Generative AI Meets Data Quality: Innovation or Risk?

Qing Chang[1] Danxia Xie[1] Longtian Zhang[2]

1. School of Social Sciences, Tsinghua University
2. School of International Trade and Economics, Central University of Finance and Economics

zhanglongtian@cufe.edu.cn

Seminar @ Beijing Institute of Mathematical Sciences and Applications

January 9, 2026

Outline

Introduction

Motivating Facts

- In recent years, the significance of data has been widely acknowledged by both academia and industry.
  - Newly emerging industries leverage data as a key factor of production and innovation.
  - The rising of Generative AI, like ChatGPT and other LLMs, is reshaping various aspects of our lives by producing an increasing volume of contents.
- Bali (PNAS, 2024): Generative AI is reshaping the social sciences by streamlining routine research tasks.
  - Writing, data cleaning, and software programming.
- The performance and reliability of AI systems critically depend on the quality of the data they utilize.

Introduction

Motivating Facts (Cont.)

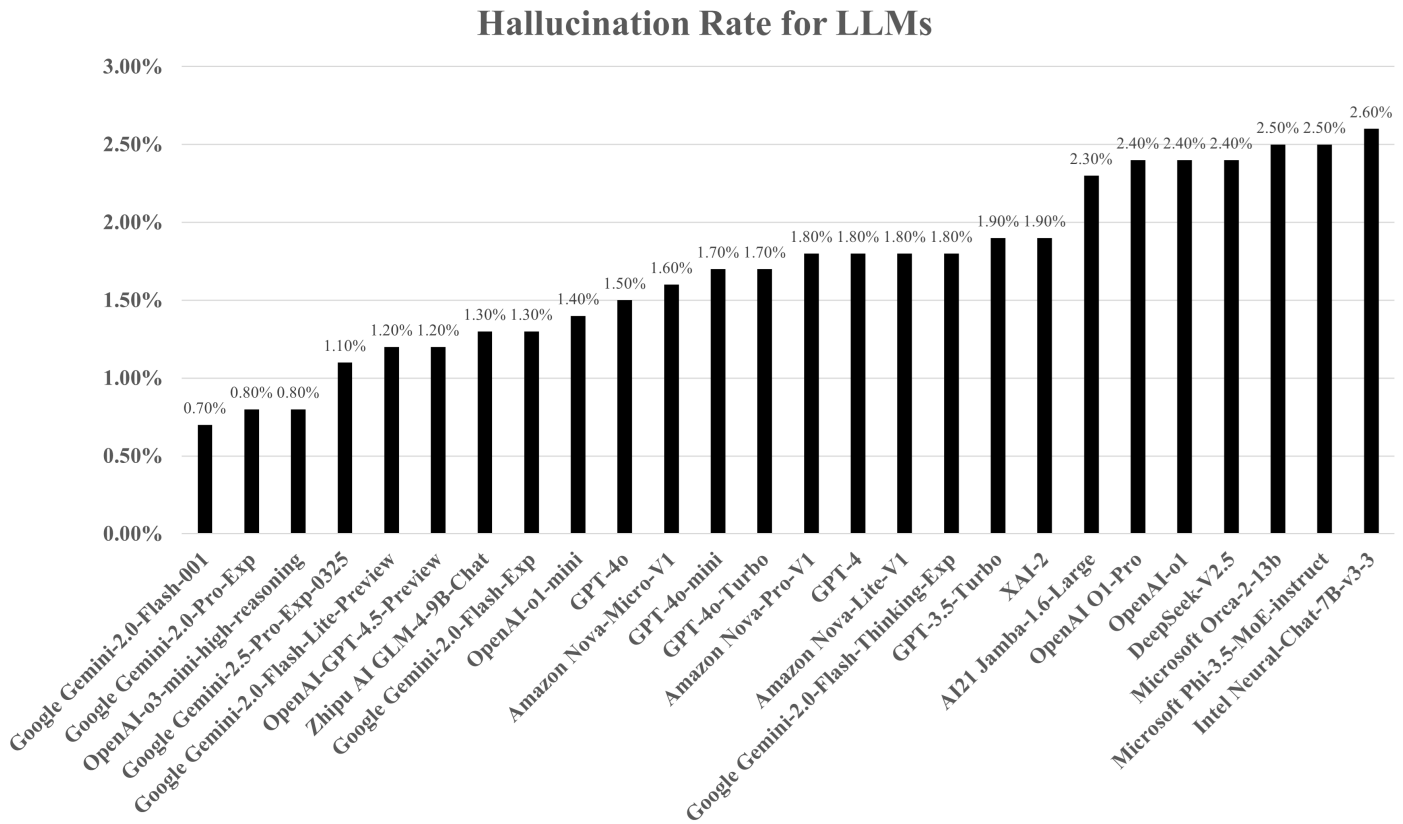

Figure: Hallucination Rate for Selected LLMs

This metric measures how frequently an LLM introduces hallucinations when summarizing a document.

Introduction

Overview

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.

- AI-generated data: derived from existing datasets, with labor introduced as the sole marginal input in the generation process.
- Data quality: determined by the ratio of the two types of data used and directly influences the error rate in production.
  - Higher data quality $\Rightarrow$ lower error rate $\Rightarrow$ higher total production
- Competitive equilibrium vs. social optimal

Producer data (real-world information) vs. AI-generated data

- Competitive equilibrium: firms may lose their business if their error rates become too high, which come from low data quality.
  - They still underutilize producer data.
- Number of AI firms
  - The optimal number of AI firms should always be ONE, while in competitive equilibrium, there may be multiple firms.
  - The equilibrium number of AI firms depends on the investment in these firms, the generation process of AI-generated data, and among others.

- Potential risks of generative AI: Acemoglu and Lensman (2024), Jones (2024).
- Research in computer science: Shumailov et al. (2024), Wenger (2024).
- Economic research in AI: Acemoglu (2024), Korinek and Vipra (2024), Brynjolfsson et al. (2025).

- Different paths to the economics of data:
  - Jones and Tonetti (2020): Horizontal nonrivalry & production process;
  - Cong, Xie, and Zhang (2021): Dynamic nonrivalry & innovation process;
  - Cong, Wei, Xie, and Zhang (2022): Vertical nonrivalry & both production process and innovation process simultaneously;
  - Xie and Zhang (2023): "producer data" lead to higher growth rate than "consumer data."
- Data do not always lead to sustained economic growth: Farboodi and Veldkamp (2020, 2021); Veldkamp (2023).

The General Model

Setup

- Representative household
  - consumption c_t, labor allocations $L_{i,t}$, $l_{P,t}$, $l_{A,t}$, and $l_{R,t}$.
- Final good producer
  - combine intermediate goods to produce output Y_t.
  - employ labor to generate producer data as byproduct
- Intermediate good producers
  - Use overall dataset and labor to produce intermediate goods.
- Generative AI firms
  - Generate data using the existing data and labor.
- Innovation sector
  - Use labor to innovate.

Utility function:

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

- Budget contraint:

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

- Two inputs (intermediate goods and labor) and two outputs (final good and producer data).
- Profit maximization problem:

$$\max_{\{Y_{i,t},L_{P,t}\}} Y_t + p_{D_P,t} D_{P,t} - \int_0^{N_t} p_{i,t} Y_{i,t} \mathrm{d}i - w_t L_{P,t},$$

where,

$$D_{P,t} = \left(\frac{Y_t}{L_t}\right)^{\theta} L_{P,t}.$$

- AI-generated data are produced by:

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

  - There are $M > 0$ homogeneous Generative AI firms in the economy, each operates in a monopolistic environment.
  - $d_{A,t}$ denotes the AI-generated data produced by a single firm.
  - The aggregate AI-generated data is given by $D_{A,t} = M d_{A,t}$.

- The optimization problem:

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

- We assume that the AI firm must incur a substantial upfront investment before it can begin providing AI-generated data to other firms.
  - Emergent capabilities of AI: $\mathcal{G}$, which is viewed as a fixed cost.
  - $\mathcal{G}$ is the threshold at which AI models become operational and begin generating new data.
  - Estimations: OpenAI's GPT-4 at \$40 million and Google's Gemini Ultra at \$30 million (Cottier et al., 2025).
- Free-entry condition of AI firms: The discounted profit of AI firms $\Pi = \mathcal{G}$.

The General Model

Intermediate good producers

- Combination of different types of data:

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

- Error rate that AI-generated data making mistakes:

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

- Data quality:

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

- Production function:

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

- Decision problem:

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

- "Loss of business" effect $\delta(e_i, t)$: 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.

- Two types of data intermediaries: one that handles producer data and another that handles AI-generated data.
- Both intermediaries operate as monopolists in their respective data markets but are constrained by free entry into data intermediation.
- One problem is:

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

subject to

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

- The innovation process is:

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

- Free entry condition:

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

Balanced Growth Path

Optimal Allocation

- Homogeneity among different varieties of intermediate goods:

$$D_{P,i,t} = D_{P,t}, \quad D_{A,i,t} = D_{A,t}, \quad e_{i,t} = e_t, \quad Q_{i,t} = Q_t$$

- Define a constant:

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

When Producer Data and AI-Generated Data Grow at Different Rates

The labor shares allocated to different sectors are given by:

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

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

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

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

- Discussions.
  - When AI-generated data dominate, I_P^{sp} shrinks to zero.
    - Although the error rate in intermediate goods production converges to one—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.
  - When producer data dominate, I_A^{sp} shrinks to zero.
    - The positive effect of AI-generated data is not large enough for usage in this case, while they still have negative effects.

Competitive Equilibrium

When $(1-\tau)\epsilon > 1$, the labor shares allocated to different sectors converge to:

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

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

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

Here, $\mathcal{C}$ is a constant which is defined by:

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

The growth rate of the economy remaining identical to that in the optimal allocations.

Balanced Growth Path

Competitive Equilibrium (Cont.)

- I_P^{dc} shrinks to zero when AI-generated data dominate, while I_A^{dc} shrinks to zero when producer data dominate.
- "Loss of business" effect
  - negligible when $g_{D_A}^{dc} \leq g_{D_P}^{dc}$
  - significant when $g_{D_A}^{dc} > g_{D_P}^{dc}$, but does not change the trend that $I_{P,t} \to 0$ (the producer data sector is still negligible).
- An important condition that ensures the model is well-behaved:

$$(1-\tau)\epsilon > 1$$

  - the sensitivity of data quality should not be excessively large
  - the elasticity of substitution between the two types of data should not be too low

Balanced Growth Path

Data quality

- The growth rates of data quality are also different in the two scenarios.

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

and

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

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:

$$M^{dc} = \left\{ \frac{(\mathcal{A}-\eta)(\sigma-1)(1-e_0)}{\sigma\eta(\rho-n)(n\chi)^{\frac{1}{\sigma-1}}\mathcal{G}} \left[\frac{\psi(1-\zeta)}{n+\delta_A(1-\zeta)}\right]^{\frac{\eta}{1-\zeta}} (I_R^{dc})^{\frac{1}{\sigma-1}} (I_A^{dc})^{1+\frac{\eta}{1-\zeta}} L_0^{1+\frac{1}{\sigma-1}+\frac{\eta}{1-\zeta}} \right\}^{\frac{1-\zeta}{1-\zeta(1-\eta)}}.$$

Numerical Examples and Further Discussions

Calibration

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 & 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

Numerical Examples and Further Discussions

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

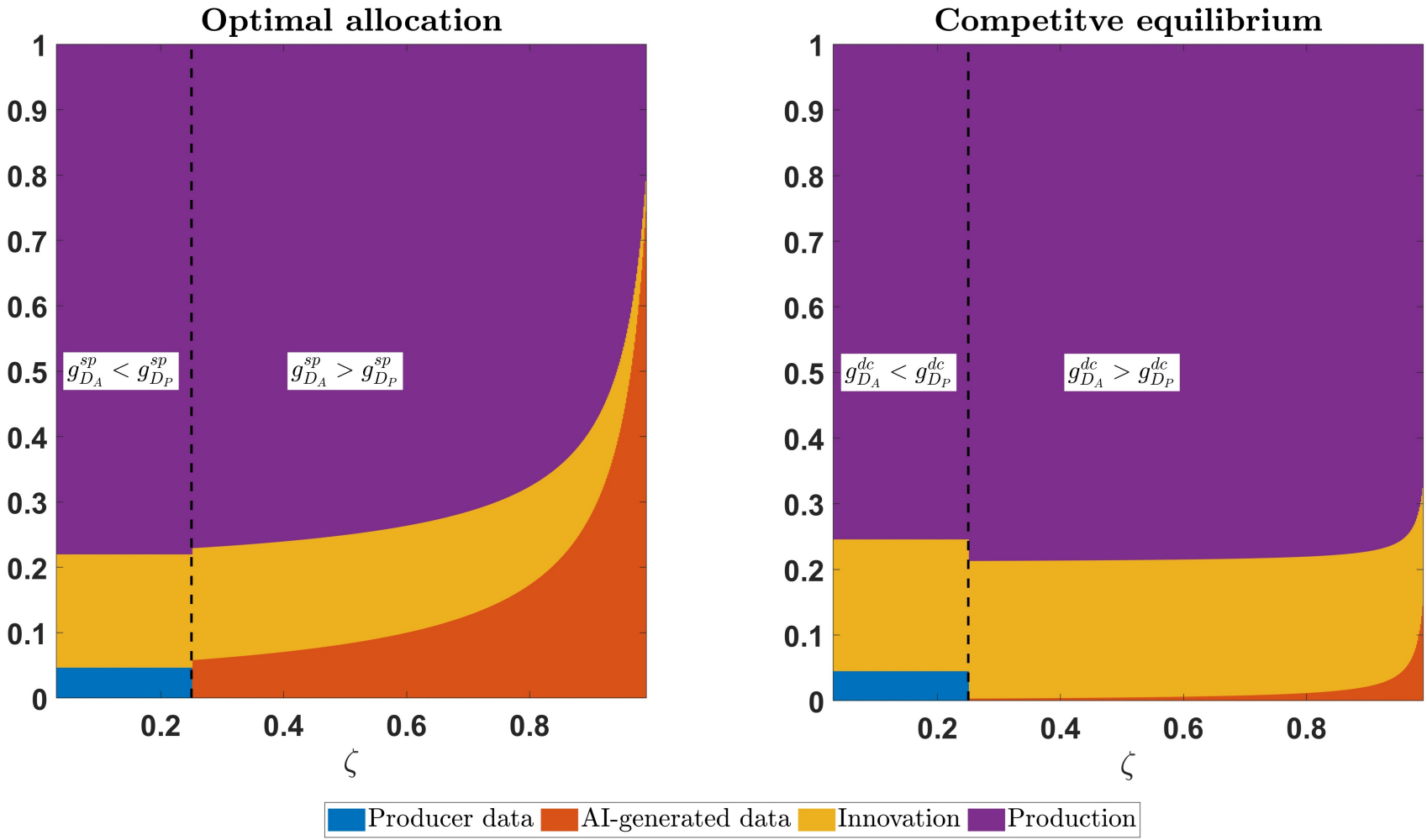

Figure: Labor allocations in the four different sectors

Numerical Examples and Further Discussions

Numerical Examples When $g_{D_A} \neq g_{D_P}$ (Cont.)

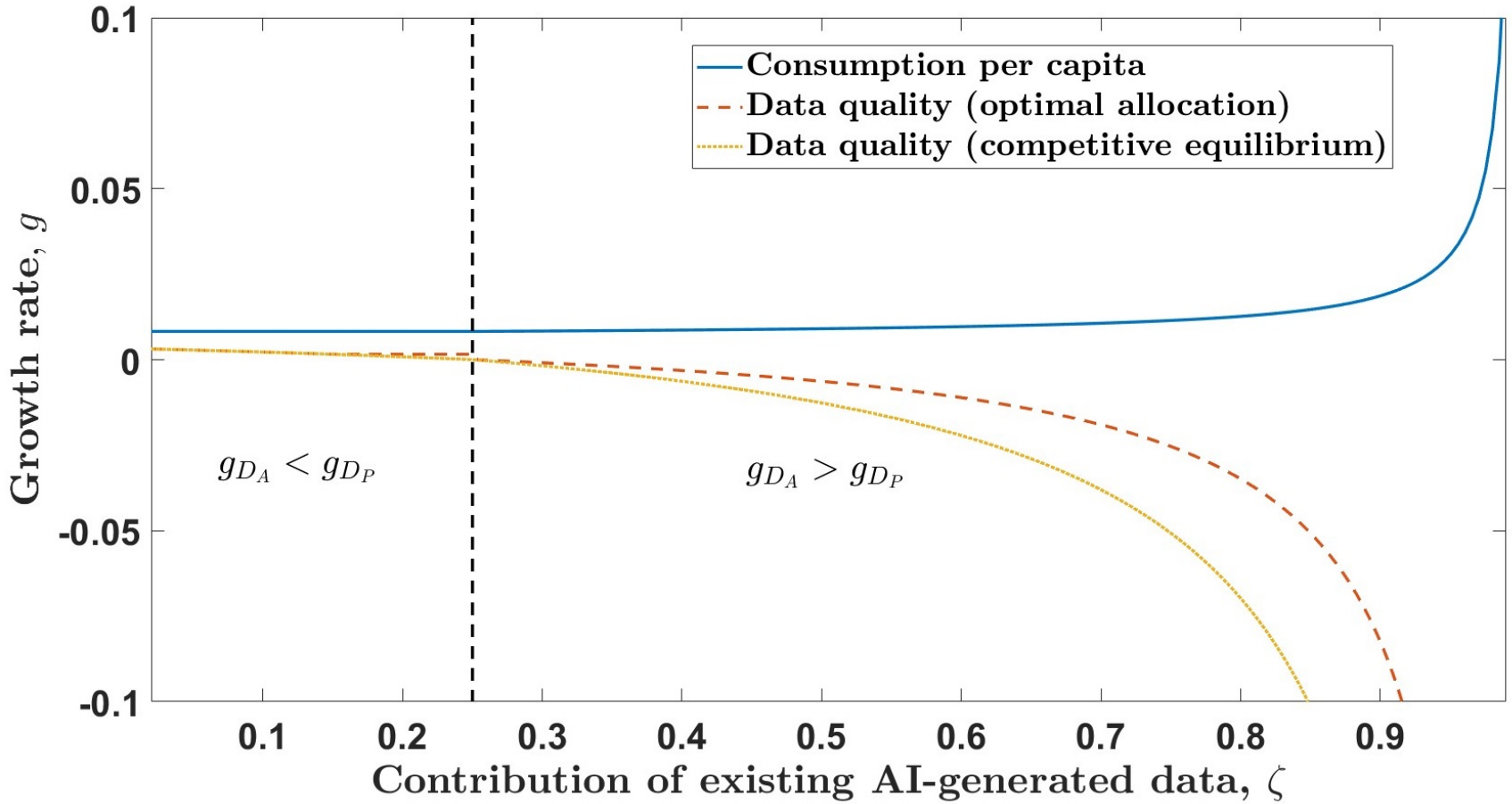

Figure: Growth rates of consumption per capita and data quality

Numerical Examples and Further Discussions

Number of Generative AI Firms

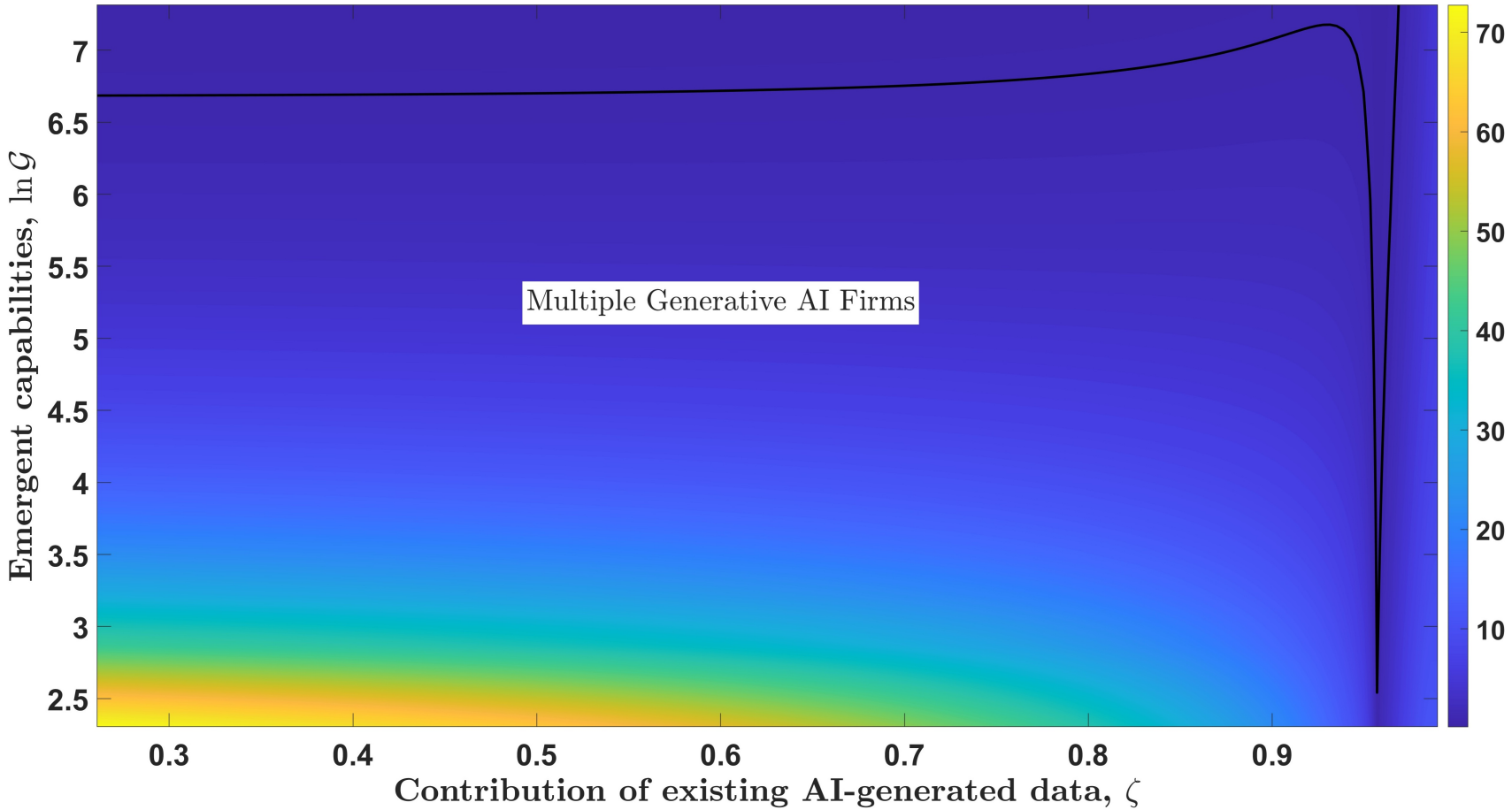

Figure: Number of Generative AI firms in Competitive Equilibrium

- 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.
  - Encourage better data composition: subsidies for high-quality producer data; limiting excessive entry in the AI sectors.
  - Support public data infrastructure that can improve baseline quality of data.
- In most cases, there are too many Generative AI firms in the market.
  - Multiple firms leads to redundant investment.
  - Governments need not be overly concerned about the concentration trend in the AI industry.

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.
  - We emphasize the importance of integrating real-world data, such as the producer data considered in our model.
  - Firms consistently underuse real-world data.
- This finding raises important policy considerations regarding whether governments should regulate the use of AI-generated data and the broader development of the AI industry.
  - Our study contributes to the literature by advancing growth theory on risks associated with AI technologies and data as a factor of production.

Thanks for your attention!