Artificial Intelligence as a Structural General-Purpose Technology: Endogenous Wage Distribution, Infrastructure Constraints, and Institutional Realignment

Rachit Lohani

February 27, 2026

Abstract

Artificial intelligence (AI) is increasingly treated as a general-purpose technology, yet much of the policy and economic discourse remains in partial equilibrium: AI is modeled as "software" that raises productivity or substitutes for labor without binding constraints or distributional feedback. This manuscript develops a general equilibrium (GE) framework in which AI capital endogenously reallocates tasks, changes the wage distribution, and interacts with physical infrastructure constraints—particularly electricity availability and data-center expansion. We embed a task-reallocation structure into a multi-factor production economy with heterogeneous labor and derive conditions under which wage polarization intensifies, even when aggregate output rises. We then propose a difference-in-differences strategy for empirical identification and construct a standardized taxonomy for AI-related capital expenditures using firm filings. Sector case studies in healthcare diagnostics, energy permitting, retail agentic commerce, and compute sovereignty illustrate second- and third-order impacts that are best understood as institutional reconfiguration rather than incremental productivity gains. We conclude with policy implications oriented toward infrastructure planning, competition policy, and labor-market insurance under AI-driven capital deepening.

1 Introduction

General-purpose technologies reorder economies because they alter production functions, factor shares, and the institutions that allocate rents. AI differs from earlier waves of automation in the domain it targets: cognitive tasks that historically anchored the wage premium for educated labor. The near-term debate is often framed in terms of job counts or model capabilities, but the deeper question is structural: how does an AI-intensive capital stock change the wage distribution, local political economy, and the governance boundary between states and firms?

Empirically, several signals point to a phase change. Data-center electricity demand is projected to rise rapidly through 2030, with AI-optimized facilities a primary driver (International Energy Agency, 2024; European Commission, 2025; Pew Research Center, 2025). At the same time, hyperscalers and frontier-model firms have escalated infrastructure spending, as evidenced in recent SEC

filings and earnings materials (Alphabet Inc., 2026; Amazon.com, Inc., 2026; Meta Platforms, Inc., 2026a; Microsoft Corporation, 2025). The macro implication is straightforward but under-modeled: AI scaling is increasingly constrained by energy, land, water, and grid interconnection, and those constraints feed back into adoption, pricing, and wage outcomes.

This paper proceeds as follows. Section 2 develops a GE model with endogenous wage dispersion under task reallocation. Section 3 introduces an infrastructure constraint that binds AI capital accumulation. Section 4 presents a standardized AI-capex taxonomy and a filings-based extraction approach. Section 5 outlines an empirical identification strategy. Section 6 provides sectoral case studies. Section 7 concludes with policy implications. Appendices provide methods, data, and a calibrated simulation exercise.

2 General Equilibrium Model with Endogenous Wage Distribution

2.1 Households

The economy is populated by a unit mass of households with heterogeneous labor endowments. Two representative skill groups are considered: high-skill labor L_H and mid-skill labor L_M. Household utility is separable over consumption C and labor supply, and labor is supplied inelastically for simplicity (extensions can add disutility and endogenous hours).

2.2 Firms and Production

A representative competitive firm produces output Y using physical capital K, AI capital A, and an effective labor aggregate $\Theta(\cdot)$:

$$Y = \left[\alpha K^{\rho} + \beta A^{\rho} + (1-\alpha-\beta)\,\Theta(L_H, L_M, A)^{\rho}\right]^{1/\rho}, \tag{1}$$

where $\rho < 1$ governs substitution across the three broad inputs.

The effective labor aggregate captures task reallocation. We model a share of mid-skill tasks as automatable by AI:

$$\Theta(L_H, L_M, A) = \left[(1-\gamma)L_H^{\sigma} + \gamma\left((1-\phi)L_M + \phi A\right)^{\sigma}\right]^{1/\sigma}, \tag{2}$$

where $\phi \in [0,1]$ is the automation intensity applied to L_M tasks, and $\sigma < 1$ captures imperfect substitution between L_H and the composite of (L_M, A). This structure is in the spirit of task-based technological change models (Autor et al., 2003; Acemoglu and Restrepo, 2018).

2.3 Factor Prices and Wage Distribution

Under perfect competition, factor prices equal marginal products. For analytic intuition, focus on the task layer in Eq. 2. Let $X \equiv (1-\phi)L_M + \phi A$. Then:

$$w_H \propto \frac{\partial \Theta}{\partial L_H} = (1-\gamma) L_H^{\sigma-1} \, \Theta^{1-\sigma}, \tag{3}$$

$$w_M \propto \frac{\partial \Theta}{\partial L_M} = \gamma X^{\sigma-1} \, \Theta^{1-\sigma} \, (1-\phi). \tag{4}$$

Eq. 4 shows that the marginal product of mid-skill labor declines with higher automation intensity ϕ, both directly through $(1-\phi)$ and indirectly through the expansion of X by A when $\sigma < 1$. Conversely, w_H rises when L_H is complementary to AI-augmented production, a condition consistent with empirical patterns in task polarization (Autor et al., 2003; Webb, 2020).

A sufficient condition for wage polarization is:

$$\frac{\partial w_H}{\partial A} > 0 \quad \text{and} \quad \frac{\partial w_M}{\partial A} < 0, \tag{5}$$

which holds broadly when $\sigma < 1$ and ϕ is sufficiently large so that AI substitutes for routine cognitive components of L_M.

We summarize wage dispersion using the skill premium w_H/w_M. In the simulation appendix we illustrate that increases in ϕ amplify the premium sharply even with constant L_H and L_M, implying that distributional impacts can be large without any aggregate employment collapse.

3 Infrastructure-Constrained AI Capital Accumulation

AI capital is compute-intensive. Let A be produced (or accumulated) via compute capacity C and data D, $A = f(C, D)$, but compute is bounded by electricity supply E:

$$C \leq \kappa E. \tag{6}$$

Therefore AI capital is indirectly bounded:

$$A \leq f(\kappa E, D). \tag{7}$$

This constraint is not theoretical; it is increasingly visible in published estimates of data-center energy use. Global data-center electricity demand is projected to rise toward *approximately* 945 TWh by 2030 in the IEA base case (International Energy Agency, 2024), and the European Commission echoes the same magnitude in its policy briefings (European Commission, 2025). In the U.S., Pew summarizes IEA estimates that data centers consumed 183 TWh in 2024 and could reach 426 TWh by 2030 (Pew Research Center, 2025). These projections imply that AI adoption is not merely a software diffusion curve; it is a physical buildout with grid, permitting, and local resource

consequences.

The GE implication is that the marginal productivity of AI capital may remain high (driving investment) even as energy constraints create local bottlenecks. That dynamic can increase rents to incumbents with preferential access to power, land, and interconnection, and it can create new forms of regional inequality.

4 A Filings-Based AI Capex Dataset and Standardized Taxonomy

A persistent empirical problem is that "AI investment" is not a standardized accounting category. We propose a taxonomy that can be consistently extracted from public disclosures and approximated in a replicable way.

AI-related investment is classified into four buckets: (i) compute infrastructure (servers, accelerators, networking, and data-center buildout), (ii) model development (AI-specific R&D attributable to training, evaluation, and safety), (iii) energy procurement and grid-related investments (PPAs, on-site generation, interconnection upgrades where disclosed), and (iv) data acquisition (licensing, collection, labeling).

The dataset is constructed using SEC filings and accompanying earnings materials, extracting (a) reported total capex, (b) disclosures attributing capex to technical infrastructure and data centers, and (c) forward guidance where available. For example, Alphabet reports capital expenditures of $52.5B (2024) and $91.4B (2025) and expects further increases focused on technical infrastructure and data centers (Alphabet Inc., 2026). Amazon reports cash capex of $77.7B (2024) and $128.3B (2025), driven primarily by technology infrastructure supporting AWS (Amazon.com, Inc., 2026). Microsoft reports additions to property and equipment of $44.477B (FY2024) and $64.551B (FY2025) (Microsoft Corporation, 2025). Meta reports 2025 capex (including finance lease principal payments) and provides 2026 capex guidance of $115–135B (Meta Platforms, Inc., 2026a,b).

Table 1 presents a minimal illustration using disclosed capex totals. In the appendix we describe extraction rules, reconciliation steps, and limitations.

Table 1: Illustrative filings-based capex totals (selected firms).

Firm	2024 capex (USD B)	2025 capex (USD B)	Primary source
Alphabet	52.5	91.4	Form 10-K (FY2025) (Alphabet Inc., 2026)
Amazon	77.7	128.3	Form 10-K (FY2025) (Amazon.com, Inc., 2026)
Microsoft	44.477	64.551	FY25 Q4 IR release (Microsoft Corporation, 2025)
Meta	(see filing)	72.22 (FY2025)	10-K / earnings exhibit (Meta Platforms, Inc., 2026a,

5 Empirical Identification Strategy

To estimate AI-induced labor and wage effects, we propose a difference-in-differences design that compares outcomes for "AI restructuring" firms to matched controls within the same industry and size bin:

$$Employment_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \beta_3 (Post_t \times Treat_i) + \Gamma X_{it} + \varepsilon_{it}. \tag{8}$$

The treatment indicator $Treat_i$ is defined based on explicit disclosures attributing workforce restructuring or productivity shifts to AI adoption, in earnings calls or filings. Controls are matched on pre-trends in revenue growth, profitability, headcount, and capex levels. Outcome variables include total employment, occupational composition (where measurable), wage dispersion proxies, and revenue per employee. The motivation aligns with empirical work on automation shocks in local labor markets (Acemoglu and Restrepo, 2020).

6 Sectoral Case Studies

Healthcare diagnostics offers a concrete setting where AI can raise throughput yet create a new liability and oversight regime. Deep neural networks can reach clinician-level performance in narrowly defined image classification tasks (Esteva et al., 2017). In practice, deployment shifts work from routine interpretation to exception handling, audit, and medico-legal accountability. This can raise measured productivity while increasing "judgment load," which interacts with fatigue and risk.

Energy permitting conflicts illustrate why AI is infrastructure-limited. Large data centers require not only megawatts but predictable interconnection timelines, water or alternative cooling, and local zoning legitimacy. Where permitting is slow or contested, firms may secure private generation or long-term PPAs, potentially crowding out other decarbonization priorities. The IEA projections and EU briefings indicate that accelerated computing is a primary driver of demand growth through 2030 (International Energy Agency, 2024; European Commission, 2025).

Retail agentic commerce changes market structure by interposing algorithmic agents between consumers and brands. When agents optimize on price, availability, and delivery time, brand loyalty becomes a feature of the agent's objective function rather than a consumer's identity. This can compress price dispersion and intensify platform power over discovery and checkout. The economic concern is not simply "targeted ads" but a reallocation of bargaining power toward whoever controls agent defaults.

Compute sovereignty—visible in EU and national initiatives and in U.S. hyperscaler buildouts—suggests that compute capacity is becoming a strategic input analogous to energy infrastructure. When AI capability depends on constrained physical resources and supply chains, states have incentives to treat compute as industrial policy.

7 Policy Implications

A policy agenda that treats AI primarily as a labor-market disruption misses the infrastructural and institutional channels through which second- and third-order effects emerge. First, grid planning must treat data centers and accelerated computing as a new class of industrial load and must rationalize interconnection queues without devolving into "fast lane" favoritism. Second, competition policy must anticipate that compute scarcity and capex scale can entrench incumbents. Third, labor policy must recognize that the shift is from production to judgment and accountability, meaning retraining should emphasize verification, systems thinking, and domain responsibility rather than generic "prompting." Fourth, AI governance should address the principal–agent problem in which firms may rationally externalize social risk if liability regimes remain ambiguous.

A Methods and Data Appendix

A.1 Data Sources

This manuscript uses three primary classes of sources. Energy and data-center demand projections are drawn from the International Energy Agency and EU energy directorate publications, supplemented by Pew Research Center summaries for U.S. context (International Energy Agency, 2024; European Commission, 2025; Pew Research Center, 2025). Corporate investment measures rely on SEC filings and investor-relations disclosures for capex totals and infrastructure attribution (Alphabet Inc., 2026; Amazon.com, Inc., 2026; Meta Platforms, Inc., 2026a; Microsoft Corporation, 2025). Labor-task theory and empirical anchors draw from the task literature in economics and AI exposure measurement (Autor et al., 2003; Acemoglu and Restrepo, 2018, 2020; Webb, 2020).

A.2 Extraction Approach for AI-Related Capex

We extract total capex and then classify subcomponents using disclosure text. The minimal replicable rule set is as follows. First, identify total capex as reported in the cash flow statement or management discussion. Second, identify mentions of "technical infrastructure," "data centers," "servers," "network equipment," or "AI infrastructure" and treat that share as compute infrastructure bucket. Third, where the firm provides explicit forward guidance for capex tied to AI efforts, record separately as guidance. Fourth, reconcile definitions ("cash capex" vs "capex including finance leases") and keep both where present, rather than forcing a single definition.

A.3 Limitations

The primary limitation is measurement: firms do not report AI capex in a standardized manner. Some report cash capex; others emphasize additions to property and equipment; others include finance leases. A second limitation is attribution: even when capex is clearly infrastructure-related, the AI share is inferred from narrative disclosures rather than line-item accounting. A third limitation is cross-country comparability because disclosure regimes vary and because infrastructure constraints differ by grid structure and permitting rules.

B Simulation Appendix: Calibration and Numerical Illustration

B.1 Calibration Philosophy

The simulation is not intended as a point forecast; it is a calibrated illustration that ties the model's comparative statics to plausible parameter ranges in the literature. The key objects are the task-substitution elasticity σ, the automation intensity ϕ, and the infrastructure constraint mapping electricity to AI capital. Task models imply that wage effects depend critically on the degree to which AI substitutes for routine tasks and complements nonroutine tasks (Autor et al., 2003; Acemoglu and Restrepo, 2018). AI exposure work suggests that AI is directed toward high-skill

tasks in a way that differs from prior automation technologies (Webb, 2020). Energy projections motivate imposing a binding constraint on A through E (International Energy Agency, 2024).

B.2 Baseline Parameter Values

Table 2 lists baseline values used in the numerical exercise. We normalize labor endowments to $L_M = 1$ and set $L_H = 0.6$ to represent a smaller high-skill workforce share. We set $\sigma = 0.7$ (imperfect substitution), $\gamma = 0.55$ to weight the mid-skill/AI task composite, and explore automation intensity ϕ in $[0.2, 0.8]$. The energy constraint maps E to effective AI capital A via $A = \min(\bar{A}, \kappa E)$.

Table 2: Baseline simulation parameters (illustrative).

Parameter	Value	Interpretation / justification
L_M	1.0	Mid-skill labor normalization
L_H	0.6	High-skill labor share (normalized)
σ	0.7	Task-substitution elasticity parameter; $\sigma < 1$ yields complementarity (Autor et al., 2003; Acemoglu and Restrepo, 2018)
γ	0.55	Weight on mid-skill/AI composite tasks
ϕ	0.2–0.8	Automation intensity applied to L_M tasks (varied)
κ	1.2	Electricity-to-compute mapping (efficiency)
$\bar{A}$	3.0	Unconstrained AI capital ceiling (normalized)

B.3 Simulation Results: Wage Polarization Under Rising Automation

We compute wages as marginal products at the task layer (Eq. 3–4) and report the skill premium w_H/w_M. Table 3 shows that the premium rises sharply as automation intensity ϕ increases, holding $A = 1$ fixed. This is the core distributional comparative static: even absent a collapse in output, AI can reprice labor by changing which tasks are scarce.

Table 3: Skill premium w_H/w_M as automation intensity rises (normalized, $A = 1$).

ϕ	0.2	0.4	0.6	0.8
w_H/w_M	1.19	1.59	2.38	4.77

B.4 Simulation Results: Infrastructure Constraint Binding

We next impose an energy constraint. Effective AI capital is $A = \min(\bar{A}, \kappa E)$. Table 4 shows that as E increases, A rises until it hits the ceiling, and the skill premium rises with A when ϕ is high.

The policy implication is that grid expansion can amplify wage dispersion if it disproportionately accelerates AI capital deepening without complementary labor-market adjustment.

Table 4: Energy constraint and effective AI capital ($\phi = 0.6$, $\kappa = 1.2$, $\bar{A} = 3$).

E	A (effective)	w_M	w_H	w_H/w_M
0.5	0.6	0.213	0.469	2.20
1.0	1.2	0.204	0.504	2.47
1.5	1.8	0.199	0.533	2.68
2.0	2.4	0.195	0.558	2.86
3.0	3.0	0.192	0.581	3.02

B.5 Reproducibility Notes (Pseudo-code)

The simulation can be reproduced in any numerical environment. The core steps are: (i) specify $(L_H, L_M, \gamma, \sigma, \phi)$, (ii) compute $X = (1-\phi)L_M + \phi A$, (iii) compute Θ from Eq. 2, and (iv) compute marginal products for w_H and w_M from Eq. 3–4. The infrastructure constraint is applied as $A = \min(\bar{A}, \kappa E)$.

References

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Acemoglu, D. and Restrepo, P. (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy*, 128(6):2188–2244.

Alphabet Inc. (2026). Annual report on form 10-k for the fiscal year ended december 31, 2025. SEC filing (PDF). Available at `https://s206.q4cdn.com/479360582/files/doc_financials/2025/q4/GOOG-10-K-2025.pdf` (accessed 2026-02-27).

Amazon.com, Inc. (2026). Annual report on form 10-k for the fiscal year ended december 31, 2025. SEC filing (HTML). Available at `https://www.sec.gov/Archives/edgar/data/1018724/000101872426000004/amzn-20251231.htm` (accessed 2026-02-27).

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118.

European Commission (2025). In focus: Data centres – an energy-hungry challenge. Policy brief/webpage. Available at `https://energy.ec.europa.eu/news/focus-data-centres-energy-hungry-challenge-2025-11-17_en` (accessed 2026-02-27).

International Energy Agency (2024). Energy and ai: Energy demand from ai. Report/webpage. Available at `https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai` (accessed 2026-02-27).

Meta Platforms, Inc. (2026a). Annual report on form 10-k for the fiscal year ended december 31, 2025. SEC filing (HTML). Available at `https://www.sec.gov/Archives/edgar/data/1326801/000162828026003942/meta-20251231.htm` (accessed 2026-02-27).

Meta Platforms, Inc. (2026b). Meta reports fourth quarter and full year 2025 results (exhibit 99.1 / ir release). SEC exhibit / investor release. Available at `https://www.sec.gov/Archives/edgar/data/1326801/000162828026003832/meta-12312025xexhibit991.htm` (accessed 2026-02-27).

Microsoft Corporation (2025). Fy25 q4 press release and webcast (cash flow table with additions to property and equipment). Investor Relations webpage. Available at `https://www.microsoft.com/en-us/investor/earnings/fy-2025-q4/press-release-webcast` (accessed 2026-02-27).

Pew Research Center (2025). What we know about energy use at u.s. data centers amid the ai boom. Short Reads/webpage. Available at `https://www.pewresearch.org/short-reads/`

2025/10/24/what-we-know-about-energy-use-at-us-data-centers-amid-the-ai-boom/ (accessed 2026-02-27).

Webb, M. (2020). The impact of artificial intelligence on the labor market. Working paper. Available at https://www.michaelwebb.co/webb_ai.pdf (accessed 2026-02-27).