Empowering the Understanding of Labor Market Mismatch with LLMs [*]

Yi Chen[†] Hanming Fang[‡] Yi Zhao[§] Zibo Zhao[¶]

November 19, 2024

Abstract

Labor market mismatch is traditionally difficult to analyze because the complex relationships between categorical variables such as industry, occupation, and major. We propose a novel method leveraging large language models (LLMs) to assess mismatch. Our approach tasks LLMs with simulating a human resources specialist's role in evaluating the fit between an applicant and a job. Our study consists of two parts. First, we validate the GPT approach with an administrative dataset from an online job platform and show that our new measure of match quality correlates positively with traditional measures and can reveal additional insights. We also demonstrate the method's applicability with survey data, where traditional measures are infeasible because of the limited sample size. In the second part, we demonstrate two applications of GPT: one reproduces known knowledge, while the other provides new insights. The first application simulates an audit study to examine how gender disclosure influences LLM assessments, finding that gendered information leads LLMs to favor women for traditionally female roles. The second application show that how GPT contributes to understanding the versatility of majors. We show that majors that can fit a broad range of occupations are unfairly penalized by traditional mismatch measures, and the GPT measure can mitigate this issue.

Keywords: Large Language Models, Categorical Variables, Labor Market Mismatch

JEL Codes: C55, J16, J24, J31

---

[*] This manuscript was previously circulated under the title "Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch" (NBER Working Paper 32327, April 2024). This paper was presented at Shanghai University of Finance and Economics, Jinan University, Shanghai Jiaotong University, the 25th Quarterly Forum of China Labor Economists Forum, the Shanghai Summit Labor Economics Seminar, the 2024 IESR Labor Conference, and the Macroeconomics Workshop of CUHK-Tsinghua Joint Research Center for Chinese Economy. We are grateful for the feedback from all the participants. Chen acknowledges financial support from the National Science Foundation of China (72422019). All remaining errors are our own.

[†] School of Entrepreneurship and Management, ShanghaiTech University. Email: chenyi1@shanghaitech.edu.cn

[‡] Department of Economics, University of Pennsylvania. Email: hanming.fang@econ.upenn.edu

[§] School of Economics and Management, Tsinghua University. Email: zhao-y21@mails.tsinghua.edu.cn

[¶] Faculty of Business and Economics, The University of Hong Kong. Email: zbzhao@hku.hk

1 Introduction

Labor market mismatch, specifically, whether a worker or an application fits a job, is traditionally difficult to measure because of it involves many categorical variables such as college major, industry, and occupation. Categorical variables are distinct from ordinal variables (e.g., self-rated health) or cardinal variables (e.g., temperature, income, or years of schooling) since there is no inherent ordering among these categories.[1] Researchers often overlook the valuable textual information contained in categorical variables. A conventional approach, recommended in classical econometric textbooks for addressing categorical variables, involves constructing a set of dummy variables for different categories, known as the fixed effect (FE) approach.[2]

However, the FE approach has two significant limitations. First, it disregards the information provided by the textual labels associated with categorical variables. For instance, consider three different occupation categories: "software test engineer," "software R&D engineer," and "sales representative." Intuitively, the first two categories are more similar to each other than "sales representative." Nonetheless, researchers typically construct three dummy variables for these categories, failing to capture subtle similarities among them. Another limitation is that analysis (such as computing sample average) based on categories with limited observations can be highly unstable. This limitation is particularly pronounced in survey data, which typically has a relatively small sample size. For example, in the 2016 and 2018 waves of the China Labor-Force Dynamic Survey (CLDS), the occupations of 2,431 employers with a college degree or above are coded in 65 detailed categories. However, half of these detailed categories have less than 10 observations, and approximately 70 percent have fewer than 30 observations. Small number of observations can make the FE approach highly unstable for analyzing jobs at such a granular level.

---

[1] See a detailed discussion on the differences between categorical, ordinal, and cardinal variables in `https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/`.

[2] Take Analysis of Variance (ANOVA) as a concrete example. Let's consider a categorical variable D_i representing race:

$$D_i = \begin{cases} 1 & \text{if White} \\ 2 & \text{if African American} \\ 3 & \text{otherwise} \end{cases}$$

The first step is to run the following regression:

$$y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \varepsilon_i,$$

where $D_{ji} = 1$ if $D_i = j$ (and equals zero otherwise). This approach effectively breaks a categorical variable (D_i) into a set of dummy variables (D_{1i}, D_{2i}, and D_{3i}).

Our research question is: Can large language models (LLMs) overcome the limitations of traditional measures and provide additional insights when exploring labor market mismatch? Existing literature presents four *direct* measures of labor market mismatch.[3] The *job switching* (JS) method examines workers' tendencies to switch jobs, assuming that this results in the loss of occupational and industrial specificity of human capital in the labor market (Kambourov and Manovskii, 2009; Sullivan, 2010). The *realized matches* (RM) method derives the match index from the actual distribution of educational or skill levels within occupations, assuming that workers self select into better-matched positions (Nieto et al., 2015; Sellami et al., 2018). This method requires a relatively large sample size for reliability. The *worker-assessment* (WA) method relies on individuals' personal opinions regarding their job match (Robst, 2007; Zhu, 2014). The *job analysis* (JA) method relies on evaluations by job analysts who define required education or skills for jobs (Guvenen et al., 2020; Lise and Postel-Vinay, 2020). Typically, this method relies on government efforts to recruit job analysts and provide job requirements. For instance, in 2021, the Ministry of Human Resources and Social Security in China hired job analysts to decide matched majors for all occupations listed in China's National Occupation Classification Code (CNOCC). Another example is the O*NET (Occupational Information Network) dataset in the United States.

The recent development of LLMs presents a novel approach to uncovering overlooked information in categorical variables. LLMs are proficient at interpreting and analyzing textual content, allowing for direct examination of the textual labels of categories. We explore the application of LLMs in analyzing labor market mismatch, which involves various categorical variables such as college major, occupation, and industry. Specifically, we employ LLMs simulate human resource (HR) specialists. We task the GPT, a model developed by OpenAI, with evaluating whether workers from a particular major can perform a specific job by inputting the titles of the major and the job into the prompt. We intentionally withhold information about other individual and job characteristics to mitigate potential confounding effects. By processing the textual labels of workers' jobs, the GPT enables us to capture similarities between "software test engineer" and "software R&D engineer." Moreover, since the GPT is pre-trained on extensive external datasets, this approach is not constrained by sample size (i.e., the GPT can assess the similarity between

---

[3] Indirect measurement relies on the model residual method, where "match quality" denotes an unobservable factor specific to the match, which determines productivity after accounting for the observable characteristics of the worker and the job. Examples of this approach include studies by Miller (1984), Moscarini (2001), Nagypál (2007), Hsieh and Klenow (2009), and Adamopoulos et al. (2022).

two categories even if each category has only one observation in the data).

Because of the black-box nature of LLMs, we cannot assume their answers are reasonable without validation. We first cross validate our GPT measure of match quality with various traditional measures and demonstrate its capacity to provide additional information. We utilize two complementary datasets for this purpose. The first dataset comprises application flow data from a leading online job board in China, consisting of 1,048,575 applications to 29,914 unique job postings. This data richness enables us to compute various traditional match quality measures. We investigate three dimensions of the match between workers and jobs: the match between workers' most recent job and applied jobs in terms of occupations and industries (referred to as occupation–occupation and industry–industry matches) and the major–occupation match. Each of these three matches is measured using our GPT method and one of the traditional methods mentioned above. We employ the job switching method for the occupation–occupation and industry–industry matches (i.e., does the position belong to the same occupation or industry as the applicant previously worked in?), and the realized matches method for the major–occupation match (i.e., to what extent does an occupation receive a disproportionately large share of applicants from a specific major?). We document strong and positive correlations between the GPT match quality measures and the traditional measures. In addition, we find that, conditional on the applicants' characteristics, the job characteristics, and the traditional match quality measures for the applicant–job pairs, our GPT measure can still strongly predict an applicant's expected wage.

We then use the CLDS survey data to demonstrate the wide applicability of our GPT method. It's important to note that the information in the CLDS data is not rich enough to implement the traditional match quality measures, in particular the realized matches method, because many categories (occupation or major) contain too few observations. However, the GPT measure remains feasible because this language model has been trained on a vast external textual database. We focus on measuring the most challenging major–occupation match. We find that our GPT measure successfully replicates the basic patterns observed in the job posting data despite the significantly smaller sample size of the CLDS.

In the next step, we demonstrate two concrete applications of our GPT method in labor economics. The first application utilizes the GPT to reproduce an already-known knowledge by simulating a classical audit studies (Goldin and Rouse, 2000), assuming that the training of GPT enables it to "think" like a human. We examine the gap in GPT's responses with two different prompts: one concealing gender information and the other revealing it. We observe that the gender gap in mismatch narrows in those occupations, industries, or majors associated with femininity, once we reveal the gender information to GPT. This provides strong evidence of the existence of gender stereotypes in the data sets that were used to train the GPT. The second application demonstrates that researchers can use GPT to gain novel insights. Specifically, we use GPT to explore the versatility of majors. Consider two extreme examples: The first major equips students with generalized skills applicable to a wide range of occupations. The second major, in contrast, does not prepare students for any specific occupation. Students from both majors may apply for many different occupations, without focusing on any particular one. Traditional mismatch measures cannot distinguish between these two majors, as students' job application behaviors appear similar. However, our GPT measure can easily differentiate them because LLMs are pretrained to "learn" from a vast corpus of textual information beyond the provided data.

This paper contributes to three strands of literature. The first explores the application of LLMs in economic research. The rapid advance of the various LLMs has led to a growing body of economic literature that utilizes them as research or teaching assistants (Cowen and Tabarrok, 2023; Korinek, 2023), natural language processors (Hansen and Kazinnik, 2023; Yang and Menczer, 2023; Lopez-Lira and Tang, 2023), and simulated agents (Argyle et al., 2023; Chen et al., 2023; Eloundou et al., 2023; Horton, 2023). We contribute to this literature by demonstrating the usefulness of GPT in simulating HR specialists to evaluate the suitability of an applicant for a job. Our study is closely related to Eloundou et al. (2023), who assessed the potential effects of LLMs on the U.S. labor market by employing both humans and GPT-4 as annotators to evaluate job tasks' exposure to LLMs. Our paper adds to the various roles that can be assigned to the GPT, specifically focusing on mismatches in the labor market and validates the GPT method in this novel application.

Our study also contributes to the literature on measuring labor market mismatches by proposing a novel method. Unlike the job switching and realized matches methods, our GPT method can recover the overlooked information in categorical variables by considering textual labels. Moreover, since GPT is pre-trained on vast datasets, our method isn't limited by sample size. Unlike the worker-assessment method, ours doesn't require data on self-assessed match quality. Compared to the job analysis method, our approach treats GPT as the job analyst and is thus more cost-effective than employing humans, especially in developing countries. It significantly reduces researchers' workload in processing text data, which is particularly valuable in developing countries where structured data sets on detailed job requirements (such as O*NET) are often unavailable.

The final strand of literature employs traditional textual analysis methods to study labor markets. Many studies utilize the bag-of-words or dictionary method to extract information from job descriptions or titles (Deming and Kahn, 2018; Atalay et al., 2020; Deming and Noray, 2020; Marinescu and Wolthoff, 2020), or to measure similarities and differences between documents using techniques like k-means clustering, word2vec, or TF-IDF (Term Frequency-Inverse Document Frequency) (Biasi and Ma, 2022; Imbert et al., 2022). Our study leverages the capabilities of recently developed LLMs, which can capture contextual nuances, semantic relationships, and diverse language patterns, to explore their application in empirical economic research.

Before proceeding, we want to emphasize that we do not claim that our GPT measure outperforms traditional measures and can replace them. Instead, the GPT measure complements traditional measures. LLMs have their own limitations. For example, it is well-known that LLMs operate as black boxes and have hallucinations (Yao et al., 2023; Zhao et al., 2024), while traditional measures are based on rigorously defined mathematical equations. The black-box and hallucination nature of LLMs makes their results more challenging to interpret. This is also the reason why our first step is to validate the GPT measure with traditional match quality measures.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce the development of LLMs; in Section 3, we provide a detailed description of our data and how to construct traditional and GPT measures; in Section 4, we demonstrate the validity and wide applicability of our GPT method and explain why it can recover the overlooked information in categorical variables; in Section 5, we demonstrate two concrete applications of our GPT method in labor economics: one reproduces existing knowledge, while the other provides new insights; finally, in Section 6 we conclude.

2 What is a Large Language Model?

A language model is an algorithm designed to understand and generate human language by predicting word sequences. The development of language models has advanced substantially in recent years, culminating in the emergence of large language models (LLMs) (Kaplan et al., 2020). LLMs utilize extensive data and parameters, enabling them to excel in comprehending and generating natural language with unparalleled proficiency. For example, while OpenAI hasn't officially disclosed the parameter count in the LLM utilized in our study (GPT-3.5-turbo), its predecessor (GPT-3) boasts 175 billion parameters and is trained on a dataset containing around 500 billion tokens (Brown et al., 2020).[4]

A notable milestone in LLMs is the Generative Pre-trained Transformers Series (GPTs). GPTs employ the powerful Transformer architecture and leverage pre-training on vast multilingual datasets. During pre-training, GPT is exposed to a large corpus of text, such as books, articles, and websites. The model learns to predict the next token in a sentence based on the context of preceding words. For instance, given the prompt "The cat is on the," GPT is trained to predict the most likely next word, like "roof" or "mat." This pre-training process allows GPTs to capture the underlying structure of language, gaining a comprehensive understanding of grammar, semantics, and context. Consequently, GPTs can adeptly grasp intricate linguistic patterns and contextual nuances, thus positioning themselves to excel in various natural language processing tasks, including text generation, translation, and answering questions.

In this study, we choose GPT-3.5-turbo developed by OpenAI over the well-known ChatGPT. The reason is that, while ChatGPT is fine-tuned and tailored specifically for conversational applications, GPT-3.5-turbo demonstrates versatility for general purposes with minimal fine-tuning requirements and comparatively lower inference cost. GPT-3.5-turbo also outperforms LLMs from other companies or institutions developed prior to the release of GPT-4, showcasing exceptional performance across a diverse range of natural language processing tasks.[5,6]

Utilizing LLMs in economic research presents several challenges. First, outcomes may vary across different models and prompts. We will provide a set of robustness check in those dimensions. Second, if the prompt is too simple, LLMs might function as black boxes, potentially limiting the

---

[4] Tokens are typically on a sub-word level. LLMs often employ the "SentencePiece" tokenizer (available at `https://github.com/google/sentencepiece`). For instance, it would tokenize the word "Powerful" into ["power", "ful"].

[5] Zheng et al. (2023) assess all available LLMs across various evaluation tasks and provide one of the most referenced metrics for LLM ranking. GPT-3.5-turbo consistently leads the leaderboard, only surpassed by models developed much later and with higher inference token costs. López Espejel et al. (2023) evaluated GPT-3.5-turbo, GPT-4, and BARD across inference and reasoning tasks. GPT-3.5-turbo achieved the highest score in deductive reasoning tasks without requiring sophisticated prompt engineering, which closely resembles the evaluation of applicant-job matching levels.

[6] We chose GPT-3.5-turbo also for practical reasons, as we needed to evaluate nearly one million applicant–job pairs. Execution costs and time were important considerations. As of November 15, 2024, the GPT-4 model is 60 times more expensive and three times slower than GPT-3.5-turbo. However, we also used a more recent LLM (Claude3-Haiku) as part of our robustness checks.

explainability of results. The prompt should also not be overly complex, as it can lower the chance of encountering similar sequences in the training data and reduce the accuracy of the model-generated conditional probability distribution. We provide a comprehensive discussion in Appendix B on why using complex prompts can also cause issues. Finally, concerns may arise regarding the replicability of results generated by LLMs. To address this, we adopt every feasible method to maximize replicability, including using fixed-version models,[7] fixing random seeds, and setting the model temperature to the lowest level to minimize randomness in the answers provided by LLMs.

3 Data and Measures of Match Quality

3.1 Data

To validate our new GPT measure, we require a comprehensive dataset that allows the construction of traditional match quality measures. Our dataset comprises 1,048,575 applications to 29,914 unique job postings on Zhaopin.com in 2013. Zhaopin.com, the third largest online job board in China at that time, predominantly caters to young, highly educated workers seeking well-paid jobs in the private sector (Kuhn and Shen, 2013). The 29,914 job postings were sampled as follows. In collaboration with the job board, we initially selected a random sample of 61,674 job seekers who initiated new job search cycles in August 2013.[8] We then tracked all their applications until November 30, 2013. Subsequently, from the pool of job postings they applied for, we sampled 29,914 unique postings (about 10%) and collected all applications submitted to those postings from January 1, 2013, to November 30, 2013. This application flow data contains a total of 1,048,575 applications from 693,748 applicants.

Each observation in our data is an applicant–posting pair and includes detailed information on both the job postings and the corresponding applicants. For each job posting, we know the job title, industry category, broad and detailed occupation category, the offered monthly wage range (if available),[9] job location at the city level, number of persons to hire, education and experience

---

[7] Specifically, we use three fixed-version models: GPT-3.5-turbo-0125 (by OpenAI), ERNIE-Bot-4.0 (by Baidu), and Claude3-Haiku (by Anthropic).

[8] Initiating a new job search cycle implies no job search activity on the platform in the past 30 days.

[9] Only 28% of applications contain information on the offered monthly wage range. This is common in online job posting data. For example, the dataset used in Kuhn and Shen (2013) comprises a total of 1,051,706 job ads, of which only 16% provide wage information.

requirements, and the size and ownership type of the hiring firms. Applicants' characteristics include demographic information such as gender and age, educational background (education level and major), marital status, employment status, working experience, and the current city of residence. Additionally, applicants report their most recent job, including the monthly wage range, industry category, as well as the broad and detailed occupation category. They also report their expectations about the next job, including the preferred job location and the expected monthly wage range. We infer the monthly expected wage as the midpoint between the minimum and maximum of the range.[10]

The job postings in our dataset are classified into 50 industry categories and 588 detailed occupation categories nested within 58 broad occupation categories. The classification of industry and occupation categories is parallel; each industry category encompasses various occupation categories, and conversely, each occupation category spans different industries. For example, as illustrated in Appendix Table A1, four related positions—"software test engineer," "game tester," "software R&D engineer," and "video algorithm engineer"—are grouped into two detailed occupation categories ("software test engineer" and "software R&D engineer") within the broad occupation category of "software personnel/internet developer/system integration staff." These positions fall within two industry categories: "computer software" and "internet business/E-commerce." These industries also employ workers in occupation categories such as "accountant" and "administrative officer/administrative assistant."

In our analysis, we focus on applicants who hold a college degree or above because those without such qualifications typically lack a major or have majors incomparable to college graduates. Additionally, applicants without any information about their major are excluded from our study. The final sample consists of 847,801 applications, comprising 80.85% of the initial sample. Furthermore, majors are categorized into 92 detailed categories nested within 12 broad categories, based on "The Undergraduate Major Catalogue of Higher Institutions (2012)" published by China's Ministry of Education.

We complement our Zhaopin.com data with the 2016 and 2018 waves of the China Labor-Force Dynamic Survey (CLDS), a national longitudinal social survey targeting the labor force in

---

[10] Approximately 20.31% of the applications lack information on the monthly wage of the expected job. We impute this information using the monthly wage from the applicant's most recent job. Our results are robust if we exclude those samples.

China.[11] The CLDS data includes 37,623 respondents, of whom 2,431 are employed and hold a college degree or above with major information.[12]

The CLDS data complements the Zhaopin.com data in three distinct ways. First, the CLDS data covers the entire labor market, while Zhaopin.com focuses exclusively on the online job market. Second, the CLDS data, like other survey data, presents realized job matches and corresponding wage outcomes, whereas the application submission data from Zhaopin.com centers on the search process and the expected wages before actual matches. Third, while the Zhaopin.com data allows us to construct traditional match quality measures using job switching and realized matches methods, it does not support the job analysis method because Zhaopin.com uses a different occupation classification from the official one. Consequently, we cannot directly utilize the recommended major–occupation correspondence provided by the Ministry of Human Resources and Social Security in China. The CLDS data adheres to the official classification systems and enables a comparison between the job analysis method and our GPT method.

3.2 Traditional Measures of Match Quality

We introduce three traditional match quality measures constructed using the Zhaopin.com data and the CLDS data: the job switching (JS) method, the realized matches (RM) method, and the job analysis (JA) method. Unfortunately, we cannot use the worker-assessment method employed by Zhu (2014) and Jiang and Guo (2022) because our data lacks information on workers' own feelings of fitness for the jobs.

We first examine the occupation–occupation and industry–industry matches using the JS method to understand the importance of occupational and industrial specificity of human capital in the labor market (Kambourov and Manovskii, 2009; Sullivan, 2010). Following Kambourov and Manovskii (2009), we adopt the JS method to construct measures of match qualities. If a job seeker applies for a job within the same occupation/industry category as their most recent job, they are considered "matched" with the applied job. This rationale stems from the idea that experience in a specific occupation/industry category helps individuals handle jobs in the same category. Specifically, the measure of occupation–occupation (industry–industry, respectively) is

---

[11] CLDS utilizes a rotating sample design where a proportion of the initial survey sample is dropped from each subsequent interview sweep and replaced with a fresh sample of equal size. As a result, the respondents in the 2016 and 2018 waves are different.

[12] To maintain consistency between the two datasets, we excluded individuals who are self-employed.

defined as a dummy variable equal to one if the last job and the applied job belong to the same detailed occupation category (industry category, respectively). However, this method overlooks category labels' information and fails to consider subtle similarities between different categories.

We then examine the major–occupation match using two different methods and datasets. With the Zhaopin.com data, we define the mismatch index using the RM method, which derives matched majors from the actual distribution of majors within occupations, following Altonji et al. (2016). Intuitively, if an occupation receives applications disproportionately from one specific major, the RM method considers this major–occupation pair a good match. Building upon the Duncan segregation index (Duncan and Duncan, 1955) and following Lemieux (2014),[13] we define the Duncan major–occupation match index for detailed major category m and detailed occupation category o as:

$$\text{Duncan match}_{m,o} = \text{Milliles}\left(\theta_{m,o} - \theta_m\right), \tag{1}$$

where $\theta_{m,o}$ is the fraction of applicants from major category m in all applications to occupation category o, and θ_m represents the fraction of major category m in the entire sample. "Milliles" is a function that divides the ratio difference into 1,000 quantiles and further scales it from 0 to 1.[14]

However, the RM method has one significant limitation: it requires a large sample size. This is because if the sample size is small, the computed ratio ($\theta_{m,o}$ and θ_m) may not be reliable. Thus, the RM method is not applicable to most survey data, including CLDS. For the same reason, even though our Zhaopin.com data has over 0.8 million observations, we cannot delve deeper beyond the detailed occupation category into the job title, thus missing important information contained in job titles, as highlighted by Marinescu and Wolthoff (2020). Later in this section, we will demonstrate how our GPT measure overcomes this limitation.

Using the CLDS data, we define an alternative major–occupation match measuring using the JA method, which relies on assessments by job analysts who determine appropriate majors for jobs. In 2021, the Ministry of Human Resources and Social Security in China employed job analysts to establish matched majors for all occupations listed in the official occupation classification

---

[13] Using data from Canada, Lemieux (2014) demonstrated the method's validity through a strong correlation between the Duncan index and workers' self-reported relationship between major and occupation.

[14] Appendix Table A2 presents examples of two occupation categories ("tour consultant" and "mechanical designer") and two major categories ("mechanical" and "tourism management"). These pairs, "mechanical–mechanical designer" and "tourism management–tour consultant," are intuitively recognized as good matches. Our match indexes align with this recognition, revealing significantly larger ratio differences and match indexes for these pairs compared to others.

(CNOCC).[15] Jobs in the CLDS data are categorized according to the CNOCC, allowing us to utilize the JA method. Specifically, the major–occupation match measure, based on the JA method, indicates whether a worker's major aligns with the recommended majors for their occupation as suggested by the job analysts. For instance, for the occupation "human resources professionals," there are three relevant majors: "human resource management," "business management," and "administrative management." Therefore, workers with one of these three majors are considered to be matched with the "human resources professionals" occupation, while those other majors are not considered a match. This is referred to as the JA major–occupation match.

3.3 GPT Measure of Match Quality

In this section, we offer a detailed explanation of how we construct our GPT measure of match quality. The GPT method closely resembles the JA method, but instead of relying on real-world experts, it utilizes GPT as the job analyst. We designate GPT-3.5-turbo to simulate a human resource specialist for assessing the compatibility between workers and jobs. For instance, in assessing the major–occupation match, we employ the following prompt:

> Pretend that you are an HR specialist. Based solely on the provided information (without considering any additional information or assumptions such as education level, working experience, previous jobs, on-the-job learning, or training), please assess whether the applicant graduated from [**major title**] is capable of performing the [**job title**]. Please respond with "Definitely can" or "Probably can" or "Probably cannot" or "Definitely cannot."

It's worth noting that we can extend our analysis beyond occupation categories to job titles because GPT can generate responses for any pair of applicant–posting pairs, regardless of the number of observations. Figure 1 provides an example of GPT responses when assessing match quality between the "management" major and a job titled "sales of automotive parts." We define the GPT measure of the major–occupation match as an indicator of whether the response is "definitely can" or "probably can."[16] We use similar prompts to develop measures for the occupation–occupation

---

[15] See the comprehensive correspondence table between majors and occupations at `http://www.mohrss.gov.cn/SYrlzyhshbzb/zcfg/SYzhengqiuyijian/202106/W020210622638208743832.pdf`.

[16] We utilize a dummy variable instead of an ordered variable to ease comparisons with traditional measures used in our study and existing literature. In a robustness check available upon request from the authors, we also employ an ordered variable, yielding similar results.

and industry–industry matches by inputting the occupation and industry category labels of the most recent job and applied job, respectively. These measures are respectively denoted as GPT major–title match, GPT occupation–occupation match, and GPT industry–industry match to distinguish them from traditional measures.

We highlight two significant advantages of the GPT method. First, it allows us to capture often overlooked information from category labels in traditional measures. Compared to the job switching method, the GPT method excels in identifying similarities between different occupation/industry categories by leveraging category label information. As demonstrated in the first block of Table 1, when evaluating applied jobs within the “software test engineer” occupation category, the same-occupation dummy only considers workers with their most recent job in the same occupation category as a good match. In contrast, GPT recognizes that individuals from the “software R&D engineer” category probably can also perform jobs in the “software test engineer” category, whereas those from the “sales representative” category probably cannot. Applying the same logic, the second block of Table 1 illustrates that GPT can identify that the “computer software” industry is more akin to “IT services” than the “computer hardware” industry.

Second, the GPT method is applicable to any category, regardless of the number of observations in that category. The RM method for measuring the major–occupation match requires a large amount of data. To generate the Duncan index in equation (1), we need to compute the extra proportion of major–occupation pairs within a major ($\theta_{m,o} - \theta_m$). This formula requires a large sample size to make the computation reliable. In contrast, because GPT utilizes external textual information during the training phase, it can evaluate the compatibility between a major and an occupation even if they appear only once in the data. This feature of no-category-size requirement significantly broadens the scope of the GPT method, making it applicable to survey data with relatively small sizes, such as the CLDS data.

3.4 Descriptive Statistics

We present the descriptive statistics in Table 2. Consistent with Kuhn and Shen (2013), applicants in Zhaopin.com data are generally young and well-educated. The applicants are on average 27 years old and have 5.7 years of work experience. Over forty percent of the applicants hold a bachelor’s degree or higher, half are female, and over seventy percent are single and unemployed. The average expected wage is 4,709 RMB per month, slightly higher than the average wage of their most recent job, which is is 4,457 RMB per month. Workers in the CLDS data are relatively older, with more work experience, higher monthly wage of current job, and a higher likelihood of being married.

Panel B of Table 2 displays the summary statistics of the match quality measures constructed from different methods and datasets. Specifically, 22% and 26% of applicants in the Zhaopin.com data applied for a job in the same occupation and industry category, respectively, as their most recent job, indicating a good match according to the JS method. Our GPT measures show a higher incidence of match as the GPT captures similarities between different occupation (industry) categories: 69% (48%, respectively) of them "probably" or "definitely" can perform the applied jobs in an occupation (industry, respectively) category. Regarding the major–occupation match, the GPT measure indicates that in the Zhaopin.com data about 54% of the applicants have a major matched with the job they applied, and 47% of the workers in the CLDS data have a major matched with their current job. In comparison, the traditional JA measure indicates that a similar 35% of the workers have a major matched with their their occupation in the CLDS data.[17]

4 Using the GPT to Construct Match Quality

4.1 Validating the GPT Method

We validate our GPT method from two perspectives. First, if the method is valid, the GPT measures should show positive correlations with traditional measures. Second, a testable implication is that if someone is considered "matched" to a position, they should expect a higher wage compared to others with similar characteristics. Therefore, if we run a Mincer regression of expected wages, we expect the coefficients for the match measures to be significantly positive after controlling for personal and job characteristics.[18]

Table 3 displays the pairwise correlations between the traditional and the GPT measures. Panel A shows the results for Zhaopin.com data. Two notable findings emerge: (1) All correlations

---

[17] Note that the incidence of major–occupation match is lower than that reported in the previous literature using the JA method (e.g., 77–83% in Nordin et al. (2010) and 55–70% in Domadenik et al. (2013)). The primary reason for this difference is that those studies usually examine matches between broader major categories and occupation categories, leading to a higher rate of matches (Sellami et al., 2018). For example, Nordin et al. (2010) consider only 38 occupation categories, in contrast to 223 occupations in the full sample of the CLDS data.

[18] For example, Sullivan (2010) find that workers in certain occupations experience a 14% (23%) increase in wages after five years of occupation (industry) specific experience. The major–occupation match has been associated with a wage premium of about 10% in the U.S. (Robst, 2007) and approximately 1% in China (Zhu, 2014).

are significantly positive; (2) The highest correlations are observed for corresponding traditional and GPT measures. For instance, the correlation between the traditional and the GPT measure of occupation–occupation match is 0.354, surpassing any pairwise correlation involving any of the two measures. Similarly, we find a stronger correlation between the traditional and the GPT measures of industry–industry (major–occupation, respectively) match, reaching as high as 0.655 (0.436, respectively).

In the next step, we augment the standard Mincer wage regressions with match quality measures to further validate our GPT method—a standard procedure to evaluate the effectiveness of traditional match quality measures (e.g., Perry et al. (2014); Guvenen et al. (2020)). We set the expected wage equation for applicant i who applies to job j in city c as follows:

$$\ln w_{i,j,c} = \beta_1 \text{SD\_Match}_{i,j} + X'_{i,j,c}\gamma + \varepsilon_{i,j,c}, \tag{2}$$

where $w_{i,j,c}$ is the expected monthly wage of the job. We standardize all match quality measures to have a mean of zero and a standard deviation of one ($\text{SD\_Match}_{i,j}$). This facilitates the comparison of coefficient estimates when using different match quality measures. $X_{i,j,c}$ incorporates applicant characteristics and job characteristics; $\varepsilon_{i,j,c}$ is the error term. Applicant characteristics consist of years of schooling, working experience (including the square term), gender, birth month dummies, type of school,[19] marital status, employment status, and detailed major FEs. Job characteristics encompass education and experience requirements, firm ownership type,[20] firm scale,[21] detailed occupation FEs, industry FEs, and city FEs of the applied job. For each dimension, we conduct three regressions. The first two regressions include the traditional measure and the GPT measure separately, while the third integrates both measures. Standard errors are clustered at the level of detailed occupation category of the applied jobs.

We present the results in Table 4. Column (1) shows that a one–SD increase in the occupation–

---

[19] We categorized schools into three groups based on the government's classification of elite universities—known as "Project 985" and "Project 211" (the last group of universities don't belong to the two programs). "Project 985," initiated in 1998, aims to establish a group of world-class universities in China. Under this initiative, significant resources were allocated to 39 universities to improve their infrastructure, faculty quality, and research capabilities. "Project 211," launched in 1995, aims to strengthen approximately 100 key universities and disciplines nationwide. It is important to note that all universities classified under "Project 985" are also designated as "Project 211" universities.

[20] There are 9 ownership types, including state-owned enterprises (SOEs) and private firms.

[21] The firm scale (number of employees) is categorized as follows: less than 20, 20-99, 100-499, 500-999, 1000-9999, and 10,000 and above.

occupation match, indicated by the same-occupation dummy, is associated with a wage increase of 1.0 percent. In comparison, our GPT occupation–occupation match indicator suggests a larger effect, increasing the wage by 1.7 percent for a one–SD increase in the indicator (column 2). For the industry–industry match, the coefficient estimate of the traditional measure is close to that for the GPT measure, as shown in columns (4) and (5). The estimated wage effect is approximately equivalent to a 2 percent increase in wages for a one–SD increase in match measures. Columns (3) and (6) show that our GPT measure is positively associated with expected wages, conditional on traditional measures using the job switching method.

Regarding the major–occupation match, column (7) of Table 4 indicates that the coefficient for the Duncan major–occupation match index implies a 0.8 percent wage increase with a one-SD increase in the match index. The GPT major-title match indicator suggests a smaller coefficient (about 0.5 percent for a one-SD increase in the match measure) for the major–occupation match (column 8), and the coefficient becomes statistically insignificant when controlling for the Duncan major–occupation match index (column 9). Taken together, columns (1)–(9) show that regarding the predictive power of applicants' expected wages, our GPT measures perform as well as or better than simple traditional measures, such as zero-or-one dummies indicating the same industry (occupation) category. However, the GPT measures do not outperform the sophisticated data-intensive measure using the realized matching method, indicating the current limitations of LLM at this stage. This limitation does not negate the usefulness of the GPT method. As previously explained, using the RM method requires a large dataset. Later in this paper, we demonstrate that the RM method performs poorly with the CLDS data, which has significantly fewer observations. In contrast, the GPT measures perform well even with a small sample.

4.2 Recovering the Overlooked Information in Categorical Variables

We then investigate the specific sources of additional information provided by the GPT measure. The traditional JS measure for the occupation–occupation (industry–industry, respectively) match is the same-occupation (industry, respectively) dummy. Therefore, the traditional JS measure cannot further distinguish occupations or industries once they don't belong to the same category. This explains why, when conditional on applying for jobs in different occupation/industry categories, the coefficients of the JS measure cannot be estimated, as illustrated in columns (1) and (3) in Table 5. However, our GPT measures still demonstrate statistically significant positive effects on the wage. A one-SD increase in the GPT occupation–occupation (industry–industry, respectively) match indicators is associated with 1.6% (0.8%, respectively) increases in expected wage. Furthermore, the magnitude of the coefficient for the GPT measure barely changes even when all other match quality measures are controlled, as indicated in columns (2) and (4). These findings suggest that the additional information contributed by GPT measure is orthogonal to other variations.

The results in Table 5 echo our discussions on the relative advantage of the GPT method over traditional methods in Section 3.3. When considering the occupation–occupation (industry–industry, respectively) match, the GPT measure captures label information associated with different occupation (industry, respectively) categories that may have been overlooked. Appendix C demonstrates that traditional natural language processing methods, such as bag-of-words (BoW) and term frequency–inverse document frequency (TF-IDF), perform poorly in establishing correlations among different categories. These methods rely on exact textual overlaps between phrases, making them more suitable for "large" textual data, such as detailed job or patent descriptions. In contrast, the GPT method performs well with small textual data because it leverages extensive pretraining on text data from the internet to establish correlations between short phrases.

4.3 Robustness Checks with Different Prompts and LLMs

We perform three sets of robustness checks regarding the prompt design and the choice of LLM in Appendix B. First, our baseline prompt asks GPT to simulate the role of an HR specialist. An alternative approach is to simulate a career advisor and evaluate job fitness from the job seekers' perspective. Appendix Figure A1 shows an example of this approach. Second, the baseline prompt asks GPT to provide a direct answer. An alternative approach is to use a more complex prompt, instructing GPT to answer step-by-step, known as "Chain of Thought" (CoT) (Wei et al., 2022). Appendix Figure A2 presents a CoT example. We explain in Appendix B why we don't use more complex prompts like CoT as our baseline, but we nevertheless perform a robustness check. Third, GPT is one of many available LLMs. We evaluate whether other LLMs can yield similar implications. We choose two alternatives: ERNIE Bot and Claude 3 Haiku. ERNIE Bot ("wenxin yiyan" in Chinese), developed by Baidu, is arguably the most recognized LLM developed by a Chinese company. We choose ERNIE Bot because it may possess more local knowledge about the Chinese labor market. Claude 3 Haiku, developed by Anthropic, is from the second-largest LLM startup (after OpenAI).[22] GPT 3.5-turbo is trained with data up to September 2021. Claude 3 Haiku, released in 2024, helps us check if our results are robust to recent LLM updates.

Appendix Tables A3–A6 present the results of the robustness checks. For practical reasons, we choose a 10% random sample from Zhaopin.com.[23] Panel A of Appendix Tables A3 and A4 show the results using two alternative prompts. While the general findings remain consistent, including the predictive power for the expected wage and the LLM's ability to provide extra information conditional on the same occupation (or industry) dummies, the predictive power of the new prompts is smaller than our baseline prompts in panel B. In particular, the more complex CoT prompting does not show stronger predictive power, suggesting that overly complicated prompts do not necessarily yield better outcomes. We discuss the reasons in Appendix B.

Panel A of Appendix Table A5 uses ERNIE Bot, a large language model developed by a Chinese company, instead of GPT. ERNIE Bot and GPT yield highly similar findings. Panel A of Appendix Table A6 uses the more recent Claude 3 Haiku. We find that the predictive power of the match measures generated by Claude 3 Haiku generally outperforms GPT-3.5-turbo. For example, a one-SD increase in the occupation–occupation match generated by Claude 3 Haiku is associated with 2.2 percentage increase in expected wage, compared to a magnitude of 1.6 percentage increase using the GPT model. Given the rapid development of LLMs, our study offers a lower-bound estimate of their usefulness in economic research.

4.4 The Wide Applicability of the GPT Method

We now utilize the CLDS data to demonstrate the broad applicability of our GPT method and offer supplementary evidence of its validity. Our focus is on assessing the most demanding major–occupation matches. Although our GPT measure of major–occupation matches does not outperform the Duncan match index (the RM method) in Table 4 with the Zhaopin.com data, the RM method has one important limitation—it requires a demanding sample size and is not applicable to data with a small sample size such as CLDS. However, our GPT method does not impose any requirement on sample size.

CLDS complements our analysis using Zhaopin.com data in two additional ways. First, Zhaopin.com only includes expected wage information because most job seekers are not currently

---

[22] `https://www.statista.com/statistics/1446568/llm-developer-funding-2023/)`

[23] To ensure the robustness does not depend on a specific random sample, we use different random samples for each check.

employed. In contrast, CLDS provides actual wage data. Second, since jobs in the CLDS data are categorized by the official occupation classification system, we can use the JA method, which equals one only if the major–occupation pair appears in the official correspondence recommended by the government, as a traditional measure of match for comparison with the GPT measure.

Table 6 presents the results of Mincer regressions examining the predictive power of various match measures on actual wages using the CLDS data.[24] The first column shows that the RM method has low predictive power in the CLDS data. This ineffectiveness arises mainly because the Duncan index is a data-driven measure, making it sensitive to sample size. Specifically, the calculation of the fraction for a particular major ($\theta_{m,o}$ and θ_m in equation 1) becomes unstable with a small sample size. In contrast, the JA method (column 2) and the GPT method (column 3) derive the major–occupation correspondence from external sources, making them immune to the small-sample limitation. The GPT measure and the JA measure are highly positively correlated, with a correlation coefficient of 0.555, as demonstrated in Panel B of Table 3. A one–SD increase in the major–occupation match measure based on the traditional JA method or the GPT method, is estimated to increase the monthly wage by approximately 5.6–5.8 percent.

While the GPT method and the JA method perform similarly in measuring match quality in CLDS data, several advantages of our GPT method should be emphasized when compared to the traditional JA method. First and foremost, the GPT method treats the GPT itself as the job analyst, which is much more cost-effective than the traditional JA method employing real job analysts. The traditional JA method often relies on government efforts to provide matched majors, fields of knowledge, or skills for occupations, such as the Ministry of Human Resources and Social Security in China, the O*NET in the U.S.,[25] and the European Commission.[26] Hence, our GPT method could be particularly valuable for comprehending labor market mismatches in developing countries, where employing an adequate number of human job analysts can be either

---

[24] Due to the different data structure, the regressions differ slightly from those using Zhaopin.com data. The main difference is that we use the information of employees' current job instead of applied job. Consequently, we don't have the information of the job's required education and experience in CLDS. In addition, the classification systems of occupations and industries are different with those in Zhaopin.com data. Given a relatively small sample size of the CLDS data, we only consider two types of firm's ownership—SOEs and others.

[25] The O*NET does not directly provide matched majors. Instead, it offers required fields of knowledge for occupations. Researchers interested in studying the major–occupation mismatch can follow Yakusheva (2010) to first construct cross-references between the majors and O*NET fields of knowledge, and then compare fields of knowledge acquired from the major to those required for the occupation.

[26] See `https://ec.europa.eu/eurostat/documents/7884615/8088533/Conversion+Table+ISCO_08_ISCED_13.pdf` for the major–occupation correspondence table.

infeasible or too costly.

Second, the GPT method can significantly reduce the workload for researchers processing text data. In cases where the official major–occupation correspondence is unavailable, researchers using the JA method must take on the demanding task of acting as job analysts, introducing potential biases. For example, Nordin et al. (2010) and Domadenik et al. (2013) established the correspondence by comparing the detailed descriptions of major categories with those of occupation categories in documents outlining standard classifications for majors and occupations. Even when such official correspondence is available, if the classification systems of majors and occupations used in the data are inconsistent with the official ones, as observed in our Zhaopin.com data, the JA method requires researchers to construct a crosswalk between these classification systems. This process is notoriously time-consuming and challenging. In contrast, the GPT method can be applied to any classification system, as demonstrated in the previous subsection.

5 Using the GPT Method to Reproduce Existing Knowledge and Discover New Insights

The previous section demonstrates that GPT can be a powerful tool for measuring labor market mismatches. In this section, we present two specific research applications that utilize GPT in labor market studies. The first application illustrates how GPT can simulate an audit study by selectively concealing or revealing gender information. This approach can uncover gender stereotypes in perceived worker–job matches. The second application showcases how GPT can measure the versatility of academic majors—a dimension often overlooked or unfairly penalized by traditional match measures.

5.1 Uncovering Gender Stereotypes with GPT

A substantial body of literature highlights the role of gender stereotypes in shaping gender disparities in labor market outcomes (for an extensive review, see Blau and Kahn (2017)). Chan and Wang (2018), utilizing both non-experimental and experimental data from an online labor platform, demonstrate that women are preferred in feminine-typed occupations, while men do not experience higher hiring probabilities in masculine-typed occupations. Similarly, Carlsson (2011) find that women in Sweden have higher callback rates for interviews in female-dominated occupations, whereas in male-dominated occupations, there is no evidence of gender disparity.

To assess the significance of gender stereotypes in perceived applicant–job matches, we employ GPT to simulate a classic audit study by Goldin and Rouse (2000). We assume that GPT's training enables it to simulate human thought processes. We investigate the disparity in GPT's responses using two different prompts: one concealing gender information (used in our primary analysis) and the other revealing it. If GPT lacks gender stereotypes in its decision-making, its responses with and without gender information should show no significant gender differences, regardless of whether the occupations, industries, or majors are perceived as feminine-typed.

Table 7 presents the results. Specifically, it examines whether GPT perceives women as "matched" to certain occupations, industries, and majors traditionally associated with femininity. We regress the disparity between the GPT measures with and without gender information on gender and the interactions of gender with standardized proportions of females in occupation, industry, and major categories. Utilizing the Zhaopin.com data, Table 7 provides strong evidence of the existence of gender stereotypes in the data sets that were used to train the GPT. For example, column (3) of Table 7 shows that if a woman applies to an occupation with a 10% higher female share, revealing her gender increases the likelihood of being considered "matched" to the position by 1.37% compared to a man.

In summary, our findings indicate that GPT can effectively replicate the behaviors of real employers in labor markets and demonstrate gender stereotypes. Therefore, employers considering the utilization of GPT for recruitment, such as resume screening, should exercise caution.[27] For researchers, its advanced natural language processing capabilities and adaptability make it a promising tool for simulating audit studies to investigate potential biases or stereotype across various contexts.[28]

---

[27] Employers in the real world are beginning to consider LLMs for recruitment. For example, an article in Forbes (`https://www.forbes.com/sites/bernardmarr/2023/03/07/the-7-best-examples-of-how-chatgpt-can-be-used-in-human-resources-hr/?sh=448412a94a82`) discusses how ChatGPT can be used in recruitment.

[28] Note that although LLMs have shown impressive potential in simulating human behavior, conducting LLM-simulated experiments still encounters challenges of endogeneity. Please refer to Gui and Toubia (2023) for technical details.

5.2 Measuring the Versatility of Majors with GPT

The education provided by a major equips students with two types of skills for the labor market. The first type is specialized skills, which are highly specific and applicable to only a few occupations but often require a deep level of expertise (e.g., medicine). The second type is general skills, which are versatile and applicable across a broad range of occupations (e.g., management science). We define a major's ability to qualify students for various occupations as its versatility. Students from versatile majors can pursue a wide range of job opportunities and are more likely to transition successfully when shifting to different occupations.

However, the versatility of a major can be unfairly penalized by traditional match measures. The job switching method interprets transitions between occupations or industries as indicators of a mismatch. Meanwhile, the realized matches method favors major–occupation pairs where graduates disproportionately apply for a specific occupation (e.g., mechanical graduates applying for mechanical designer jobs, as shown in Appendix Table A2). As a result, if students from a major apply evenly across various occupations, the major is unlikely to achieve a high match score under the RM measure.

Appendix Table A7 uses an artificial example to illustrate how GPT can reveal the overlooked value of a versatile major. Consider three majors and four occupations: the first is a specialized major, with students disproportionately applying to a single occupation; the second is a versatile major, with students applying evenly across all four occupations; and the third is an unprepared major, which does not equip students for the labor market, resulting in evenly distributed but poorly aligned applications. Using the Duncan Index formula (equation 1), the indices for these majors are 0.919, 0.4, and 0.4, respectively. Although the second and third majors share the same index due to similar application patterns, they differ fundamentally in versatility. GPT, leveraging extensive external knowledge, can differentiate between these cases by identifying that the versatile major aligns well with all four occupations, whereas the unprepared major aligns with none, highlighting the unique adaptability of the versatile major.

Using real-world data from Zhaopin.com, Figure 2 illustrates the relationship between the Duncan Index and the GPT measure, aggregated at both the broad and detailed major levels. In Panel A, the medicine major appears in the upper-right corner, indicating that students' applications are concentrated in a limited set of occupations, and GPT identifies those applications as strong matches. In contrast, the philosophy major is located in the lower-left corner, reflecting both diversified applications and lower perceived suitability by GPT. Of particular interest are majors in the lower-right corner, such as management science. These majors demonstrate highly diversified application patterns compared to specialized fields like medicine, yet GPT assesses their graduates as broadly qualified for most jobs they apply for. Those majors are what we referred as "versatile majors."

To understand why versatile majors are unfairly penalized by traditional RM measures, column (5) of Table 5 incorporates an interaction term between the Duncan Index and the GPT measure, revealing a significantly negative coefficient. Since versatile majors typically exhibit a combination of low Duncan Index values and high GPT scores, this new specification predicts higher expected wages for such majors.[29] Figure 3 illustrates the gap between predicted wages from the two specifications, with darker colors indicating stronger penalties under the original model, which excludes GPT-based versatility information. The results show that majors in the lower-right corner are disproportionately punished by traditional RM measures when versatility data from GPT is omitted.

6 Conclusion and Discussion

The recent development in artificial intelligence (AI), highlighted by large language models, enables economists to uncover textual information that was previously challenging to capture. One such example is the textual information in categorical variables. Traditional econometric methods typically use a fixed-effect approach to handle categorical variables, overlooking the textual label information associated with different categories that could capture "similarities" among them. The emergence of LLMs provides a viable approach to address this issue.

We employ LLMs in the context of labor market mismatch. Specifically, we task a large language model with simulating the role of an HR specialist to assess the suitability of an applicant with specific characteristics for a given job. Our empirical analysis, utilizing both administrative data from an online job posting platform and typical survey data, justifies the GPT as a potential measure of match quality and highlights its advantage over traditional measures of match quality. We emphasize three main findings. First, our GPT measure highly correlates with traditional

---

[29] All match measures are standardized in the regression; thus, a low Duncan Index corresponds to a negative standardized score.

approaches to defining match quality, including the job switching method, the realized matches method, and the job analysis method. Second, the GPT approach can provide additional information beyond the traditional methods by utilizing the textual information in the category labels. For example, while the job switching method only considers whether the previous job and the applied job belong to the same occupation/industry, the GPT method takes into account the "similarity" between the two jobs. Last, the GPT method is easy and inexpensive to apply. Unlike the job switching method and the realized method, which require either high-quality panel data or large-sample administrative data, our GPT method is applicable to survey data (e.g., CLDS) with only 2,431 observations. Unlike the job analysis method, which traditionally employs real-world human experts and is usually too expensive for researchers, the GPT method employs large language models as simulated experts and is significantly more affordable.

The first application uses GPT to simulate an audit study, uncovering the presence of gender stereotypes. Traditional audit studies typically rely on fictitious résumé. We replicate this approach by varying the inclusion of gender information in the prompts provided to large language models (LLMs). We find that once we reveal the gender information to GPT, it would consider females better matches to those positions traditionally considered feminine, indicating that GPT also possesses gender stereotypes. This is not surprising, as GPT is trained on real-world textual data and mimics human language expression. In the second application, we demonstrate how GPT can assist in measuring the adaptability of academic majors, a task that traditional methods struggle to address. When a major equips students with skills applicable to a wide range of occupations, these individuals tend to work in diverse positions and can afford frequent transition between jobs. Traditional match measures interpret such behavior as indicative of poor worker–job match. The GPT approach mitigates this issue by bypassing reliance on observed behavioral data. Instead, it leverages external information, including extensive textual data from the entire internet.

In summary, by utilizing GPT to analyze labor market mismatches, we effectively overcome several limitations associated with traditional measures. GPT's ability to process and interpret textual data allows for more nuanced labor market analysis, which is particularly relevant in developing economies with limited access to detailed datasets or informative official documents on the labor market. Our research marks an initial step in integrating AI technology with economic analysis, opening possibilities for future research to refine and expand upon our methodology. This integration of GPT into labor market analysis not only demonstrates its usefulness in handling complex datasets and categorical variables but also highlights its potential to provide new insights into other economic inquiries.

References

Adamopoulos, T., L. Brandt, J. Leight, and D. Restuccia (2022). Misallocation, selection, and productivity: A quantitative analysis with panel data from China. *Econometrica 90*(3), 1261–1282.

Altonji, J. G., L. B. Kahn, and J. D. Speer (2016). Cashier or consultant? Entry labor market conditions, field of study, and career success. *Journal of Labor Economics 34*(S1), S361–S401.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis 31*(3), 337–351.

Atalay, E., P. Phongthiengtham, S. Sotelo, and D. Tannenbaum (2020). The evolution of work in the United States. *American Economic Journal: Applied Economics 12*(2), 1–34.

Biasi, B. and S. Ma (2022). The education-innovation gap. Working Paper 29853, National Bureau of Economic Research.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Ramesh, P. Shyam, M. S. Chintala, P. Clark, et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Carlsson, M. (2011). Does hiring discrimination cause gender segregation in the Swedish labor market? *Feminist Economics 17*(3), 71–102.

Chan, J. and J. Wang (2018). Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science 64*(7), 2973–2994.

Chen, Y., T. X. Liu, Y. Shan, and S. Zhong (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences 120*(51), e2316205120.

Cowen, T. and A. T. Tabarrok (2023). How to learn and teach economics with large language models, including GPT. Working Paper.

Deming, D. and L. B. Kahn (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics 36*(S1), S337–S369.

Deming, D. J. and K. Noray (2020). Earnings dynamics, changing job skills, and STEM careers. *Quarterly Journal of Economics 135*(4), 1965–2005.

Domadenik, P., D. Farcnik, and F. Pastore (2013). Horizontal mismatch in the labour market of graduates: The role of signalling. IZA discussion paper.

Duncan, O. D. and B. Duncan (1955). A methodological analysis of segregation indexes. *American Sociological Review 20*(2), 210–217.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2311.05232.

Goldin, C. and C. Rouse (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review 90*(4), 715–741.

Gui, G. and O. Toubia (2023). The challenge of using LLMs to simulate human behavior: A causal inference perspective. arXiv preprint arXiv:2312.15524.

Guvenen, F., B. Kuruscu, S. Tanaka, and D. Wiczer (2020). Multidimensional skill mismatch. *American Economic Journal: Macroeconomics 12*(1), 210–244.

Hansen, A. L. and S. Kazinnik (2023). Can ChatGPT decipher Fedspeak? Working Paper.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics 124*(4), 1403–1448.

Imbert, C., M. Seror, Y. Zhang, and Y. Zylberberg (2022). Migrants and firms: Evidence from China. *American Economic Review 112*(6), 1885–1914.

Jiang, S. and Y. Guo (2022). Reasons for college major-job mismatch and labor market outcomes: Evidence from China. *China Economic Review*, 101822.

Kambourov, G. and I. Manovskii (2009). Occupational specificity of human capital. *International Economic Review 50*(1), 63–115.

Kaplan, J., S. McCandlish, T. Henighan, et al. (2020). Scaling laws for neural language models. *CoRR abs/2001.08361*.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature 61*(4), 1281–1317.

Kuhn, P. and K. Shen (2013). Gender discrimination in job ads: Evidence from China. *Quarterly Journal of Economics 128*(1), 287–336.

Lemieux, T. (2014). Occupations, fields of study and returns to education. *Canadian Journal of Economics 47*(4), 1047–1077.

Lise, J. and F. Postel-Vinay (2020). Multidimensional skills, sorting, and human capital accumulation. *American Economic Review 110*(8), 2328–2376.

Lopez-Lira, A. and Y. Tang (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. Working Paper.

López Espejel, J., E. H. Ettifouri, M. S. Yahaya Alassan, E. M. Chouham, and W. Dahhane (2023). GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal 5*, 100032.

Marinescu, I. and R. Wolthoff (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics 38*(2), 535–568.

Miller, R. A. (1984). Job matching and occupational choice. *Journal of Political Economy 92*(6), 1086–1120.

Moscarini, G. (2001). Excess worker reallocation. *Review of Economic Studies 68*(3), 593–612.

Nagypál, É. (2007). Learning by doing vs. learning about match quality: Can we tell them apart? *Review of Economic Studies 74*(2), 537–566.

Nieto, S., A. Matano, and R. Ramos (2015). Educational mismatches in the EU: Immigrants vs natives. *International Journal of Manpower 36*(4), 540–561.

Nordin, M., I. Persson, and D.-O. Rooth (2010). Education–occupation mismatch: Is there an income penalty? *Economics of Education Review 29*(6), 1047–1059.

Perry, A., S. Wiederhold, and D. Ackermann-Piek (2014). How can skill mismatch be measured? New approaches with PIAAC. *Methods, Data, Analyses 8*(2), 137–174.

Robst, J. (2007). Education and job match: The relatedness of college major and work. *Economics of Education Review 26*(4), 397–407.

Sellami, S., D. Verhaest, and W. Van Trier (2018). How to measure field-of-study mismatch? A comparative analysis of the different methods. *Labour 32*(4), 141–173.

Sullivan, P. (2010). Empirical evidence on occupation and industry specific human capital. *Labour Economics 17*(3), 567–580.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, 24824–24837.

Yakusheva, O. (2010). Return to college education revisited: Is relevance relevant? *Economics of Education Review 29*(6), 1125–1142.

Yang, K.-C. and F. Menczer (2023). Large language models can rate news outlet credibility. Working Paper.

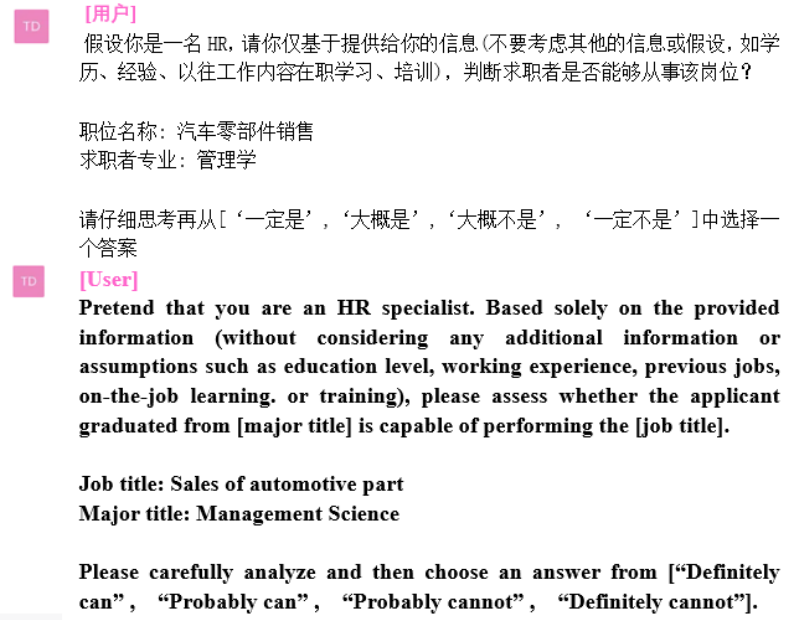
Yao, J.-Y., K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan (2023). LLM lies: Hallucinations are not bugs, but features as adversarial examples. arXiv preprint arXiv:2310.01469.

Zhao, H., F. Yang, H. Lakkaraju, and M. Du (2024). Opening the black box of large language models: Two views on holistic interpretability. arXiv preprint arXiv:2402.10688.

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

Zhu, R. (2014). The impact of major–job mismatch on college graduates' early career earnings: Evidence from China. *Education Economics 22*(5), 511–528.

Figure 1: An Example of Constructing the “Major–Occupation” Match Measure using GPT



[用户]

假设你是一名 HR，请你仅基于提供给你的信息（不要考虑其他的信息或假设，如学历、经验、以往工作内容在职学习、培训），判断求职者是否能够从事该岗位？

职位名称：汽车零部件销售
求职者专业：管理学

请仔细思考再从[‘一定是’，‘大概是’，‘大概不是’， ‘一定不是’]中选择一个答案

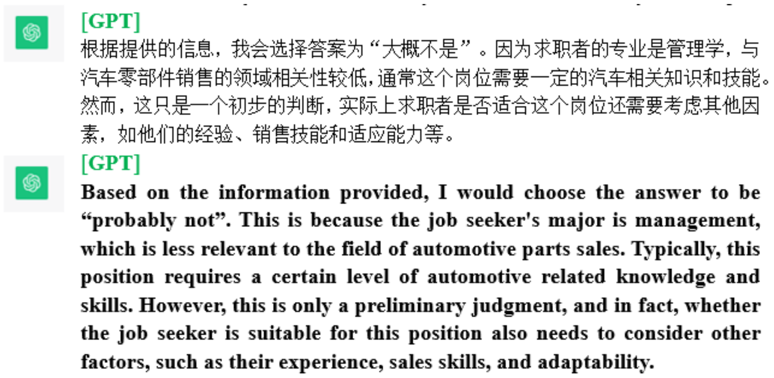
[User]

Pretend that you are an HR specialist. Based solely on the provided information (without considering any additional information or assumptions such as education level, working experience, previous jobs, on-the-job learning. or training), please assess whether the applicant graduated from [major title] is capable of performing the [job title].

Job title: Sales of automotive part
Major title: Management Science

Please carefully analyze and then choose an answer from [“Definitely can”, “Probably can”, “Probably cannot”, “Definitely cannot”].

(a) Prompt of the Question



[GPT]

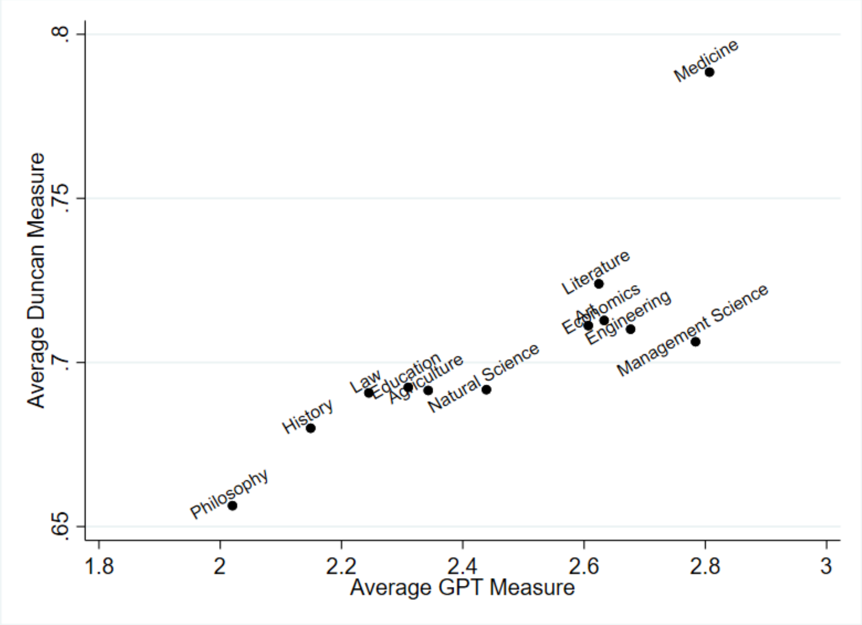
根据提供的信息，我会选择答案为“大概不是”。因为求职者的专业是管理学，与汽车零部件销售的领域相关性较低，通常这个岗位需要一定的汽车相关知识和技能。然而，这只是一个初步的判断，实际上求职者是否适合这个岗位还需要考虑其他因素，如他们的经验、销售技能和适应能力等。

[GPT]

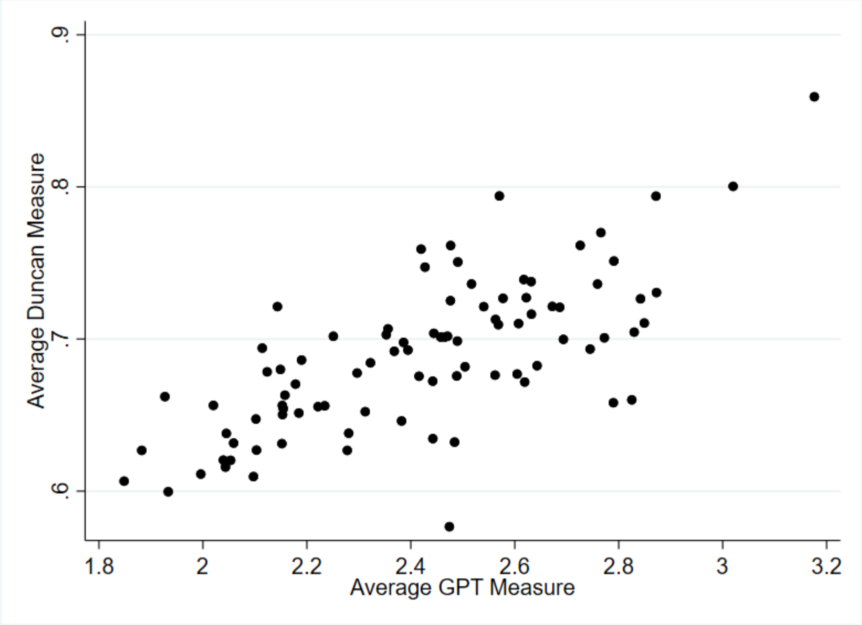
Based on the information provided, I would choose the answer to be “probably not”. This is because the job seeker's major is management, which is less relevant to the field of automotive parts sales. Typically, this position requires a certain level of automotive related knowledge and skills. However, this is only a preliminary judgment, and in fact, whether the job seeker is suitable for this position also needs to consider other factors, such as their experience, sales skills, and adaptability.

(b) Answer

Figure 2: Average “Major–Occupation” Match Measures by Major Categories



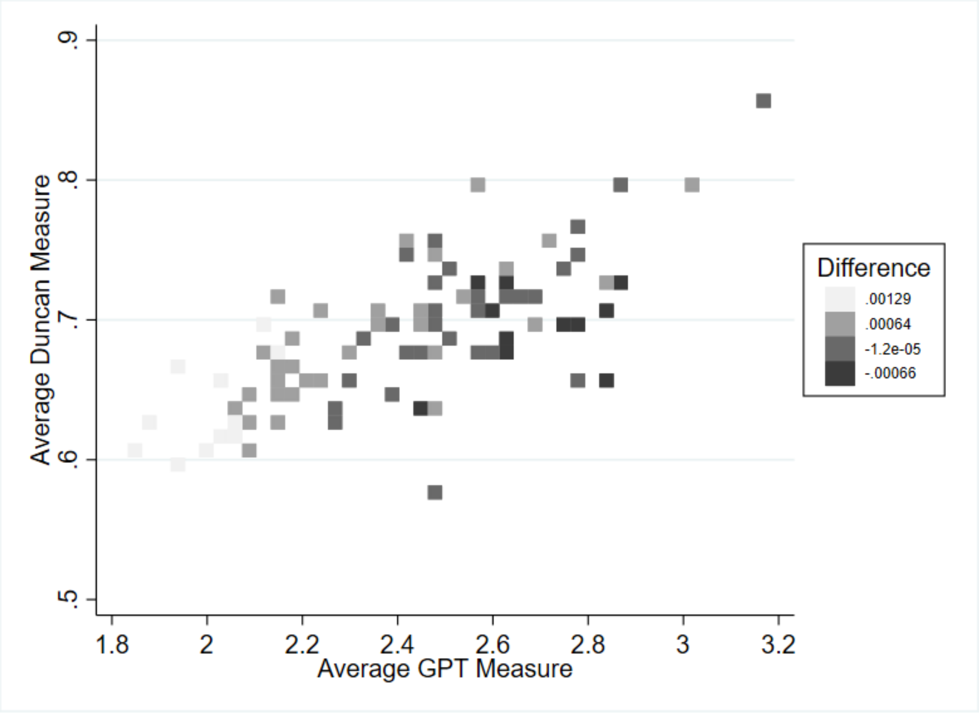
(a) By Broad Major Categories



(b) By Detailed Major Categories

Note: Panels (a) and (b) of this figure presents the scatter plot of the average Duncan index and the GPT major–occupation measures for 12 broad major categories and 92 detailed categories, respectively.

Figure 3: Differences in Predicted Wages When Omitting the Orthogonal Information on Match Quality Provided by GPT



Notes: This figure presents the average differences in predicted wages when using only the Duncan index for major–occupation match (equation 2), compared to results that also control for both the GPT major–title measure and the interaction between the two measures, across 92 detailed major categories.

Table 1: Examples of Comparing GPT Measures with Traditional Measures (Zhaopin.com Data)

Detailed Occupation Category of Applied Job	**Detailed Occupation Category of Current Job**	**Same-occupation Dummy**	**GPT Response**	**GPT Occupation-occupation Match**
Software test engineer	Software test engineer	1	Probably can	1
	Software R&D Engineer	0	Probably can	1
	Sales representative	0	Probably cannot	0
Industry Category of Applied Job	**Industry Category of Current Job**	**Same-industry Dummy**	**GPT Response**	**GPT Industry-industry Match**
IT services	IT services	1	Probably can	1
	Computer software	0	Probably can	1
	Computer hardware	0	Probably cannot	0

Table 2: Summary Statistics

Data	Zhaopin.com			CLDS Data		
	Mean (1)	S.D. (2)	Obs. (3)	Mean (4)	S.D. (5)	Obs. (6)
Panel A: Individual Characteristics						
Female	0.48	0.50	847,801	0.51	0.50	2,431
Age	26.90	4.74	847,801	35.68	9.64	2,428
Married	0.28	0.45	847,801	0.71	0.45	2,431
Bachelor degree or above	0.44	0.50	847,801	0.54	0.50	2,431
Years of schooling	15.53	0.76	847,801	15.69	0.90	2,431
Working experience	5.70	3.60	847,801	19.99	9.73	2,428
Monthly wage of the most recent job	4,457	3,076	846,535	5,055	3,837	2,206
Monthly wage of expected job	4,709	3,243	846,740			
Unemployed	0.73	0.44	847,801			
Panel B: Match Measures						
Same-occupation dummy	0.22	0.42	847,801			
GPT occupation–occupation match	0.69	0.46	843,296			
Same-industry dummy	0.26	0.44	847,801			
GPT industry–industry match	0.48	0.50	773,203			
Duncan major-occupation match	0.71	0.33	816,161	0.60	0.31	2,431
JA major–occupation match				0.35	0.48	2,382
GPT major–title match	0.54	0.50	832,623	0.47	0.50	2,303

Notes: Panels A and B present the means and standard deviations of individual characteristics and match quality measures, respectively. Columns (1)–(3) and columns (4)–(6) show the results for the Zhaopin.com data and the CLDS data, respectively. For the CLDS data, we begin by excluding individuals with zero wages, followed by excluding those with wages below the 1st percentile or above the 99th percentile.

Table 3: Pairwise Correlations between the Traditional and GPT Measures

Panel A: Zhaopin.com Data	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
Panel B: CLDS Data	**Duncan major-occupation match**	**JA major-title match**	**GPT major-title match**			
Duncan major-occupation match	1					
JA major-occupation match	0.507***	1				
GPT major-title match	**0.429*****	**0.555*****	1			

Table 4: Wage Premium of the Applicant–Posting Match (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Same-occupation dummy	0.010***		0.004							0.007**		0.003
	(0.003)		(0.003)							(0.003)		(0.003)
GPT occupation-occupation match		0.017***	0.016***								0.015***	0.014***
		(0.003)	(0.003)								(0.003)	(0.003)
Same-industry dummy				0.022***		0.017***				0.020***		0.015***
				(0.002)		(0.002)				(0.002)		(0.002)
GPT industry-industry match					0.018***	0.008***					0.016***	0.007***
					(0.002)	(0.001)					(0.001)	(0.001)
Duncan major-occupation match							0.008***		0.008***	0.007***		0.005***
							(0.002)		(0.001)	(0.002)		(0.001)
GPT major-title match								0.005***	0.002		0.003**	0.001
								(0.002)	(0.001)		(0.001)	(0.001)
Basic control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	810,886	810,886	810,886	743,568	743568	743,568	801,107	801,107	801,107	727,060	727,060	727,060
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0340	0.1111	0.1165	0.1430	0.1031	0.1553	0.0222	0.0076	0.0232	0.1771	0.1930	0.2465

Notes: This table presents the wage premium of the applicant–posting match index across all three dimensions for Zhaopin.com data: the occupation–occupation match, industry–industry match, and major–occupation match. For each dimension, we conduct three Mincer regressions. The first two regressions include the traditional measure and the GPT measure separately, while the last one incorporates both measures. Columns (10)–(12) control for all traditional measures, all GPT measures, and all traditional and GPT measures, respectively. The baseline control variables comprise individual characteristics (gender, dummies for birth month, years of schooling, type of school, marital status, employment status, working experience, and its square term) and posting characteristics (education and experience requirements, and ownership type and scale of the hiring firms). Standard errors are clustered at the detailed occupation category level. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: Sources of the Extra Information of the GPT Measures (Zhaopin.com Data)

Dependent Variable:	Monthly Wage of Expected Job (Log)				
Regressions Conditional on:	Applied job in a different occupation category		Applied job in a different industry category		Interaction between Duncan and GPT Measures
	(1)	(2)	(3)	(4)	(5)
Same-occupation dummy	Omitted	Omitted		0.001	
				(0.003)	
GPT occupation-occupation match	**0.016*** **	0.013***		0.016***	
	(0.003)	(0.003)		(0.003)	
Same-industry dummy		0.018***	Omitted	Omitted	
		(0.002)			
GPT industry-industry match		0.007***	**0.008*** **	0.006***	
		(0.001)	(0.001)	(0.001)	
Duncan major-occupation match		0.004**		0.007***	0.007***
		(0.001)		(0.001)	(0.002)
GPT major-title match		0.002		0.001	0.002
		(0.001)		(0.001)	(0.001)
Duncan major-occupation match× GPT major-title match					-0.002**
					(0.001)
Basic control	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes
Observations	628,686	559,901	531,498	519,340	801,107

Notes: This table investigates whether the GPT measure can still provide extra information when controlling for traditional measures by imposing various restrictions. Columns (1) and (2), ((3) and (4), respectively) focus on a subsample that apply to jobs in a different occupation (industry, respectively) category as the previous one. Column (5) presents the estimated interaction effects between the Duncan index and the GPT major–title match. The baseline control variables comprise individual characteristics (gender, dummies for birth month, years of schooling, type of school, marital status, employment status, working experience, and working experience squared) and posting characteristics (education and experience requirements, and ownership type and scale of the hiring firms). Standard errors are clustered at the detailed occupation category level. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: Wage Premium of the Worker–Occupation Match (CLDS Data)

Dependent Variable	Monthly Wage of Current Job (Log)			
	(1)	(2)	(3)	(4)
Duncan major-occupation match	0.025			-0.010
	(0.016)			(0.019)
JA major-occupation match		0.056***		0.037**
		(0.015)		(0.017)
GPT major-title match			0.058***	0.048**
			(0.019)	(0.019)
Basic control	Yes	Yes	Yes	Yes
Survey year FE	Yes	Yes	Yes	Yes
City FE	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes
Occupation category FE	Yes	Yes	Yes	Yes
Industry category FE	Yes	Yes	Yes	Yes
Observations	2,035	2,002	2,035	2,002
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0759	0.4509	0.4403	0.6730

Notes: This table shows the wage premium of the major–occupation match measured by the traditional and GPT methods using the CLDS data. The basic control variables includes gender, years of schooling, working experience (including the square term), dummies for month of birth, marital status, type of school, and firm's scale and ownership type of the current job. Standard errors are clustered at the detailed occupation category level. * significant at 10%; ** significant at 5%; *** significant at 1%.

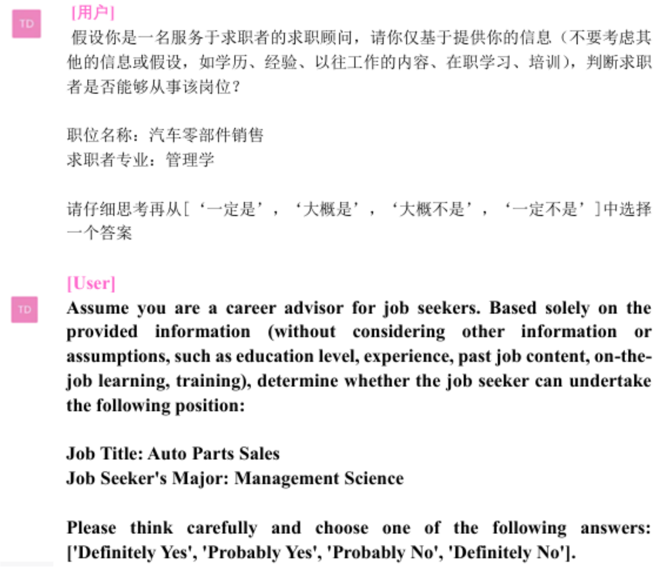
Table 7: Female Share in Industry/Occupation/Major and Gender Stereotype of the GPT (Zhaopin.com Data)

Dependent Variable	Diff in GPT Measures w/ and w/o Gender Info		
	GPT occupation-occupation match (1)	GPT industry-industry match (2)	GPT major-title match (3)
Female	0.014***	0.012***	-0.021***
	(0.005)	(0.004)	(0.004)
Female×Female% by occupation category	0.128***	0.012	0.137***
	(0.023)	(0.010)	(0.022)
Female×Female% by industry category	0.007	0.244***	-0.006
	(0.029)	(0.024)	(0.026)
Female×Female% by major category	0.019	0.032***	0.117***
	(0.014)	(0.007)	(0.023)
Basic control	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes
Observations	609,317	730,396	794,237
R-squared	0.053	0.034	0.017

Notes: This table presents further evidence of the existence of gender stereotype using Zhaopin.com data. The baseline control variables comprise dummies for birth month, education and experience requirements of applied jobs, and ownership type and scale of the hiring firms. Standard errors are clustered at the detailed occupation category level. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A: Additional Tables and Figures

Appendix Figure A1: An Example of Constructing the “Major–Occupation” Match Measure Using GPT with Prompts from Job Seekers’ Perspective



[用户]

假设你是一名服务于求职者的求职顾问，请你仅基于提供你的信息（不要考虑其他的信息或假设，如学历、经验、以往工作的内容、在职学习、培训），判断求职者是否能够从事该岗位？

职位名称：汽车零部件销售
求职者专业：管理学

请仔细思考再从[‘一定是’，‘大概是’，‘大概不是’，‘一定不是’]中选择一个答案

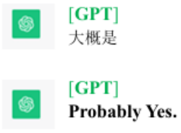
[User]

Assume you are a career advisor for job seekers. Based solely on the provided information (without considering other information or assumptions, such as education level, experience, past job content, on-the-job learning, training), determine whether the job seeker can undertake the following position:

Job Title: Auto Parts Sales
Job Seeker's Major: Management Science

Please think carefully and choose one of the following answers: ['Definitely Yes', 'Probably Yes', 'Probably No', 'Definitely No'].

(a) Prompt of the Question



[GPT]

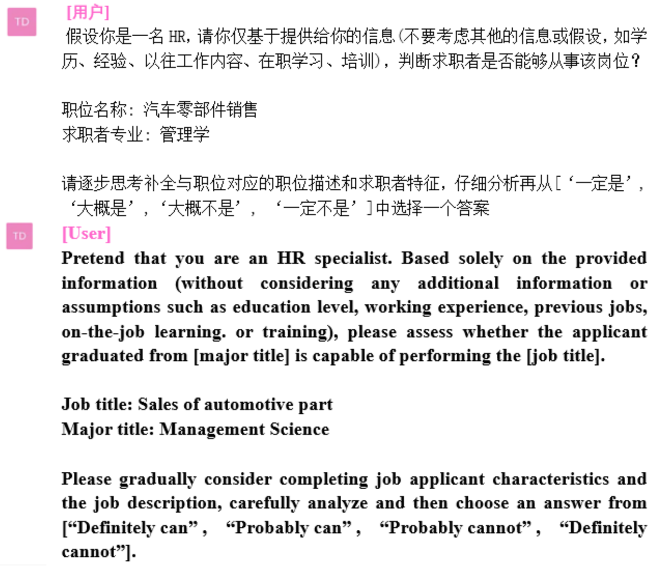
大概是

[GPT]

Probably Yes.

(b) Answer

Appendix Figure A2: An Example of Constructing the “Major–occupation” Match Measure Using GPT with CoT Prompting



[用户]

假设你是一名 HR，请你仅基于提供给你的信息(不要考虑其他的信息或假设，如学历、经验、以往工作内容、在职学习、培训)，判断求职者是否能够从事该岗位？

职位名称：汽车零部件销售
求职者专业：管理学

请逐步思考补全与职位对应的职位描述和求职者特征，仔细分析再从[‘一定是’，‘大概是’，‘大概不是’，‘一定不是’]中选择一个答案

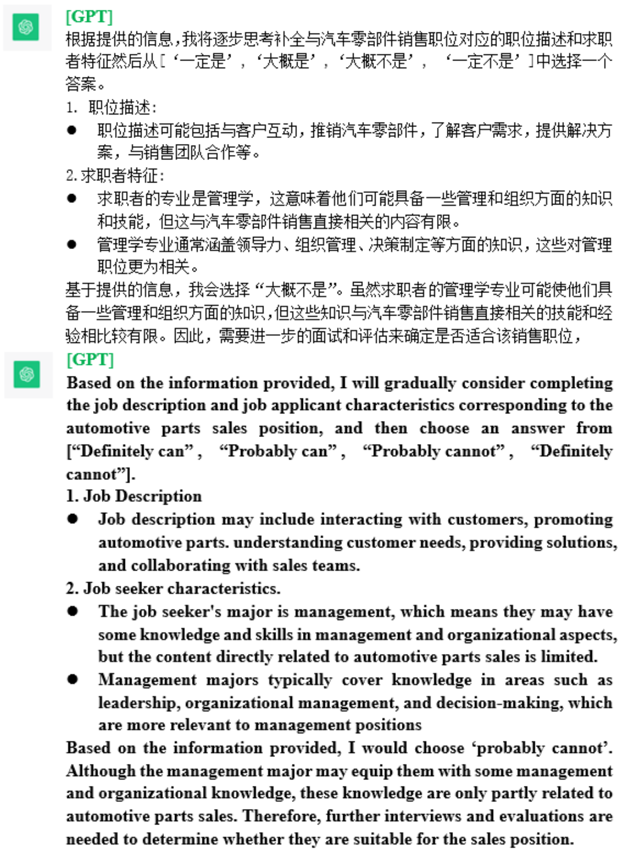
[User]

Pretend that you are an HR specialist. Based solely on the provided information (without considering any additional information or assumptions such as education level, working experience, previous jobs, on-the-job learning. or training), please assess whether the applicant graduated from [major title] is capable of performing the [job title].

Job title: Sales of automotive part
Major title: Management Science

Please gradually consider completing job applicant characteristics and the job description, carefully analyze and then choose an answer from [“Definitely can”, “Probably can”, “Probably cannot”, “Definitely cannot”].

(a) Prompt of the Question



[GPT]

根据提供的信息，我将逐步思考补全与汽车零部件销售职位对应的职位描述和求职者特征然后从[‘一定是’，‘大概是’，‘大概不是’，‘一定不是’]中选择一个答案。

1. 职位描述：
- 职位描述可能包括与客户互动，推销汽车零部件，了解客户需求，提供解决方案，与销售团队合作等。

2. 求职者特征：
- 求职者的专业是管理学，这意味着他们可能具备一些管理和组织方面的知识和技能，但这与汽车零部件销售直接相关的内容有限。
- 管理学专业通常涵盖领导力、组织管理、决策制定等方面的知识，这些对管理职位更为相关。

基于提供的信息，我会选择“大概不是”。虽然求职者的管理学专业可能使他们具备一些管理和组织方面的知识，但这些知识与汽车零部件销售直接相关的技能和经验相比较有限。因此，需要进一步的面试和评估来确定是否适合该销售职位，

[GPT]

Based on the information provided, I will gradually consider completing the job description and job applicant characteristics corresponding to the automotive parts sales position, and then choose an answer from [“Definitely can”, “Probably can”, “Probably cannot”, “Definitely cannot”].

1. Job Description
- **Job description may include interacting with customers, promoting automotive parts. understanding customer needs, providing solutions, and collaborating with sales teams.**

2. Job seeker characteristics.
- **The job seeker's major is management, which means they may have some knowledge and skills in management and organizational aspects, but the content directly related to automotive parts sales is limited.**
- **Management majors typically cover knowledge in areas such as leadership, organizational management, and decision-making, which are more relevant to management positions**

Based on the information provided, I would choose ‘probably cannot’. Although the management major may equip them with some management and organizational knowledge, these knowledge are only partly related to automotive parts sales. Therefore, further interviews and evaluations are needed to determine whether they are suitable for the sales position.

(b) Answer

Appendix Table A1: Illustrating the Industrial and Occupational Classification Systems in the Zhaopin.com Data

Job Title	Detailed Occupation Category	Broad Occupation Category	Industry Category
Software test engineer	Software test engineer	Software personnel/Internet developer/ System integration staff	Computer software
Game tester	Software test engineer	Software personnel/Internet developer/ System integration staff	Internet business/E-commerce
Software R&D engineer	Software R&D engineer	Software personnel/Internet developer/ System integration staff	Computer software
Video algorithm engineer	Software R&D engineer	Software personnel/Internet developer/ System integration staff	Internet business/E-commerce
Accountant	Accountant	Financial personnel/Auditors/ Taxation staff	Computer software
Human resources specialist	Administrative officer/ Administrative assistant	Administrative staff/Logistics personnel/ Secretarial staff	Computer software
Accountant	Accountant	Financial personnel/Auditors/ Taxation staff	Internet business/E-commerce
Human resources specialist	Administrative officer/ Administrative assistant	Administrative staff/Logistics personnel/ Secretarial staff	Internet business/E-commerce

Appendix Table A2: Examples of Duncan Major–Occupation Match Index

Detailed Occupation Category	Detailed Major Category (Proportion of Applicants in Major Category in the Data, %)	Proportion of Applicants in Major Category within Occupation Category (%)	Proportion Difference (%)	Duncan Index
Tour consultant	Mechanical (9.17)	2.11	-7.06	0.016
Tour consultant	Tourism management (2.38)	30.56	28.18	0.99
Mechanical designer	Mechanical (9.17)	82.85	73.68	1
Mechanical designer	Tourism management (2.38)	0.04	-2.33	0.082

Notes: This table presents examples of the Duncan major–occupation match index for two occupation categories ("Tour consultant" and "Mechanical designer") and two major categories ("Mechanical" and "Tourism management").

Appendix Table A3: Robust Check I—Using Prompts from Job Seekers' Perspective (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: GPT with Alternative Prompt												
Same-occupation dummy	0.007*		0.002							0.005		0.001
	(0.004)		(0.005)							(0.004)		(0.005)
GP occupation-occupation match		0.012***	0.011*								0.010**	0.009
using alternative prompt		(0.004)	(0.006)								(0.004)	(0.006)
Same-industry dummy				0.019***		0.024***				0.018***		0.022***
				(0.004)		(0.004)				(0.004)		(0.004)
GPT industry-industry match					0.017***	-0.005					0.016***	-0.005
using alternative prompt					(0.004)	(0.004)					(0.004)	(0.003)
Duncan major-occupation match							0.005*		0.004	0.004		0.002
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match								0.005*	0.004		0.005*	0.004
using alternative prompt								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0236	0.0508	0.0526	0.1104	0.078	0.1118	0.0062	0.0072	0.0109	0.1273	0.1228	0.1544
Panel B: GPT with Baseline Prompt												
Same-occupation dummy	0.007*		0.001							0.005		-0.000
	(0.004)		(0.004)							(0.004)		(0.004)
GPT occupation-occupation match		0.016***	0.015***								0.015***	0.014***
		(0.004)	(0.004)								(0.004)	(0.004)
Same-industry dummy				0.019***		0.017***				0.018***		0.016***
				(0.004)		(0.003)				(0.004)		(0.004)
GPT industry-industry match					0.016***	0.003					0.014***	0.003
					(0.004)	(0.003)					(0.004)	(0.003)
Duncan major-occupation match							0.005*		0.005*	0.004		0.003
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match								0.001	-0.001		-0.001	-0.002
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0236	0.0902	0.0907	0.1104	0.0715	0.1117	0.0062	0.0002	0.0063	0.1273	0.1489	0.1858
Observations	96,432	96,432	96,432	96,432	96,432	96,432	96,432	96,432	96,432	96,432	96,432	96,432
Baseline control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Refer to Table 4.

Appendix Table A4: Robust Check II—Using CoT Prompting (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: GPT with CoT Prompt												
Same-occupation dummy	0.011***		0.009**							0.010**		0.009**
	(0.004)		(0.004)							(0.004)		(0.004)
GPT occupation-occupation match using CoT prompt		0.011***	0.008**								0.010***	0.007**
		(0.003)	(0.004)								(0.003)	(0.003)
Same-industry dummy				0.017***		0.017***				0.016***		0.016***
				(0.004)		(0.004)				(0.004)		(0.004)
GPT industry-industry match using CoT prompt					0.008**	-0.001					0.008**	-0.001
					(0.004)	(0.003)					(0.004)	(0.003)
Duncan major-occupation match							0.007**		0.006**	0.005**		0.005*
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match using CoT prompt								0.003	0.002		0.003	0.002
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0544	0.0408	0.0769	0.0847	0.0207	0.0848	0.0101	0.0031	0.0115	0.1374	0.0618	0.1568
Panel B: GPT with Baseline Prompt												
Same-occupation dummy	0.011***		0.008*							0.010**		0.007*
	(0.004)		(0.004)							(0.004)		(0.004)
GPT occupation-occupation match		0.014***	0.011***								0.013***	0.010**
		(0.004)	(0.004)								(0.004)	(0.004)
Same-industry dummy				0.017***		0.014***				0.016***		0.013***
				(0.004)		(0.004)				(0.004)		(0.004)
GPT industry-industry match					0.015***	0.004					0.014***	0.004
					(0.004)	(0.004)					(0.004)	(0.004)
Duncan major-occupation match							0.007**		0.007**	0.005**		0.005
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match								0.002	0.000		0.001	-0.000
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0544	0.0647	0.0879	0.0847	0.0594	0.0868	0.0101	0.0014	0.0101	0.1374	0.1163	0.1652
Observations	101,141	101,141	101,141	101,141	101,141	101,141	101,141	101,141	101,141	101,141	101,141	101,141
Baseline control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Refer to Table 4.

Appendix Table A5: Robust Check III—Using ERNIE Bot (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: ERNIE Bot with Baseline Prompt												
Same-occupation dummy	0.008**		0.004							0.007*		0.003
	(0.004)		(0.004)							(0.004)		(0.004)
ERNIE Bot occupation-occupation match		0.012***	0.010***								0.012***	0.009***
		(0.003)	(0.004)								(0.003)	(0.003)
Same-industry dummy				0.017***		0.016***				0.016***		0.014***
				(0.004)		(0.004)				(0.004)		(0.004)
ERNIE Bot industry-industry match					0.012***	0.002					0.011***	0.002
					(0.004)	(0.003)					(0.003)	(0.003)
Duncan major-occupation match							0.006**		0.006*	0.005*		0.004
							(0.003)		(0.003)	(0.003)		(0.003)
ERNIE Bot major-title match								0.003	0.002		0.002	0.001
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0277	0.0524	0.0579	0.0828	0.0402	0.0837	0.0093	0.0027	0.0101	0.1095	0.0877	0.1331
Panel B: GPT with Baseline Prompt												
Same-occupation dummy	0.008**		0.005							0.007*		0.005
	(0.004)		(0.004)							(0.004)		(0.004)
GPT occupation-occupation match		0.012***	0.010***								0.011***	0.008**
		(0.003)	(0.003)								(0.003)	(0.003)
Same-industry dummy				0.017***		0.016***				0.016***		0.015***
				(0.004)		(0.003)				(0.004)		(0.004)
GPT industry-industry match					0.011***	0.002					0.010***	0.001
					(0.004)	(0.003)					(0.004)	(0.003)
Duncan major-occupation match							0.006**		0.007**	0.005*		0.005
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match								0.001	-0.001		0.000	-0.002
								(0.002)	(0.002)		(0.002)	(0.002)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0277	0.0441	0.054	0.0828	0.0381	0.0833	0.0093	0.0004	0.0095	0.1095	0.0759	0.1298
Observations	100,260	100,260	100,260	100,260	100,260	100,260	100,260	100,260	100,260	100,260	100,260	100,260
Baseline control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Refer to Table 4.

Appendix Table A6: Robust Check IV—Using Claude 3 Haiku (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Claude 3 Haiku with Baseline Prompt												
Same-occupation dummy	0.005		-0.001							0.004		-0.002
	(0.004)		(0.004)							(0.003)		(0.003)
Claude 3 Haiku occupation-occupation match		0.022***	0.022***								0.021***	0.021***
		(0.003)	(0.003)								(0.003)	(0.003)
Same-industry dummy				0.044***		0.040***				0.043***		0.037***
				(0.005)		(0.005)				(0.005)		(0.005)
Claude 3 Haiku industry-industry match					0.012***	0.004**					0.011***	0.004**
					(0.003)	(0.002)					(0.003)	(0.002)
Duncan major-occupation match							0.004		0.002	0.003		0.001
							(0.003)		(0.003)	(0.003)		(0.003)
Claude 3 Haiku major-title match								0.007***	0.006**		0.006**	0.007**
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0120	0.2085	0.209	0.1477	0.0545	0.1541	0.0046	0.0153	0.0170	0.1586	0.2689	0.3579
Panel B: GPT with Baseline Prompt												
Same-occupation dummy	0.005		0.000							0.004		-0.001
	(0.004)		(0.003)							(0.003)		(0.003)
GPT occupation-occupation match		0.016***	0.016***								0.016***	0.015***
		(0.004)	(0.004)								(0.004)	(0.004)
Same-industry dummy				0.044***		0.044***				0.043***		0.042***
				(0.005)		(0.005)				(0.005)		(0.005)
GPT industry-industry match					0.011**	-0.000					0.010**	-0.000
					(0.005)	(0.002)					(0.005)	(0.002)
Duncan major-occupation match							0.004		0.003	0.003		0.001
							(0.003)		(0.003)	(0.003)		(0.003)
GPT major-title match								0.003	0.002		0.002	0.002
								(0.003)	(0.003)		(0.003)	(0.003)
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0120	0.1104	0.1104	0.1477	0.0310	0.1477	0.0046	0.0036	0.0062	0.1586	0.1388	0.2469
Observations	90,780	90,780	90,780	90,780	90,780	90,780	90,780	90,780	90,780	90,780	90,780	90,780
Baseline control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Refer to Table 4.

Appendix Table A7: An Artificial Example Illustrating How GPT Provides Additional Information to the Traditional Duncan Index Measure of Major–Occupation Mismatch

Majors	Occupations Applied For	Proportion of Applicants (%)	Duncan Major-Occupation Index	GPT Major-Title Match
Specialized major	O1	91	1	1
	O2	3	0.1	0
	O3	3	0.1	0
	O4	3	0.1	0
Versatile major	O1	25	0.4	1
	O2	25	0.4	1
	O3	25	0.4	1
	O4	25	0.4	1
Unprepared major	O1	25	0.4	0
	O2	25	0.4	0
	O3	25	0.4	0
	O4	25	0.4	0

Appendix B: Why We Don't Use More Complex Prompts

In our main analysis, we employ a relatively simple prompt to produce different matching measures, akin to approaches used in existing literature (e.g., Eloundou et al. (2023)). We refrain from using more complex prompting strategies, like multivariable prompting and Chain of Thought (CoT) prompting, due to the tendency of Language Models (LLMs) to either focus solely on one dimension or demonstrate indiscriminate agreement with more complex prompts (Huang et al., 2023).

Complex prompting strategies appear to encourage the GPT to think more comprehensively in ways akin to humans. However, they may introduce biases into the GPT's responses. Multivariable prompting involves incorporating additional information about individual characteristics (such as work experience, education level) and job characteristics (such as industry category, occupation category, ownership type of hiring firms) into the prompts. Providing more detailed information about workers and jobs helps humans obtain more accurate answers. Nonetheless, we have observed that in our experiment, the GPT tends to generate responses heavily reliant on a single dimension, particularly working experience.

CoT prompting is an approach aimed at opening the black box of LLMs by requiring the GPT to address questions step by step. In Appendix Figure A2, we illustrate an example of constructing the major–occupation match measure using CoT prompting. Before offering an answer regarding whether a worker who graduated in "Management science" is capable of performing the job "Sales of automotive parts," the GPT is prompted to first complete specific job applicant characteristics from the worker's major and job description from the job title. The GPT concludes that the worker "probably cannot" be capable for that job because the knowledge from the management science major is only partially related to automotive parts sales. In this example, CoT prompting demonstrates effectiveness. However, we have found that the GPT tends to provide inconsistent answers in very similar scenarios. This inconsistency arises because CoT prompts the GPT to first generate a multivariable setting (e.g., adding information on possible knowledge an applicant might possess), which consequently leads to inconsistent answers because of the same reason as multivariable prompting.

Using a conceptual framework of how LLMs work as a generating forward-looking sequences, we illustrate potential issues associated with complex prompting strategies. When evaluating the match between workers and jobs, the GPT predicts the likelihood of a sequence of words appearing in a given context, drawing from a vast corpus of text data. We conceptualize a sequence in the latent space as an ordered array $[x_1, x_2, ..., x_k]$, originating from a learned concept set $X_1, ..., X_n$. Within this framework, the core function of a model is to predict the subsequent token x_{k+1}, based on a prior sequence $[x_1, x_2, ..., x_k]$. This prediction mechanism is quantitatively expressed as estimating the probability

$$P(x_{k+1} = X_i | x_1, x_2, ..., x_k).$$

A generative model's training aims to minimize the vector distance between the empirically observed distribution $\hat{P}(x_{k+1} = X_i | x_1, x_2, ..., x_k)$ and the model's generated distribution $P(x_{k+1} = X_i | x_1, x_2, ..., x_k)$. An efficient model under this framework excels at minimizing this distance, thus proficiently generating tasks where inference sequences resemble those encountered in its training. Conversely, the model's effectiveness tends to diminish when faced with scenarios where the inference sequences significantly deviate from the training data. The first task is named as "interpolation tasks," and the second as "extrapolation tasks."

The framework's architecture justifies our decision to avoid complex prompting strategies such as multivariable prompting or CoT prompting. Introducing additional conditioning elements, whether through CoT prompting or augmented conditioning variables, potentially reduces the likelihood of encountering analogous sequences in the training data. This increases the risk of the model being applied to extrapolation tasks rather than interpolation tasks, thereby impairing the precision of the model-generated conditional probability distribution. Adding more conditions may lead to a decrease in model performance since the data used by the model for generating $P(x_3 = X_i|x_1, x_2)$ is likely to be smaller than the data used for generating $P(x_3 = X_i|x_1)$ and $P(x_3 = X_i|x_2)$. Therefore, it would be preferable for us to calculate $P(x_3 = X_i|x_1, x_2)$ with additional constraints using the generated $P(x_3 = X_i|x_1)$ and $P(x_3 = X_i|x_2)$, rather than directly eliciting $P(x_3 = X_i|x_1, x_2)$ from the model. Note that the structure of neural autoregressive generative models does not inherently imply that they compute $P(x_3 = X_i|x_1, x_2)$ by using information from $P(x_3 = X_i|x_1)$ and $P(x_3 = X_i|x_2)$. Instead, their generated $P(x_3 = X_i|x_1, x_2)$ results from minimizing the distance between the predicted $\hat{P}(x_3 = X_i|x_k)$ and the label $P(x_3 = X_i|x_k)$ during training.

One of our robustness checks supports our argument against using complex prompts. Appendix Table A4 shows that the CoT prompting performs worse than our baseline simple prompt. Panel B column (11) shows that the three GPT measures generated with simple prompts yield an additional R-squared of 0.1163 in accounting for the variation in the predicted wage. The additional R-squared of the CoT prompting is only half that of the simple prompt (0.0618).

Appendix C: More Details on Traditional NLP Methods We Use

In this appendix, we describe the methodology used to construct alternative measures of match quality using traditional text analysis methods: the Bag-of-Words (BoW) method and the Term Frequency-Inverse Document Frequency (TF-IDF) method. These methods provide continuous measures of mismatch between workers and jobs by leveraging the textual information embedded in our dataset. By comparing these methods with our GPT-based measure and traditional categorical measures, we aim to demonstrate their effectiveness and limitations in capturing the nuances of labor market mismatch.

We use the same dataset from Zhaopin.com as described in the main text. The Bag-of-Words (BoW) method is a fundamental text analysis technique that represents text data as numerical vectors based on word frequency counts. In this method, each text is converted into a vector where each element corresponds to the count of a particular word in the text. The similarity between two texts is then calculated using cosine similarity between their corresponding vectors. And the Term Frequency-Inverse Document Frequency (TF-IDF) method enhances the BoW approach by weighting the word counts by the inverse frequency of the words across all documents. This reduces the influence of common words and highlights words that are more unique to a particular text, potentially capturing more meaningful similarities.

In tf-idf vectorization, we then vectorize the

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

where: $\text{TF}(t, d)$ (Term Frequency) is the number of times term t appears in document d divided by the total number of terms in d. And $\text{IDF}(t)$ (Inverse Document Frequency) is calculated as:

$$\text{IDF}(t) = \log\left(\frac{N}{n_t}\right)$$

Computing Similarity:

we compute the cosine similarity between the TF-IDF and the BoW vectors of the applicant's variable and the job's variable using the formula below as the proximated matching measure as it represent the semantic similarities.

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

While BoW and TF-IDF do not capture the relative positional relationship among words (e.g. representation of 'researcher assistant' and 'assistant researcher' are the same according to TF-IDF and BoW), transformer embeddings and LLM prompting (inference from embedding) do capture the relative positional relationship. LLM prompting has further advantages of conditioning ability, making it perform better with personality conditioning than vanilla query [cite persona hub paper here].

Appendix Table C1: Pairwise Correlations between the Traditional, GPT, and Other Textual Analysis Measures

Panel A: Occupation-occupation Match	Same-occupation dummy (1)	GPT occupation-occupation match (2)	TF-IDF occupation-occupation match (3)	BoW occupation-occupation match (4)
Same-occupation dummy	1			
GPT occupation-occupation match	0.355***	1		
TF-IDF occupation-occupation match	0.989***	0.369***	1	
BoW occupation-occupation match	0.969***	0.379***	0.992***	1
Panel B: Industry-industry Match	Same-industry dummy	GPT industry-industry match	TF-IDF industry-industry match	BoW industry-industry match
Same-industry dummy	1			
GPT industry-industry match	0.655***	1		
TF-IDF industry-industry match	0.999***	0.656***	1	
BoW industry-industry match	0.999***	0.656***	1.000***	1
Panel C: Major-occupation Match	Duncan major-occupation match	GPT major-title match	TF-IDF major-title match	BoW major-title match
Duncan major-occupation match	1			
GPT major-title match	0.436***	1		
TF-IDF major-title match	0.035***	0.044***	1	
BoW major-title match	0.036***	0.044***	0.993***	1

Appendix Table C2: Wage Premium of the Major-Occupation Match Based on Various Textual Analysis Methods (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)				
	(1)	(2)	(3)	(4)	(5)
Duncan major-occupation match	0.008***		0.008***		0.008***
	(0.002)		(0.002)		(0.002)
TF-IDF major-title match		-0.000	-0.000		
		(0.000)	(0.000)		
BoW major-title match				-0.000	-0.001
				(0.000)	(0.000)
Basic control	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes
Occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes
Observations	801,107	801,107	801,107	801,107	801,107
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0222	0.0000	0.0222	0.0001	0.0223

Additional References

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2311.05232.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu (2023). A survey on hallucination in Large Language Models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232.