Measuring Labor Market Match Quality with Large Language Models *

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

March 31, 2025

Abstract

We propose a novel method to assess the quality of labor market matches using large language models (LLMs). Our approach tasks LLMs with the role of a human resources specialist to evaluate the fit between an applicant and a job. Our study consists of two parts. First, we validate the LLM approach using an administrative dataset from an online job platform and show that our new measure of match quality correlates positively with traditional measures, yet it can reveal additional information. We also demonstrate the wide applicability of the method with survey data, where traditional measures are often infeasible due to the limited sample size. Second, we provide a concrete example of how our approach generates novel insights by demonstrating how LLMs can be used to assess the versatility of college majors. We show that college majors that can fit a broad range of occupations are unfairly penalized by traditional measures of match quality, which the LLM measure can mitigate.

Keywords: Large Language Models, Categorical Variables, Labor Market Match Quality

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 the 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, University of Hong Kong. Email: zbzhao@hku.hk

1 Introduction

The quality of the match in the labor market, specifically whether a worker or an applicant fits a job, is traditionally difficult to measure because it involves many categorical variables such as college majors, occupations, and industries. Categorical variables are different from ordinal variables (e.g. self-rated health status) or cardinal variables (e.g. temperature, income, or years of schooling), since there is no inherent ordering among these categories. A conventional approach, recommended in classical econometric textbooks to handle categorical variables, involves constructing a set of dummy variables for different categories, known as the fixed-effects approach (FE).

However, the FE approach has two significant limitations. First, it disregards the valuable information contained in the textual labels associated with categorical variables. For example, consider three different occupation categories: “software engineer,” “system tester,” and “sales representative.” Intuitively, the first two categories are more similar to each other than the “sales representative.” However, researchers typically construct three dummy variables for these categories, failing to capture subtle similarities among them. For a concrete example, Papageorgiou (2022) investigates whether larger cities help workers find better matched jobs by using job switching as an indicator of match quality. This dummy variable equals one if the job category in the current period matches that of the previous period. Therefore, his approach does not differentiate“different but similar” jobs from “totally different” jobs. 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 the survey data, which typically have a relatively small sample size. For example, the 2016 and 2018 waves of the China Labor-Force Dynamic Survey (CLDS) have 2,431 employers with a college degree or higher, yet their occupations are coded in 65 detailed categories. As a result, half of these detailed categories have less than 10 observations, and approximately 70 percent have fewer than 30 observations. The small number of observations in each category can make the FE approach highly unstable when analyzing jobs at such a granular level.

In this paper, we assess whether large language models (LLMs) can overcome the limitations of traditional measures and provide additional insights when exploring the quality of the labor market matches. The existing literature presents four *direct* measures of labor market match qualities.[1] The *job-switching* (JS) method examines workers' tendencies to switch jobs, assuming that this results in loss of occupation- and/or industry-specific human capital in the labor market (Kambourov and Manovskii, 2009; Sullivan, 2010; Papageorgiou, 2022). 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; Altonji et al., 2016; 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 is based on evaluations by job analysts that define the 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 enumerate job requirements. The best known example is the O*NET (Occupational Information Network) dataset in the United States. In 2021, the Ministry of Human Resources and Social Security in China also hired job analysts to decide matched majors for all occupations listed in China's National Occupation Classification Code (CNOCC).

The recent development of LLMs presents a novel approach to recovering 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 measuring labor market match quality, which involves various categorical variables such as college majors, occupations, and industries. Specifically, we use LLMs to 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 on other characteristics of the individual and the job to mitigate possible confounders. By processing the textual labels of workers' jobs, the GPT enables us to capture similarities between "software engineer" and "system tester." Moreover, since the GPT is pre-trained on extensive external data sets, this approach is not constrained by sample size (i.e., the GPT can assess the similarity between two categories even if the data have only few observations in each category).

Due to the black-box nature of LLMs, we cannot assume that their answers are reasonable

[1] 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 Miller (1984), Moscarini (2001), Nagypál (2007), Hsieh and Klenow (2009), and Adamopoulos et al. (2022).

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 data sets for this purpose. The first data set comprises application flow data from a leading online job board in China (Zhaopin.com), consisting of 1,048,575 applications to 29,914 unique job postings. This data richness enables us to compute various traditional measures of match quality. 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), as well as the major–occupation match. Each of these three dimensions of match quality is measured using our GPT method in addition to one of the traditional methods mentioned above. Specifically, 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 measures of match quality and the corresponding traditional measures. In addition, we find that, conditional on the characteristics of the applicants and the jobs as well as the traditional measures of match quality for the applicant–job pairs, our GPT measure can still strongly predict an applicant's expected wage.

Our exercise using the Zhaopin.com data reveals a limitation of the JA method. The official major–occupation correspondence in China, recommended by the Ministry of Human Resources and Social Security, is listed in the official code system. However, researchers' data do not always follow this system, and establishing a correspondence between two code systems with hundreds of categories can be a daunting task. Since the Zhaopin.com data do not use the official code system, we cannot directly compare the GPT method with the JA method using this dataset. Therefore, we perform the GPT–JA cross-validation using the second dataset (CLDS), which follows the official code system.

The exercise with the CLDS survey data also demonstrates the wide applicability of our GPT measure of match quality. It is important to note that the information in the CLDS data is not rich enough to implement the traditional measures of match quality, particularly the realized matches method, because many categories (occupation or college major) contain too few observations.

However, the GPT measure remains feasible because the large 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. This exercise broadens the usefulness of the GPT method in survey data. Although high-quality, large administrative data is not always available to researchers, particularly in developing countries, LLMs are more accessible, and many are open source, such as Llama and DeepSeek.

In the next step, we present a concrete application of our GPT measure of match quality in labor economics that researchers can use GPT to gain novel insights. Specifically, we use GPT to explore the versatility of college majors. Consider two extreme hypothetical examples: The first major equips students with generalized skills that can be applied to a wide range of occupations. The second major, in contrast, does not prepare students for any specific occupation. Students in both majors may apply for many different occupations, without focusing on any particular one. Traditional measures of match quality cannot distinguish between these two majors, as students' job application behaviors appear similar. However, our GPT measure can easily differentiate them because LLMs are pre-trained 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 fast-growing literature in economics 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; Horton, 2023; Eloundou et al., 2024). 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. (2024), who evaluated the potential effects of LLMs on the US labor market by employing both humans and GPT-4 as annotators to evaluate the exposure of job tasks to LLMs. Our paper adds to the various roles that can be assigned to the GPT, specifically focusing on measuring the match quality in the labor market, and validates the GPT method in this novel application.

Our study also contributes to the literature on measuring labor market match quality 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 data sets, our method is not constrained by sample size, unlike the realized matches method. Unlike the worker-assessment method, ours does not require data on self-assessed match quality. Compared to the job-analysis method, our approach treats GPT as a job analyst and is thus more cost-effective than employing humans, especially in developing countries. It significantly reduces researchers' workload in processing textual data, which is particularly valuable in developing countries where structured data sets on detailed job requirements (such as O*NET) are often unavailable. Our exercise using GPT as a job analyst costs only $120.[2]

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 such as 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 can make the results more difficult to interpret. This is also the reason why our first step is to validate the GPT measure with traditional measures of match quality.

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 of match quality; in Section 4, we demonstrate the validity and wide applicability of our GPT measure and explain why it can recover overlooked information in categorical variables; in Section 5, we present a concrete application of our GPT

[2] A detailed cost calculation is provided in Section 3.3.

method in labor economics; 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 has not 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 that contains around 500 billion tokens (Brown et al., 2020). Tokens are typically on a sub-word level. LLMs often employ the "SentencePiece" tokenizer (available at `https://github.com/google/sentencepiece`). For example, it would tokenize the word "Powerful" into ["power", "ful"].

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 data sets. During pre-training, the 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 example, given the prompt "The cat is on the,", 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 the 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 in a wide range of natural language processing tasks.[3,4] There is no doubt that AI technology will continue to evolve rapidly. Our study provides a lower bound estimate of the potential usefulness of LLMs in economic research.

The use of LLMs in economic research presents several challenges. First, the answers given by the LLMs may vary between different models and prompts. We will provide a set of robustness check in these dimensions. Second, if the prompt is too simple, LLMs might function as black boxes, potentially limiting the explainability of results. The prompt should also not be overly complex as it can reduce 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[5] 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 measures of match quality. Our data set 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 and highly educated workers looking for 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 selected a random sample of

[3]Zheng et al. (2023) assess all available LLMs in 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 on 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.

[4]We choose GPT-3.5-turbo also for practical reasons, as we need to evaluate nearly one million applicant–job pairs. Execution costs and time were important considerations. As of March 3, 2025, the GPT-4 model is 60 times more expensive and three times slower than the GPT-3.5-turbo. However, we will use a more recent LLM (Claude3-Haiku) as part of our robustness checks.

[5]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).

61,674 job seekers who initiated new job search cycles in August 2013.[6] 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),[7] job location at the city level, number of persons to hire, education and experience 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 college major), marital status, employment status, work experience, and the current city of residence. In addition, 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 location of the job and the expected monthly wage range. We infer the monthly expected wage as the midpoint between the minimum and maximum of the range.[8]

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"

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

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

[8]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.

and “administrative officer/administrative assistant.”

In our analysis, we focus on applicants who hold a college degree or higher because those without such qualifications typically lack a major or have majors incomparable to college graduates. In addition, applicants without 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 classified into 92 detailed categories nested within 12 broad categories, based on “The Undergraduate Major Catalog 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 China. The CLDS data includes 37,623 respondents, of whom 2,431 are employed and hold a college degree or above with major information.[9]

The CLDS data complement the Zhaopin.com data in three distinct ways. First, the CLDS data cover the entire labor market, while Zhaopin.com focuses exclusively on the online job market. Second, the CLDS data, like other survey data, present the realized job matches and the corresponding wage outcomes, whereas the application submission data from Zhaopin.com center on the search process and the expected wages before actual matches. Third, while the Zhaopin.com data allow us to construct traditional measures of match quality 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 adhere to the official classification systems and thus enable a comparison between the job-analysis method and our GPT method.

3.2 Traditional Measures of Match Quality

We introduce three traditional measures of match quality 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 lack information on workers’ self-assessment of their fit for the jobs.

[9]To maintain consistency between the two data sets, we excluded individuals who are self-employed.

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; Papageorgiou, 2022). 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 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 the textual information in the category labels and fails to consider subtle similarities between different categories.

We then examine the major–occupation match using two different methods and data sets. With Zhaopin.com data, we define the match index using the RM method, which derives the matched majors from the actual distribution of the majors within occupations, following Lemieux (2014). 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),[10] 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 difference in the ratio into 1,000 quantiles and scales it further from 0 to 1.[11]

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,

[10]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.

[11]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 indices align with this recognition, revealing significantly larger ratio differences and match indices for these pairs compared to others.

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 to the job title, thus missing important information contained in the job titles, as highlighted by Marinescu and Wolthoff (2020).

Using the CLDS data, we define an alternative major–occupation match measure 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 (CNOCC).[12] The jobs in the CLDS data are categorized according to CNOCC, allowing us to use the JA method. Specifically, the measure of major–occupation matches, 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 example, for the occupation "human resource professionals," there are three relevant majors: "human resource management," "business management," and "administrative management." Therefore, workers with one of these three majors are considered 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 to assess 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 information provided (without considering additional information or assumptions such as educational level, work experience, previous jobs, on-the-job learning or training), assess whether the applicant graduated from [**major title**] is capable of performing [**job title**]. Please respond with "Definitely can" or "Probably can" or "Probably cannot" or "Definitely cannot."

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

It is worth noting that we can extend our analysis beyond occupation categories to job titles because GPT can generate responses for any applicant–posting pair, regardless of the number of observations. Figure 1 provides an example of the responses of the GPT when assessing the quality of the match between the "management" major and a job titled "sales of automotive parts."

We deliberately require GPT to respond with one of four verbal options rather than using a continuous numeric scale (e.g., 0–100). In the following, we provide brief intuitions justifying this choice, as the technical details are beyond the scope of this paper. Because LLMs are trained on a vast corpus of textual data, they excel at capturing associations among textual information but struggle with precise numerical reasoning. A vivid example is their frequent failure to correctly answer simple mathematical questions such as "Which number is higher: 9.11 or 9.9?"[13] As a result, the vast majority of industry applications that use LLMs for evaluation purposes, such as those released on Hugging Face,[14] adopt categorical rather than numerical frameworks (Huang et al., 2024). Further supporting this approach, a recent study that uses GPT-4 to assess the exposure of job tasks to LLMs (Eloundou et al., 2024) also opts for a categorical framework. Notably, two of the four authors of that study are affiliated with OpenAI and are likely well aware of the strengths and limitations of the LLMs developed by their company, which reinforces the validity of our approach.

We define the GPT measure of the major–occupation match as an indicator of whether the response is "definitely can" or "probably can." In the reminder of this paper, we consistently use a dummy variable instead of an ordered variable to ease comparisons with traditional measures used in our study and the existing literature.[15] We use similar prompts to develop measures for the occupation–occupation 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 three significant advantages of the GPT method. First, it allows us to capture textual information from category labels that is often overlooked in traditional measures. Com-

[13] `https://towardsdatascience.com/9-11-or-9-9-which-one-is-higher-6efbdbd6a025/`, accessed March 3, 2025.

[14] Hugging Face is a leading machine learning and data science platform and community, known for its open-source tools like the Transformers library and the Hugging Face Hub, which enable users to build, share, and deploy machine learning models, particularly in natural language processing.

[15] In a later robustness check, we evaluate the intensive margin from "probably can" to "definitely can."

pared 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 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 "system tester" category probably can also perform jobs in the "software 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 the "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 infrequently in the data. This feature of the no-category size requirement significantly broadens the scope of the GPT method, making it applicable to survey data with relatively small sizes, such as CLDS data.

Third, the GPT method is cost-effective, especially compared to the JA method. With the rapid advancement of AI technology, the cost of using LLMs has been rapidly decreasing and is expected to decrease further. For our primary LLM (GPT-3.5-Turbo), the current API price is \$0.50 per million input tokens and \$1.50 per million output tokens.[16] Our method asks GPT to evaluate 60,646 unique occupation–occupation pairs, 2,465 industry–industry pairs, and 311,944 major–title pairs. The total cost is approximately \$120.[17]

[16] `https://platform.openai.com/docs/pricing`, accessed March 3, 2025.

[17] The number of tokens per query typically does not exceed 200 for both input and output. Constructing the occupation–occupation match measures consumes 10.05 million input tokens and 7.37 million output tokens, costing \$16.08. The industry–industry match measure costs \$0.83 (0.415 million input tokens and 0.416 million output tokens), and the major—title match measure costs \$97.71 (51.81 million input tokens and 47.54 million output tokens).

3.4 Descriptive Statistics

We present the descriptive statistics in Table 2. Consistent with Kuhn and Shen (2013), applicants in the 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 4,457 RMB per month. The workers in the CLDS data are relatively older, with more work experience, a higher monthly wage in their 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 data sets. Specifically, 22% and 26% of the 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 match between major and 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 workers have a major that is matched with their occupation in the CLDS data.[18]

4 Using the GPT to Measure 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 the traditional measures. Second, a testable

[18]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 these studies usually examine matches between the broader major 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.

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 measures to be significantly positive after controlling for the characteristics of the individuals and the jobs.[19]

Table 3 displays the pairwise correlations between the traditional and GPT measures. Panel A shows the results for the Zhaopin.com data. Two notable findings emerge: (1) all correlations are significantly positive; (2) the highest correlations are observed for the 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 measures of match quality to further validate our GPT method—a standard procedure to evaluate the effectiveness of traditional measures of match quality (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 measures of match quality 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 measures of match quality.[20] $X_{i,j,c}$ incorporates the characteristics of the applicants and the jobs; $\varepsilon_{i,j,c}$ is the error term. The characteristics of the applicant consist of years of schooling, work experience (including the square term), gender, birth month dummies, type of school,[21] marital status, employment status and detailed major

[19]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 US (Robst, 2007) and approximately 1% in China (Zhu, 2014).

[20]Without standardization, the coefficient $\beta_1/100$ represents the wage increase associated with a one-percentage-point increase in the corresponding measure of match quality. However, the significance of a one-percentage-point change varies because the average values differ across measures. For example, Table 2 shows that 22% of applications are directed to the same occupation category as the previous job, while the average GPT match is 69%. In this case, a one percentage-point change is less significant for the GPT measure than for the traditional measure.

[21]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. In this initiative, significant

FEs. The characteristics of the job include the educational and experience requirements, the type of firm ownership,[22] firm scale,[23] detailed occupation FE, industry FE, and city FE of the applied job. For each dimension, we perform three regressions. The first two regressions include the traditional measure and the GPT measure separately, while the third integrates both measures. Standard errors are two-way clustered at the levels of the detailed occupation category and the industry category of the applied jobs.

We present the results in Table 4. Column (1) shows that a one standard-deviation increase in the occupation–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 standard-deviation 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 standard-deviation increase in measures of match quality. 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 standard-deviation increase in the match index. The GPT major-title match indicator suggests a significant but smaller coefficient (about 0.5 percent for a one standard-deviation increase in the the measures of match quality) for the major–occupation match (column 8), and the coefficient becomes statistically insignificant when controlling for the Duncan major–occupation match index (column 9).

Appendix Table A3 further validates GPT's responses by partially examining the intensive margin. The baseline specification defines an applicant–job pair as "matched" if GPT deems that the applicant "definitely can" or "probably can" perform the job. To assess the intensive margin, we split the matching dummy variable into two separate dummies—one indicating that

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.

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

[23]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.

an application "definitely can" perform the job and the other that it "probably can." Appendix Table A3 shows that for occupation–occupation and industry–industry matches, the coefficients for "definitely can" are larger than those for "probably can" (columns 1–2). For the major–occupation match, the two coefficients are statistically similar (column 3).

Taken together, Table 4 and Appendix Table A3 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 explained previously, 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 standard-deviation 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 measures of match quality are controlled, as indicated in columns (2) and (4). These findings suggest that the additional information provided by the 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 by the traditional measures.

Next, we show that traditional natural language processing methods, such as bag-of-words (BoW) and term frequency–inverse document frequency (TF-IDF) is not useful in recovering the overlooked association among the labels of different categories. Appendix C shows that traditional methods perform poorly in establishing correlations among different categories (see Appendix Table C2). 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 takes advantage of extensive pre-training on textual data from the Internet to establish correlations between short phrases.

In summary, the key advantage of our GPT method lies in its ability to capture conceptual similarities across labels in different categories. Therefore, its best application is for categorical variables with many similar labels. A corollary of this reasoning is that the added value of GPT increases with the number of categories, as a larger number of categories leads to smaller average conceptual gaps between them. Appendix Table A4 shows that our GPT method performs similarly to the job-switching method when using broad occupation categories but outperforms it when using detailed categories (Table 4). However, this finding does not undermine the usefulness of the GPT method, as the appropriate level of categorization depends on the specific application.

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. 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 LLMs available. 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 have more local knowledge about the Chinese labor market. Claude 3 Haiku, developed by Anthropic, is the second largest LLM startup (after OpenAI).[24] 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 A5–A8 present the results of the robustness checks. For practical reasons, we choose a 10% random sample from Zhaopin.com.[25] Panel A of the Appendix Tables A5 and A6 shows 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, more complex CoT prompting does not show stronger predictive power, suggesting that overly complicated prompts do not necessarily yield better outcomes.

Panel A of Appendix Table A7 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 A8 uses the more recent Claude 3 Haiku. We find that the predictive power of the measures generated by Claude 3 Haiku generally outperforms the GPT-3.5-turbo. For example, a one standard-deviation 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 CLDS data to demonstrate the broad applicability of our GPT measure of match quality and offer supplementary evidence of its validity. Our focus is on assessing the most demanding major–occupation matches. Although our GPT measure of the major–occupation matches does not outperform the Duncan match index (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 measure does not impose any requirements on the sample size.

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

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

Table 6 presents the results of the Mincer regressions that examine the predictive power of various measures of match quality on actual wages using the CLDS data.[26] 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 the 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 standard-deviation increase in the measure of major–occupation match based on the traditional JA method or the GPT method is estimated to increase the monthly wage by approximately 5.6–5.8 percent. When both JA and GPT measures are included in column (4), the coefficient for the traditional JA measure decreases and becomes insignificant, while the GPT measure remains positive and stable. These findings further validate the GPT measure and highlight its broad applicability across different datasets and contexts.

In addition to overall better performance in wage prediction, other advantages of our GPT method should be emphasized compared to the traditional JA method. First, the GPT measure treats the GPT itself as the job analyst, which is much more cost-effective than the traditional JA method that employs 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.,[27] and the European Commission.[28] Therefore, our GPT measure of match quality could be particularly valuable for comprehending labor market matches in developing countries, where employing an adequate number of human job analysts can be either infeasible or too costly.

Second, the GPT method can significantly reduce the workload of researchers processing tex-

[26]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 the current job of the employees instead of the applied job. Consequently, we do not have the information about the required education and experience of the job in CLDS. In addition, the classification systems for occupations and industries are different from those of the Zhaopin.com data. Given a relatively small sample size of the CLDS data, we only consider two types of firm ownership: SOEs and others.

[27]The O*NET does not directly provide matched majors. Instead, it provides information on the fields of knowledge required for each occupation. Those interested in studying the major–occupation matches can follow Yakusheva (2010) to first construct crosswalks 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.

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

tual 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. For example, Nordin et al. (2010) and Domadenik et al. (2013) established the correspondence by comparing detailed descriptions of major categories with those of occupation categories in documents that outline 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 easily applied to any classification system.

5 Application: Measuring the Versatility of Majors with GPT

The previous section demonstrates that the GPT can be a powerful tool to measure match quality in the labor market. In this section, we demonstrate a concrete application of our GPT method in generating novel insights for economic research. We show how GPT can measure the versatility of academic majors, a dimension often overlooked or unfairly penalized by traditional measures of match quality.

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 with 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 measures of match quality. The job-switching method interprets transitions between occupations or industries as indications 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 A9 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,[29] 0.4, and 0.4, respectively. Although the second and third majors share the same Duncan 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 the broad and detailed major levels. In Panel A, the medicine major appears in the upper right corner, indicating that the students' applications are concentrated in a limited set of occupations, and the 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 our GPT measure. Of particular interest are majors in the lower right corner, such as management science. On the one hand, these majors demonstrate highly diversified application patterns compared to specialized fields such as medicine, yet GPT assesses their graduates as broadly qualified for most jobs they apply for. On the other hand, people who major in management science display a similar degree of diversification in their applications compared to majors such as law, education, and agriculture. However, GPT considers management science graduates to be a better fit for their applications than graduates from the other three majors. Management science is one example what we referred as "versatile major." Our classification aligns with the findings of Leighton and Speer (2020), who use a more complex approach to measuring the specificity of the major.[30] Their study shows that

[29]This number averages the Duncan index at the major level. Specifically, $0.919 = 1 \times 91\% + 0.1 \times 3\% + 0.1 \times 3\% + 0.1 \times 3\%$.

[30]They define major specificity as the transferability of graduates' skills across occupations. In versatile majors, the performance gap between the best- and worst-suited occupations is relatively small. To measure this transferability, they calculate a modified Gini coefficient for each major using data on the average incomes of its graduates in different occupations.

accounting, often considered a "specific" subfield within management science, actually provides graduates with highly versatile skills.

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.[31] Figure 3 illustrates the gap between the 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 are 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-effects approach to handle categorical variables, thus overlooking the textual information associated with different categories that could capture "similarities" between them. The emergence of LLMs provides a viable approach to address this issue.

We employ LLMs to measure the quality of matches in the labor market. Specifically, we task a large language model with 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. We emphasize three main findings. First, our GPT measure is highly correlated with traditional 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 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

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

the same occupation/industry, the GPT method takes into account the "similarity" between the two jobs. Lastly, the GPT method is easy and inexpensive to apply. Unlike the job-switching method and the realized matches 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.

We provide a concrete example of the use of the GPT method to study the labor market. We demonstrate how GPT can assist in measuring the versatility 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 measures of match quality interpret such behavior as indicative of poor worker–job match. The GPT approach mitigates this issue by bypassing the reliance on observed behavioral data. Instead, it leverages external information, including extensive textual data from the entire Internet.

In summary, using GPT to analyze labor market match quality, we effectively overcome several limitations associated with traditional measures of match quality. 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 data sets or informative official documents on the labor market. Our research marks an initial step in the integration of 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 data sets 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.

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.

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 (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science 384*(6702), 1306–1308.

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.

Huang, J.-t., W. Wang, E. J. Li, M. H. Lam, S. Ren, Y. Yuan, W. Jiao, Z. Tu, and M. R. Lyu (2024). Who is ChatGPT? benchmarking LLMs' psychological portrayal using psychobench. arXiv preprint arXiv:2310.01386.

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.

Leighton, M. and J. D. Speer (2020). Labor market returns to college major specificity. *European Economic Review 128*, 103489.

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.

Papageorgiou, T. (2022). Occupational matching and cities. *American Economic Journal: Macroeconomics 14*(3), 82—132.

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 Measure of Major–Occupation Match using GPT

[用户]
假设你是一名HR专家，仅根据提供的信息（不考虑其他信息或假设，例如教育水平、工作经验、以前的工作、在职学习或培训），评估是否‘管理学’毕业的申请人能够胜任‘汽车零部件销售’。请回答‘一定是’或‘大概是’或‘大概不是’或‘一定不是’。

[User]
Pretend that you are an HR specialist. Based solely on the information provided (without considering additional information or assumptions such as educational level, work experience, previous jobs, on-the-job learning or training), assess whether the applicant graduated from ‘Management’ is capable of performing ‘Automotive Parts Sales’. Please respond with “Definitely can” or “Probably can” or “Probably cannot” or “Definitely cannot”.

A. Prompt of the Question (Chinese and English Translation)

[GPT]
大概是

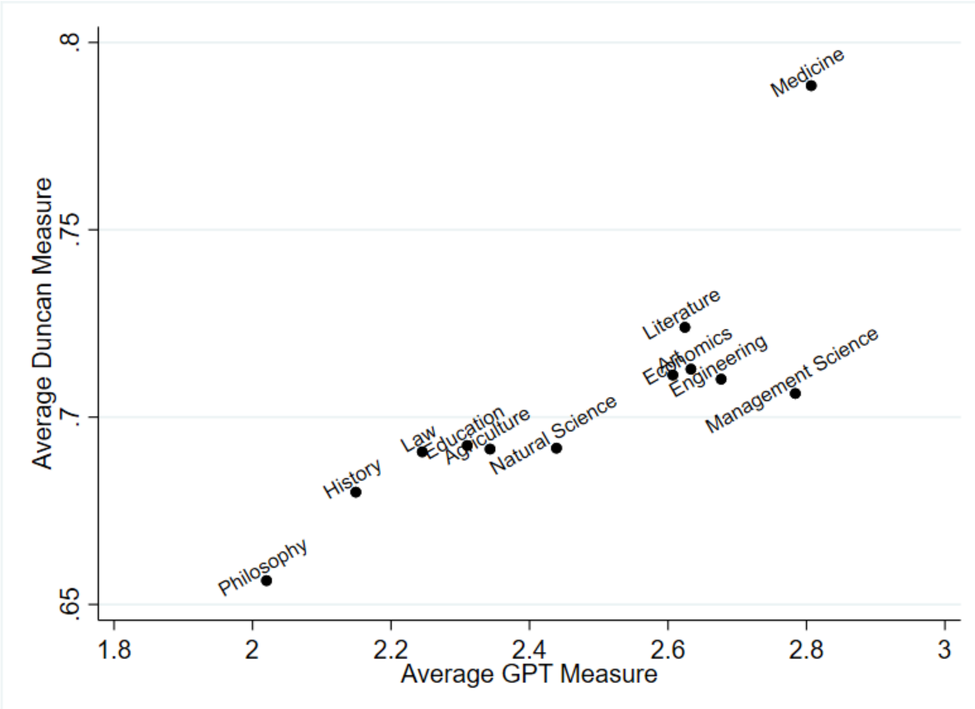
理由：
管理学毕业的申请人通常具备一定的组织管理、沟通和市场分析能力，这些技能在汽车零部件销售中可能会有用。然而，能否完全胜任还需要结合其具体的行业知识、产品理解和销售经验。

[GPT]
Probably Can

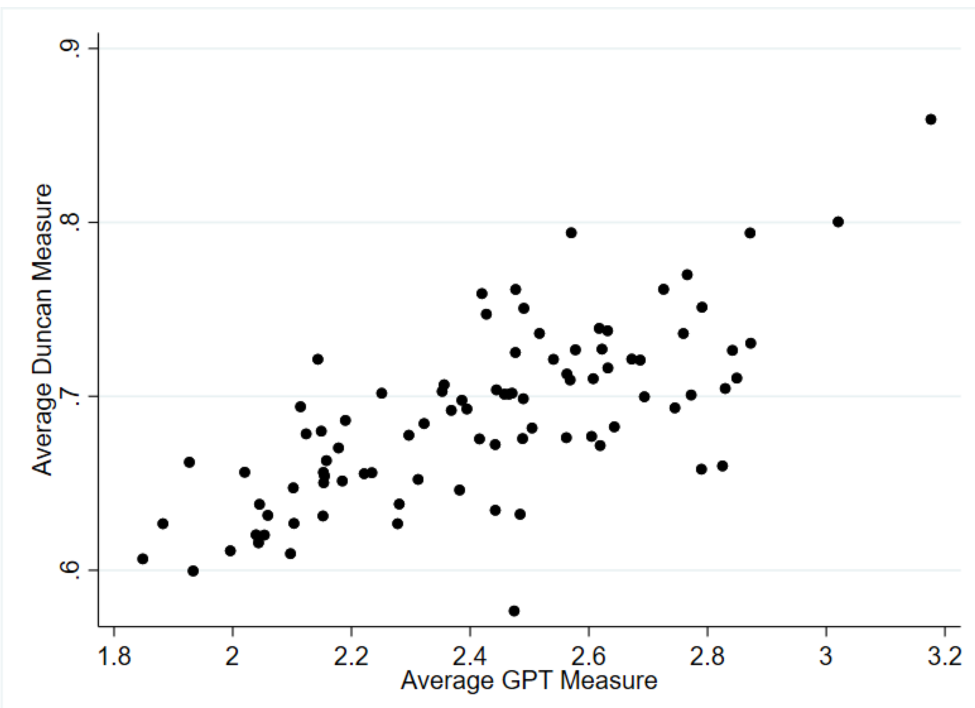
Reason:
Graduates in management typically possess skills in organization, communication, and market analysis, which could be useful in automotive parts sales. However, their ability to fully perform the role would also depend on specific industry knowledge, product understanding, and sales experience.

B. Answer (Chinese and English Translation)

Figure 2: Average Measures of Major–Occupation Matches by Major Categories



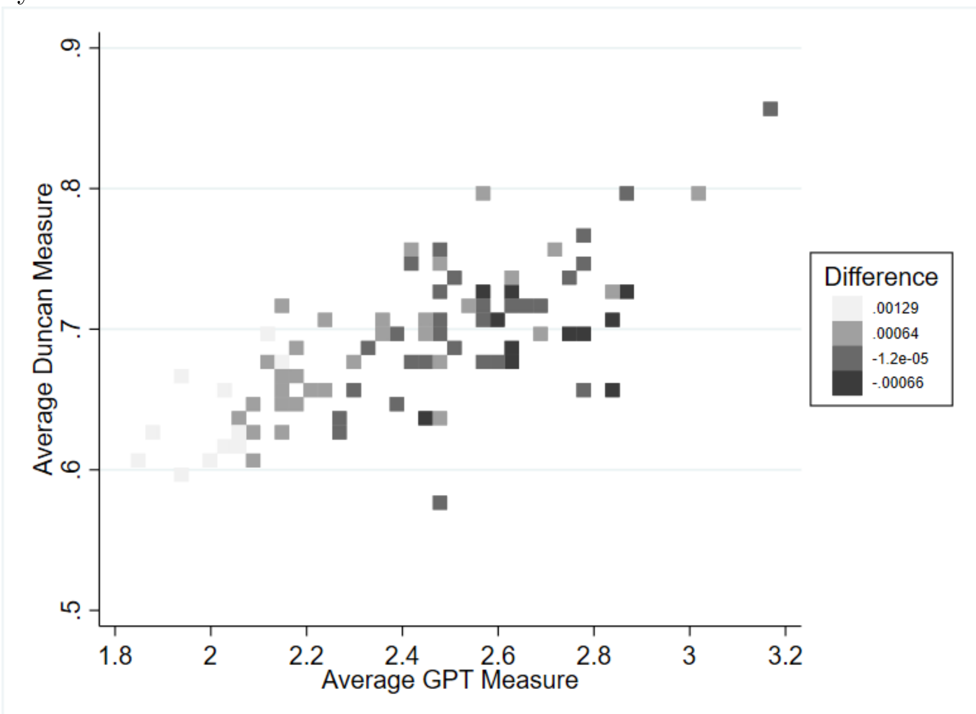
A. By Broad Major Categories



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 measures of major–occupation matches for 12 broad major categories and 92 detailed categories, respectively.

Figure 3: Differences in Predicted Wages When Omitting the Versatility Information of Majors Provided by GPT



Notes: This figure presents the average differences in predicted wages when using only the Duncan index of major–occupation match (equation 2), compared to results that also control for both the GPT measure of major–title match 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 engineer	Software engineer	1	Probably can	1
	System tester	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	S.D.	Obs.	Mean	S.D.	Obs.
	(1)	(2)	(3)	(4)	(5)	(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 measures of match quality, 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 focus on the working subsample, excluding individuals with zero wages as well as those with wages below the 1st percentile or above the 99th percentile.

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

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

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 Quality (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
	(0.004)		(0.004)							(0.003)		(0.004)
GPT occupation-occupation match		0.017***	0.016***								0.015***	0.013***
		(0.003)	(0.003)								(0.003)	(0.003)
Same-industry dummy				0.022***		0.017***				0.020***		0.015**
				(0.005)		(0.006)				(0.005)		(0.006)
GPT industry-industry match					0.018***	0.008***					0.016***	0.007**
					(0.003)	(0.003)					(0.003)	(0.003)
Duncan major-occupation match							0.008***		0.008***	0.006***		0.005***
							(0.002)		(0.002)	(0.002)		(0.002)
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	794,461	794,461	794,461	743,568	743,568	743,568	801,107	801,107	801,107	712,586	712,586	712,586
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0312	0.1039	0.1089	0.1428	0.1030	0.1553	0.0222	0.0076	0.0232	0.1744	0.1860	0.2397

Notes: This table presents the wage premium of the applicant–posting match quality 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 three traditional measures, three GPT measures, and all six measures, respectively. The baseline control variables comprise individual characteristics (gender, dummies for birth month, years of schooling, type of school, marital status, employment status, work experience, and its square term) and posting characteristics (education and experience requirements, and ownership type and scale of the hiring firms). Standard errors are two-way clustered at the levels of detailed occupation category and industry category of the applied jobs. * 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***		0.015***	
	(0.003)	(0.003)		(0.003)	
Same-industry dummy		0.018**	Omitted	Omitted	
		(0.007)			
GPT industry-industry match		0.007**	**0.008*** **	0.006**	
		(0.003)	(0.003)	(0.003)	
Duncan major-occupation match		0.004**		0.007***	0.007***
		(0.002)		(0.001)	(0.002)
GPT major-title match		0.002		0.001	0.002
		(0.001)		(0.001)	(0.001)
Duncan major-occupation match×					-0.002***
GPT major-title match					(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	612,717	545,836	531,498	508,099	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, work experience, and work experience squared) and posting characteristics (education and experience requirements, and ownership type and scale of the hiring firms). Standard errors are two-way clustered at the levels of detailed occupation category and industry category of the applied jobs. * 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
	(0.009)			(0.017)
JA major-occupation match		0.056***		0.037
		(0.017)		(0.021)
GPT major-title match			0.058***	0.048**
			(0.016)	(0.018)
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 match quality of major–occupation matches measured by the traditional and GPT methods using the CLDS data. The basic control variables includes gender, years of schooling, work 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 two-way clustered at the levels of detailed occupation category and industry category of current jobs. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix A: Additional Tables and Figures

Appendix Figure A1: An Example of Constructing the Measure of Major–Occupation Match 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，仅根据提供的信息（不考虑其他信息或假设，例如教育水平、工作经验、以前的工作、在职学习或培训），评估是否‘管理学’毕业的申请人能够胜任‘汽车零部件销售’。请回答‘一定是’或‘大概是’或‘大概不是’或‘一定不是’。

[User]
Assume you are a career advisor for job seekers. Based solely on the information provided (without considering additional information or assumptions such as educational level, work experience, previous jobs, on-the-job learning or training), assess whether the applicant graduated from 'Management' is capable of performing 'Automotive Parts Sales'. Please respond with "Definitely can" or "Probably can" or "Probably cannot" or "Definitely cannot".

A. Prompt of the Question (Chinese and English Translation)

[GPT]
大概是

理由：管理学毕业的申请人通常具备组织管理、市场营销、沟通协调和商业策略等方面的知识，这些能力对汽车零部件销售非常有帮助。然而，能否完全胜任还可能取决于具体岗位的技术要求（如对汽车零部件的专业知识）。

[GPT]
Probably Can.

Reason: Management graduates typically possess knowledge in areas such as organizational management, marketing, communication, and business strategy, which are highly relevant to automotive parts sales. However, their ability to fully meet the requirements may also depend on the specific technical demands of the position (e.g., specialized knowledge of automotive parts).

B. Answer (Chinese and English Translation)

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

[用户]
假设你是一名HR专家，仅根据提供的信息（不考虑其他信息或假设，例如教育水平、工作经验、以前的工作、在职学习或培训），评估是否‘管理学’毕业的申请人能够胜任‘汽车零部件销售’。请逐步思考补全与职位对应的职位描述和求职者特征，仔细分析再回答‘一定是’或‘大概是’或‘大概不是’或‘一定不是’。

[GPT]

1. 职位描述分析：汽车零部件销售

- 核心职责：
1. 了解汽车零部件产品的技术特点及应用场景。
2. 面向汽车制造商、修理厂或经销商推广、销售产品。
3. 建立并维护客户关系，确保销售目标的达成。
4. 分析市场需求和竞争动态，制定销售策略。
5. 跟踪订单执行，协调交付及售后问题。
- 所需技能和特质：
- 技术理解力：能够快速学习并理解汽车零部件相关知识。
- 沟通与谈判能力：清晰传递价值主张并与客户达成协议。
- 市场分析能力：评估行业趋势并制定有效的销售策略。
- 客户导向：以客户需求为中心，维护长久合作。
- 基础商业知识：包括定价、成本核算及供应链管理。

2. 求职者背景分析：管理学毕业

- 管理学的典型学习内容：
- 商业管理基础：市场营销、企业战略、运营管理、供应链管理。
- 沟通与协调能力：组织行为学、人力资源管理、谈判技巧。
- 数据分析：市场分析、财务报表解读等。
- 实用知识：项目管理、时间管理、客户关系管理。
- 匹配度分析：
1. 技术理解力：大概不是
- 管理学专业课程通常不包含汽车零部件的技术知识。申请人可能需要额外学习产品知识。
2. 沟通与谈判能力：大概是
- 管理学的学习内容通常重视沟通技巧和客户管理，符合销售岗位的需求。
3. 市场分析能力：大概是
- 管理学课程中涉及市场营销及竞争分析，这与销售岗位要求的市场分析能力匹配。
4. 客户导向：一定是
- 客户关系管理课程培养了与客户沟通和满足客户需求的能力，符合岗位需求。
5. 基础商业知识：一定是
- 管理学教育通常包括定价、成本分析等知识，与销售职责一致。

3. 综合判断

虽然申请人缺乏技术背景，但管理学专业提供了足够的商业与沟通技能，可以满足岗位的核心要求（尤其是销售策略、客户管理和市场分析）。如果申请人愿意通过自学或入职培训补充汽车零部件知识，胜任岗位的可能性较高。

4. 最终结论：

大概是

A. Prompt of the Question and Answer in Chinese

[User]
Pretend that you are an HR specialist. Based solely on the information provided (without considering additional information or assumptions such as educational level, work experience, previous jobs, on-the-job learning or training), assess whether the applicant graduated from 'Management' is capable of performing 'Automotive Parts Sales'. Please think through the position description and the candidate's characteristics step by step, analyze carefully, and then answer with "Definitely can" or "Probably can" or "Probably cannot" or "Definitely cannot".

[GPT]
Job Description Analysis: Automotive Parts Sales
Core Responsibilities:

1. Understand the technical features and application scenarios of automotive parts.
2. Promote and sell products to automotive manufacturers, repair shops, or dealers.
3. Build and maintain customer relationships to ensure sales targets are met.
4. Analyze market demand and competitive dynamics to develop sales strategies.
5. Monitor order execution, coordinate delivery, and address after-sales issues.

Required Skills and Attributes:

- **Technical Understanding:** Ability to quickly learn and comprehend knowledge related to automotive parts.
- **Communication and Negotiation Skills:** Clearly convey value propositions and reach agreements with customers.
- **Market Analysis Skills:** Evaluate industry trends and develop effective sales strategies.
- **Customer Orientation:** Focus on customer needs and maintain long-term cooperation.
- **Basic Business Knowledge:** Including pricing, cost calculation, and supply chain management.

Candidate Background Analysis: Management Graduate
Typical Curriculum in Management Studies:

- **Business Management Fundamentals:** Marketing, corporate strategy, operations management, supply chain management.
- **Communication and Coordination Skills:** Organizational behavior, human resources management, negotiation techniques.
- **Data Analysis:** Market analysis, financial statement interpretation.
- **Practical Knowledge:** Project management, time management, customer relationship management.

Matching Analysis:

1. **Technical Understanding:** Likely not.
   - Management courses generally do not cover the technical aspects of automotive parts. Additional learning might be needed.
2. **Communication and Negotiation Skills:** Likely yes.
   - Management studies emphasize communication skills and customer management, aligning with the role's requirements.
3. **Market Analysis Skills:** Likely yes.
   - Courses in marketing and competitive analysis align well with the role's need for market analysis.
4. **Customer Orientation:** Definitely yes.
   - Customer relationship management courses develop skills in addressing customer needs and fostering collaboration.
5. **Basic Business Knowledge:** Definitely yes.
   - Management education often includes pricing, cost analysis, and related knowledge, aligning with the job's duties.

Comprehensive Evaluation:
Although the candidate lacks a technical background, a management degree provides strong business and communication skills, which meet the core requirements of the role (especially in sales strategy, customer management, and market analysis). If the candidate is willing to learn automotive parts knowledge through self-study or on-the-job training, they are likely to perform well in this role.
Final Conclusion:
Probably can

B. Prompt of the Question and Answer (English Translation)

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

Appendix Table A3: Validating the GPT Method Through Intensity Margin (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)		
	(1)	(2)	(3)
GPT occupation-occupation match	0.052***		
(Definitely can)	(0.008)		
GPT occupation-occupation match	0.028***		
(Probably can)	(0.006)		
GPT industry-industry match		0.048***	
(Definitely can)		(0.012)	
GPT industry-industry match		0.026***	
(Probably can)		(0.006)	
GPT major-title match			0.008**
(Definitely can)			(0.004)
GPT major-title match			0.011***
(Probably can)			(0.003)
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	794,461	743,568	801,107

Notes: This table differs from Table 4 in that it splits the matching dummy variable into two separate dummy variables—one indicating that an application "definitely can" perform the job and the other that it "probably can." The empirical specification remains the same. Standard errors are two-way clustered at the levels of detailed occupation category and industry category of the applied jobs. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A4: Wage Premium of the Applicant-Posting Match Using Broad Occupation Category (Zhaopin.com Data)

Dependent Variable	Monthly Wage of Expected Job (Log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Same-occupation dummy	0.015***		0.007			
	(0.005)		(0.010)			
GPT occupation-occupation match		0.015***	0.010			
		(0.006)	(0.010)			
Duncan major-occupation match				0.012***		0.011***
				(0.003)		(0.002)
GPT major-title match					0.006**	0.001
					(0.003)	(0.002)
Basic control	Yes	Yes	Yes	Yes	Yes	Yes
Major category FE	Yes	Yes	Yes	Yes	Yes	Yes
Broad occupation category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry category of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes
City of applied job FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	794,075	794,075	794,075	801,108	801,108	801,108
Additional R-squared for controlling match measure ($\times 10^{-2}$)	0.0722	0.0798	0.0878	0.0455	0.0101	0.0459

Notes: This table differs from Table 4 in that it uses broad categories (58 in total) instead of detailed categories (588 in total), while maintaining the same specification. Standard errors are two-way clustered at the levels of broad occupation category and industry category of the applied jobs. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix Table A5: 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)		(0.006)							(0.004)		(0.006)
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***
				(0.003)		(0.007)				(0.003)		(0.007)
GPT industry-industry match					0.017***	-0.005					0.016***	-0.005
using alternative prompt					(0.003)	(0.006)					(0.003)	(0.006)
Duncan major-occupation match							0.005**		0.004*	0.004		0.002
							(0.002)		(0.002)	(0.002)		(0.002)
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***
				(0.003)		(0.005)				(0.003)		(0.005)
GPT industry-industry match					0.016***	0.003					0.014***	0.003
					(0.003)	(0.004)					(0.002)	(0.004)
Duncan major-occupation match							0.005**		0.005**	0.004		0.003
							(0.002)		(0.002)	(0.002)		(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 A6: 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)		(0.005)							(0.004)		(0.005)
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)		(0.005)				(0.004)		(0.005)
GPT industry-industry match using CoT prompt					0.008*	-0.001					0.008*	-0.001
					(0.004)	(0.005)					(0.004)	(0.005)
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
								(0.002)	(0.002)		(0.002)	(0.002)
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 A7: 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
	(0.005)		(0.006)							(0.005)		(0.006)
ERNIE Bot occupation-occupation match		0.012***	0.010***								0.012***	0.009**
		(0.003)	(0.004)								(0.003)	(0.004)
Same-industry dummy				0.017***		0.016***				0.016***		0.014***
				(0.003)		(0.004)				(0.003)		(0.004)
ERNIE Bot industry-industry match					0.012***	0.002					0.011***	0.002
					(0.003)	(0.004)					(0.003)	(0.004)
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
								(0.002)	(0.002)		(0.002)	(0.002)
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
	(0.005)		(0.005)							(0.005)		(0.005)
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**
				(0.003)		(0.006)				(0.003)		(0.006)
GPT industry-industry match					0.011***	0.002					0.010***	0.001
					(0.003)	(0.005)					(0.003)	(0.005)
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)	(0.001)		(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 A8: 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)							(0.004)		(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***
				(0.003)		(0.004)				(0.003)		(0.004)
Claude 3 Haiku industry-industry match					0.012*	0.004*					0.011*	0.004*
					(0.007)	(0.003)					(0.006)	(0.002)
Duncan major-occupation match							0.004		0.002	0.003		0.001
							(0.003)		(0.003)	(0.003)		(0.004)
Claude 3 Haiku major-title match								0.007***	0.006**		0.006***	0.007**
								(0.002)	(0.002)		(0.002)	(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)							(0.004)		(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***
				(0.003)		(0.004)				(0.003)		(0.003)
GPT industry-industry match					0.011	-0.000					0.010	-0.000
					(0.009)	(0.002)					(0.008)	(0.002)
Duncan major-occupation match							0.004		0.003	0.003		0.001
							(0.003)		(0.003)	(0.003)		(0.004)
GPT major-title match								0.003	0.002		0.002	0.002
								(0.002)	(0.003)		(0.002)	(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 A9: An Artificial Example Illustrating How GPT Provides Information on the Versatility of a Major

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 Do We Not Use More Complex Prompts?

In our main analysis, we employ a relatively simple prompt to produce different matching measures, similar to the approaches used in the existing literature (e.g., Eloundou et al. (2024)). 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 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 the characteristics of individuals (such as work experience, education level) and jobs (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. However, we have observed that in our experiment the GPT tends to generate responses heavily reliant on a single dimension, particularly work experience.

CoT prompting is an approach that aims to open the black box of LLMs by requiring the GPT to address the questions step by step. In Appendix Figure A2, we illustrate an example of constructing the measure of major–occupation match 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 the specific characteristics of the job applicant from the worker's major and the job description from the job title. The GPT concludes that the worker "probably cannot" be capable of this job because the knowledge from the management science major is only partially related to sales of automotive parts. 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 textual data. We conceptualize a sequence in the latent space as an ordered array $[x_1, x_2, ..., x_k]$, originating from a set of learned concepts $X_1, ..., X_n$. 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 generated distribution of the model $P(x_{k+1} = X_i | x_1, x_2, ..., x_k)$. 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, either through CoT prompting or augmented conditioning variables, potentially reduces the likelihood of encountering analogous sequences in the training data. This increases the risk that the model is applied to extrapolation tasks rather than interpolation tasks, thus impairing the precision of the model-generated conditional probability distribution. Adding more conditions may lead to a decrease in model performance, as the data used by the model to generate $P(x_3 = X_i|x_1, x_2)$ 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 A6 shows that CoT prompting performs worse than our baseline simple prompt. Column (11) in Panel B shows that the three GPT measures generated with simple prompts produce 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

In this appendix, we provide additional details on constructing alternative measures of match quality using traditional NLP methods. We focus on two commonly used approaches in the literature: the bag-of-words (BoW) method and the term frequency–inverse document frequency (TF-IDF) method. Both methods utilize the textual information in our dataset, measuring the quality of the match based on the similarities between the textual labels of the categorical variables.

In the BoW method, we convert text into numerical vectors based on word frequencies. For example, in analyzing major–occupation matches, the textual labels of majors and occupations are transformed into numerical vectors, where each element represents the count of a specific word in the corresponding text. We then compute the cosine similarity between these vectors to measure match quality.[1] The formula is:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|},$$

where $\mathbf{A}$ and $\mathbf{B}$ represent the corresponding vectors of two texts. A cosine similarity of 1 indicates that the texts are identical in their word distribution, signifying a strong match between the major and occupation. This method heavily favors major–occupation pairs with highly similar textual labels, such as "Accounting" (*Kuaiji xue* in Chinese) and "Accountant" (*Kuaiji* in Chinese), while strongly penalizing pairs with less similarity in their labels.

The TF-IDF method improves on the BoW method by weighting word frequencies based on their significance within the entire corpus. It assigns higher weights to rare words, calculated as the logarithm of the inverse of their frequency, while reducing the impact of common words, such as stopwords. This adjustment emphasizes words that are unique to a specific text, allowing the TF-IDF method to identify more meaningful and distinctive similarities between texts. However, both the BoW and TF-IDF methods have limitations in capturing contextual meanings and word relationships. For example, they cannot effectively differentiate between phrases such as "researcher assistant" and "assistant researcher."

The BoW (BoW) and TF-IDF methods perform poorly in assessing the quality of the match. As shown in the Appendix Table C1, the correlations between the measures of match quality derived from NLP methods and traditional measures (JS and RM methods) are either near one or close to zero. This occurs because NLP methods rely solely on exact text equivalence and do not account for the underlying similarity of meanings in textual labels. Consequently, when analyzing major–occupation matches, the cosine similarities of the two NLP methods are only approximately 0.002, as exact text matches are rare. This explains why NLP measures do not offer predictive power in Mincer regressions, as indicated in the Appendix Table C2.

[1] Cosine similarity is the cosine of the angle between two vectors.

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 (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science 384*(6702), 1306–1308.

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.