PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

PIER

Discussion
Measuring Labor Market Match Quality with Large Language Models by Yi Chen, Hanming Fang, Yi Zhao, and Zibo Zhao

Nada Wasi
Puey Ungphakorn Institute for Economic Research
May 2025

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

Summary

- Propose using large language model (LLMs) to assess job match quality
- Task LLMs (GPT-3.5-turbo) to evaluate whether an applicant fits with a job using two datasets : Chinese online job platform and labor force survey
- The new GPT measure
  - positive correlates with traditional measures
  - can explain wage differentials after controlling for workers and jobs characteristics
- Discuss GPT measure advantages & applications
  - works with small sample sizes
  - simulates how gender disclosure influences LLM assessment
  - does not penalize “match quality” for versatile majors

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

Focus on three traditional measures

Traditional measures		
1. Job Switching (JS)	Dummy = 1 if same occupation Dummy = 1 if same industry	Previous occupation Previous industry
2. Realized Match (RM)	Duncan index: actual distribution of majors within occupation	Major-occupation
3. Job Analyst	Evaluation by job analysts that defines required education or skills for jobs	

Two datasets

	Online job platform Zhaopin.com	China Labor Force Dynamic Survey (CLDS)
	- job applications - applicants' expected wage (or imputed from previous jobs)	- realized job matches - realized wage outcome
Period	Jan-Nov 2013	2016 and 2018 repeated cross-section
No. of applicants	847,801 applications of college+ applicants	2,431 college+, employed respondents
No. of job posting	29,914	N/A
Occupation	58 broad & 588 detailed categories	Standard Chinese classification
Industry	50 categories	
Major	12 broad & 92 detailed categories	
Traditional measures	[1] Same-occupation dummy (detailed?) [2] Same-industry dummy [3] Duncan major-occupation match (detailed?)	[3] Duncan major-occupation match [4] Job Analyst major-occupation match
GPT measures	Occupation-occupation match Industry-industry match Major-title match	Major-title match

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

#1: GPT captures additional information→ work well with detailed categories?

Detailed Occupation Category of Applied Job	Detailed Occupation Category of Current Job	Same-occupation Dummy	GPT Response	GPT Occupation-occupation Match
Software engineer	Software engineer	1	Probably can	1
	System tester	0	Probably can	1
	Sales representative	0	Probably cannot	0

- This example suggests that GPT is more useful for cases with more detailed category, unstructured texts.
- For broad occupation/industry categories or clean data, would the results be similar? but for the first row – shouldn't we expect GPT to answer "Definitely can"?

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

#1: GPT captures additional information→ work well with detailed categories?

- Higher correlation between industry-industry match because the industry categories are cleaner (50 categories) than the occupation categories (588 categories)?

Table 3: Pairwise Correlations between the Traditional and GPT Measures of Match Quality

Panel A: Zhaopin.com	Same-occupation dummy (1)	GPT occupation-occupation match (2)	Same-industry dummy (3)	GPT industry-industry match (4)	Duncan major-occupation match (5)	GPT major-title match (6)
Same-occupation dummy	1					
GPT occupation-occupation match	**0.354*** **	1				
Same-industry dummy	0.130***	0.100***	1			
GPT industry-industry match	0.117***	0.100***	**0.655*** **	1		
Duncan major-occupation match	0.103***	0.103***	0.075***	0.089***	1	
GPT major-title match	0.098***	0.078***	0.081***	0.088***	**0.436*** **	1

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

#2: Can we learn more from GPT answers?

- GPT were tasked to respond with
  "Definitely can" "Probably can" "Probably cannot" or "Definitely cannot"
  The current version codes 1/0 as "match" if "definitely can" or "probably can".

- Would "Definitely can" explain wage differential more than "Probably can"?
  Previous studies consider different levels of match.
  same occupation as previous jobs
  - Kambourov & Manovskii (2009) 3-digit occupation vs. 1- or 2-digit
  major-occupation match
  - Lemieux (2014) direct (engineering → engineer), related (business → managerial role)
  - Altonji et al. (2016) Top 5 occupations in that major

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

#3 Does Duncan index measure "match quality" or occupational segregation?

The paper pointed out 2 drawbacks of the Duncan index (realized major-occupation match)

[1] versatility major (work in different occupations) is penalized.
low Duncan index value, high GPT match score

[2] the proportion estimates become unstable for small sample size (e.g., CLDS)
GPT derive such match from external sources → no small-sample limitation

Questions

- Did the Duncan Index measure the degree of occupational segregation (e.g., by men and women, or by field of study) or measure "match quality"?
  (Lemieux 2014 and Altonji et al. 2016 didn't use it to measure match quality.)
- Small sample size: if we have another large external dataset to estimate the RM major-occupation match, can't we use them to predict match for CLDS?
- The Duncan index was also calculated for the job platform data → not realized match?

PUEY UNGPHAKORN INSTITUTE FOR ECONOMIC RESEARCH

PIER

#4: Any cautions regarding the use of GPT measures?

- When GPT give different answers from traditional measures, is there any case they get it wrong?
- In what dimension GPT-4o or GPT-4.5 can improve over the current GPT 3.5 version?

- Other uses?
  - can we input “major”, “recent industry”, “recent occupation” simultaneously (or the whole CV) and ask GPT to rank the candidates?
  - on the versatility issue, can GPT look at (typical) transcripts? for example, a computer science major graduate may also take a management class
  - did GPT statistically discriminate applicants? women did perform better in certain jobs?