Digital Skill Premia, Structural Polarization, and Skill Complementarity

Evidence from Chinese Listed-Firm Job Postings, 2016–2025

Xi Xiang

Department of Economics, Nanjing University

2026.4.3

Outline

Outline

Motivation

- **Central question**: Does digital technology raise wages, and for whom?
- Classic SBTC framework (Katz & Murphy 1992; Acemoglu 2002): technology raises demand for abstract/complementary skills
- Task-based models (Autor et al. 2003; Acemoglu & Autor 2011): technology may also *polarize* — substituting routine middle-skill tasks

Why China?

1. Exceptionally rapid tech adoption 2016–2025 (cloud $\rightarrow$ big data $\rightarrow$ LLMs)
2. Sharp distributional tensions: massive low-skill supply + booming college workforce
3. Listed-firm job postings: high-frequency employer-side wage signal

Three Contributions

1. **Time-varying premium dynamics**
   Annual regressions reveal a hump-shaped AI/ML premium trajectory (35% $\rightarrow$ 55% $\rightarrow$ 25%), interpreted as demand shock followed by lagged supply response

2. **Skill complementarity**
   All pairwise skill interactions are negative (sub-additive): holding two digital skills pays less than the sum of individual premia

3. **Structural polarization & wage inequality**
   Industry digital-skill demand *diffused* (Gini: 0.20 $\rightarrow$ 0.14), yet upper-tail wage inequality widened; quantile regression unpacks why

Outline

Data

Sample

- 6.93 million job postings from Chinese listed firms, 2016–2025
- *Wage analysis sample*: 820,807 obs. with valid salary
- 57,460 distinct firms; 3,824 city×year cells
- Salary standardized to monthly midpoint, log form

Skill Classification

- Structured 13-category NLP dictionary
- Core digital skills:
  - AI / Machine Learning (1.3%)
  - Programming (10.6%)
  - Data Engineering (4.8%)
  - Cloud / DevOps (3.5%)
  - Data Analysis (21.6%)
- *HasDigital* composite: 79.8%

Baseline Specification

OLS Wage Regression

$$\ln w_{icft} = \alpha + \sum_k \beta_k S^k_{icft} + \gamma_1 Edu_{it} + \gamma_2 Exp_{it} + \lambda_{ct} + \mu_j + \varepsilon_{icft}$$

- λ_{ct}: **city×year fixed effects** (3,824 groups) absorb local labor-market shocks
- μ_j: **industry FE** (21 categories)
- Standard errors **clustered at firm level** ($G = 57{,}460$)

Estimation: City×year FE absorbed via iterative within-transformation (Mundlak–Frisch–Waugh, 4 rounds) — equivalent to full dummy inclusion but $O(N)$ per round

Extensions

Annual regressions

Estimate $\hat{\beta}_k(t)$ year by year to trace premium dynamics

Interaction model

$$+\sum_{k<k'} \theta_{kk'} S^k S^{k'}$$

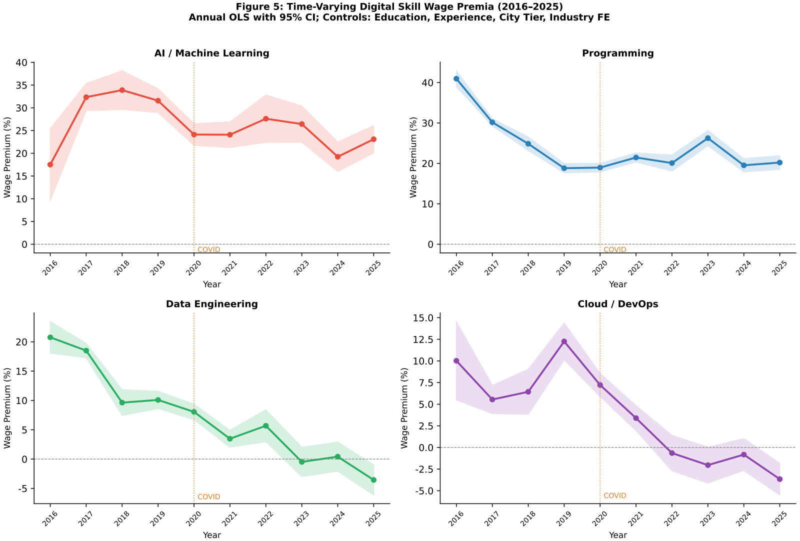
$\theta_{kk'} < 0$: sub-additivity
$\theta_{kk'} > 0$: super-additivity

Quantile regression

Estimate at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ to characterize distributional effects

Outline

Annual Wage Premia, 2016–2025



Key Findings: Premium Dynamics

- **AI/ML — hump-shaped**:
  - $\approx$35% in 2016 $\rightarrow$ peak $\approx$55% in 2018–2019 $\rightarrow$ $\approx$25% by 2022–2023
  - Lagged supply response: AI program graduates entered market $\approx$2020
  - Post-2023 partial recovery: generative AI demand wave
- **Programming — gradual moderation**: 30% $\rightarrow$ 20% (steady supply growth)
- **Data Engineering — U-shaped**: Hadoop era (2016–17) $\rightarrow$ trough $\rightarrow$ stream-processing era (Flink/Kafka post-2021) recovery
- **COVID-19 (2020)**: Temporary dip; premia recover within 2 years

Policy Implication

Cross-sectional studies sampled near a peak vs. trough yield very different conclusions — temporal dynamics matter

Outline

Interaction Model: Universal Sub-Additivity

Pairwise Interaction	$\hat{\theta}_{kk'}$	SE
AI $\times$ Programming	-0.048^{***}	(0.014)
AI $\times$ Data Eng.	-0.106^{***}	(0.016)
Programming $\times$ Data Eng.	-0.229^{***}	(0.009)
Programming $\times$ Cloud	-0.095^{***}	(0.009)
AI $\times$ Cloud	-0.097^{***}	(0.021)

Implied joint premia:

- AI + Programming: additive 61.1% $\rightarrow$ realized 57.1%
- AI + Data Eng.: additive 62.3% $\rightarrow$ realized 46.1%
- Prog. + Data Eng.: additive 55.5% $\rightarrow$ realized 32.7%

Mechanisms:

1. *Task substitutability*: overlapping ETL/query tasks
2. *Employer signaling skepticism*: multi-skill claims discounted

Implication: Depth over Breadth

O-ring theory prediction (Kremer 1993)

Super-additivity: simultaneous competence in multiple domains should multiply output

Our finding: Sub-additivity

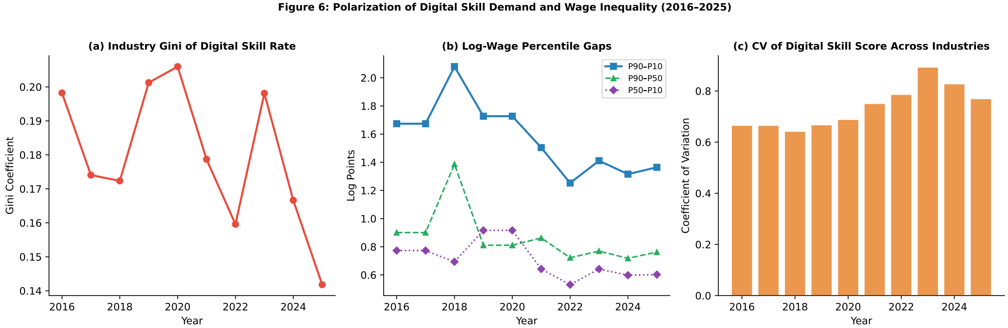
All five pairwise interactions are **negative and significant**
Largest gap: Programming $\times$ Data Engineering (-22.9 pp)

Curriculum design implication:

- Mastering *one* high-premium skill (e.g., AI/ML) dominates spreading effort across multiple domains
- Optimal skill portfolio: concentrated, not diversified

Outline

Diffusion of Digital Skill Demand



(a) Industry Gini of digital-skill demand rate. (b) P90–P10/P90–P50/P50–P10 log-wage gaps. (c) CV of digital-skill score across industries.

The Polarization Paradox

Demand side: diffusion

- Industry Gini fell from 0.20 (2016) to 0.14 (2025) — -30%
- Digital skills spread from tech-intensive sectors to the broader economy

Expected implication:
Broader demand $\Rightarrow$ compressed inequality

Wage side: upper-tail widening

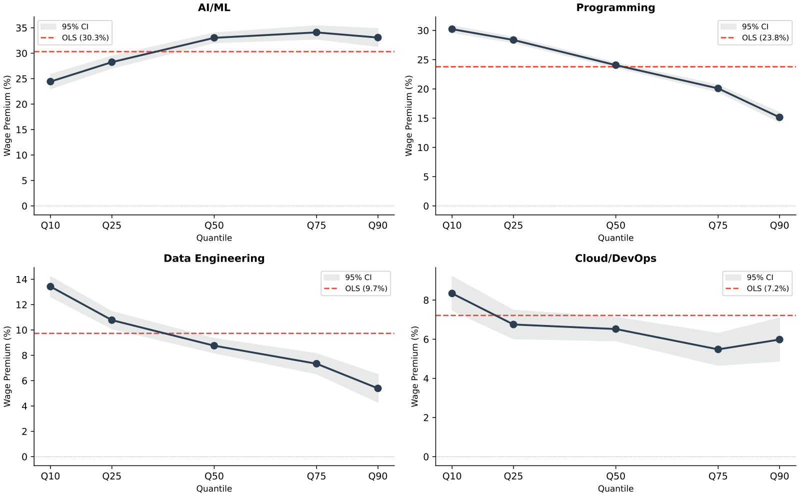
- P90–P10 log-wage gap: 1.67 (2016) $\rightarrow$ peak 2.08 (2018) $\rightarrow$ $>$1.25 (2025)
- P90–P50 gap dominates; P50–P10 gap stable
- Upper-tail polarization, not symmetric spreading

Paradox: diffusion $\neq$ compressed inequality

Diffusion of demand does not guarantee compression of wages when returns are heterogeneous across the distribution

Quantile Regression: Resolving the Paradox

Figure 9: Quantile Regression — Skill Wage Premia Across the Wage Distribution
(Shaded area = 95% CI; dashed line = OLS benchmark)



Heterogeneous Quantile Patterns

AI / ML

- Q10: 25.3% → Q75: 34.8% → Q90: 33.1%
- Inverted-U: premium concentrates in high-tier technical roles
- OLS ≈ median estimate

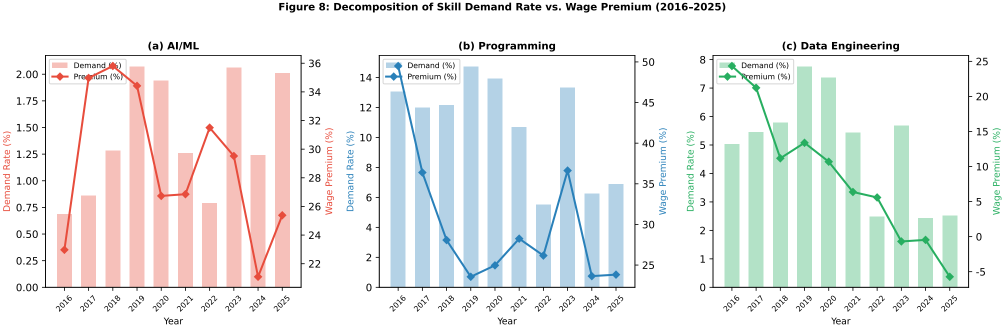
Programming

- Q10: 31.2% → Q90: 16.1% (monotone decline)
- Diffused into mid-to-lower-wage positions
- Scarcity rent at the bottom: rare skill in low-skill context

Explanation

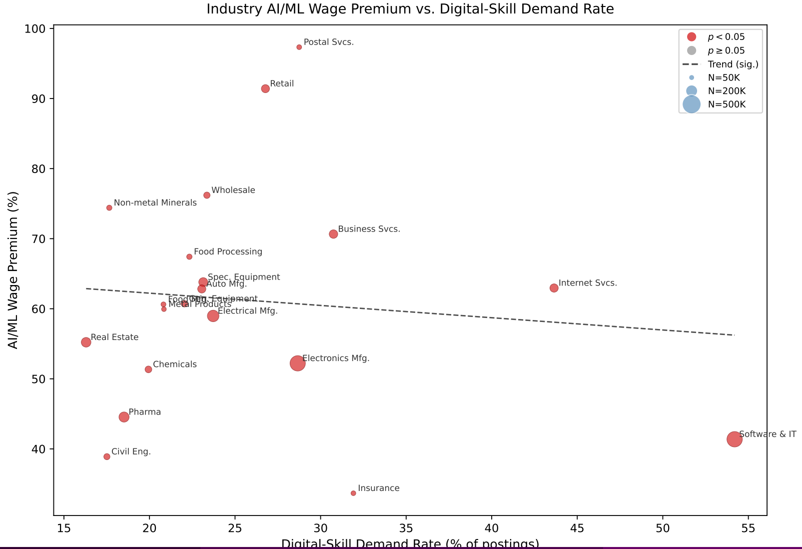
Extensive-margin diffusion of skills co-exists with intensive-margin amplification at specific wage quantiles — both mechanisms operate simultaneously

Demand–Premium Decomposition



Bars (left): annual demand rate. Line (right): conditional wage premium. AI/ML: demand grows monotonically while premium peaks then retreats — classic lagged supply response.

Industry Heterogeneity



Outline

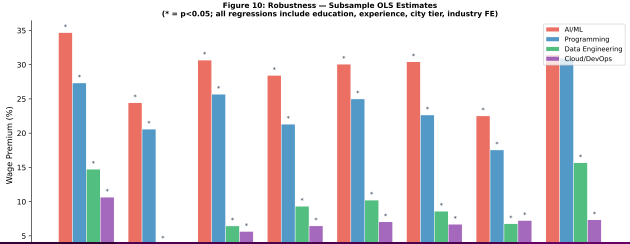
Robustness: Subsample Stability

Subsample	N	AI/ML	Programming	Data Eng.
Full baseline	820,807	26.0%***	24.8%***	9.9%***
Pre-2020 (2016–19)	433,018	29.7%***	27.6%***	13.6%***
Post-2020 (2020–25)	387,789	23.7%***	20.4%***	4.2%***
Manufacturing	378,413	26.7%***	27.1%***	8.9%***
Services	285,803	24.9%***	21.8%***	8.9%***
Tier-1 cities	285,151	26.3%***	25.7%***	10.5%***
Tier-2/3 cities	535,656	25.6%***	23.2%***	8.6%***
High edu ($\geq$ BA)	397,319	19.2%***	18.0%***	6.8%***
Low edu ($<$ BA)	423,488	25.7%***	33.3%***	16.6%***

- All premia positive and significant in every subsample
- Data engineering and cloud compress sharply post-2020; AI and programming persist
- Programming premium *higher* for low-education workers: scarcity rents

Salary Trim	N	AI/ML	Programming	Data Eng.
No trim [0%, 100%]	820,807	26.0%***	24.8%***	9.9%***
[1%, 99%]	804,975	25.8%***	23.8%***	9.9%***
[5%, 95%]	739,391	20.2%***	23.4%***	9.7%***
[10%, 90%]	671,137	14.9%***	21.9%***	9.0%***

- AI premia decline more steeply under trimming: partially driven by high-paying outliers
- Programming and data engineering premia more stable
- OLS lies close to median quantile estimate for AI/ML



Outline

Three Stylized Facts

1. **Time-varying premia**
   AI/ML premium is hump-shaped (2016–2025), consistent with demand shock + lagged supply response. Cross-sectional studies are sensitive to sampling timing.

2. **Sub-additive skill combinations**
   All pairwise digital-skill interactions are negative; largest for Programming $\times$ Data Eng. (-22.9 pp). Invest *deeply* in one high-value skill.

3. **Diffusion $\neq$ equality**
   Industry digital demand diffused (-30% Gini), but upper-tail wage inequality widened. AI/ML premia peak at Q75; programming premia fall from Q10 to Q90.

Policy Implications

(i) Curriculum timing

Higher-education systems adapt with 3–5 year lags. Risk: overproducing workers for skills with already-compressing premia

(ii) Depth over breadth

Sub-additivity supports focused mastery over broad digital literacy in skill-formation policy

(iii) Redistribution alongside diffusion

Broader digital-skill demand does not automatically narrow wages. Complementary redistributive instruments are needed

Thank You!

Xi Xiang
Department of Economics, Nanjing University

Comments welcome