,t}} = (1 - e_{i,t}) D_{i,t}^{\eta} = \frac{Y_{i,t}}{L_{i,t}},$$

$$\frac{\partial Y_{i,t}}{\partial D_{P,i,t}} = \xi \tau e_{i,t} D_{i,t}^{\eta} L_{i,t} D_{P,i,t}^{\tau-1} D_{A,i,t}^{-\tau} + (1 - e_{i,t}) \beta \eta D_{i,t}^{\eta-1+\frac{1}{\varepsilon}} D_{P,i,t}^{-\frac{1}{\varepsilon}} L_{i,t},$$

and

$$\frac{\partial Y_{i,t}}{\partial D_{A,i,t}} = -\xi \tau e_{i,t} D_{i,t}^{\eta} L_{i,t} D_{P,i,t}^{\tau} D_{A,i,t}^{-\tau-1} + (1 - e_{i,t})(1 - \beta) \eta D_{i,t}^{\eta-1+\frac{1}{\varepsilon}} D_{A,i,t}^{-\frac{1}{\varepsilon}} L_{i,t}.$$

The first-order conditions with respect to $L_{i,t}$, $D_{P,i,t}$, and $D_{A,i,t}$ are given by

$$\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\frac{Y_{i,t}}{L_{i,t}}=w_t, \tag{B.7}$$

$$\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\frac{\partial Y_{i,t}}{\partial D_{P,i,t}}=p^d_{D_P,t}-2\xi\tau\delta_0 e_{i,t}^2 D_{P,i,t}^{\tau-1}D_{A,i,t}^{-\tau}V_{i,t}, \tag{B.8}$$

and

$$\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\frac{\partial Y_{i,t}}{\partial D_{A,i,t}}=p^d_{D_A,t}+2\xi\tau\delta_0 e_{i,t}^2 D_{P,i,t}^{\tau}D_{A,i,t}^{-\tau-1}V_{i,t}. \tag{B.9}$$

Combining equations (B.4) and (B.5) and considering symmetry, we obtain

$$\begin{aligned}
& N_t^{\frac{1}{\sigma-1}}\left[1+\theta w_t\left(\frac{L_t}{Y_t}\right)^{\theta}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right]=p_{i,t}\\
\Rightarrow\quad & N_t^{\frac{1}{\sigma-1}}\left(1+\theta w_t\frac{L_{P,t}}{Y_t}\right)=p_{i,t}\\
\Rightarrow\quad & N_t^{\frac{1}{\sigma-1}}\left[1+\theta\left(1-\frac{1}{\sigma}\right)p_{i,t}\frac{Y_{i,t}}{L_{i,t}}\frac{L_{P,t}}{Y_t}\right]=p_{i,t}\\
\Rightarrow\quad & N_t^{\frac{1}{\sigma-1}}+\theta\left(1-\frac{1}{\sigma}\right)p_{i,t}\frac{L_{P,t}}{N_tL_{i,t}}=p_{i,t}\\
\Rightarrow\quad & p_{i,t}=\frac{N_t^{\frac{1}{\sigma-1}}}{1-\theta\left(1-\frac{1}{\sigma}\right)\frac{L_{P,t}}{N_tL_{i,t}}}\\
\Rightarrow\quad & p_{i,t}=\frac{L_{R,t}^{\frac{1}{\sigma-1}}}{(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_{P,t}}{1-l_{P,t}-l_{A,t}-l_{R,t}}\right]}.
\end{aligned} \tag{B.10}$$

Here, the third equation follows from substituting equation (B.7) into the expression. Combining equations (B.8) and (B.9), we obtain

$$\begin{aligned}
& \left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)\eta(1-e_{i,t})D_{i,t}^{\eta}L_{i,t}=D_{P,i,t}p^d_{D_P,t}+D_{A,i,t}p^d_{D_A,t}\\
\Rightarrow\quad & \eta\left(1-\frac{1}{\sigma}\right)\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t}Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)Y_{i,t}=D_{P,i,t}p^d_{D_P,t}+D_{A,i,t}p^d_{D_A,t}.
\end{aligned}$$

Substituting equation (B.7) into the above equation, we obtain

$$\eta w_t L_{i,t} = D_{P,i,t} p^d_{D_P,t} + D_{A,i,t} p^d_{D_A,t}. \tag{B.11}$$

When $g_{D_A} \geq g_{D_P}$, along the BGP, equation (B.9) can be further derived as

$$p_{D_A,t} \approx \left(1 - \frac{1}{\sigma}\right) p_{i,t}(1 - e_0)(1 - \beta)\eta D^{\eta-1}_{A,i,t} L_{i,t} N_t \equiv \mathcal{E} d^{\eta-1}_{A,t} L_t^{1+\frac{1}{\sigma-1}}. \tag{B.12}$$

Here, $\mathcal{E}$ is a constant. This equation comes from the fact that $p_{i,t} \propto L_t^{\frac{1}{\sigma-1}}$ and $N_t \propto L_t$.

From the Generative AI firm problem defined by equation (28), and plug equation (B.12) into the equation, the current-value Hamiltonian equation is given by

$$\mathcal{H}(d_{A,t}, L_{A,t}, \mu_t) = \mathcal{E} d^{\eta}_{A,t} L_t^{1+\frac{1}{\sigma-1}} - w_t \frac{L_{A,t}}{M} + \mu_t \left(\psi d^{\zeta}_{A,t} \frac{L_{A,t}}{M} - \delta_A d_{A,t}\right),$$

where μ_t is the co-state variable. The first-order conditions with respect to $d_{A,t}$ and $L_{A,t}$ are

$$\frac{\partial \mathcal{H}}{\partial d_{A,t}} = \eta p_{D_A,t} + \mu_t \left(\psi \zeta d^{\zeta-1}_{A,t} \frac{L_{A,t}}{M} - \delta_A\right) = -\dot{\mu}_t + r_t \mu_t \tag{B.13}$$

and

$$\frac{\partial \mathcal{H}}{\partial L_{A,t}} = -w_t + \mu_t \psi d^{\zeta}_{A,t} = 0. \tag{B.14}$$

Combining equations (B.13) and (B.14), we obtain

$$\begin{aligned} & \psi d^{\zeta-1}_{A,t} L_{A,t} \left(\frac{\eta p_{D_A,t} d_{A,t}}{w_t L_{A,t}} + \frac{\zeta}{M}\right) - \delta_A = -\frac{\dot{\mu}_t}{\mu_t} + r_t \\ \Rightarrow \quad & \psi M^{-\zeta} D^{\zeta-1}_{A,t} L_{A,t} \left(\frac{\eta p_{D_A,t} D_{A,t}}{w_t L_{A,t}} + \zeta\right) - \delta_A = -\frac{\dot{\mu}_t}{\mu_t} + r_t. \end{aligned} \tag{B.15}$$

We will come back to the discussion on the free-entry condition after we derive the profit of the firm.

From the data intermediary problem defined by equations (30) and (31), and considering the symmetry among different varieties of intermediate goods, the prices of data can be readily derived as:

$$p^d_{D_P,t} = \frac{p_{D_P,t}}{N_t} \tag{B.16}$$

and

$$p^d_{D_A,t} = \frac{p_{D_A,t}}{N_t}. \tag{B.17}$$

These two equations ensure the non-zero profit condition for the data intermediary. For further derivations, please refer to the appendix in Jones and Tonetti (2020).

Finally, from the free entry condition of the innovation sector defined by equation (32), and by symmetry, we obtain

$$\begin{aligned} \chi w_t &= V_{i,t} + \frac{N_t \delta_0 e_{i,t}^2 V_{i,t}}{\frac{1}{\chi} L_{R,t}} \\ \Rightarrow \quad \chi w_t &= V_{i,t} \left(1 + \frac{\delta_0 e_{i,t}^2}{n}\right). \end{aligned} \tag{B.18}$$

Here, the second equation follows from the result derived in equation (A.13) along the balanced growth path.

B.1 Growth Rates Along the Balanced Growth Path

Obviously, most of the growth rates of key variables are identical to those derived in the optimal allocation, which can be found in Appendix A.1. However, in the competitive equilibrium, since households cannot determine the usage of the two types of data and intermediate good producers do not account for household utility, we expect that the labor share allocated to AI-generated data production, $l_{A,t}$, may shrink when $g_{D_P} > g_{D_A}$, and the labor share allocated to producer data generation, $l_{P,t}$, may also shrink when $g_{D_P} < g_{D_A}$. Denoting g_{l_A} and g_{l_P} as the respective growth rates when these two labor shares are no longer constant, the growth rates of the two types of data are now given by

$$g_{D_A} = \begin{cases} \dfrac{n}{1-\zeta}, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \dfrac{g_{l_A} + n}{1-\zeta}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \tag{B.19}$$

and

$$g_{D_P} = \begin{cases} \left[1 + \theta\left(\dfrac{1}{\sigma - 1} + \dfrac{\eta}{1-\zeta}\right)\right] n + g_{l_P}, & \text{if} \quad g_{D_A} > g_{D_P}, \\ \left[1 + \dfrac{\theta}{1-\theta\eta}\left(\dfrac{1}{\sigma-1} + \eta\right)\right] n, & \text{if} \quad g_{D_A} \leq g_{D_P}. \end{cases}$$

Additionally, in this subsection, we derive the growth rates of prices along the balanced growth path.

First, from the labor market clearing condition shown in Table 1, it is evident that $L_{P,t}$, $L_{A,t}$, $L_{R,t}$, and the integral $\int_0^{N_t} L_{i,t} \mathrm{d}i$ all grow at the same rate as the total population L_t along the balanced growth path, which is n. By symmetry, the growth rate of $N_t L_{i,t}$ is also

n. Additionally, since the growth rate of N_t is n, it follows that $L_{i,t}$ remains constant along the balanced growth path. Thus, the denominator of equation (B.10) remains constant, and the growth rate of $p_{i,t}$ is given by

$$g_{p_i} = \frac{1}{\sigma - 1} n.$$

Next, noting that $e_{i,t}$ remains constant along the balanced growth path (see equation (A.16)), from equation (B.7) we obtain

$$\begin{aligned} & w_t = p_{i,t} \left(1 - \frac{1}{\sigma}\right) (1 - e_{i,t}) D_t^{\eta} && \text{(B.20)} \\ \Rightarrow \quad & g_w = g_{p_i} + \eta g_D \\ \Rightarrow \quad & g_w = \begin{cases} \left(\dfrac{1}{\sigma - 1} + \dfrac{\eta}{1 - \zeta}\right) n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \dfrac{1}{1 - \theta\eta} \left(\dfrac{1}{\sigma - 1} + \eta\right) n, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \end{aligned}$$

Meanwhile, considering equation (B.8), when $g_{D_A} < g_{D_P}$, this equation can be rewritten as

$$\left(1 - \frac{1}{\sigma}\right) \eta\beta p_{i,t} L_{i,t} D_{i,t}^{\frac{1}{\varepsilon} + \eta - 1} D_{P,i,t}^{-\frac{1}{\varepsilon}} = p_{D_P,t}^d.$$

Rewriting the equation in terms of growth rates, we obtain

$$\begin{aligned} & g_{p_i} + \left(\frac{1}{\varepsilon} + \eta - 1\right) g_D - \frac{1}{\varepsilon} g_{D_P} = g_{p_{D_P}} - g_N \\ \Rightarrow \quad & g_{p_{D_P}} = g_{p_i} + (\eta - 1) g_{D_P} + n \\ \Rightarrow \quad & g_{p_{D_P}} = \frac{1 - \theta}{1 - \theta\eta} \left(\frac{1}{\sigma - 1} + \eta\right) n. && \text{(B.21)} \end{aligned}$$

Similarly, when $g_{D_A} \geq g_{D_P}$, equation (B.8) can be rewritten as

$$\left(1 - \frac{1}{\sigma}\right) p_{i,t} \xi\tau D_{i,t}^{\eta} L_{i,t} D_{P,i,t}^{\tau - 1} D_{A,i,t}^{-\tau} = p_{D_P,t}^d - 2\xi\tau\delta_0 D_{P,i,t}^{\tau - 1} D_{A,i,t}^{-\tau} V_{i,t}. \quad \text{(B.22)}$$

On the left-hand side of equation (B.22), the growth rate of the first term, $p_{i,t} D_{i,t}^{\eta} L_{i,t} D_{P,i,t}^{\tau - 1} D_{A,i,t}^{-\tau}$, is given by

$$\frac{n}{\sigma - 1} + \eta g_D + (\tau - 1) g_{D_P} - \tau g_{D_A} = \frac{n}{\sigma - 1} + (\eta - 1) g_{D_A} + (1 - \tau)(g_{D_A} - g_{D_P}).$$

Similarly, the growth rate of the second term, $p_{i,t}D_{i,t}^{\eta-1+\frac{1}{\varepsilon}}D_{P,i,t}^{-\frac{1}{\varepsilon}}L_{i,t}$, is given by

$$\frac{n}{\sigma-1}+\left(\eta-1+\frac{1}{\varepsilon}\right)g_D-\frac{1}{\varepsilon}g_{D_P}=\frac{n}{\sigma-1}+(\eta-1)g_{D_A}+\frac{1}{\varepsilon}(g_{D_A}-g_{D_P}).$$

Thus, the growth rate of the left-hand side depends on whether $(1-\tau)\varepsilon$ is greater than 1. On the right-hand side, it is straightforward to see that the growth rate of the second term is given by $g_V+(\tau-1)g_{D_P}-\tau g_{D_A}$. From equation (B.18), we know that $g_V=g_w=g_c=\eta g_{D_A}+n/(\sigma-1)$, so this growth rate matches that of the first term on the left-hand side.

Since we assume the elasticity of substitution between the two types of data, ε, to be large, it is typically expected that $(1-\tau)\varepsilon>1$. Consequently, the growth rate of the left-hand side depends primarily on the first term. Furthermore, we have

$$\begin{aligned}
& g_{p_{D_P}}-g_N\leq g_{p_i}+(\eta-1)g_{D_A}+(1-\tau)(g_{D_A}-g_{D_P}) \\
\Rightarrow\quad & g_{p_{D_P}}\leq\frac{1}{\sigma-1}n+(\eta-1)\frac{1}{1-\zeta}n+(1-\tau)\left[\frac{1}{1-\zeta}-1-\theta\left(\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)\right]n+n-(1-\tau)g_{l_P} \\
\Rightarrow\quad & g_{p_{D_P}}\leq\left[\frac{\eta-1+(1-\theta\eta)(1-\tau)}{1-\zeta}+\frac{1-(1-\tau)\theta}{\sigma-1}+\tau\right]n-(1-\tau)g_{l_P}. \qquad \text{(B.23)}
\end{aligned}$$

Meanwhile, from equation (B.5), we can also express $g_{p_{D_P}}$ as

$$g_{p_{D_P}}=g_w-\theta g_c=(1-\theta)\left(\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)n. \qquad \text{(B.24)}$$

Combining equations (B.23) and (B.24), we derive the growth rate of $l_{P,t}$ as

$$\begin{aligned}
& (1-\tau)g_{l_P}\leq\left[\frac{\eta-1+(1-\theta\eta)(1-\tau)}{1-\zeta}+\frac{1-(1-\tau)\theta}{\sigma-1}+\tau-(1-\theta)\left(\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)\right]n \\
\Rightarrow\quad & (1-\tau)g_{l_P}\leq\left[\frac{\tau\theta}{\sigma-1}-\frac{\tau(1-\theta\eta)}{1-\zeta}+\tau\right]n \\
\Rightarrow\quad & g_{l_P}|_{g_{D_A}>g_{D_P}}\leq\frac{\tau}{1-\tau}\left(\frac{\theta}{\sigma-1}-\frac{1-\theta\eta}{1-\zeta}+1\right)n. \qquad \text{(B.25)}
\end{aligned}$$

Thus, the growth rate of $D_{P,t}$ in this regime is given by

$$g_{D_P}|_{g_{D_A}>g_{D_P}}\leq\frac{1}{1-\tau}\left(\frac{\theta\eta-\tau}{1-\zeta}+\frac{\theta}{\sigma-1}+1\right)n.$$

Notably, in this case, we require $g_{l_P} < 0$, implying that

$$\frac{1-\theta\eta}{1-\zeta} > \frac{\theta}{\sigma-1} + 1,$$

which always holds when $g_{D_A} > g_{D_P}$.

Last, we determine the growth rate of the price of AI-generated data. From equation (B.3), we derive that the interest rate r_t remains constant along the balanced growth path, given by

$$r_t = g_c + \rho. \tag{B.26}$$

From equation (B.14), we derive the growth rate of the shadow price μ_t as

$$\begin{aligned}
& \mu_t = \frac{M^\zeta w_t}{\psi D_{A,t}^\zeta} \\
\Rightarrow \quad & \frac{\dot{\mu}_t}{\mu_t} = \frac{\dot{w}_t}{w_t} - \zeta \frac{\dot{D}_{A,t}}{D_{A,t}} \\
\Rightarrow \quad & g_\mu = \begin{cases} \left(\frac{1}{\sigma-1} + \frac{\eta-\zeta}{1-\zeta}\right) n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \left[\frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right) - \frac{\zeta}{1-\zeta}\right] n - \frac{\zeta}{1-\zeta} g_{l_A}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases}
\end{aligned} \tag{B.27}$$

Substituting equations (B.26), (B.27), and the dynamic equation of AI-generated data (10) into (B.15), we obtain

$$\left(g_{D_A} + \delta_A\right)\left(\frac{\eta p_{D_A,t} D_{A,t}}{w_t L_{A,t}} + \zeta\right) - \delta_A = -\frac{\dot{\mu}_t}{\mu_t} + g_c + \rho. \tag{B.28}$$

Given the constancy of the terms in the above equation, we derive

$$\begin{aligned}
& g_{p_{D_A}} = g_w + g_{l_A} + n - g_{D_A} \\
\Rightarrow \quad & g_{p_{D_A}} = \begin{cases} \left(\frac{1}{\sigma-1} + \frac{\eta-1}{1-\zeta} + 1\right) n, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \left[\frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right) + 1 - \frac{1}{1-\zeta}\right] n - \frac{\zeta}{1-\zeta} g_{l_A}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases}
\end{aligned} \tag{B.29}$$

When $g_{D_A} < g_{D_P}$, equation (B.9) simplifies to

$$\begin{aligned}
& \left(1 - \frac{1}{\sigma}\right) \beta^{\eta-1+\frac{1}{\varepsilon}} (1-\beta) \eta p_{i,t} D_{P,t}^{\eta-1+\frac{1}{\varepsilon}} D_{A,t}^{-\frac{1}{\varepsilon}} L_{i,t} = \frac{p_{D_A,t}}{N_t} \\
\Rightarrow \quad & g_{p_{D_A}} = g_{p_i} + g_N + \left(\eta - 1 + \frac{1}{\varepsilon}\right) g_{D_P} - \frac{1}{\varepsilon} g_{D_A}
\end{aligned}$$

$$\Rightarrow \quad g_{p_{D_A}} = \left\{\frac{1}{\sigma-1}+1+\left(\eta-1+\frac{1}{\varepsilon}\right)\left[1+\frac{\theta}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)\right]-\frac{1}{\varepsilon}\frac{1}{1-\zeta}\right\}n-\frac{1}{\varepsilon}\frac{g_{l_A}}{1-\zeta}$$
$$\Rightarrow \quad g_{p_{D_A}} = \left\{\left[1+\left(\eta-1+\frac{1}{\varepsilon}\right)\frac{\theta}{1-\theta\eta}\right]\frac{1}{\sigma-1}-\frac{1}{\varepsilon}\frac{1}{1-\zeta}+\left(\eta-\theta\eta+\frac{1}{\varepsilon}\right)\frac{1}{1-\theta\eta}\right\}-\frac{1}{\varepsilon}\frac{g_{l_A}}{1-\zeta}. \tag{B.30}$$

Combining equations (B.29) and (B.30), we derive the growth rate of the labor share l_A as

$$\left[\frac{1}{1-\theta\eta}\left(\frac{1}{\sigma-1}+\eta\right)+1-\frac{1}{1-\zeta}\right]n-\frac{\zeta}{1-\zeta}g_{l_A} = \left\{\left[1+\left(\eta-1+\frac{1}{\varepsilon}\right)\frac{\theta}{1-\theta\eta}\right]\frac{1}{\sigma-1}-\frac{1}{\varepsilon}\frac{1}{1-\zeta}+\left(\eta-\theta\eta+\frac{1}{\varepsilon}\right)\frac{1}{1-\theta\eta}\right\}n-\frac{1}{\varepsilon}\frac{g_{l_A}}{1-\zeta}$$
$$\Rightarrow \quad \left(\zeta-\frac{1}{\varepsilon}\right)\frac{g_{l_A}}{1-\zeta} = \left[\left(1-\frac{1}{\varepsilon}\right)\frac{\theta}{1-\theta\eta}\frac{1}{\sigma-1}-\left(1-\frac{1}{\varepsilon}\right)\frac{1}{1-\zeta}+\left(1-\frac{1}{\varepsilon}\right)\frac{1}{1-\theta\eta}\right]n$$
$$\Rightarrow \quad g_{l_A} = \frac{\varepsilon-1}{\zeta\varepsilon-1}\left[\frac{1-\zeta}{1-\theta\eta}\left(\frac{\theta}{\sigma-1}+1\right)-1\right]n. \tag{B.31}$$

Substituting (B.31) into (B.19), (B.27), and (B.29), respectively, we obtain the growth rates of $D_{A,t}$, μ_t, and p_{D_A} as

$$g_{D_A}|_{g_{D_A}<g_{D_P}} = \frac{1}{\zeta\varepsilon-1}\left[\frac{\varepsilon-1}{1-\theta\eta}\left(\frac{\theta}{\sigma-1}+1\right)-\varepsilon\right]n,$$

$$g_{\mu}|_{g_{D_A}<g_{D_P}} = \left[\frac{1}{1-\theta\eta}\left(1-\frac{\zeta\theta(\varepsilon-1)}{\zeta\varepsilon-1}\right)\frac{1}{\sigma-1}+\varepsilon\zeta+\frac{1}{1-\theta\eta}\left(\eta-\frac{\zeta(\varepsilon-1)}{\zeta\varepsilon-1}\right)\right]n,$$

and

$$g_{p_{D_A}}|_{g_{D_A}<g_{D_P}} = \left\{\left[1-\frac{\zeta\theta(\varepsilon-1)}{\zeta\varepsilon-1}\right]\frac{1}{1-\theta\eta}\frac{1}{\sigma-1}+\frac{1}{\zeta\varepsilon-1}+\frac{1}{1-\theta\eta}\left(\eta-\theta\eta-\frac{1-\zeta}{\zeta\varepsilon-1}\right)\right\}n.$$

Since we require $g_{l_A} < 0$, this implies the following condition:

$$\begin{cases}\dfrac{\theta}{\sigma-1}+1<\dfrac{1-\theta\eta}{1-\zeta}, & \text{if} \quad \zeta\varepsilon>1,\\ \dfrac{\theta}{\sigma-1}+1>\dfrac{1-\theta\eta}{1-\zeta}, & \text{if} \quad \zeta\varepsilon<1.\end{cases}$$

However, considering that when $g_{D_A} < g_{D_P}$, we have

$$\frac{\theta}{\sigma-1}+1>\frac{1-\theta\eta}{1-\zeta}.$$

Thus, g_{l_A} is negative only when $\zeta\varepsilon < 1$. Otherwise, we obtain the corner solution: $g_{l_A} = 0$,

$l_{A,t} = 0$, and $D_{A,t} \to 0$.

By now, we derive all the growth rates of the variables.

B.2 Labors Employed in Different Sectors

First, from the labor market clearing condition shown in Table 1, we have

$$\begin{aligned}
&\int_0^{N_t} L_{i,t}\mathrm{d}i = L_t - L_{P,t} - L_{A,t} - L_{R,t} \\
\Rightarrow \quad & N_t L_{i,t} = L_t - L_{P,t} - L_{A,t} - L_{R,t} \\
\Rightarrow \quad & \frac{N_t}{L_t} L_{i,t} = 1 - l_{P,t} - l_{A,t} - l_{R,t} \\
\Rightarrow \quad & L_{i,t} = n\chi \frac{1 - l_{P,t} - l_{A,t} - l_{R,t}}{l_{R,t}}.
\end{aligned} \tag{B.32}$$

Here, the last line follows from substituting the result derived in equation (A.13). Meanwhile, when $g_{D_A} \geq g_{D_P}$, we further refine equation (B.28) as

$$\begin{aligned}
&\frac{\eta p_{D_A,t} D_{A,t}}{w_t L_{A,t}} = \frac{-g_\mu + g_c + \rho + \delta_A}{g_{D_A} + \delta_A} - \zeta \\
\Rightarrow \quad & \frac{p_{D_A,t} D_{A,t}}{w_t L_{A,t}} = \frac{\zeta n + (\rho + \delta_A)(1 - \zeta)}{\eta\left[n + \delta_A(1 - \zeta)\right]} - \frac{\zeta}{\eta} \equiv \frac{\mathcal{A}}{\eta}.
\end{aligned} \tag{B.33}$$

Next, combining equations (B.5) and (B.33) with (B.11), we obtain

$$\begin{aligned}
& \eta w_t N_t L_{i,t} = D_{P,t} p_{D_P,t} + D_{A,t} p_{D_A,t} \\
\Rightarrow \quad & \eta w_t N_t L_{i,t} = w_t L_{P,t} + \frac{\mathcal{A}}{\eta} w_t L_{A,t} \\
\Rightarrow \quad & \eta L_{i,t} \frac{N_t}{L_t} = l_{P,t} + \frac{\mathcal{A}}{\eta} l_{A,t} \\
\Rightarrow \quad & \eta L_{i,t} \frac{l_{R,t}}{n\chi} = l_{P,t} + \frac{\mathcal{A}}{\eta} l_{A,t} \\
\Rightarrow \quad & \eta(1 - l_{P,t} - l_{A,t} - l_{R,t}) = l_{P,t} + \frac{\mathcal{A}}{\eta} l_{A,t} \\
\Rightarrow \quad & (1 + \eta) l_{P,t} + \left(\frac{\mathcal{A}}{\eta} + \eta\right) l_{A,t} + \eta l_{R,t} - \eta = 0.
\end{aligned} \tag{B.34}$$

Here, the fourth line follows from the result derived in equation (A.13). Thus, we obtain the first key equation for determining the labor shares.

Next, from the intermediate good producer problem defined by equation (B.6), we obtain

$$
\begin{aligned}
V_{i,t} =& \frac{\left(\frac{Y_t}{Y_{i,t}}\right)^{\frac{1}{\sigma}}\left(1+\theta p_{D_P,t} Y_t^{\theta-1}\frac{L_{P,t}}{L_t^{\theta}}\right)Y_{i,t} - w_t L_{i,t} - p^d_{D_P,t} D_{P,i,t} - p^d_{D_A,t} D_{A,i,t}}{r_t + \delta(e_{i,t}) - \frac{\dot{V}_t}{V_t}} \\
=& \frac{p_{i,t}Y_{i,t} - (1+\eta) w_t L_{i,t}}{g_c + \rho + \delta_0 e_{i,t}^2 - g_V} \\
=& \frac{\left(\frac{\sigma}{\sigma-1} - 1 - \eta\right) w_t L_{i,t}}{g_c + \rho + \delta_0 e_{i,t}^2 - g_w} \\
=& \frac{\left(\frac{\sigma}{\sigma-1} - 1 - \eta\right) w_t L_{i,t}}{\rho + \delta_0 e_{i,t}^2}.
\end{aligned}
$$

Then, combining the above equation with equation (B.18), we obtain

$$
\begin{aligned}
L_{i,t} =& \frac{\chi(\rho + \delta_0 e_{i,t}^2)}{\left(1 + \frac{\delta_0 e_{i,t}^2}{n}\right)\left(\frac{\sigma}{\sigma-1} - 1 - \eta\right)} \\
=& \begin{cases} \dfrac{\chi(\rho + \delta_0 e_0^2)}{\left(1 + \frac{\delta_0 e_0^2}{n}\right)\left(\frac{\sigma}{\sigma-1} - 1 - \eta\right)}, & \text{if} \quad g_{D_A} > g_{D_P}, \\ \dfrac{\chi\rho}{\frac{\sigma}{\sigma-1} - 1 - \eta}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \\
\equiv& C.
\end{aligned}
$$

Substituting into equation (B.32), we obtain the second key equation for determining the labor shares:

$$
n\chi l_{P,t} + n\chi l_{A,t} + (n\chi + C) l_{R,t} - n\chi = 0. \tag{B.35}
$$

Combining equations (B.34) and (B.35), we derive the labor shares across different sectors. When $g_{D_A} > g_{D_P}$, we have $l_P^{dc}|_{g_{D_A} > g_{D_P}} \to 0$, while the other labor shares are given by

$$
l_A^{dc}|_{g_{D_A} > g_{D_P}} = \frac{C\eta^2}{C(\mathcal{A} + \eta^2) + n\chi\mathcal{A}}
$$

and

$$l_R^{dc}|_{g_{D_A}>g_{D_P}} = \frac{n\chi\mathcal{A}}{C(\mathcal{A}+\eta^2)+n\chi\mathcal{A}}.$$

Conversely, when $g_{D_A} < g_{D_P}$, we have $l_A^{dc}|_{g_{D_A}<g_{D_P}} \to 0$, while the remaining labor shares are given by

$$l_P^{dc}|_{g_{D_A}<g_{D_P}} = \frac{C\eta}{C(1+\eta)+n\chi}$$

and

$$l_R^{dc}|_{g_{D_A}<g_{D_P}} = \frac{n\chi}{C(1+\eta)+n\chi}.$$

Finally, we derive the labor shares when $g_{D_A} = g_{D_P}$, which corresponds to the condition $(1-\eta\theta)/(1-\zeta) = 1+\theta/(\sigma-1)$. In this regime, the ratio of the two types of data converges to a nonzero and finite constant. Denoting the data quality in this regime as $\bar{Q}$, we obtain

$$\begin{aligned}
\bar{Q}^{\frac{1}{\tau}} =& \frac{D_{P,t}}{D_{A,t}} = \frac{\left(\frac{Y_t}{L_t}\right)^{\theta} L_{P,t}}{D_{A,t}} \\
=& \frac{\left(\frac{L_{R,t}}{n\chi}\right)^{\frac{\sigma\theta}{\sigma-1}} L_t^{-\theta} L_{P,t} \left[\left[1-e_0\exp\left(-\xi\bar{Q}\right)\right] D_t^{\eta} L_{i,t}\right]^{\theta}}{D_{A,t}} \\
=& \frac{\left[1-e_0\exp\left(-\xi\bar{Q}\right)\right]^{\theta} \left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\theta\varepsilon}{\varepsilon-1}} (n\chi)^{\frac{\sigma\theta}{1-\sigma}} L_{P,t} D_{A,t}^{\eta\theta} L_{R,t}^{\frac{\sigma\theta}{\sigma-1}} L_{i,t}^{\theta} L_t^{-\theta}}{D_{A,t}} \\
=& \left[1-e_0\exp\left(-\xi\bar{Q}\right)\right]^{\theta} \left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\theta\varepsilon}{\varepsilon-1}} (n\chi)^{\frac{\sigma\theta}{1-\sigma}} \left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta\theta-1}{1-\zeta}} \\
& l_{P,t} l_{A,t}^{\frac{\eta\theta-1}{1-\zeta}} l_{R,t}^{\frac{\sigma\theta}{\sigma-1}} L_{i,t}^{\theta}.
\end{aligned} \tag{B.36}$$

Here, the third line follows from the fact that in this regime, we have

$$\begin{aligned}
D_{i,t} = D_t =& \left[\beta D_{P,t}^{\frac{\varepsilon-1}{\varepsilon}} + (1-\beta) D_{A,t}^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon-1}{\varepsilon}} \\
=& \left[\beta\left(\frac{D_{P,t}}{D_{A,t}}\right)^{\frac{\varepsilon-1}{\varepsilon}} + 1 - \beta\right]^{\frac{\varepsilon}{\varepsilon-1}} D_{A,t} \\
=& \left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}} + 1 - \beta\right)^{\frac{\varepsilon}{\varepsilon-1}} D_{A,t},
\end{aligned}$$

and the last line follows from the result derived in equation (3). Given $l_{P,t}$, $l_{A,t}$, $l_{R,t}$, and $L_{i,t}$, we can determine $\bar{Q}$ using the equation above. Furthermore, from equation (B.8), we

obtain

$$\frac{w_t L_{i,t}}{Y_{i,t}}\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}} D_{A,t}^{\eta-1}L_{i,t}\left\{\xi\tau e_0\exp\left(-\xi\bar{Q}\right)\bar{Q}^{\frac{\tau-1}{\tau}}+\beta\eta\left[1-e_0\exp\left(-\xi\bar{Q}\right)\right]\right.$$
$$\left.\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{-1}\bar{Q}^{-\frac{1}{\tau\varepsilon}}\right\}=w_t\left(\frac{L_t}{Y_t}\right)^{\theta}\frac{1}{N_t}-\frac{2\xi\tau\delta_0 n\chi\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}{n+\delta_0\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}\frac{w_t}{D_{A,t}}\bar{Q}^{\frac{\tau-1}{\tau}}$$

$$\Rightarrow\quad\left[\frac{\xi\tau e_0\exp\left(-\xi\bar{Q}\right)}{1-e_0\exp\left(-\xi\bar{Q}\right)}\bar{Q}^{\frac{\tau-1}{\tau}}+\beta\eta\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{-1}\bar{Q}^{-\frac{1}{\tau\varepsilon}}\right]L_{i,t}=\left(\frac{L_t}{Y_t}\right)^{\theta}\frac{D_{A,t}}{N_t}-$$
$$\frac{2\xi\tau\delta_0 n\chi\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}{n+\delta_0\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}\bar{Q}^{\frac{\tau-1}{\tau}}$$

$$\Rightarrow\quad\left[\frac{\xi\tau e_0\exp\left(-\xi\bar{Q}\right)}{1-e_0\exp\left(-\xi\bar{Q}\right)}\bar{Q}^{\frac{\tau-1}{\tau}}+\beta\eta\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{-1}\bar{Q}^{-\frac{1}{\tau\varepsilon}}\right]L_{i,t}=\left[1-e_0\exp\left(-\xi\bar{Q}\right)\right]^{-\theta}$$
$$(n\chi)^{1+\frac{\sigma\theta}{\sigma-1}}\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{-\frac{\varepsilon\eta\theta}{\varepsilon-1}}\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{1-\eta\theta}{1-\zeta}}l_{A,t}^{\frac{1-\eta\theta}{1-\zeta}}l_{R,t}^{-1-\frac{\sigma\theta}{\sigma-1}}L_{i,t}^{-\theta}-$$
$$\frac{2\xi\tau\delta_0 n\chi\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}{n+\delta_0\left[e_0\exp\left(-\xi\bar{Q}\right)\right]^2}\bar{Q}^{\frac{\tau-1}{\tau}}. \tag{B.37}$$

Given $\bar{Q}$, the labor shares can be determined using equations (B.32), (B.34), (B.35), and the equation above.

B.3 Levels of the Key Variables

Given the labor shares l_P^{dc}, l_A^{dc}, and l_R^{dc} derived in the previous subsection, we now turn to determine the levels of other key variables. Similar to the optimal allocation results presented in Appendix A.3, the variety of intermediate goods N_t, per capita consumption c_t, per capita output y_t, AI-generated data $D_{A,t}$, and producer data $D_{P,t}$ can be expressed as follows:

$$N_t^{dc}=\frac{1}{n\chi}l_R^{dc}L_t.$$

The per capita consumption and output are given by:

$$c_t^{dc} = y_t^{dc} = \begin{cases} \frac{(1-\beta)^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(1-e_0)\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}(l_A^{dc})^{\frac{\eta}{1-\zeta}}(l_R^{dc})^{\frac{1}{\sigma-1}}\dots \\ (1-l_P^{dc}-l_A^{dc}-l_R^{dc})L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}, & \text{if } g_{D_A} > g_{D_P}, \\ \frac{1-e_0\exp(-\xi\bar{Q})}{(n\chi)^{\frac{1}{\sigma-1}}}\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}}\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}\dots \\ (l_A^{dc})^{\frac{\eta}{1-\zeta}}(l_R^{dc})^{\frac{1}{\sigma-1}}(1-l_P^{dc}-l_A^{dc}-l_R^{dc})L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}, & \text{if } g_{D_A} = g_{D_P}, \\ \left[\frac{\beta^{\frac{\eta\varepsilon}{\varepsilon-1}}}{(n\chi)^{\frac{1}{\sigma-1}}}(l_P^{dc})^{\eta}(l_R^{dc})^{\frac{1}{\sigma-1}}(1-l_P^{dc}-l_A^{dc}-l_R^{dc})\right]^{\frac{1}{1-\theta\eta}} L_t^{\frac{1}{1-\theta\eta}(\eta+\frac{1}{\sigma-1})}, & \text{if } g_{D_A} < g_{D_P}. \end{cases}$$

The levels of AI-generated data and producer data are determined as follows:

$$D_{A,t}^{dc} = \begin{cases} \left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{1}{1-\zeta}}(l_A^{dc})^{\frac{1}{1-\zeta}}L_t^{\frac{1}{1-\zeta}}, & \text{if} \quad g_{D_A} \geq g_{D_P}, \\ \ll D_{P,t}^{dc}, & \text{if} \quad g_{D_A} < g_{D_P}, \end{cases}$$

and

$$D_{P,t}^{dc} = \begin{cases} \ll D_{A,t}^{dc}, & \text{if} \quad g_{D_A} > g_{D_P}, \\ = (y_t^{dc})^{\theta} l_P^{dc} L_t, & \text{if} \quad g_{D_A} \leq g_{D_P}. \end{cases}$$

The above derivations provide a comprehensive characterization of key economic variables in equilibrium, depending on the relative growth dynamics of AI-generated data and producer data.

Besides, from equation (B.7), we can derive the wage as follows:

$$\begin{aligned} w_t &= \left(1-\frac{1}{\sigma}\right)p_{i,t}(1-e_{i,t})D_{i,t}^{\eta} \\ &= \begin{cases} \left(1-\frac{1}{\sigma}\right)(1-\beta)^{\eta}p_{i,t}(1-e_0)(D_{A,t}^{dc})^{\eta}, & \text{if} \quad g_{D_A} > g_{D_P}, \\ \left(1-\frac{1}{\sigma}\right)p_{i,t}(1-e_{i,t})\left(\beta\bar{Q}^{\frac{\varepsilon-1}{\tau\varepsilon}}+1-\beta\right)^{\frac{\eta\varepsilon}{\varepsilon-1}}(D_{A,t}^{dc})^{\eta}, & \text{if} \quad g_{D_A} = g_{D_P}, \\ \left(1-\frac{1}{\sigma}\right)\beta^{\eta}p_{i,t}(D_{P,t}^{dc})^{\eta}, & \text{if} \quad g_{D_A} < g_{D_P}. \end{cases} \end{aligned}$$

Furthermore, from equation (B.33), the price of AI-generated data is given by:

$$p_{D_A,t} = \frac{\mathcal{A}}{\eta} w_t l_A^{dc} (D_{A,t}^{dc})^{-1} L_t. \tag{B.38}$$

Similarly, from equation (B.5), the price of producer data can be expressed as:

$$p_{D_P,t} = w_t (c_t^{dc})^{-\theta}.$$

B.4 Free-entry of the Generative AI Firms

The instantaneous profit of a Generative AI firm is derived as

$$\begin{aligned}\pi_A &= p_{D_A,t} d_{A,t} - w_t \frac{L_{A,t}}{M} \\ &= \frac{1}{M}\left(p_{D_A,t} D_{A,t} - w_t L_{A,t}\right) \\ &= \frac{\mathcal{A}-\eta}{M\eta} w_t l_A^{dc} L_t, \end{aligned} \tag{B.39}$$

where the third line follows from substituting equation (B.38). We focus on the case in which $g_{D_A} > g_{D_P}$, so from equation (B.20) we obtain

$$\begin{aligned} w_t &= p_{i,t}\left(1-\frac{1}{\sigma}\right)(1-e_0) D_{A,t}^{\eta} \\ &= \frac{\left(1-\frac{1}{\sigma}\right)(1-e_0)\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_R^{dc})^{\frac{1}{\sigma-1}} (l_A^{dc})^{\frac{\eta}{1-\zeta}}}{(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]} L_t^{\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}. \end{aligned} \tag{B.40}$$

Substituting this into the expression for instantaneous profit yields

$$\pi_A = \frac{(\mathcal{A}-\eta)\left(1-\frac{1}{\sigma}\right)(1-e_0)\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_R^{dc})^{\frac{1}{\sigma-1}} (l_A^{dc})^{1+\frac{\eta}{1-\zeta}}}{M\eta(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]} L_t^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}. \tag{B.41}$$

The discounted sum of profits that a Generative AI firm earns is then given by

$$\Pi = \frac{(\mathcal{A}-\eta)\left(1-\frac{1}{\sigma}\right)(1-e_0)\left[\frac{\psi M^{-\zeta}(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (l_R^{dc})^{\frac{1}{\sigma-1}} (l_A^{dc})^{1+\frac{\eta}{1-\zeta}}}{M\eta(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]} L_0^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}} \int_0^{\infty} e^{-r^*t+\left(1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}\right)nt} \mathrm{d}t$$

$$= \frac{(\mathcal{A}-\eta)\left(1-\frac{1}{\sigma}\right)(1-e_0)\left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}(l_R^{dc})^{\frac{1}{\sigma-1}}(l_A^{dc})^{1+\frac{\eta}{1-\zeta}}}{\eta(\rho-n)(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]} M^{\frac{\zeta(1-\eta)-1}{1-\zeta}} L_0^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}. \quad \text{(B.42)}$$

Applying the free-entry condition $\Pi = \mathcal{G}$, the number of Generative AI firms in the competitive equilibrium is derived as

$$M = \left\{ \frac{(\mathcal{A}-\eta)\left(1-\frac{1}{\sigma}\right)(1-e_0)\left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}}(l_R^{dc})^{\frac{1}{\sigma-1}}(l_A^{dc})^{1+\frac{\eta}{1-\zeta}} L_0^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}}}{\eta(\rho-n)(n\chi)^{\frac{1}{\sigma-1}}\left[1-\theta\left(1-\frac{1}{\sigma}\right)\frac{l_P^{dc}}{1-l_P^{dc}-l_A^{dc}-l_R^{dc}}\right]\mathcal{G}} \right\}^{\frac{1-\zeta}{1-\zeta(1-\eta)}}, \quad \text{(B.43)}$$

which differs from the result obtained under the optimal allocation.

C. Quantitative Analysis

C.1 Preparation Work

We begin by analyzing the optimal allocation. From equations (A.33) and (A.36), given the value of l_P^{sp}, the values of l_R^{sp} and l_A^{sp} can be derived as follows:

$$l_R^{sp} = \frac{(1-\mathcal{A})l_P^{sp} + \mathcal{A}}{(\eta+\mathcal{A})\left[1+\frac{\rho(\sigma-1)}{n}\right]-\eta} \quad \text{(C.1)}$$

and

$$\begin{aligned} l_A^{sp} =& 1 - l_P^{sp} - \left(1+\frac{\rho(\sigma-1)}{n}\right) l_R^{sp} \\ =& -\frac{(1+\eta)\left[1+\frac{\rho(\sigma-1)}{n}\right]-\eta}{(\eta+\mathcal{A})\left[1+\frac{\rho(\sigma-1)}{n}\right]-\eta} l_P^{sp} + \frac{\eta\frac{\rho(\sigma-1)}{n}}{(\eta+\mathcal{A})\left[1+\frac{\rho(\sigma-1)}{n}\right]-\eta}. \end{aligned} \quad \text{(C.2)}$$

To ensure that the labor shares satisfy $0 \le l_R^{sp} \le 1$ and $0 \le l_A^{sp} \le 1$, the feasible range of l_P^{sp} must satisfy:

$$-\frac{\mathcal{A}}{1-\mathcal{A}} \le l_P^{sp} \le \frac{\rho(\sigma-1)(\eta+\mathcal{A})}{n(1-\mathcal{A})}$$

and

$$-\frac{\mathcal{A}\left[n+\rho(\sigma-1)\right]}{\rho(1+\eta)(\sigma-1)+n} \leq l_P^{sp} \leq \frac{\eta\rho(\sigma-1)}{\rho(1+\eta)(\sigma-1)+n}.$$

By combining the above two conditions, we obtain the final range for l_P^{sp}:

$$0 \leq l_P^{sp} \leq \min\left\{\frac{\rho(\sigma-1)(\eta+\mathcal{A})}{n(1-\mathcal{A})}, \frac{\eta\rho(\sigma-1)}{\rho(1+\eta)(\sigma-1)+n}\right\}. \tag{C.3}$$

Finally, substituting (C.1) and (C.2) into (A.40) and (A.39), respectively, and considering the feasible range of l_P^{sp} derived in equation (C.3), we numerically solve for l_P^{sp} and $\bar{Q}^{sp}$.

Similarly, we now turn to the analysis of the competitive equilibrium. From equations (B.34) and (B.35), given the value of l_P^{dc}, the values of l_R^{dc} and l_A^{dc} can be derived as follows:

$$l_R^{dc} = \frac{(\eta-\mathcal{A})l_P^{dc}+\mathcal{A}}{(\eta^2+\mathcal{A})\left(1+\frac{C}{n\chi}\right)-\eta^2}$$

and

$$\begin{aligned} l_A^{dc} =& 1-l_P^{dc}-\left(1+\frac{C}{n\chi}\right)l_R^{dc} \\ =& -\frac{(\eta+\eta^2)\left(1+\frac{C}{n\chi}\right)-\eta^2}{(\eta^2+\mathcal{A})\left(1+\frac{C}{n\chi}\right)-\eta^2}l_P^{dc} + \frac{\eta^2\frac{C}{n\chi}}{(\eta^2+\mathcal{A})\left(1+\frac{C}{n\chi}\right)-\eta^2}. \end{aligned}$$

To ensure that the labor shares remain within valid bounds, the feasible range of l_P^{dc} is derived from equations (B.34) and (B.35) as follows:

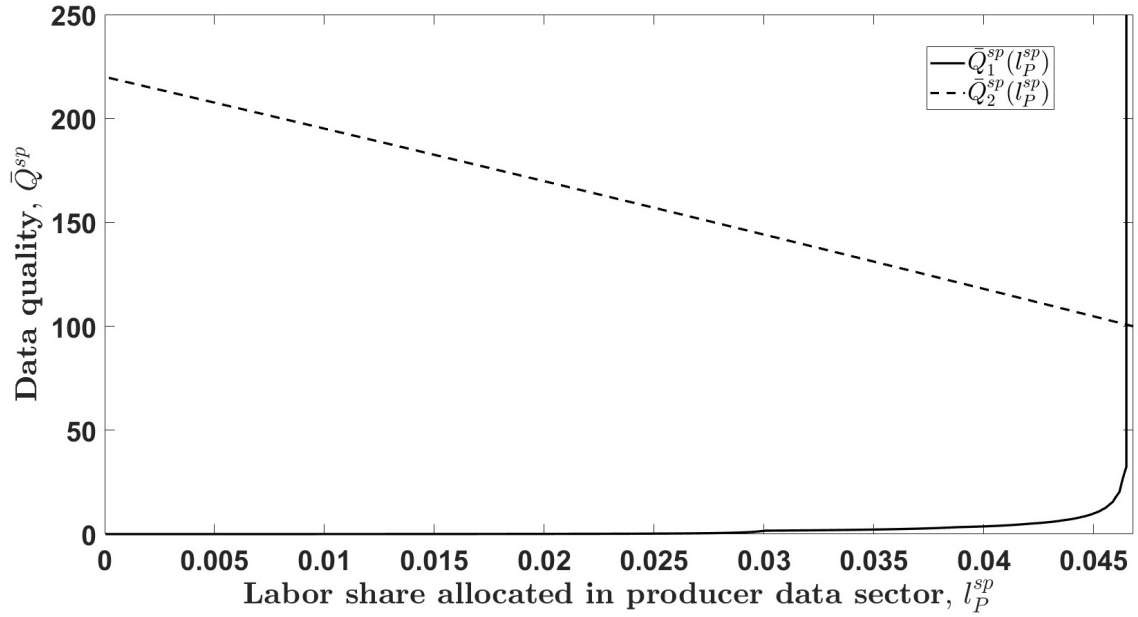
$$0 \leq l_P^{dc} \leq \min\left\{\frac{(\mathcal{A}+\eta^2)C}{n\chi(\eta-\mathcal{A})}, \frac{C\eta}{(1+\eta)C+n\chi}\right\}.$$

Finally, substituting the above results into equations (B.36) and (B.37), we numerically solve for l_P^{dc} and $\bar{Q}^{dc}$.

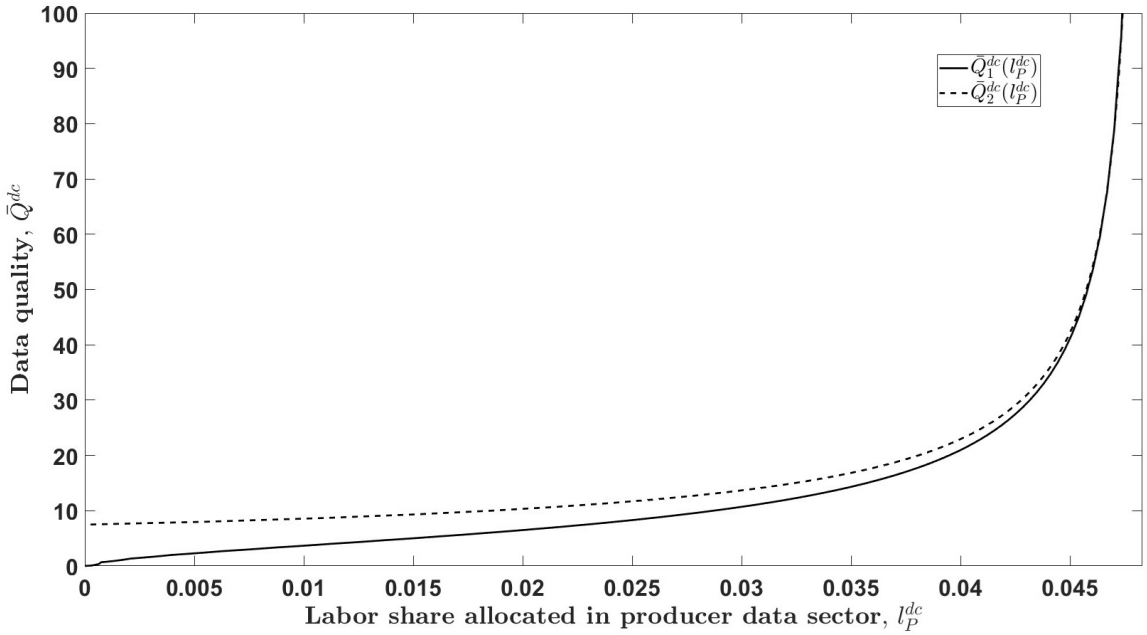
When τ, κ, and δ_0 all take their baseline parameter values, the two functions in both the optimal allocation and the competitive equilibrium are shown in Figure C.1.

The overall welfare of the economy is given by the integral:

$$\int_0^\infty e^{-(\rho-n)t}\left(\ln c_t + \kappa \ln Q_t\right) \mathrm{d}t$$



(a) Optimal allocation



(b) Competitive equilibrium

Figure C.1: Implicit functions determining l_P and $\bar{Q}$

Note. These figures display the implicit functions $\bar{Q}_1(l_P)$ (solid line) and $\bar{Q}_2(l_P)$ (dashed line) in both the optimal allocation and the competitive equilibrium. We utilize the equations $l_A(l_P)$ and $l_R(l_P)$ to refine the feasible range of l_P.

$$
\begin{aligned}
&= \int_0^\infty e^{-(\rho-n)t} \left[\ln\left(\tilde{c}e^{g_c t}\right) + \kappa \ln \bar{Q}\right] \mathrm{d}t \\
&= g_c \int_0^\infty e^{-(\rho-n)t} t \mathrm{d}t + \left(\ln \tilde{c} + \kappa \ln \bar{Q}\right) \int_0^\infty e^{-(\rho-n)t} \mathrm{d}t \\
&= -\frac{g_c}{\rho-n} \int_0^\infty t \mathrm{d}e^{-(\rho-n)t} + \frac{\ln \tilde{c} + \kappa \ln \bar{Q}}{\rho-n} \\
&= -\frac{g_c}{\rho-n} \left(t e^{-(\rho-n)t} \Big|_0^\infty - \int_0^\infty e^{-(\rho-n)t} \mathrm{d}t\right) + \frac{\ln \tilde{c} + \kappa \ln \bar{Q}}{\rho-n} \\
&= \frac{g_c}{(\rho-n)^2} + \frac{\ln \tilde{c} + \kappa \ln \bar{Q}}{\rho-n}.
\end{aligned}
$$

In this expression, $\tilde{c}$ represents the constant component of c_t^{sp} and c_t^{dc} under the respective settings. By computing the welfare levels under both the optimal allocation and the competitive equilibrium, we can derive the welfare ratio between the two settings.

C.2 Quantitative Analysis When $g_{D_A} = g_{D_P}$

We now present numerical results for the case where the two types of data grow at the same rate, as explicit solutions could not be derived in Sections 4 and 5. First, we outline the process for determining labor shares and data quality in this regime. Then, we conduct comparative statics on the three key parameters: τ, κ, and δ_0.

The process of solving the model numerically. We take the solution of the optimal allocation as an example. The three labor shares, l_P^{sp}, l_A^{sp}, and l_R^{sp}, along with data quality, $\bar{Q}^{sp}$, are derived from equations (A.33), (A.36), (A.39), and (A.40). Since equations (A.33) and (A.36) are linear, we can easily express l_A^{sp} and l_R^{sp} as functions of l_P^{sp}, and substitute these into the remaining two equations, (A.40) and (A.39), to obtain two functions, $\bar{Q}_1^{sp}\left(l_P^{sp}\right)$ and $\bar{Q}_2^{sp}\left(l_P^{sp}\right)$.[11] The derivations are detailed in Appendix C. The solution in the competitive equilibrium is derived similarly, yielding two corresponding functions, $\bar{Q}_1^{dc}\left(l_P^{dc}\right)$ and $\bar{Q}_2^{dc}\left(l_P^{dc}\right)$.

Implicit functions for solving l_P and $\bar{Q}$. When τ, κ, and δ_0 all take their baseline parameter values, the two functions in both the optimal allocation and the competitive equilibrium are shown in Figure C.1. To simplify notation, we omit superscripts when representing variables in both settings. From the figure, we determine the solutions for l_P and $\bar{Q}$ in the corresponding regime by identifying the intersections of these functions. Additionally,

[11] We also determine the feasible range of l_P^{sp} by ensuring that $0 \leq l_A^{sp}\left(l_P^{sp}\right) \leq 1$ and $0 \leq l_R^{sp}\left(l_P^{sp}\right) \leq 1$, thereby narrowing the range of this variable.

we observe that both equations, $\bar{Q}_1(l_P)$ and $\bar{Q}_2(l_P)$, in the two settings are monotonic within the feasible range of relevant variables, and that $\bar{Q}_2^{dc}\left(l_P^{dc}\right)$ always grows faster than $\bar{Q}_1^{dc}\left(l_P^{dc}\right)$. This ensures that the solutions in both settings are unique, and that when parameters take alternative values, the equations can still be solved without the concern of multiple solutions. Finally, in this particular regime, we find that data quality in the optimal allocation is significantly higher than in the competitive equilibrium. This finding reinforces our main argument that firms do not internalize the externality associated with excessive reliance on AI-generated data.

Quantitative analysis of welfare. We then compare welfare outcomes between the competitive equilibrium and the optimal allocation. The parameters with the highest degree of uncertainty are τ, κ, and δ_0, so we examine the model's behavior across a wide range of values for these three parameters while holding all other parameters at their calibrated values.[12] We compute the ratio of welfare in the two settings, denoted as $\lambda = W^{dc}/W^{sp}$. Welfare in both settings is given by:

$$W = \frac{g_c}{(\rho - n)^2} + \frac{\ln \tilde{c} + \kappa \ln \bar{Q}}{\rho - n}, \tag{C.4}$$

where $\tilde{c}$ represents the constant term in c_t^{sp} and c_t^{dc} in each setting. In other words, per capita consumption can be expressed as $c_t = \tilde{c} L_t^{g_c t}$. By holding one of the three parameters fixed, we compute λ while varying the other two, with results presented in Figure C.2. Overall, we find that this ratio remains relatively stable across parameter variations, supporting the robustness of our calibration.

From the model presented in this paper, we observe that δ_0 does not influence the competitive equilibrium significantly compared with the other two parameters κ and τ. Although this parameter is necessary in the competitive equilibrium, its impact in welfare calculation is negligible, which is consistent with the analysis when the two types of data grow differently. From Figure C.2(a), we find that κ has a much stronger impact on the welfare ratio λ than δ_0, and W^{dc} converges to W^{sp} as κ decreases. However, δ_0 still influences welfare, albeit to a lesser extent, as indicated in the figure. Likewise, from Figure C.2(b), we observe that the effect of τ is significantly greater than that of δ_0. Furthermore, Figure C.2(c) reveals that the effect of κ diminishes as τ increases, and the welfare ratio λ increases when κ decreases and τ increases. As κ increases, the additional utility coming from data quality becomes more important, thereby amplifying the externality present in the competitive equilibrium.

[12] Specifically, we vary κ and δ_0 from 0.01 to 0.99, and τ from 0.2 to 0.99, as smaller values of τ lead to irregular solutions. The model remains well-behaved within the studied parameter range.

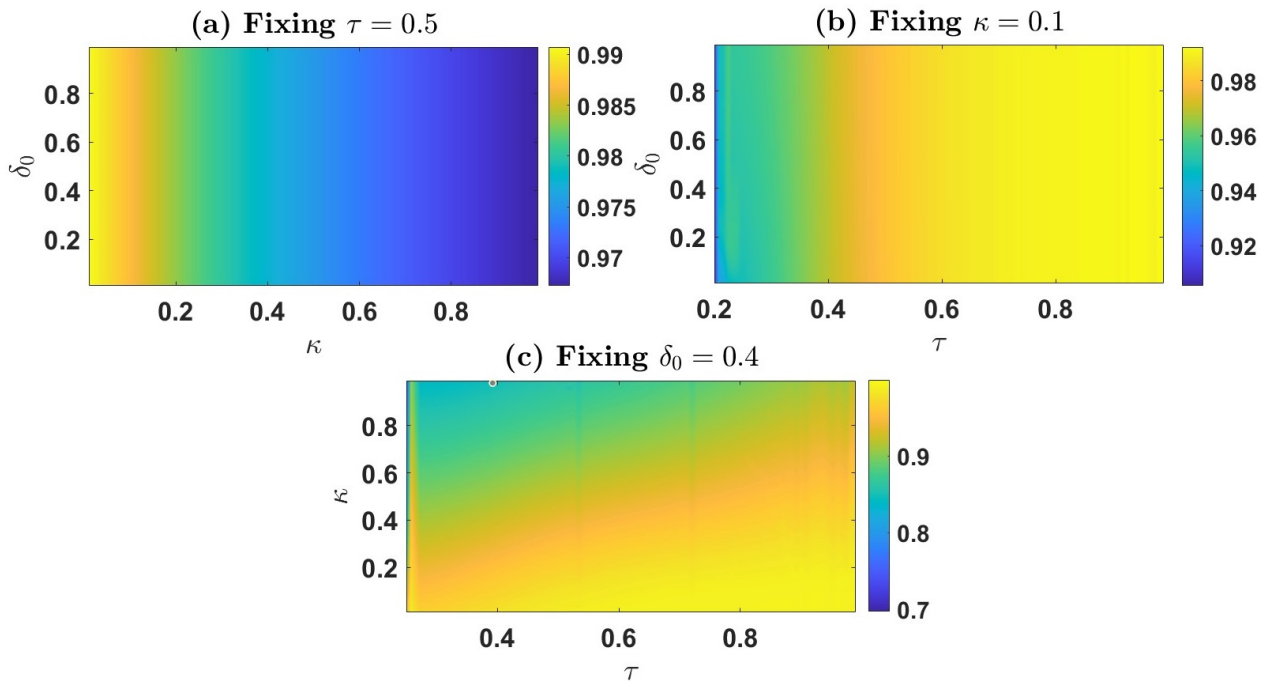


Figure C.2: Quantitative analysis of welfare

Note. These figures present the quantitative analysis of welfare across a wide range of values for τ, κ, and δ_0. The yellow and light-colored regions correspond to regimes where the welfare ratio λ is close to 1, while the blue and dark-colored regions indicate the opposite extreme.

To summarize, the sensitivity analysis presented above confirms that our model remains well-behaved even when the two types of data grow at the same rate, complementing the solutions derived in Propositions 1 and 2. In this regime, data quality is consistently greater than 1, indicating that the use of producer data is economically advantageous. In the competitive equilibrium, the weak influence of δ_0 on welfare demonstrates that the "loss of business" effect is trivial compared with other externalities. Firms lack the intension of allocating more producer data to mitigate this potential loss, thus lead to lower data quality in the competitive equilibrium.