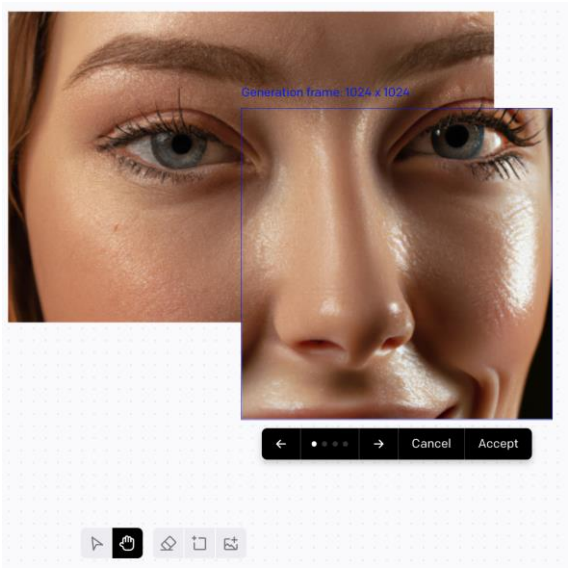
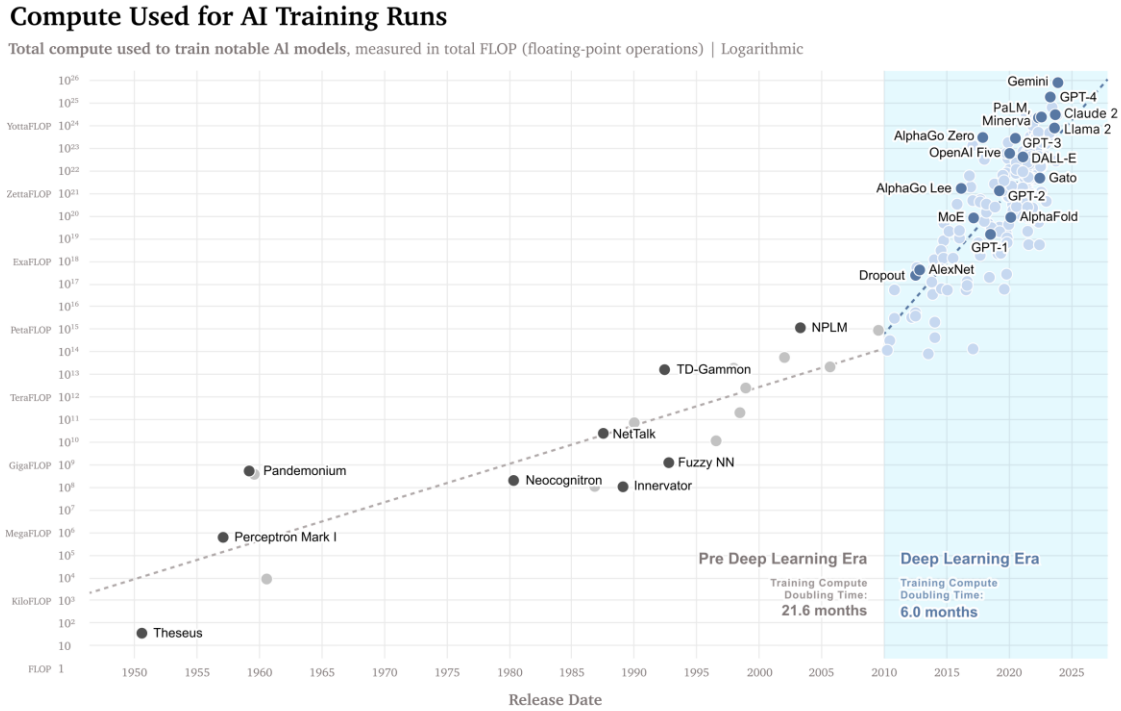
AMEC Symposium – U.S. Productivity Growth: Looking Ahead

Daniel Rock
University of Pennsylvania
rockdi@wharton.upenn.edu

Where we are vs. where we're going



Scaling in compute has hit an accelerated stride



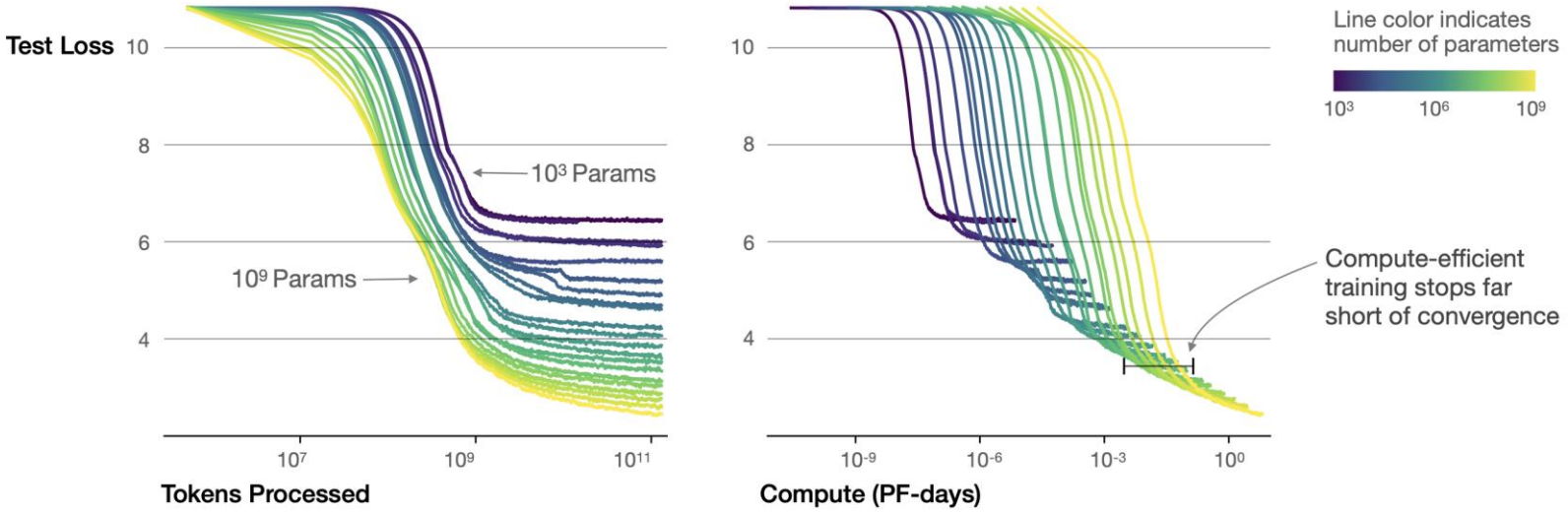
(a) Pre-2010 Trend. *Compute usage for training AI systems before 2010 doubled every 1.8 months. This tracks Moore's Law-esque improvements in compute price-performance (doubling every two years).*

Figure 5: The importance of compute AI in a historical context. *(Data from Epoch (2023) and Sevilla, Heim, A. Ho, et al. (2022).)*

Sastry et al. (2024)

Why have these technologies improved?

- **Scale**
- Taste



Source: (Kaplan et al., 2020)

GPTs are GPTs (w/Tyna Eloundou, Sam Manning, and Pamela Mishkin)

- Developed new rubric for exposure of tasks given different capabilities of LLMs
- New rubric for augmentation vs. automation to LLMs
  - **Generated labels with human contractors and using LLMs themselves**
  - **Methodological Contribution: Use GPT for social science research**
- Validating with public data on where we see LLM use
  - Adoption vs. exposure
- Mapping exposure of new jobs and skills around LLMs to LLM automation
  - Prompt engineer, HITL, etc.
- Exposure across demographics, wages, etc.

- *What hypothesis are we testing?*
  - Are LLMs general-purpose technologies?
    - Pervasive
    - Improve over time
    - Spawn complementary innovations
  - NOT “are the algos going to take all of our jobs?”

General-Purpose Technology Criteria:

- Pervasive
    - Check: Do lots of occupations have exposure?
- Improves over time
    - Check: Developer activity, model improvements, we’re going to take it for granted too
- Spawns complementary innovation
    - Check: Is occupational exposure also contingent on building with other systems?

What did we do?

NONCONFIDENTIAL // EXTERNAL

Task ID	Occupation Title	DWAs	Task Description
14675	Computer Systems Engineers/Architects	Monitor computer system performance to ensure proper operation.	Monitor system operation to detect potential problems.
18310	Acute Care Nurses	Operate diagnostic or therapeutic medical instruments or equipment. Prepare medical supplies or equipment for use.	Set up, operate, or monitor invasive equipment and devices, such as colostomy or tracheotomy equipment, mechanical ventilators, catheters, gastrointestinal tubes, and central lines.
4668.0	Gambling Cage Workers	Execute sales or other financial transactions.	Cash checks and process credit card advances for patrons.
15709	Online Merchants	Execute sales or other financial transactions.	Deliver e-mail confirmation of completed transactions and shipment.
6529	Kindergarten Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.
6568	Elementary School Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.

Table 1: Sample of occupations, tasks, and Detailed Work Activities from the O*NET database. We see that aggregating over activities alone is imprecise, as evidenced by the fact that we'd expect Gambling Cage Workers to complete the given DWA in person, using some physicality while we'd expect Online Merchants to complete the same activity solely with a computer.

NONCONFID

Summary of exposure rubric

No exposure (E0) if:

- using the described LLM results in no or minimal reduction in the time required to complete the activity or task while maintaining equivalent quality[a] or
- using the described LLM results in a decrease in the quality of the activity/task output.

Direct exposure (E1) if:

- using the described LLM via ChatGPT or the OpenAI playground can decrease the time required to complete the DWA or task by at least half (50%).

LLM+ Exposed (E2) if:

- access to the described LLM alone would not reduce the time required to complete the activity/task by at least half, but
- additional software could be developed on top of the LLM that could reduce the time it takes to complete the specific activity/task with quality by at least half. Among these systems, we count access to image generation systems.[b]

---

[a] Equivalent quality means that a third party, typically the recipient of the output, would not notice or care about LLM assistance.

[b] In practice, as can be seen in the full rubric in Appendix A.1, we categorize access to image capabilities separately (E3) to facilitate annotation, though we combine E2 and E3 for all analyses.

Is LLM exposure pervasive?

Occupation Level Exposure

	Human		GPT-4	
	mean	std	mean	std
$E1$	0.14	0.14	0.14	0.16
$E1 + 0.5 * E2$	0.30	0.21	0.34	0.22
$E1 + E2$	0.46	0.30	0.55	0.34

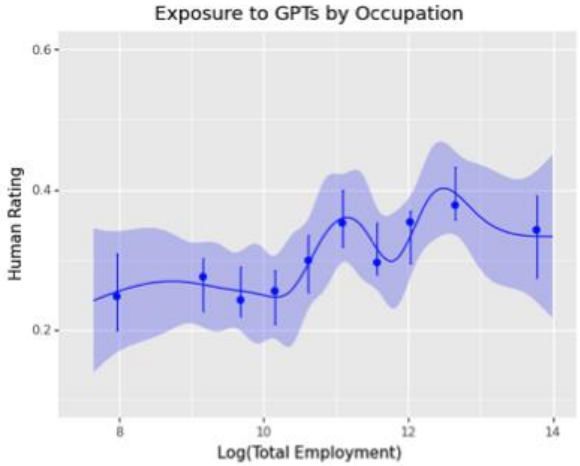
Task Level Exposure

	Human		GPT-4	
	mean	std	mean	std
$E1$	0.15	0.36	0.14	0.35
$E1 + 0.5 * E2$	0.31	0.37	0.35	0.35
$E1 + E2$	0.47	0.50	0.56	0.50

Group	Occupations with highest exposure	% Exposure
Human $E1$	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human $E1 + 0.5 * E2$	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human $E1 + E2$	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	Humans labeled 15 occupations as "fully exposed."	
Model $E1$	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model $E1 + 0.5 * E2$	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model $E1 + E2$	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	The model labeled 86 occupations as "fully exposed."	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6

There are many important caveats to this analysis (internal to the study)

- **Subjective human judgments**: Labelers understand LLM capabilities, but don't know these roles deeply.
- **Measuring LLMs with GPT-4**: Brittle rubrics and arbitrary thresholds. Iteration leads to slightly different results.
- **Validity of the task-based framework**: What is the atomic unit of work? These task lists are one instrument with many imperfections. Some tasks are {up, down}stream of others. Big assumption here that this dataset reflects some organization of work.
- **Lack of human annotator expertise**: Annotators unaware of occupations <> activities. (Collected more labels in some cases)
- **Forward-looking and subject to change**: This is an ongoing effort and the equilibrium is very hard to predict.
- **Disagreement between humans and GPT-4**: Humans and GPT-4 are differentially aware of context. This can change results and makes outputs of the model sensitive to prompting (among other concerns).
- **Saying nothing about social, legal/regulatory, political considerations**: Technical feasibility is only one part of the process.
- **Arbitrary focus on software**: Robots are starting to use LLMs…

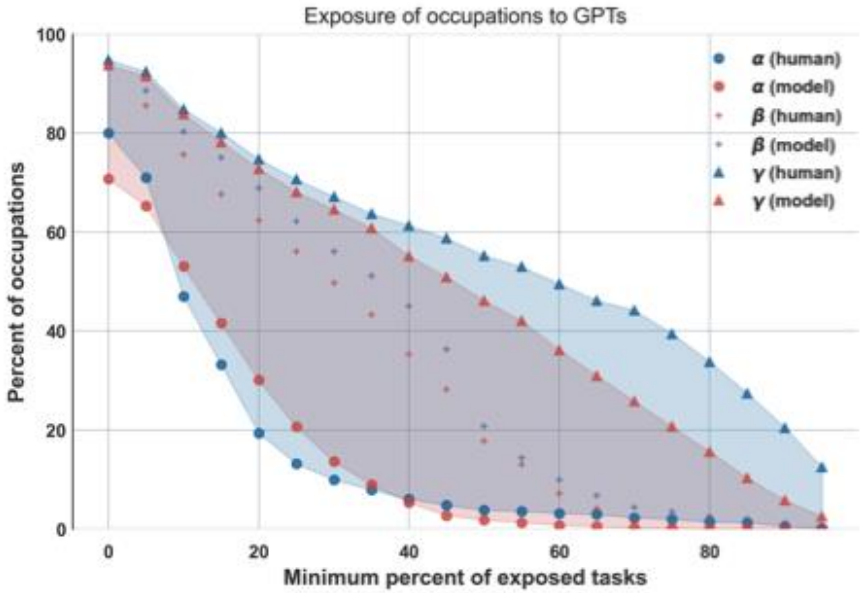

Exposure to GPTs by Occupation
Human Rating
0.6
0.4
0.2
8
10
12
14
Log(Total Employment)

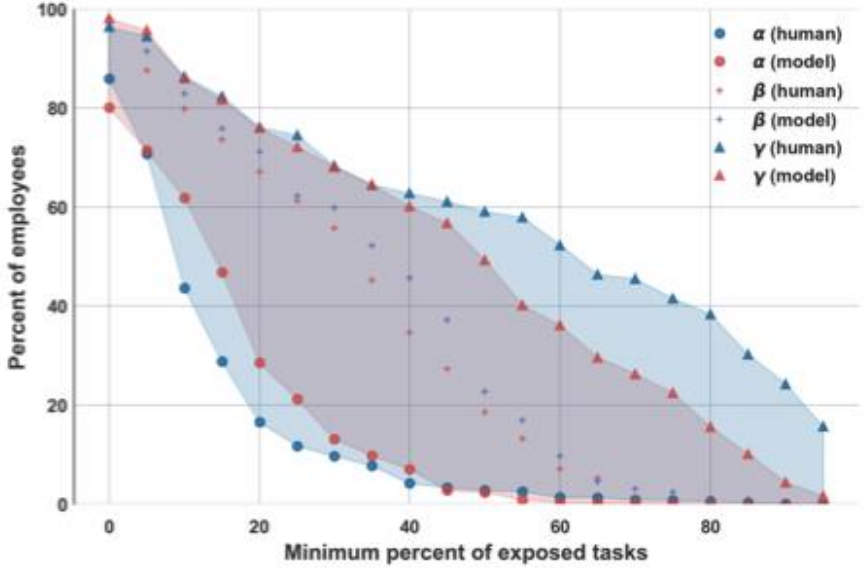

Exposure to GPTs by Occupation
GPT-4 Rating
0.6
0.4
0.2
8
10
12
14
Log(Total Employment)


Exposure to GPTs by Occupation
Human Rating
0.5
0.4
0.3
0.2
0.1
0
10
10.5
11
11.5
12
Log(Annual Wage)


Exposure to GPTs by Occupation
GPT-4 Rating
0.6
0.4
0.2
0
10
10.5
11
11.5
12
Log(Annual Wage)

Is deploying LLMs going to require or generate complementary investment?





More exposed roles typically have greater barriers to entry

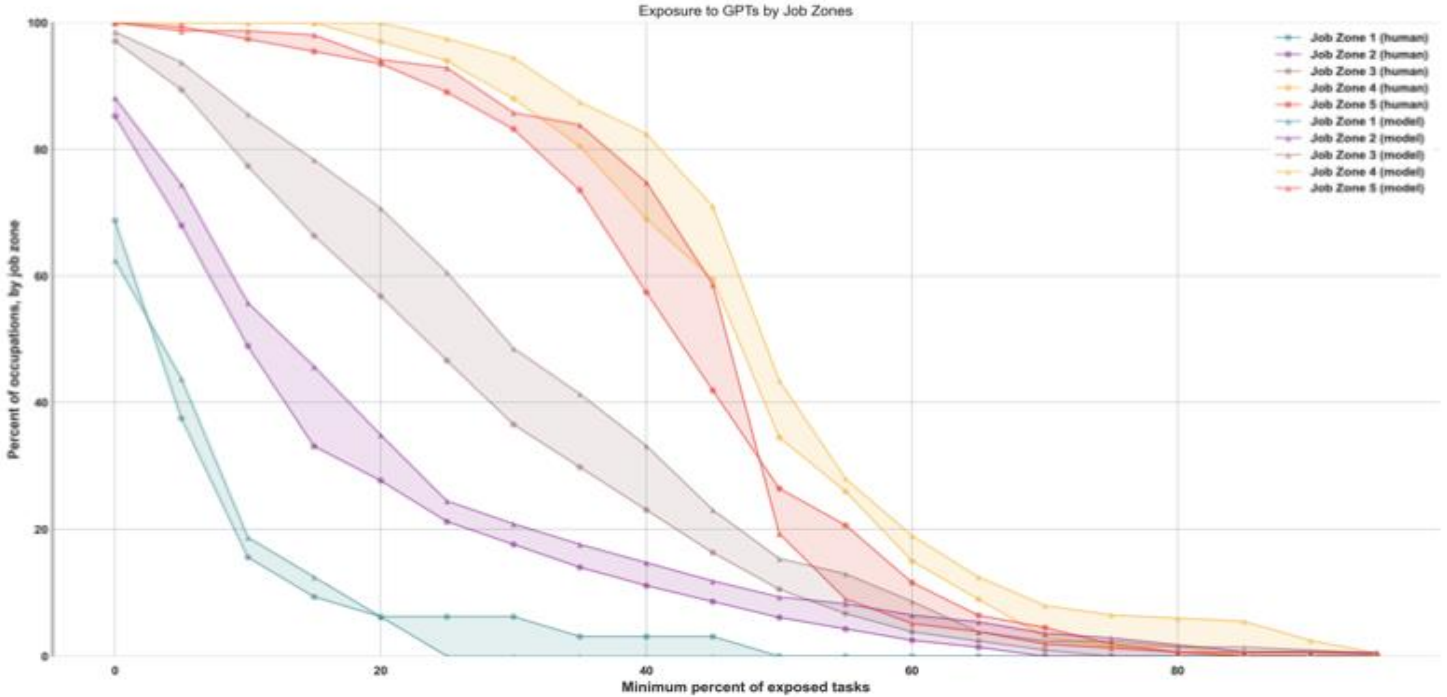
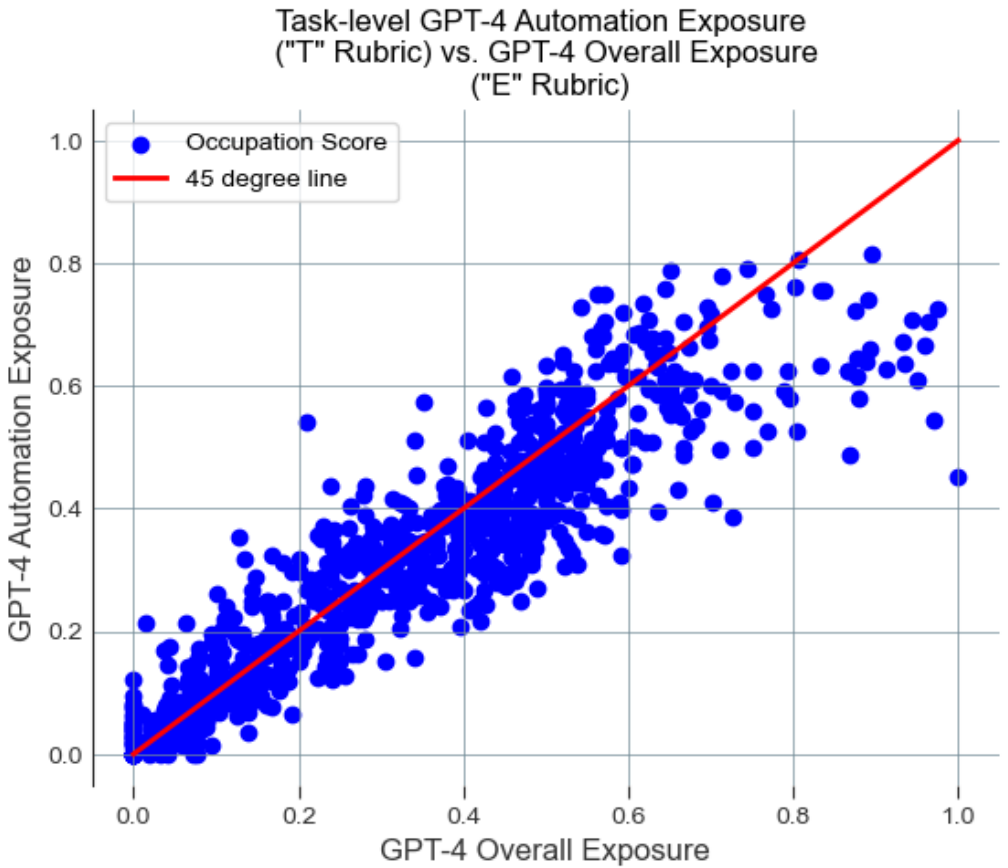


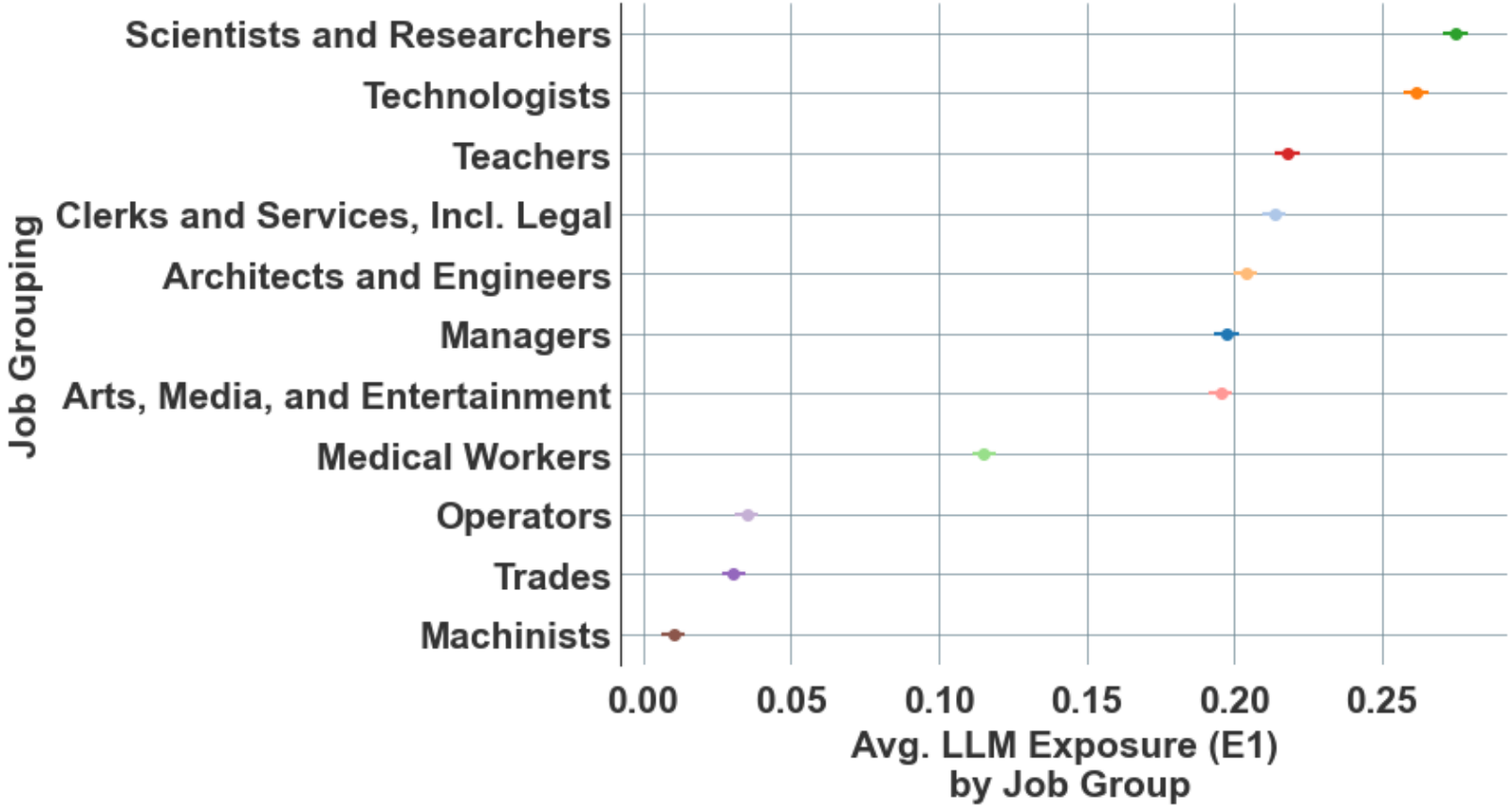
Figure 5: β exposure ratings of occupations in the five Job Zones, which are groups of similar occupations that are classified according to the level of education, experience, and on-the-job training needed to perform them.

Okay so what about automation?

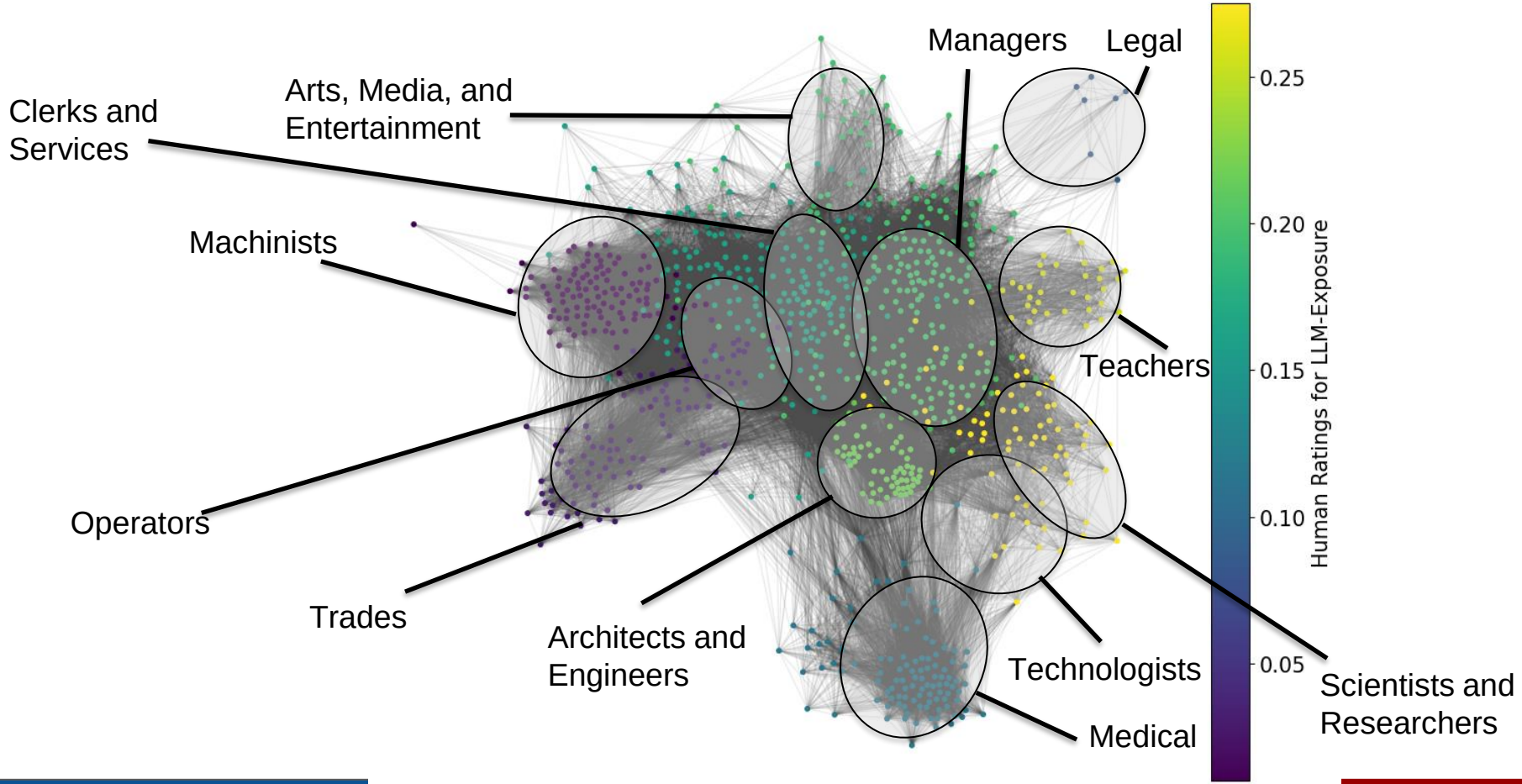
Title
Telephone Operators
Telemarketers
Word Processors and Typists
Credit Authorizers, Checkers, and Clerks
Order Clerks
Bookkeeping, Accounting, and Auditing Clerks
Desktop Publishers
Payroll and Timekeeping Clerks
Insurance Claims and Policy Processing Clerks
Brokerage Clerks
Insurance Underwriters
Travel Agents
Statistical Assistants
Medical Records Specialists
Tellers
Legal Secretaries and Administrative Assistants
Billing and Posting Clerks
Proofreaders and Copy Markers
Medical Transcriptionists
Loan Interviewers and Clerks



Clustering exposure (just to LLMs) by job ”archetypes” shows the pattern



Researchers and developers rank amongst the most exposed groups



What are the key takeaways so far?

- GPTs are GPTs!
  - Pervasive
  - Improving over time
  - Will probably require complementary innovation
  - **Takeaway: The equilibrium for a general-purpose technology is hard to know in advance**.

- But we do know where to look first. This set of scores and methods can help provide some answers.
  - 80% of occupations have around 10% of their tasks exposed.
  - **Takeaway: *Tasks* and *Systems* are the right units of analysis. Locate potential for change!**

- **What we did not find:** AI is coming for all of the jobs. There isn't evidence that's happening.

The Productivity J-Curve (Brynjolfsson, Rock, and Syverson 2021)

How do intangibles affect productivity measurement?

$$Productivity = \frac{Output}{Input}$$

- Intangible capital would be an unmeasured input
  - Will cause productivity to be overstated
- However, intangible capital is also an output (measured as investment flow)
  - Will cause productivity to be understated
- Net effect on productivity measurement depends on relative timing of input vs. output mismeasurement

Intangible Growth Accounting

Standard production function: $Y = AF(K, L)$

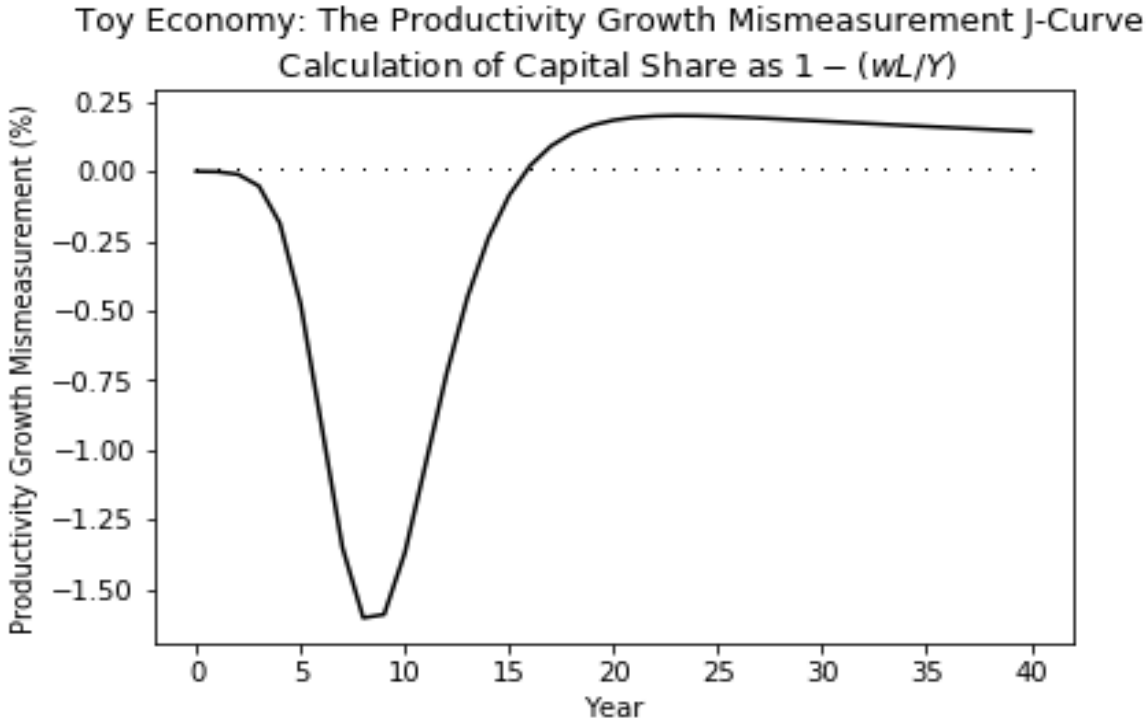
Standard Solow residual TFP: $g_A = g_Y - \left(\frac{rK}{Y}\right) g_K - \left(\frac{wL}{Y}\right) g_L$

Intangible (U)-augmented production: $Y + \phi I_U = A^* F^*(K, U, L)$

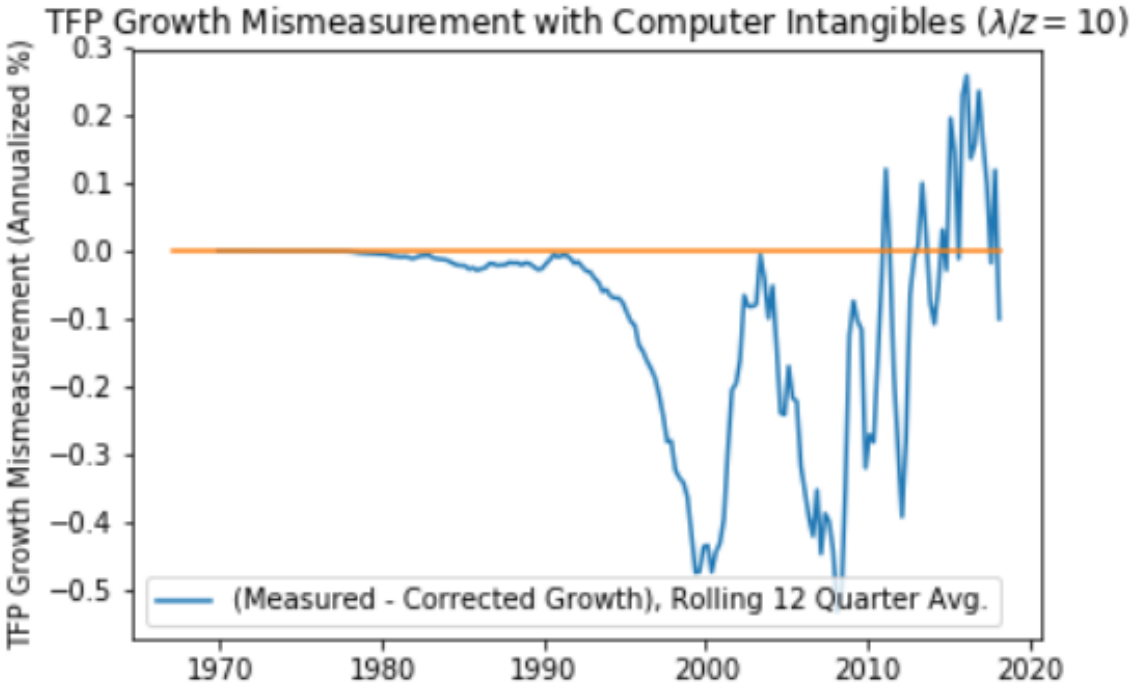
Intangible-augmented TFP growth:

$$g_{A^*} = \left(\frac{Y}{Y + \phi I_U}\right)\left(g_Y - \left(\frac{rK}{Y}\right) g_K - \left(\frac{wL}{Y}\right) g_L - \left(\frac{r_U U}{Y}\right) g_U\right) + \left(\frac{\phi I_U}{Y + \phi I_U}\right) g_{I_U}$$

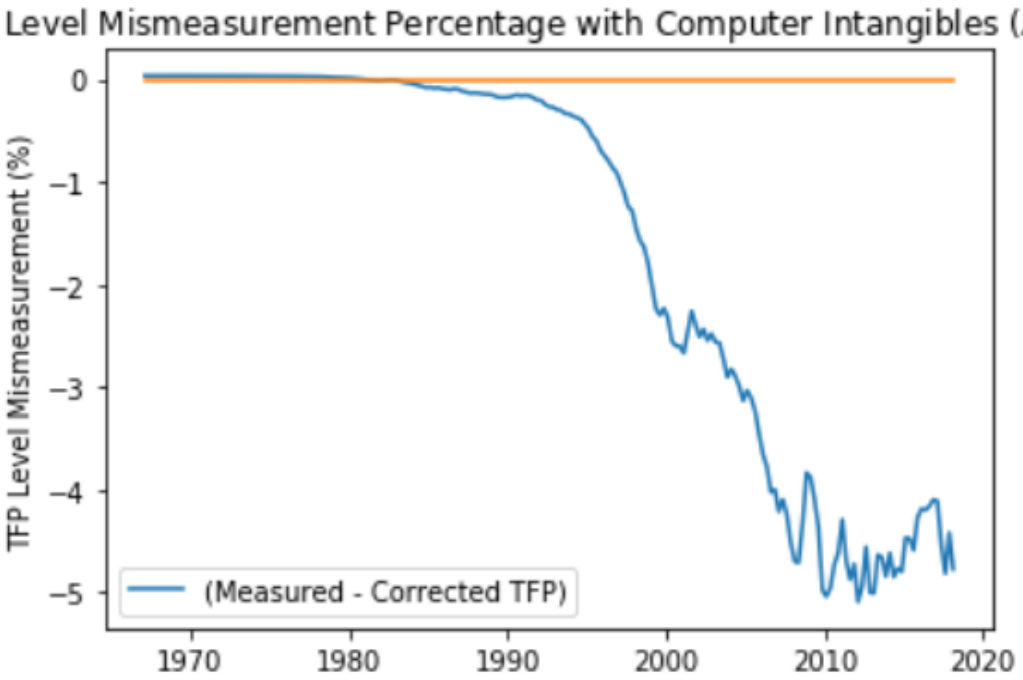
The J-Curve


Toy Economy: The Productivity Growth Mismeasurement J-Curve
Calculation of Capital Share as 1 − (wL/Y)
Productivity Growth Mismeasurement (%)
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
−1.50
0
5
10
15
20
25
30
35
40
Year

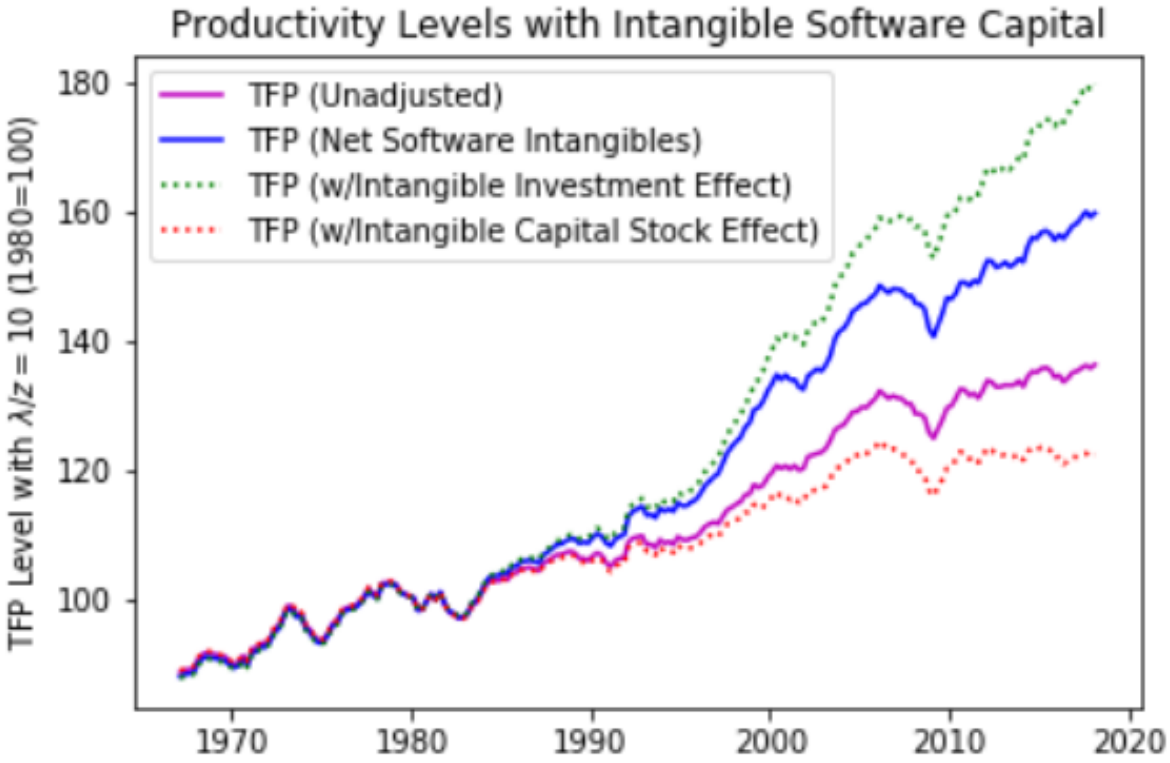
TFP Growth Mismeasurement by Year: IT Hardware



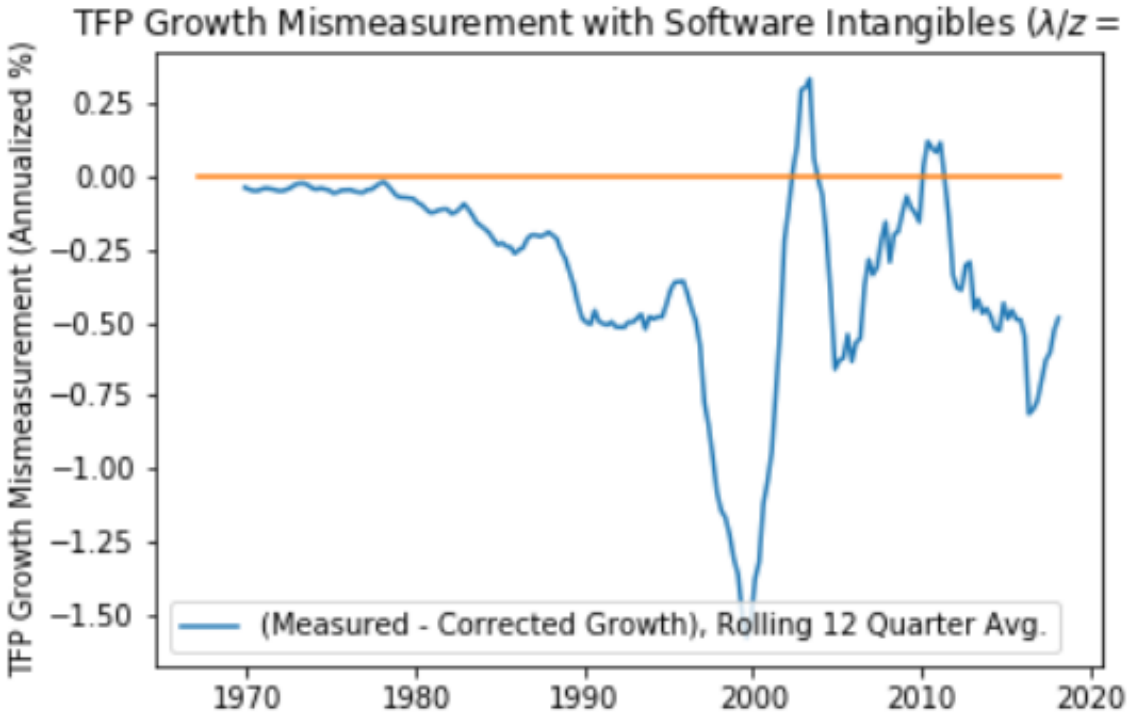
TFP Accumulated Level Mismeasurement: IT Hardware



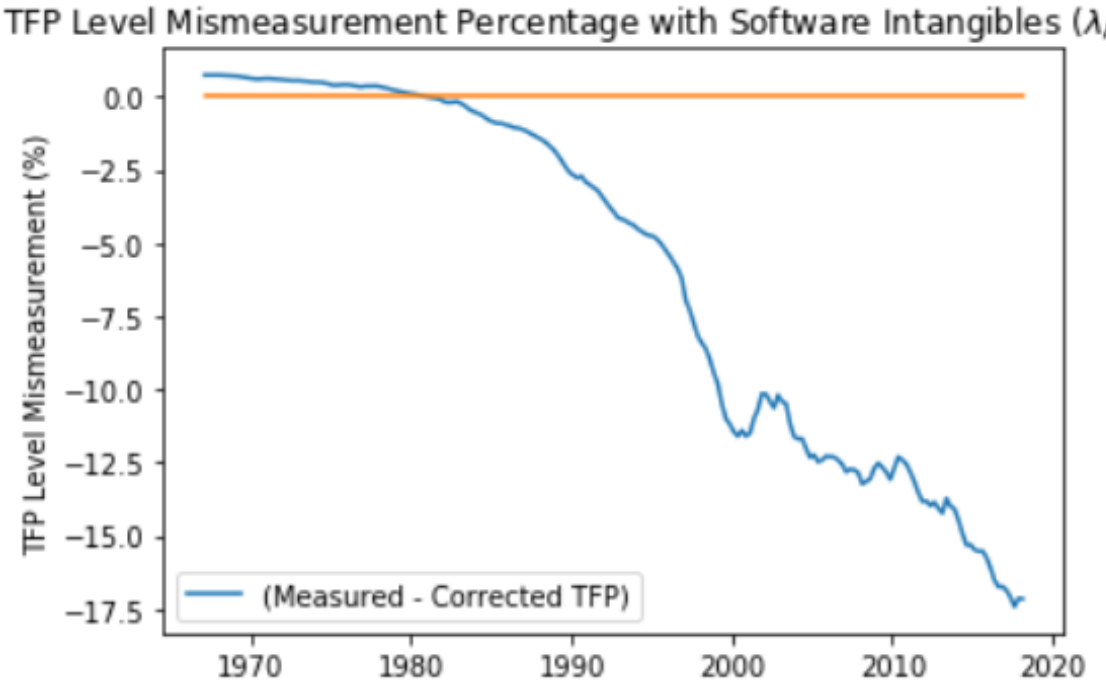
Adjusted TFP Levels: IT Software



TFP Growth Mismeasurement by Year: IT Software



TFP Accumulated Level Mismeasurement: IT Software



Does This Explain the Post-2004 Productivity Slowdown?

No; implied slowdown actually larger

A mismeasurement explanation for the slowdown doesn't require just mismeasurement; it requires a *change* in mismeasurement (in a particular direction and around 2004)

Period	Measured Annual TFP Growth (%)	Implied Annual TFP Growth (%)	Implied – Measured
1995-2004	1.63	2.20	0.57
2005-2017	0.40	0.71	0.31
Slowdown	1.23	1.49	0.26

If AI is a GPT, the full effects may take a long time to play out

- Pervasive, Improving over time, Spawning complementary innovation
- Productivity gains from technologies like this:
    - Require intangible capital (historically up to $10-12 of intangible investment per tangible dollar invested)
    - Gains are not immediate, but some investments are up front
    - May affect productivity measurement in general (i.e. contents of the Solow Residual are different)
- Early advances are promising with potentially fast-changing task structure