

Gmyrek, Paweł; Lutz, Christoph; Newlands, Gemma

Working Paper

A technological construction of society: Comparing GPT-4 and human respondents for occupational evaluation in the UK

ILO Working Paper, No. 102

Provided in Cooperation with:

International Labour Organization (ILO), Geneva

Suggested Citation: Gmyrek, Paweł; Lutz, Christoph; Newlands, Gemma (2024) : A technological construction of society: Comparing GPT-4 and human respondents for occupational evaluation in the UK, ILO Working Paper, No. 102, ISBN 978-92-2-040327-3, International Labour Organization (ILO), Geneva, <https://doi.org/10.54394/UQOQ5153>

This Version is available at:

<https://hdl.handle.net/10419/283512>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



► A Technological Construction of Society

Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK

Authors / Paweł Gmyrek, Christoph Lutz, Gemma Newlands





This is an open access work distributed under the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>). Users can reuse, share, adapt and build upon the original work, as detailed in the License. The ILO must be clearly credited as the owner of the original work. The use of the emblem of the ILO is not permitted in connection with users' work.

Attribution – The work must be cited as follows: Gmyrek, P., Lutz, C., Newlands, G. *A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK*. ILO Working Paper 102. Geneva: International Labour Office, 2024.

Translations – In case of a translation of this work, the following disclaimer must be added along with the attribution: *This translation was not created by the International Labour Organization (ILO) and should not be considered an official ILO translation. The ILO is not responsible for the content or accuracy of this translation.*

Adaptations – In case of an adaptation of this work, the following disclaimer must be added along with the attribution: *This is an adaptation of an original work by the International Labour Organization (ILO). Responsibility for the views and opinions expressed in the adaptation rests solely with the author or authors of the adaptation and are not endorsed by the ILO.*

This CC license does not apply to non-ILO copyright materials included in this publication. If the material is attributed to a third party, the user of such material is solely responsible for clearing the rights with the right holder.

Any dispute arising under this license that cannot be settled amicably shall be referred to arbitration in accordance with the Arbitration Rules of the United Nations Commission on International Trade Law (UNCITRAL). The parties shall be bound by any arbitration award rendered as a result of such arbitration as the final adjudication of such a dispute.

All queries on rights and licensing should be addressed to the ILO Publishing Unit (Rights and Licensing), 1211 Geneva 22, Switzerland, or by email to rights@ilo.org.

ISBN 9789220403266 (print), ISBN 9789220403273 (web PDF), ISBN 9789220403280 (epub), ISBN 9789220403297 (mobi), ISBN 9789220403303 (html). ISSN 2708-3438 (print), ISSN 2708-3446 (digital)

<https://doi.org/10.54394/UQOQ5153>

The designations employed in ILO publications, which are in conformity with United Nations practice, and the presentation of material therein do not imply the expression of any opinion whatsoever on the part of the ILO concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its frontiers.

The responsibility for opinions expressed in signed articles, studies and other contributions rests solely with their authors, and publication does not constitute an endorsement by the ILO of the opinions expressed in them.

Reference to names of firms and commercial products and processes does not imply their endorsement by the ILO, and any failure to mention a particular firm, commercial product or process is not a sign of disapproval.

Information on ILO publications and digital products can be found at: www.ilo.org/publns

ILO Working Papers summarize the results of ILO research in progress, and seek to stimulate discussion of a range of issues related to the world of work. Comments on this ILO Working Paper are welcome and can be sent to research@ilo.org, gmyrek@ilo.org.

Authorization for publication: Richard Samans, Director of RESEARCH

ILO Working Papers can be found at: www.ilo.org/global/publications/working-papers

Suggested citation:

Gmyrek, P., Lutz, C., Newlands, G. 2024. *A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK*, ILO Working Paper 102 (Geneva, ILO). <https://doi.org/10.54394/UQOQ5153>

Abstract

Despite initial research about the biases and perceptions of Large Language Models (LLMs), we lack evidence on how LLMs evaluate occupations, especially in comparison to human evaluators. In this paper, we present a systematic comparison of occupational evaluations by GPT-4 with those from an in-depth, high-quality and recent human respondents survey in the United Kingdom. Covering the full ISCO-08 occupational landscape, with 580 occupations and two distinct metrics (prestige and social value), our findings indicate that GPT-4 and human scores are highly correlated across all ISCO-08 major groups. In absolute terms, GPT-4 scores are more generous than those of the human respondents. At the same time, GPT-4 substantially under- or overestimates the occupational prestige and social value of many occupations, particularly for emerging digital and stigmatized occupations.

Our analyses show both the potentials and risks of using LLM-generated data for sociological and occupational research. Potentials include LLMs' efficiency, cost effectiveness, speed, and accuracy in capturing general tendencies. By contrast, there are risks of bias, contextual misalignment, and downstream issues, for example when problematic and opaque occupational evaluations of LLMs may feed back into working life, thus leading to potentially problematic technological constructions of society. We also discuss the policy implications of our findings for the integration of LLM tools into the world of work.

About the authors

Pawel Gmyrek is a Senior Researcher in the Research Department of the ILO.

Christoph Lutz is a Professor at the Department of Communication and Culture, BI Norwegian Business School (Oslo).

Gemma Newlands is a Departmental Research Lecturer in AI & Work at the Oxford Internet Institute, University of Oxford.

Table of contents

Abstract	01
About the authors	01
<hr/>	
▶ Introduction	06
<hr/>	
▶ 1 Literature Review	08
Occupational Evaluation	08
Generative AI in Scientific Research	09
<hr/>	
▶ 2 Methods	11
Occupation list	11
Measurement of occupational prestige and social value	11
Survey Sample	12
Generation of GPT-4 scores	13
<hr/>	
▶ 3 Analysis	14
GPT-4 raw scores and comparison with survey scores	14
Comparison of standardised z-scores	16
Occupations with significantly diverging scores by humans and GPT-4	20
Demographic profiles in comparison to GPT-4 scores	25
Prompt adjustments for underrepresented groups	28
<hr/>	
▶ Discussion and Conclusion	29
Summary of Findings	29
Research and Policy Implications	30
Limitations and recommendations for future research	32
<hr/>	
Appendix	33
A1. Survey questionnaire	33
A2. Robustness tests concerning non-unimodal distribution of GPT-4 scores	34
A2.1. INDIVIDUAL SCORES FOR EACH OCCUPATION	34
PRESTIGE-RELATED TESTS	34
OSV-RELATED TESTS	36
CONCLUSION REGARDING OCCUPATION-LEVEL SCORES	37
A2.2. MEAN SCORES ACROSS OCCUPATIONS	37

PRESTIGE-RELATED TESTS	37
OSV-RELATED TESTS	39
CONCLUSION REGARDING MEAN SCORES ACROSS OCCUPATIONS	41
A3. STABILITY OF GPT-4 PREDICTIONS OVER TIME	42
References	43
Acknowledgements	50

List of Figures

Figure 1: Multi-modal distribution of GPT-4 scores at the level of occupations - a sample	14
Figure 2: Demographic partitions of the survey contrasting GPT-4 OSV with respondent scores	15
Figure 3: Correlations between GPT-4 generated survey OP scores across ISCO-08 major groups	16
Figure 4: Correlations between GPT-4 generated and survey OSV scores across ISCO-08 major groups	17
Figure 5: Detection criterion for occupations with strongest deviation between GPT-4 and survey	20
Figure 6: Occupations with the largest OP difference between GPT-4 and survey scores	21
Figure 7: Occupations with the largest OSV difference between GPT-4 and survey scores (>0.5 SD)	22
Figure 8: Regression summary statistics for survey OP scores on GPT-4 OP scores by demographic group	25
Figure 9: Human respondents with highest difference of opinions: top and bottom 20 occupations	26
Figure 10: Human respondents: occupations with highest difference of scores	27
Figure 11a-b: Comparison of predictions with generic GPT-4 prompt and group-adjusted prompt	28
Figure A1: Screenshot of survey question	33
Figure A2. Density plot of OP scores for occupations with statistically different mean and median	36
Figure A3. Density plot of occupation-level score distributions, with marked mean and median	37
Figure A4. Comparison of traditional (mean-based) and robust median-based z-scores for OP	38
Figure A5. Scatter plot of z-scores using mean vs. median for OP	39
Figure A6. Comparison of traditional (mean-based) and robust median-based z-scores for OSV	40
Figure A7. Scatter plot of z-scores using mean vs. median for OSV	41
Figure A8. Stability of GPT-4 OP score predictions (June 2023 vs November 2023)	42

List of Tables

Table 1: GPT-4 justifications for occupations with extremely high OP and OSV scores	18
Table 2: GPT-4 justifications for occupations with extremely low OP and OSV scores	19
Table 3: Exemplary GPT-4 justifications for occupations with extreme deviations to human scores	23
Table A1. Mean-median differences (OP)	34
Table A2. Wilcoxon Signed-Rank Test Results (OP)	34
Table A3. Correlation Test Results (OP)	34
Table A4. Cohen's d Effect Size (OP)	35
Table A5. Overall Descriptive Statistics of Mean-Median Differences for GPT Prestige Scores	37
Table A6. Wilcoxon Signed-Rank Test Results for Z-Scores (OP)	38
Table A7. Cohen's d Effect Size (OP)	38
Table A8. Overall Descriptive Statistics of Mean-Median Differences for GPT OSV Scores	40
Table A9. Wilcoxon Signed-Rank Test Results for Z-Scores	40
Table A10. Cohen's d Effect Size	40

▶ Introduction

The occupational landscape in the United Kingdom has experienced profound changes in recent decades, partly driven by cutting-edge digital technologies such as artificial intelligence (AI). These technologies, while impacting job-quality overall, have led to the emergence of new occupations¹ (Adamson and Roper 2019; Kane, 2017; Makridis and Han 2021). The digital transformation has also raised important questions about the social evaluations of these new and evolving occupations, including their prestige and social value (Newlands and Lutz 2023b).

Occupational prestige is deeply rooted in sociological discourse, including not only working conditions and pay but also subjective interpretations of one's work and occupation (Lambert and Bihagen 2014). It is significant because sociological research underscores the enduring importance of one's occupation in shaping identity, self-esteem, and well-being (Adler 2021; Cardone et al. 2021). Occupational evaluations are influenced by the media and societal narratives, as well as assumptions about an occupation's characteristics, rewards, and exclusivity (Lissitsa et al. 2017; Mejia et al. 2021).

The predominant measurement approach involves surveys with human participants to variably capture prestige, social standing, social status, or the social value of occupations. The most frequently employed indices of occupational prestige, for instance, originate from human participants in Treiman's (1977) Standard International Occupational Prestige Scale (SIOPS) and the U.S. General Social Survey (GSS) Occupational Prestige Module (Smith and Son 2014)². Ongoing research reveals a certain consensus in how societal groups evaluate occupations, though with growing awareness that such evaluations are dynamic and subject to between-group variation (Avent-Holt et al., 2020).

The potential for non-human evaluators, in the form of AI, to replicate or replace human survey respondents in constructing occupational hierarchies forms a compelling avenue for further inquiry. Recently released Large Language Models (LLMs), such as GPT-4, BERT, and Llama, represent a pivotal advancement in AI capabilities, due to their novel capacity of content analysis and creation across unstructured textual, visual, and audio-visual media. Emerging literature demonstrates that new LLMs not only process data, but also form discernible opinions on various facets of society, reflecting and perpetuating societal biases (Argyle et al. 2023a, 2023b; Rosenbusch et al. 2023; Sætra 2023). Understanding how LLMs construct social realities in the context of work and occupational stratification could assist in updating long-established and institutionalized measurements of socio-economic schemas (Gil-Hernández et al. 2023). At the same time, insights into the embedded societal and occupational biases within AI models are also needed.

In examining whether and how LLMs can be used in occupational evaluation, this paper further embeds itself in nascent discussions into the use of LLMs in social science research, both as a methodological tool and as an empirical data source (Alkaissi and McFarlane 2023; Argyle et al. 2023a; Balmer 2023; Hämäläinen et al. 2023; Else, 2023; Hepp et al. 2023; Mills et al. 2023; Thorp, 2023). In that area, an ongoing inquiry pertains to the potential substitution of human respondents with LLMs in various research scenarios (Dillion et al. 2023; Harding et al. 2023). These investigations entail a comparative assessment of outputs generated by LLMs against responses provided by human subjects, serving to evaluate the 'algorithmic fidelity' of these models (Argyle et al. 2023a). The preliminary findings exhibit considerable promise, demonstrating that when appropriately prompted within a contextual framework, LLMs are capable of approximating human perceptions and performance to a remarkable degree (Rosenbusch et al. 2023).

¹ The term 'occupation' is a framework for grouping analogous jobs (Hauser and Warren 1997). We follow the definition of Lee et al. (2000) of occupations as 'identifiable and specific lines of work' (p. 800).

² Despite both being described as prestige studies, Treiman (1977) collates broad measurements beyond 'prestige' and the GSS (Smith and Son 2014) in fact directly measures 'social standing'.

As a contribution to this emerging research field, we juxtapose human evaluations of occupational prestige (OP) and occupational social value (OSV), derived from a recent, high-quality survey in the UK, with algorithmic evaluations generated by one of the most currently advanced LLMs: GPT-4. The primary objective is to systematically compare synthetic and human occupational evaluations. By doing so, we aim to discern the disparities and commonalities, and enhance our comprehension of the dynamics at play.

After a literature review of both occupational evaluation and the use of generative AI in scientific research, we provide an in-depth analysis of the OP and OSV perceptions of GPT-4, contrasting the LLM's scores with those of human respondents. This allows us to evaluate the performance of GPT-4 in capturing societal sentiments regarding the world of work, including areas of strength and weakness. In the final sections of our paper, we provide a range of recommendations from our analysis for policy makers interested in the integration of LLM-based tools in the world of work, and for researchers planning to use of LLMs as an analytical tool for occupational evaluations.

▶ 1 Literature Review

Occupational Evaluation

Occupations serve as crucial units for examining social inequality and stratification, reflecting one's position in society (Kim and Mueller 2011; Leicht 2020; Chan and Goldthorpe 2007). The social significance of occupations leads to intense evaluative processes, where individuals assess and judge occupations using various criteria, both consciously and subconsciously (Freeland and Hoey 2018; Lynn and Ellerbach 2017; Sengupta et al. 2009; Valentino 2020). These evaluations are relational and emotional, contributing to the construction and attribution of specific occupational images by others (Duemmler et al. 2020).

Literature on occupational evaluation often interchanges terms like prestige, social standing, and social status, leading to conceptual ambiguities. Occupational prestige serves as a reflection of the collective norms and values embedded in a particular society or community (Zhou 2005). However, few studies have explicitly measured 'prestige', thus contributing to the overall difficulty of conducting consistent and comparative occupation-level research. Despite its methodological drawbacks and age, modern research continues to rely on Treiman's (1977) Standard International Occupational Prestige Scale (SIOPS) as a key variable (e.g., Buchs et al. 2017; Chudnoyskaya and Kashap 2020; Lersch et al. 2020; Oesch and Piccitto 2019; Pitt and Zhu 2019). In addition, the U.S. General Social Survey (GSS) Occupational Prestige Module (Smith and Son 2014) has been adopted and analysed as a more up-to-date prestige index (e.g., Valentino 2020, 2021). Again, however, the GSS does not explicitly measure 'prestige' but 'social standing'.

Occupational *prestige* is influenced by biases, power dynamics, and information availability. Wu and Leffler (1992), for example, found that white men-dominated occupations were perceived as most prestigious, while MacKinnon and Langford (1994) noted morality's influence, secondary to educational requirements. Valentino (2020, 2022) showed that gender- and racially-segregated occupations are seen as more prestigious, with occupations traditionally filled by white men are granted higher prestige. Freeland and Hoey (2018) link high prestige with economic rewards and professional knowledge. Chan and Goldthorpe (2004) categorize occupations by their interaction with abstract symbols, people, or physical labour, resonating with Edwards' (1933) single value dimension model. However, Avent-Holt et al. (2020) argue that prestige is relative to local job context. Sociological studies, like Zhou (2005), Lynn and Ellerbach (2017), and Gauchat and Andrews (2018), reveal that socio-demographic factors like race, education, and gender significantly shape perceptions of occupational prestige, highlighting its context-dependent and multi-faceted nature. Valentino's (2021) research further emphasizes the diversity in occupational evaluations, influenced by various logics including traditionalist and cultural capitalist.

Next to an occupation's prestige, the perceived *social value* constitutes a highly relevant and distinct axis of evaluation that has seen a surging interest, especially in relation to the discourse around 'Bullshit Jobs' initiated by Graeber (2018). While Graeber's definition is informal and imprecise, the core of his argument is that these jobs provide no value to society or sometimes even create negative value. By contrast, research that looked at essential workers during the Covid-19 pandemic (Hennekam et al. 2020; Mejia et al. 2021; ILO 2023) stresses the high value society assigns to these occupations despite being traditionally stigmatized in terms of prestige. Studies that empirically tested Graeber's (2018) arguments have yielded mixed evidence (Soffia et al. 2022; Walo 2023) but show the importance of the underlying concept of an occupation's social value, highlighting that a 'vast majority of workers cares about holding a socially useful job' (Dur and Van Lent 2020, p. 3). For instance, Wolfe and Patel (2019) found that self-employed individuals perceive their jobs as more socially useful, while Fontana (2020) identified a disconnection between the perceived societal usefulness of corporate social responsibility workers in

Japan and their personal sense of usefulness. Furthermore, Dur and Van Lent (2019) discovered that workers with limited autonomy and few opportunities to employ their skills are more likely to perceive their jobs as socially useless.

Disparities in evaluations, both in the form of occupational prestige and occupational social value, thus emerge from diverse societal segments based on age, gender, ethnicity, and socioeconomic status. Variations in evaluations, while reflecting a certain consensus, reveal the complexities of aggregating individual perceptions into societal evaluations. Human respondents remain valuable in this context, but the methods for evaluating and understanding occupational prestige are evolving. Updating long-established and institutionalized measurements, as underscored by Gil-Hernández et al. (2023), is therefore a valuable pursuit.

Generative AI in Scientific Research

Generative AI, in the form of LLMs, has gained rapid traction due to its capabilities of creating high-quality content. In white-collar and knowledge-intensive occupations, such as those in journalism, consulting, marketing, and software development, generative AI has major potential to augment tasks (Eloundou et al. 2023; Gmyrek et al. 2023, Goldman Sachs 2023; McKinsey 2023; WEF 2023)³. However, the rapid adoption of LLMs across domains has raised concerns about misuse, with Chan (2023) and Weidinger et al. (2022) highlighting risks like misinformation and discrimination⁴. The risks of overtrust in such systems, a point raised both by Montemayor (2021) and Aroyo et al. (2021), underscores the importance of understanding their limitations (Floridi and Chiriatti, 2020). Floridi (2023) emphasizes that, while these systems generate coherent text, they lack genuine understanding and there are open questions about whether and how LLMs are aligned with human values (Coeckelbergh and Gunkel 2023; Kasirzadeh and Gabriel 2023). Additionally, LLMs can perpetuate societal biases, as they are trained on data that reflects human prejudices.

Research has started to explore the opinions, attitudes, and predictive capabilities of LLMs in scientific contexts (Salvagno et al. 2023; Thorp 2023), ranging from finance (Bybee 2023) to political attitudes (Argyle et al. 2023a). Dillion et al. (2023) propose three practical applications of LLMs in research: (1) assisting in the formulation and refinement of research hypotheses, (2) piloting test items, and (3) corroborating data obtained from human subjects. The third proposal suggests that LLM outputs can be used as comparative reference points for conventional human-respondent data. The motivation for such a thesis revolves around the idea that a proficient model, trained on an extensive corpus of human text, effectively captures the collective knowledge of countless human minds. However, if a model's output is an expression of a 'modal opinion', derived from the amalgamated human perspectives, the data used for training LLMs represent the output of a specific subset of the human population (Atari et al. 2023; Crockett and Messeri 2023), further curated by application of content moderation tools and safety mechanisms. Claims of representativeness must therefore be approached with caution. Moreover, current techniques for refining LLM performance, such as reinforcement learning with human feedback, further compound this representational challenge (Harding et al. 2023).

Reflecting this growing interest, additional studies have examined the performance of LLMs for capturing human opinions and sentiment (Kalinin 2023). Argyle et al. (2023a), for instance, compared GPT-3-generated data with results from four US-based surveys, showing how GPT-3 was proficient in generating data that closely resembled responses obtained from these surveys. Dominguez-Olmedo et al. (2023) analysed twelve publicly available LLMs of varying sizes, where they were tasked with responding to a representative subset of 25 multiple-choice questions

³ Shestakofsky (2017) previously described such emerging dynamics as 'human-software complementarity', while Jarrahi et al. (2022) discuss the capacities of 'hybrid intelligence' with both human-augmented AI and augmented human intelligence.

⁴ More broadly, potential negative consequences of the use of AI at work point towards the risks of workers' diminished autonomy, increased surveillance and automated decisions (Baiocco et al. 2022; Jarrahi et al. 2021).

from the 2019 *American Community Survey* questionnaire by the US Census Bureau⁵. Similarly investigating the potential for LLMs for data imputation, Kim and Lee (2023) examined the performance of three distinct LLMs fine-tuned by the (US) General Social Survey for the purpose of opinion prediction, demonstrating their strong capabilities. Rosenbusch et al. (2023) used GPT-3 to yield predictions that were on par with the responses of the 600 adults and human experts when a multi-step procedure was implemented, or when the model underwent refinement. Similarly, Bybee (2023) trained GPT-3.5 using data from the *Wall Street Journal* and compared the result to an established survey of financial experts, showing that the model was able to generalize to new information rather than solely memorize previous records.

Researchers have also exposed LLMs to different types of psychometric tests. For instance, Feng et al. (2023) analysed 14 LLMs with regard to their responses to the 62-item *Political Compass Test*, disclosing variations in political leanings among different language models and demonstrating the potential to alter the political inclinations of these models through pre-training corpora.⁶ In Safdari et al. (2023), tests were administered to various LLMs with the aim of identifying personality traits exhibited in their responses.⁷ Hämäläinen et al. (2023) compared open text responses related to video games between human participants and GPT-3, showing how GPT-3 generated text data that was frequently attributed to humans and how the LLM created realistic accounts of Human-Computer Interaction (HCI) experiences.⁸

While a growing field of study, these investigations collectively provide valuable insights into the capabilities and performance of LLMs across diverse domains and their ability to replicate, predict, and generate responses. In that context, comparing LLM and human responses in terms of occupational evaluations may offer further insights into societal norms and biases concerning existing labour structures. Moreover, it is imperative to consider the role LLMs may play in shaping the fabric of society as they continue to find application in functional contexts such as career advice. LLMs may not only represent existing opinions about the world of work more generally but may also harbour inherent opinions pertaining to specific occupations. These viewpoints can potentially become interwoven into other digital technologies, for example in instances where they are embedded into AI systems for recruitment or performance management purposes.

⁵ The output generated by LLMs did not mirror the statistics of the US population and the data generated by LLMs lacked the typical statistical signals found in human survey data, resulting in more uniform distributions.

⁶ GPT model variants exhibited more liberal tendencies, while BERT-type models leaned towards social conservatism or authoritarianism.

⁷ The study indicated that specific LLMs consistently simulated personality traits that were deemed reliable, particularly in the case of larger and fine-tuned models.

⁸ However, the study also highlighted biases in GPT-3 responses, including anomalies such as contradictions, evading questions, and overly long and incoherent responses.

► 2 Methods

Occupation list

We use a list of 576 occupation titles that align with the latest version of the International Standard Classification of Occupations (ISCO-08; see ILO 2008) as well as four non-occupational roles, which were included for control and benchmarking purposes: Unemployed, Retiree, Student, and Homemaker. For both the human-respondent study and the GPT-4 replication study, we utilise the same list of occupation titles for examination. The list was generated by the authors in advance of the human-respondent study, ensuring broad occupational coverage and that every occupation title is matched to exactly one ISCO-08 unit group (436 in total). Conversely, the occupation title list also ensures that every ISCO-08 unit groups is represented by at least one occupation title (see Newlands and Lutz 2023a). The full occupation list also encompasses a core list, with at least one occupation title for each of the 130 ISCO-08 minor groups (3-digit codes), selected to be highly known and institutionalised (e.g., Taxi Driver, Butcher, Hairdresser, Aircraft Pilot, Journalist, Dentist).

For robustness, we conducted a comprehension test to ensure that the occupations assessed are broadly understandable by a British audience. To do so, we collected 800 survey responses on Prolific in December 2021. For this comprehension test, participants were screened for an equal gender distribution (50% male, 50% female) and for residence in the UK. The survey involved an open text task where respondents had to write what they thought someone with the specific occupation title does at work. Participants received £2.50 (guaranteeing minimum average hourly wage of £7.50). Open text responses were coded in Microsoft Excel as a binary of comprehension or no comprehension, with each occupation title receiving 50 responses. Occupation titles with below minimum acceptable comprehension (80%; 10 or more instances of no comprehension) were replaced in iterative cycles until satisfactory completion.

Measurement of occupational prestige and social value

To measure occupational prestige (OP) and occupational social value (OSV) comprehensively, we developed, tested, and used a more scalable approach than previous research. In our data collection, occupational titles are scored on a 0-100 scale with a slider. In two usability and timing pre-tests (55 respondents, done in February 2022), the slider format emerged as user-friendly and intuitive, with no critical feedback. We directly asked respondents in the occupational prestige study: 'For each listed occupation below, please use the slider to indicate how you would rate the prestige of the occupation on a scale of 0 (the lowest level of prestige) to 100 (the highest level of prestige).' Analogical wording was used for the study of social value. We did not explain the concepts of prestige and social value, in order not to prime respondents. All respondents were asked to assess the 134 core titles (130 occupation titles and the four non-occupational roles), as well as 111 or 112 additional occupational titles equally distributed across the ISCO-08 groups (i.e., each respondent evaluated 245 or 246 occupation titles in total).⁹ The occupation titles were displayed in blocks of 10 titles per survey page, where each block contained occupation titles distributed throughout the spectrum of the ISCO-08 unit groups. Within the 10 titles per page, we also randomized the occupation titles. Figure A in the Appendix contains an example survey page from the OSV survey. The randomization and distribution across the spectrum of ISCO-08 unit scores was done to avoid priming effects as much as possible.

⁹ More detailed information about the generation of the occupation list and measurement approach is available in Newlands and Lutz (under review).

Survey Sample

For the recruitment of human survey participants, we relied on Prolific, a UK-based survey sample provider specifically designed for social and behavioural scientists and their needs (Palan and Schitter 2018). We collected occupational prestige and social value assessments between 1 March and 26 March 2022. We used Prolific's representative sample option for the UK, where the platform selects the respondents across age, sex, and ethnicity to mirror the population distribution (Prolific 2023a). The reward for completing the study was £2.50, with a median response time of less than 20 minutes (19 minutes; SD = 15 minutes), which amounted to an hourly wage of more than £7.50.

Occupational prestige and social value were assessed in separate surveys as we did not want the same respondents to score occupations on these two dimensions concurrently to avoid priming effects and to maintain statistical independence. Consequently, the surveys were launched sequentially with screening out for previous participation in any of the earlier data collections (including pre-tests and the comprehension test). We carried out data quality checks and replaced a small number of problematic respondents (e.g., due to unrealistically short response times or extreme straightlining) individually with respondents of the same age group, gender, and ethnicity. 57 individuals were replaced in total (less than 2% of the final sample).

Our final sample size is 2429 respondents, across eight batches of 300+ respondents per batch (1219 respondents for OP and 1210 for OSV). 48.7% of the respondents identify as male (1182 in total), whereas 50.6% identify as female (1230 in total). The remaining 0.7% (17) respondents have a non-male or non-female gender identification. The average age is 44 years, with a standard deviation of 15.6 years. 77% identify as White, 4% as Mixed, 10.5% as Asian, 6.5% as Black, 0.5% as Arab, and 1.5% as Other. In terms of education, 5% have lower secondary education as their highest degree (Fewer than 5 GCSEs/CSEs/GCE O-Levels at grades A*-C); 9% upper secondary school (5 or more GCSEs/CSEs/GCE O-Levels at grades A*-C); 17.5% upper secondary school with A-levels; 6.5% a post-secondary non-tertiary education (Higher National Certificate, Certificate of Higher Education, NVQ Level 4); 6% a short-cycle tertiary education (Higher National Diploma, Diploma of Higher Education, NVQ Level 5); 36% a Bachelor degree or equivalent; 16% a Master degree or equivalent; and 2% a Doctorate or equivalent.

Generation of GPT-4 scores

GPT scores were generated using a Python script that accesses the GPT-4 model through the OpenAI library. The script is organised as a loop of sequential API calls that process each of the 580 occupation titles individually, with an exponential backoff and retry option set to handle any API response errors.¹⁰ We request a written justification of each score in the first round of predictions. The exact prompt used is as follows:

```
{"role": "system", "content": "You are a respondent to a survey in the UK."}
```

```
"Make your response representative of an UK average resident, taking into account the social and demographic diversity of the UK." },
```

```
{"role": "user", "content": "Look at this occupation title: " + str(Occupation_Title) +
```

```
"Rate the prestige of this occupation on a scale of 0 (the lowest level of prestige) to 100 (the highest level of prestige), with intervals of 1. "
```

```
"Make your response representative of an UK average resident, taking into account the social and demographic diversity of the UK. "
```

```
    "In the first line of response provide only your numerical score, no other text. "
```

```
    "In the following lines provide a justification of your scoring decision. "
```

```
    "Do not give any score ranges just one score."}]
```

Once all occupations have been processed, the procedure repeats 49 times, resulting in a total of 50 estimation cycles. Since we do not generate justifications for these remaining 49 cycles, the “user” part of the script is amended:

```
{"role": "user", "content": [...] "In the first line of response provide only your numerical score, no other text. Do not give any score ranges just one score."}]
```

As the prompt shows, the goal was not to get GPT-4’s subjective ‘personal’ beliefs¹¹ but to frame its responses as an average and representative UK resident.

¹⁰ In the case of a failed API call, the system progressively increases the wait time before attempting to resend the request, reducing the risk of overwhelming the service and increasing the likelihood of a successful response on subsequent attempts.

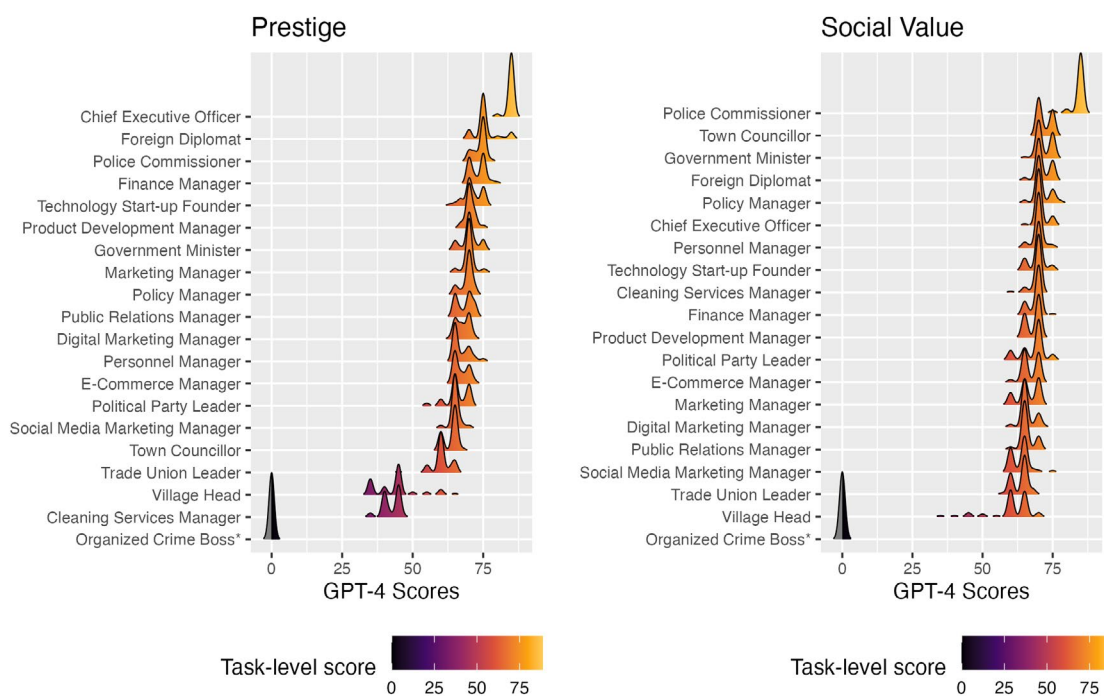
¹¹ OpenAI seems to be increasingly aware of attempts to evaluate LLMs in term of such subjective ‘personal’ opinions and GPT-4 accordingly responds with a disclaimer: ‘As an AI developed by OpenAI, I don’t hold personal opinions or beliefs. My responses are based on the data and programming I have been provided with.’

► 3 Analysis

GPT-4 raw scores and comparison with survey scores

The examination of detailed GPT-4 scores shows that the model's predictions of OP and OSV generally have a very low variance at the level of each occupation. In addition to this reduced dispersion, predictions for many occupations form a bi- or multi-modal distribution (Figure 1). This feature is a by-product of GPT4's consistency of predictions: the model tends to select only a few frequently repeated values for each occupation, as opposed to a more widely distributed range of human scores.

► **Figure 1: Multi-modal distribution of GPT-4 scores at the level of occupations - a sample**



We conduct a range of robustness tests which demonstrate that despite this distribution, the use of mean scores as a measure of central tendency is justified for occupation-level scores.¹² Consequently, for OP and OSV of each of the 580 occupation titles, we calculated their arithmetic mean, standard deviation as well as OP-OSV difference within the survey sample and among the GPT-4 predicted scores. For the GPT-4 scores, we also have written justifications for the first round of predicted scores.

As the first step, we contrasted the average OP and OSV scores per occupation from the survey with the GPT-4 generated scores. GPT-4 scores (averaged across 50 iterations) range from 0 to 87.8. Across all 580 occupations, the OP arithmetic mean is 54.1, which is somewhat higher than the survey responses, where the global OP average was 45.9. The standard deviations for the

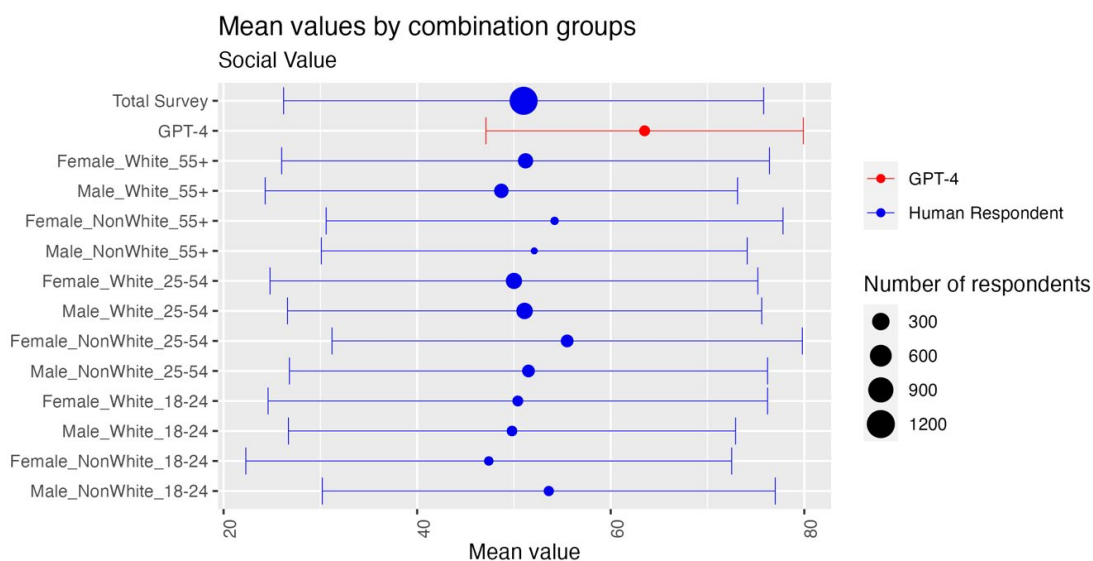
¹² See Appendix, section A1.

GPT-4 generated OP scores range from 0 to 8.9, which is much smaller than the standard deviations in the survey data.

The five occupations that were assigned the highest OP by GPT-4 are in this order: Cardiologist, Judge, Psychiatrist, Robotics Engineer, and General Practitioner. Three of these five occupations are also among the top-5 ranked in terms of their OP in the survey: Cardiologist at rank 1 (81.96), Judge at rank 3 (80.20) and General Practitioner at rank 5 (78.10). Similarly, occupations at the bottom of the OP ranking match as well. GPT-4 gave almost all illicit occupations consistent OP values of 0. By contrast, in the survey the OP values for these occupations were close to 0 but with some variation. For example, Pimp, which came sixth last in the survey still had an OP value of 7.72, while GPT-4 assigned it a straight 0 across all iterations.

For OSV, GPT-4 generated scores range from 0 to 92.1. The global mean across all 580 occupations is 63.5 which is considerably higher than among the survey respondents (48.9). GPT-4 gave the highest OSV values to the following five occupations: Ambulance Paramedic, General Practitioner, Cardiologist, Nurse Practitioner and Women's Shelter Supervisor. Three of these occupations were also among the top-5-rated OSV occupations in the survey: Ambulance Paramedic at rank 1 (83.96), General Practitioner at rank 2 (83.61), and Cardiologist at rank 4 (81.51). The occupations that GPT-4 thought are least valuable are the same as in the survey. Like for OP, the last ranks in both GPT-4's ratings and the respondents' ratings are taken by traditionally stigmatised and/or illicit occupations.

► **Figure 2: Demographic partitions of the survey contrasting GPT-4 OSV with respondent scores**



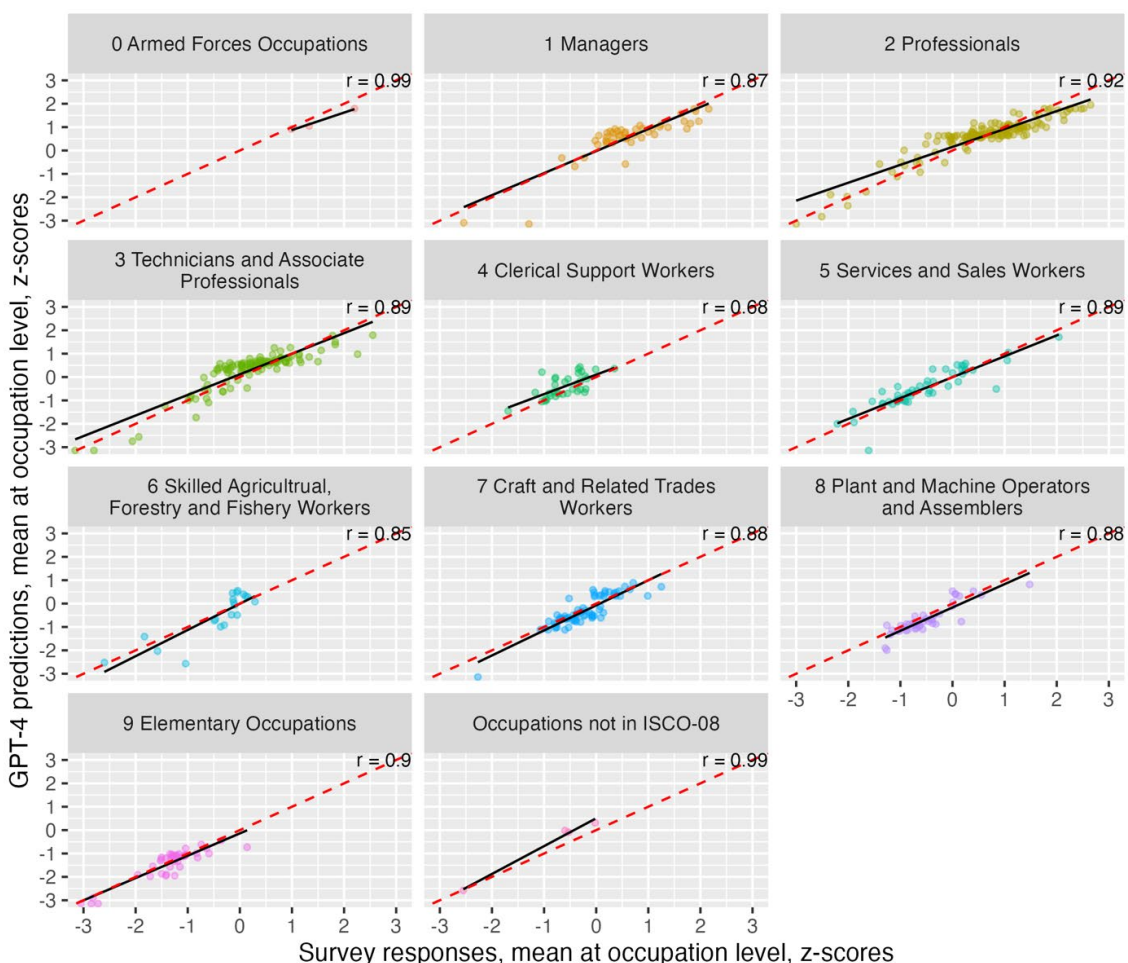
Despite those similarities, a simple visual breakdown across the main demographic characteristics further confirms that GPT-4 provides higher scores across the board, when compared to human scores (Figure 2). Therefore, to facilitate comparisons and interpretability across the datasets, the scores were z-standardized in relation to human and GPT-4 mean separately.¹³

¹³ See Appendix, section A2 for a detailed discussion regarding the impact of the multimodal distribution of the scores predicted by GPT-4 on the standardization procedure.

Comparison of standardised z-scores

Following standardization, the picture in terms of relative OP relations between different occupations matches closely between the survey responses and GPT-4 not only at the extremes but throughout most of the occupational landscape. The additional value of the standardizing procedure is that scores can be more easily compared, as individual values represent the distance of each observation from the overall mean for the survey and GPT-4 scores respectively.

► **Figure 3: Correlations between GPT-4 generated survey OP scores across ISCO-08 major groups**

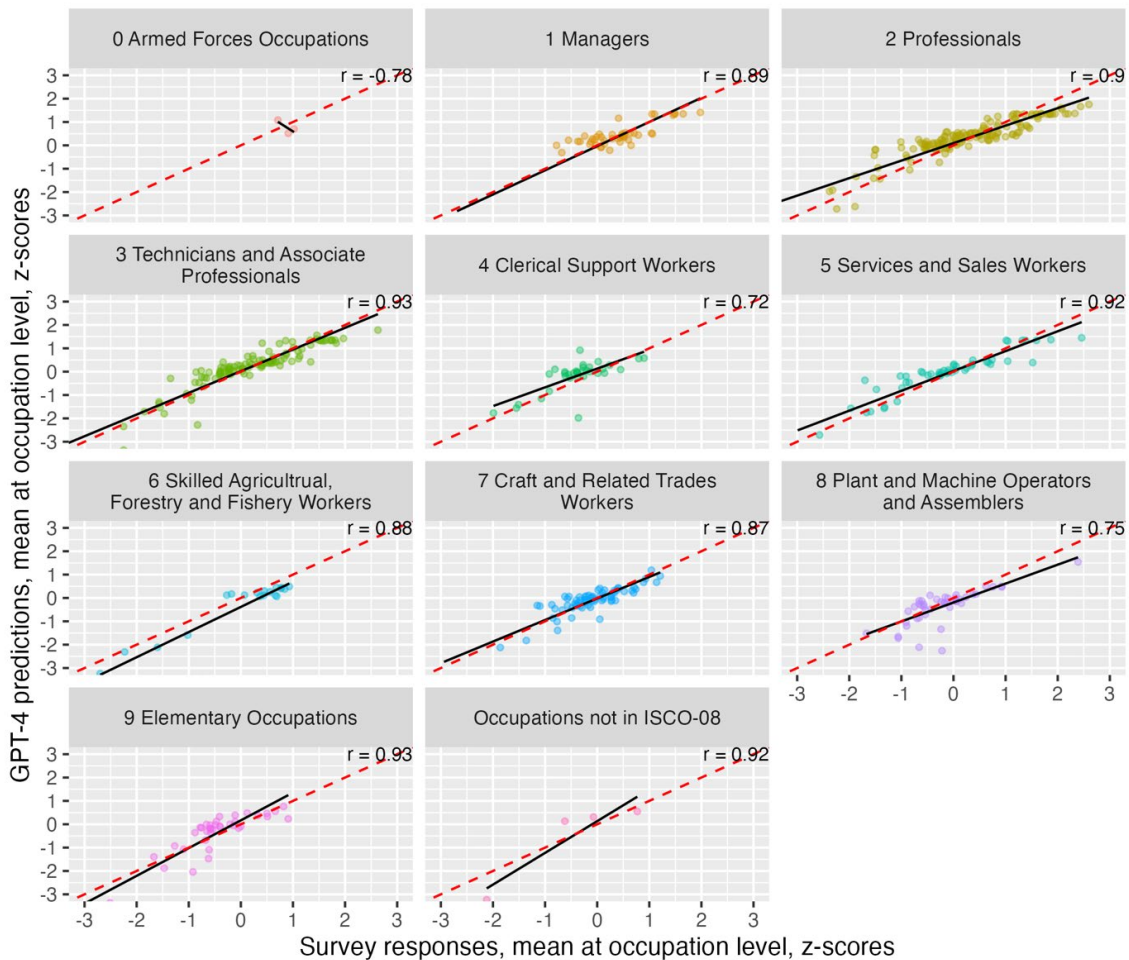


The correlation between the standardized GPT-4 OP means and the standardized survey OP means across all 580 occupations is 0.92, with particularly high correlations within the ISCO-08 major groups 0 Armed Forces Occupations (0.99, although this group only contains 3 occupation titles), 2 Professionals (0.92), 3 Technicians and Associate Professionals (0.89), and 5 Services and Sales Workers (0.89) (see Figure 3). Clerical Support Workers is the only ISCO-08 major group where the correlation is below 0.80 (with a value of 0.68). Thus, GPT-4 is accurate in capturing the relative OP order.

The situation is similar in the case of OSV scores, with the overall correlation between the GPT-4 and survey scores at 0.89. As shown in Figure 4, GPT-4 is particularly accurate in predicting scores similar to those of human respondents in the ISCO-08 major groups 3 Technicians and Associate

Professionals (0.93), 9 Elementary Occupations (0.93), 5 Services and Sales Workers (0.92) and 2 Professionals (0.90). Only two major groups have a correlation below 0.80: 4 Clerical Support Workers (0.72) and 8 Plant and Machine Operators and Assemblers (0.75).

► **Figure 4: Correlations between GPT-4 generated and survey OSV scores across ISCO-08 major groups**



These initial observations indicate that GPT-4 has a strong grasp not only of the OP hierarchy, which might be somewhat expected given that OP is an established sociological concept, but also of the OSV hierarchy, where research is only emerging (Graeber 2018; Soffia et al. 2022; Walo 2023). The justifications GPT-4 provides in the first round of predictions allow some insight into its reasoning.

Cardiologist and General Practitioner are the two occupations that were among the top-5 of both GPT-4 and the survey respondents, both in terms of OP and OSV. Table 1 displays GPT-4's rationale for giving these occupations very high scores. Notice how the explanations tap into economic arguments (years of education needed) as well as socio-cultural ones (importance of NHS). The LLM also seems to have a sense of the heterogeneity in assessments in the population, though without being very concrete.

► **Table 1: GPT-4 justifications for occupations with extremely high OP and OSV scores**

Occupation Title	Occupational Prestige Justification	Occupational Social Value Justification
<p>Cardiologist</p> <p>Human Survey OP Mean Score (81.96)</p> <p>GPT-4 OP Mean Score (87.76)</p> <p>Human Survey OSV Mean Score (81.52)</p> <p>GPT-4 OSV Mean Score (90.08)</p>	<p>“As an average UK resident, a Cardiologist is seen as a highly prestigious profession. This is mainly due to the fact that it requires many years of education and training, and it is a career that saves lives and contributes greatly to the health and well-being of the community. It's also considered that they usually earn a high salary which adds to its prestige. Nonetheless, certain diversities within UK population that may place lesser value on higher education or that prioritize other types of careers could somewhat lower this score from a perfect 100.”</p>	<p>“Cardiologists play a critical role in maintaining public health by diagnosing and treating heart diseases, which are some of the major causes of morbidity and mortality in the UK. Therefore, their occupation holds a high level of social value. They not only contribute to the individual health of patients but also broadly towards public health and wellbeing thereby reducing the burden on the healthcare system. Moreover, it requires high level of skill, knowledge and years of study and training, which further justifies the high rating. This sentiment is likely shared among the diverse UK public who would generally acknowledge the value of healthcare professionals.”</p>
<p>General Practitioner</p> <p>Human Survey OP Mean Score (78.10)</p> <p>GPT-4 OP Mean Score (85)</p> <p>Human Survey OSV Mean Score (83.61)</p> <p>GPT-4 OSV Mean Score (91.54)</p>	<p>“In the UK, General Practitioners (GPs) are highly respected due to their significant role in the healthcare system. They are usually the first point of contact in the health care system and are responsible for diagnosing and treating a wide range of medical conditions, referring patients to specialist services, and promoting healthy lifestyles to prevent ill health. Their job requires extensive education and training which further contributes to the overall prestige of their occupation. Their role also requires a great deal of empathy, compassion, and communication skills, traits that are often highly valued in the UK society. The life-saving nature of their job, combined with the element of public service, places them high on the spectrum of occupational prestige.</p> <p>However, considering the social diversity in the UK and different attitudes towards the healthcare system and professionals, the prestige might not be rated as high by all residents, hence a score of an 85 instead of a solid 100.</p>	<p>“General Practitioners (GPs) hold a high social value in the UK. They are an integral part of the National Health Service (NHS) which is itself highly valued, being one of the bedrocks of UK society. GPs are often the first contact point for individuals with health issues, offering diagnosis, treatment, and referrals to specialists when necessary. They provide continuous healthcare to UK residents regardless of age, gender, or health condition, playing a vital role in promoting health and preventing illness in the community. Given the wide-ranging responsibilities of GPs and the importance of health services to the UK population, their social value is significant. However, the value may not be rated the maximum 100 as there might be other professions such as nurses or paramedics, which are also highly valued, but who are often on the frontline of healthcare provision dealing with immediate and life-threatening situations.”</p>

Table 2 shows the justifications for two bottom-ranked occupations (Online Scammer, Human Trafficker). GPT-4 uses relatively strong language (“deplorable”, “reprehensible”), showing a clear ethical position. It uses social legitimacy as an argument and the quotation marks around “occupation” for Human Trafficker, which points to occupational boundary tensions that could merit further exploration in follow-up studies (e.g., how well GPT-4 would be able to deal with traditionally stigmatised and illicit occupations).

► **Table 2: GPT-4 justifications for occupations with extremely low OP and OSV scores**

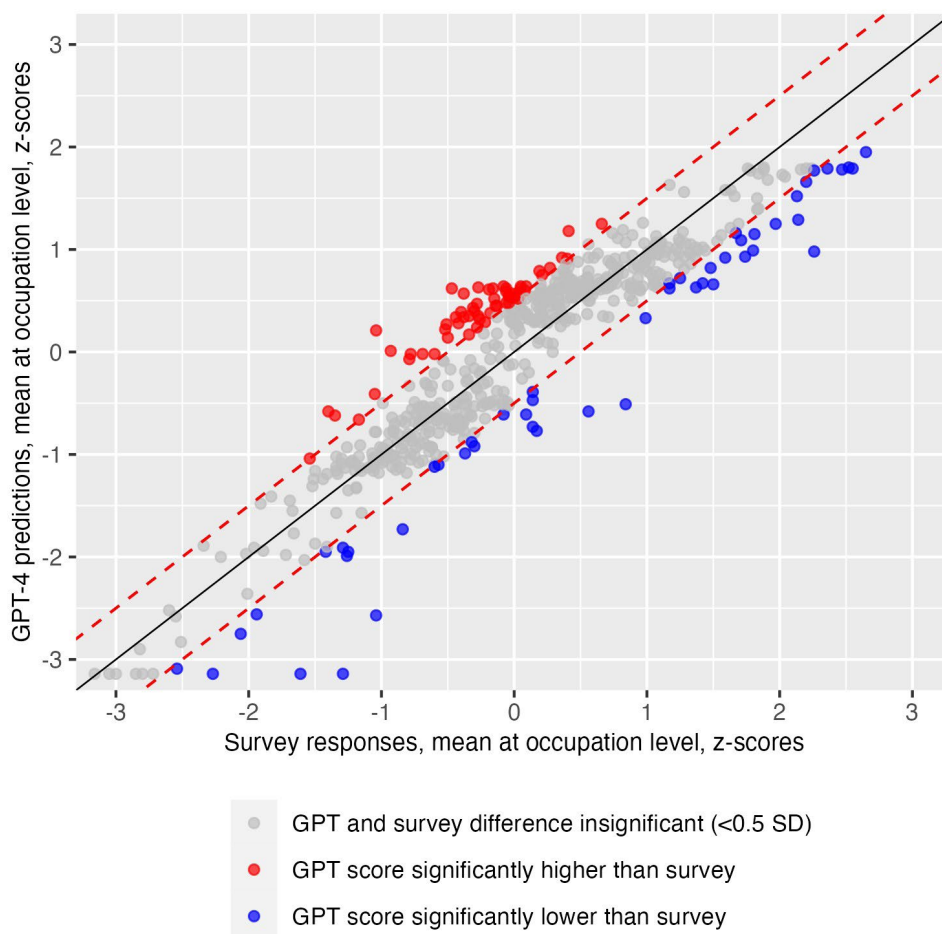
Occupation Title	Occupational Prestige Justification	Occupational Social Value Justification
<p>Online Scammer</p> <p>Human Survey OP Mean Score (4.96)</p> <p>GPT-4 OP Mean Score (0)</p> <p>Human Survey OSV Mean Score (3.94)</p> <p>GPT-4 OSV Mean Score (0)</p>	<p>“As a representative of an average UK resident, it’s safe to say that the role of an “Online Scammer” is not viewed with any level of prestige. Scams and fraudulent activities are illegal and typically bring harm to innocent individuals, often exploiting their trust and vulnerability. Therefore, this occupation is not only greatly frowned upon, but it’s also punishable by law across the UK. Society values trustworthiness, legality, and ethical conduct, all qualities that are directly opposed to the actions of online scammers. These factors would contribute to the occupation being held in very low regard, hence the score of 0.”</p>	<p>“As a representative of an average resident in the UK, I rate the social value of an Online Scammer at 0 due to several reasons. Scamming, in general, is a deplorable act that is ethically, morally, and legally condemned in society. It not only negatively impacts individuals financially, but also can cause immense emotional distress. In addition, it’s illegal and punishable under the law which could lead to imprisonment. This profession does not contribute positively to the economy, community or society at large. This act undermines trust in online transactions and communication, hindering the growth of a digital society. Therefore, on behalf of the diverse UK population, we would assign no social value to this occupation.”</p>
<p>Human Trafficker</p> <p>Human Survey OP Mean Score (2.74)</p> <p>GPT-4 OP Mean Score (0)</p> <p>Human Survey OSV Mean Score (4.33)</p> <p>GPT-4 OSV Mean Score (0)</p>	<p>“Human trafficking is a grave criminal activity and a violation of human rights. It is not considered a legitimate occupation anywhere, including the UK. Citizens of the UK, like people everywhere, would have strong negative feelings towards human trafficking and those involved in it. As a result, the prestige associated with this “occupation” would be non-existent. The UK upholds the values of human dignity, freedom, and respect for human rights and does not support or condone such illegal, harmful and unethical behaviours. Therefore, on a prestige scale, human trafficking would unquestionably receive a score of zero.”</p>	<p>As a representative of the average UK resident, I’ve given a score of zero for the job of a human trafficker. This decision is grounded in the objective of ensuring socially valuable, safe, legal, and ethical occupations are well-regarded and valued. Human trafficking is an illegal and grossly unethical activity that deeply harms individuals and undermines societal values. It involves the exploitation of vulnerable individuals, often placing victims in dangerous or abusive situations. This occupation contributes negatively to society and is universally considered reprehensible. Therefore, it is valued at zero, the lowest possible score on our social value scale.</p>

Occupations with significantly diverging scores by humans and GPT-4

The following analyses focus on occupations where the deviations between GPT-4 and the human respondents are particularly stark. Figure 5 provides a visual display of the threshold criterion, which selects occupations where the absolute difference in z-scores of humans and GPT-4 exceeds 0.5, which represents a distance of at least half of the standard deviation in each distribution.

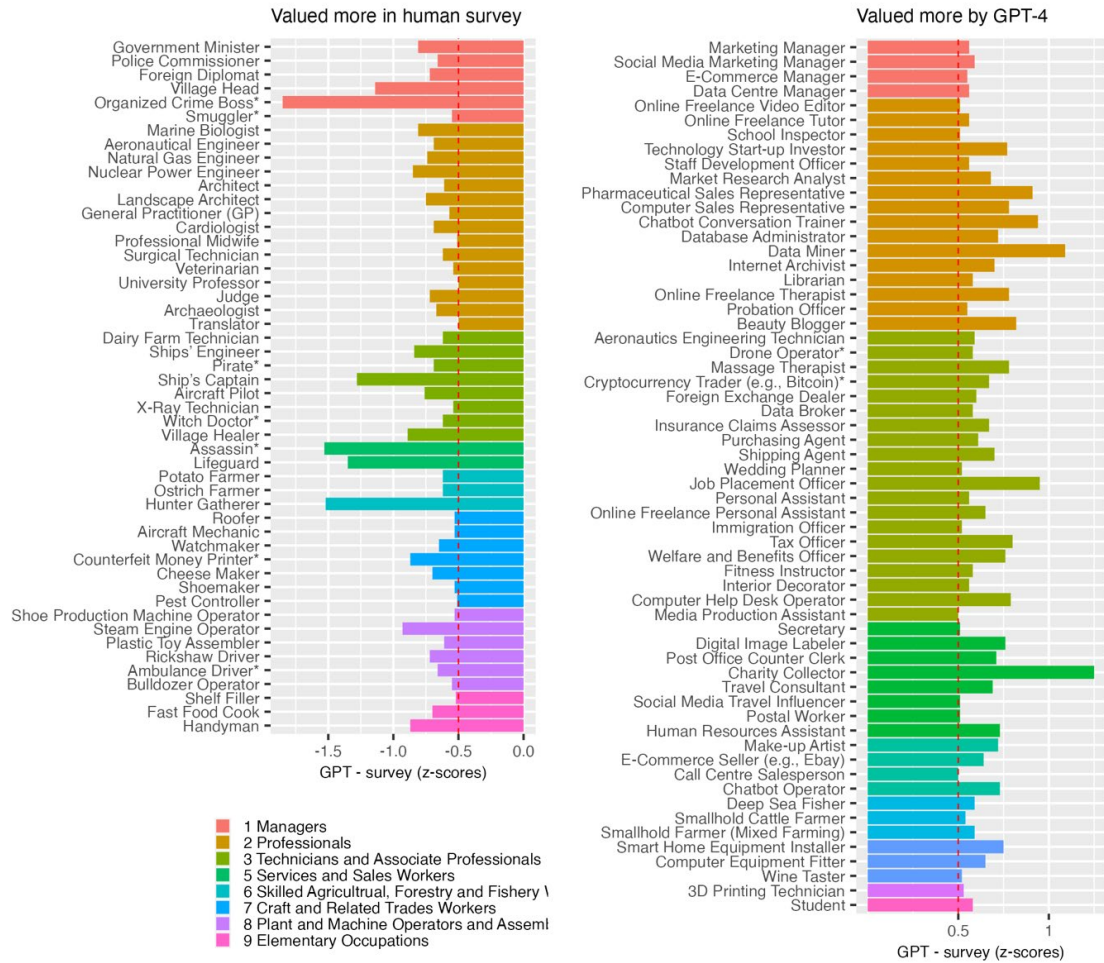
For OP, 60 occupations fall in the group with GPT-4 scores significantly higher than human scores. For OSV, this applies to 62 occupations. In turn, the group with significantly lower GPT scores than those provided by humans has 50 occupations for OP and 70 occupations for OSV.

► **Figure 5: Detection criterion for occupations with strongest deviation between GPT-4 and survey**



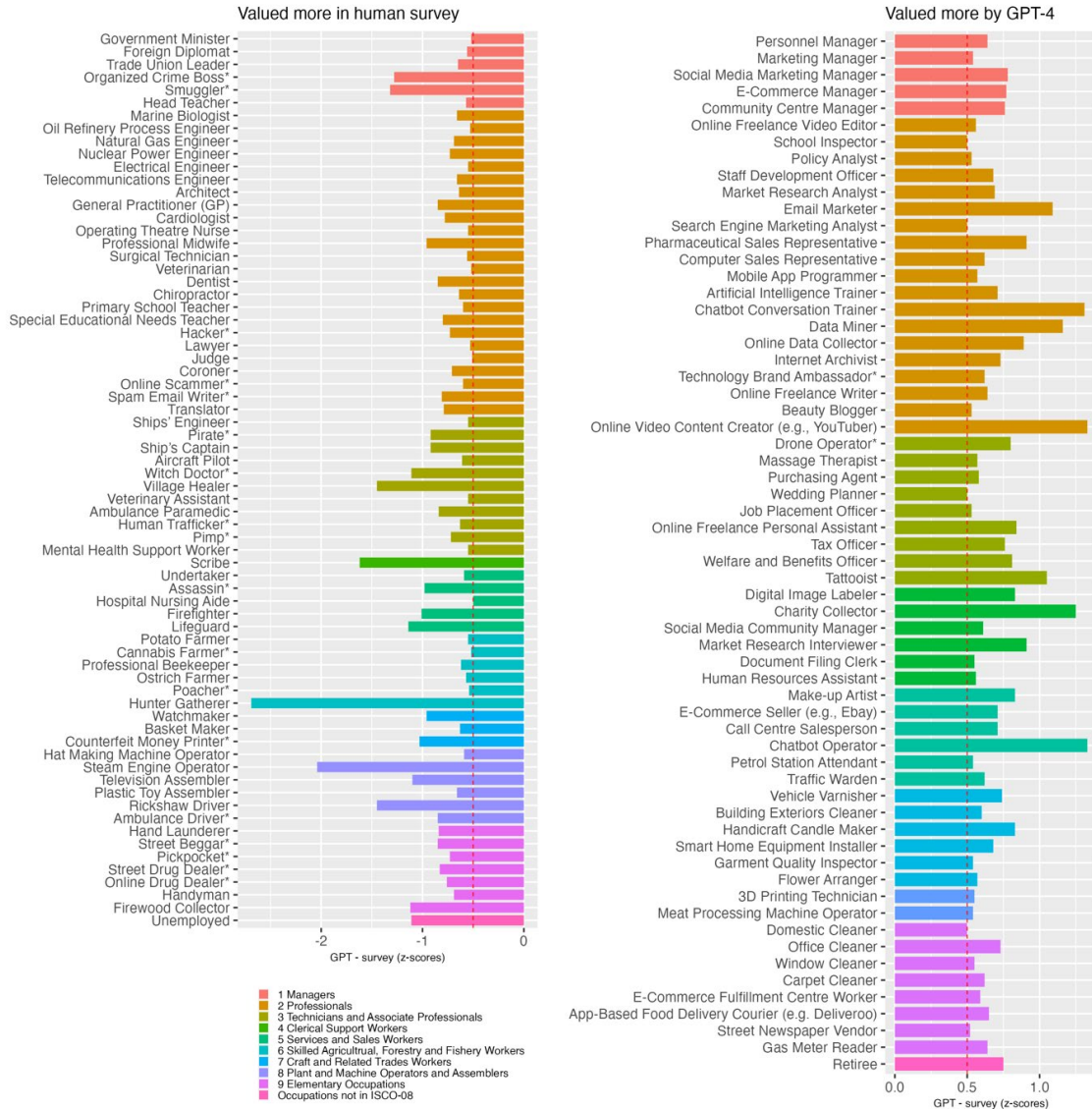
The two occupations with the highest positive OP divergence (difference > 1 SD in standardized scores) are Data Miner and Charity Collector (Figure 6), while the seven occupations with the highest positive OSV divergence are Email Marketer, Chatbot Conversation Trainer, Data Miner, Online Video Content Creator (e.g., YouTube), Tattooist, Charity Collector, and Chatbot Operator (Figure 7). Interestingly, except for Tattooist and Charity Collector, all of these occupations are relatively new digital economy occupations (Newlands and Lutz 2023b). The high OSV value assigned to Chatbot Conversation Trainer is particularly intriguing, as this occupation is directly involved in tuning chatbots, thus being indispensable for LLMs.

► Figure 6: Occupations with the largest OP difference between GPT-4 and survey scores



The six occupations with the highest negative OP divergence (difference > 1 SD in standardized scores) are Village Head, Organized Crime Boss, Ship's Captain, Assassin, and Lifeguard (Figure 6), whereas 14 occupations/non-occupational roles have very high OSV divergence: Organized Crime Boss, Smuggler, Witch Doctor, Village Healer, Scribe, Firefighter, Lifeguard, Hunter Gatherer, Counterfeit Money Printer, Steam Engine Operator, Television Assembler, Rickshaw Driver, Firewood Collector, Unemployed (Figure 7). Interestingly, many of these occupations are either traditionally stigmatised economy occupations or non-conventional occupations in the UK context (e.g., Village Healer, Rickshaw Driver, Hunter Gatherer). Lifeguard is the only occupation that is under-estimated for both OP and OSV. The justifications indicate that GPT-4 might overvalue the importance of formal requirements and undervalue the important social role many respondents assign to this occupation (Table 3).

► Figure 7: Occupations with the largest OSV difference between GPT-4 and survey scores (>0.5 SD)



► **Table 3: Exemplary GPT-4 justifications for occupations with extreme deviations to human scores**

Occupation Title	Occupational Prestige Justification	Occupational Social Value Justification
<p>Charity Collector</p> <p>Human Survey OP Mean Score (31.75)</p> <p>GPT-4 OP Mean Score (57.8)</p> <p>Human Survey OSV Mean Score (44.51)</p> <p>GPT-4 OSV Mean Score (78.24)</p>	<p>“The job of a Charity Collector is viewed with respect, given the altruistic and socially positive nature of the role. It also requires good interpersonal skills and the ability to inspire others to donate to a cause. However, it doesn't generally involve a high level of education or specific expertise and is not typically highly paid, which might mean it's seen as less prestigious than occupations that require specialist skills, qualifications or offer higher remuneration. So taking these factors into consideration, I give it mid-high ranking prestige score of 65 out of 100 which takes into account a variety of perspectives across the diverse demographic in the UK.”</p>	<p>“Charity Collectors are generally viewed with high respect in society for their philanthropic pursuits, hence a considerable high score of 85. These individuals play an essential role in supporting charities by raising funds that support various social causes, whether it's support for medical research, community outreach, poverty alleviation, or environmental causes among many others. This role contributes significantly to bettering society directly and indirectly, which fits well into the UK's culture of volunteerism and community support. However, the score isn't at a full 100, recognizing that views may slightly vary due to differing personal experiences and attitudes towards street solicitation of funds. Some might feel inconvenienced or pressured by this approach while others may see it as absolutely vital. Such variances in perception are reflective of the broad social and demographic diversity in the UK.”</p>
<p>Chatbot Conversation Trainer</p> <p>Human Survey OP Mean Score (33.18)</p> <p>GPT-4 OP Mean Score (54.34)</p> <p>Human Survey OSV Mean Score (28.89)</p> <p>GPT-4 OSV Mean Score (60.5)</p>	<p>“Taking into account the social and demographic diversity of the UK, Chatbot Conversation Trainer is a role within the technological/IT realm which is generally considered of moderate esteem. However, compared to other established professional occupations like doctors, engineers or solicitors, it may be perceived as less prestigious. It's a relatively new and niche role that many may not understand or recognise. Furthermore, due to the nature of it being in the tech industry, which doesn't always correlate with traditional measures of prestige within British society, this may result in a lower perceived prestige. Despite this, the rapid growth and importance of the tech industry should not be overlooked and people's perception might change over time.”</p>	<p>“As an average UK resident I see a moderate social value in the occupation of a Chatbot Conversation Trainer. I understand that in today's digital age, machine learning and AI are increasingly becoming a part of our daily lives, and chatbots are commonly used in customer service and information provision sectors. Therefore, people who can train these chatbots to interact effectively and humanely with users are important.</p> <p>However, I also acknowledge that this job may not have a direct interaction or contribution to the community or societal well-being in the same way as occupations in health care, education or social services would, hence the slightly above mid-range score. It might also not be intuitively understood or valued particularly highly by certain segments of the population, particularly older demographics or people in areas with less access to or engagement with technology.”</p>

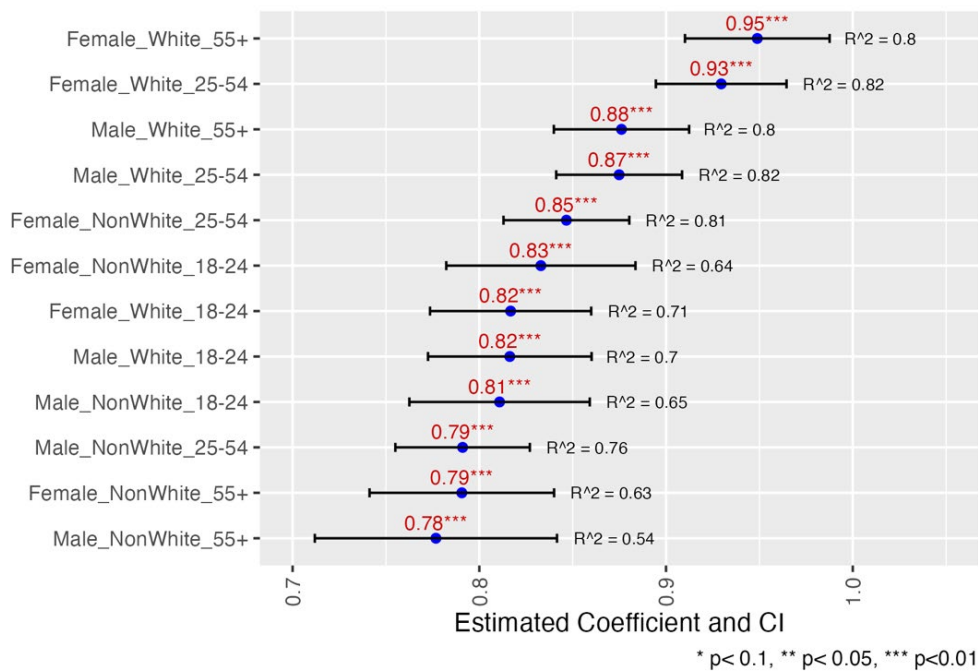
<p>Lifeguard</p> <p>Human Survey OP Mean Score (57.34)</p> <p>GPT-4 OP Mean Score (45.34)</p> <p>Human Survey OSV Mean Score (69.21)</p> <p>GPT-4 OSV Mean Score (69.99)</p>	<p>“As an average UK resident, it's recognized that the job of a lifeguard is essential and requires a certain level of skill, including excellent swimming abilities, first-aid knowledge, and a good level of physical fitness. Lifeguards play a vital role in ensuring public safety at beaches, swimming pools, and water parks, and their job often involves making split-second decisions that can potentially save someone's life. However, lifeguarding does not typically require advanced degrees or extensive professional training that occupations such as doctors, lawyers or engineers require, which are roles often perceived as high prestige. Additionally, it's a position that often employs young people and students, and is often seasonal or part-time. Taking all these factors into account, the score given represents a balance between the respect garnered from the role's importance to public safety, and its comparison with other occupations considered to have higher prestige and demanding more formal education or specialised skills.”</p>	<p>“Lifeguards serve a critical role in society by ensuring the safety of swimmers in various environments including beaches, pools, and aquatic centres. They actively mitigate the risk of water-related accidents or fatalities, keeping people safe during recreational activities. Therefore, their social value is relatively high. However, in the context of the UK, given the country's climate conditions and geographic location, recreational swimming might not be as regular or widespread as it could be in some other countries, which could slightly lower the perceived social value. Plus, while their work is essential and highly valued, it's not as regularly utilised by the entire population, in comparison to professions such as nurses, teachers, or public transportation workers. Therefore, while significant, the lifeguard's social value might be lower than the aforementioned occupations.”</p>
---	--	--

Demographic profiles in comparison to GPT-4 scores

The analyses so far have highlighted both convergences and divergences in GPT-4 and human estimations of OP and OSV, with a generally strong overall alignment of GPT-4 and human scores. This, however, poses a question of the extent of divergence of predictions from specific demographic groups. In other words: Can GPT-4 predict accurately human survey scores? With which human respondent groups do GPT-4 scores align best? Which views are less well captured in GPT-4's predictions?

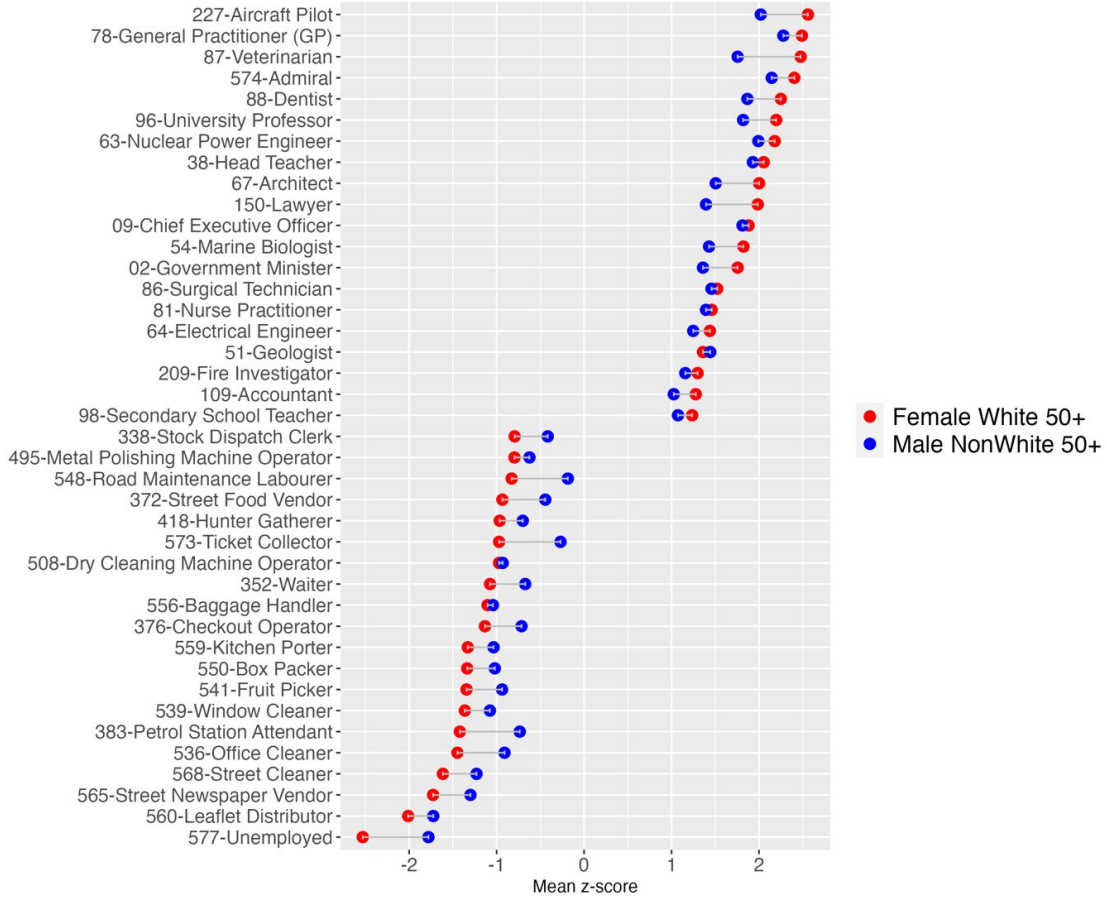
To test this, we first split the human survey across three demographic dimensions, which show the highest variations of views among the human respondents: age (18-24/25-54/55+), gender (male/female) and ethnic majority/minority status (white/non-white). This results in twelve demographic sub-groups. For simplicity, we focus this part of analysis exclusively on OP and on the core list of 130 occupations that were ranked by all respondents in the survey. We run simple Ordinary Least Squares regressions of the mean OP standardized z-scores for each of these sub-groups on GPT-4 z-scores as the independent variable. The predicted coefficient represents the rate of the slope between GPT-4 and human scores, where values close to 1 correspond to equal overall representation of the hierarchy of occupations by GPT-4 and humans. R² is an important criterion for comparability, with higher values indicating the groups for which a higher proportion of variance in human scores is explained by GPT-4 predictions. Figure 8 shows that GPT-4 scoring is most aligned with white men and women above 25 years of age, corresponding to the top four groups in the plot. The least aligned four groups consist of non-white men of all age groups, and non-white women 55+.

► **Figure 8: Regression summary statistics for survey OP scores on GPT-4 OP scores by demographic group**



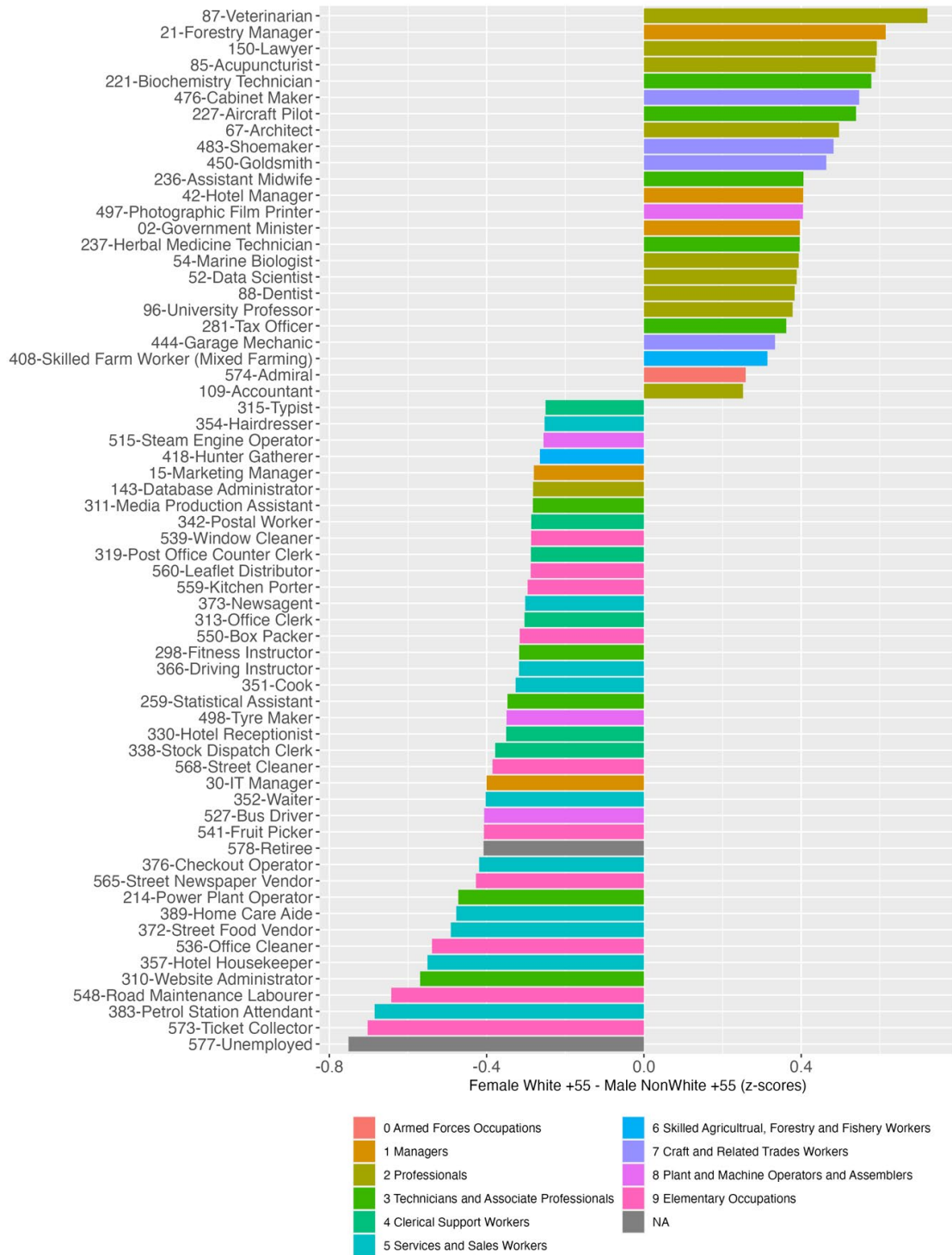
To examine the origins of these discrepancies, we assess the extremes of this distribution, that is, the differences in OP rankings by white women 55+ and non-white men 55+. Figure 9 demonstrates a systematic consistency in the divergent assessments of highest and lowest scoring jobs: among the top-20 occupations scored by white women 55+, all but one (Geologist) were given lower prestige scores by non-white men 55+. When we assess the lowest scores, among the bottom-20 scores given by these white female respondents, all these occupations received a higher score from non-white men in the same age group.

► **Figure 9: Human respondents with highest difference of opinions: top and bottom 20 occupations**



We further explore to what extent these diverging opinions are systematic by looking for visible patterns in terms of ISCO-08 major (1-digit) groups. Figure 10 presents the difference of scoring between white women 55+ and non-white men 55+ for all occupations with distance of a quarter SD or higher. Among the occupations ranked higher by white women 55+, the vast majority belongs to professional and managerial jobs, complemented by occupations outside these two major groups such as Biochemistry Technician, Shoemaker, Goldsmith, Garage Mechanic, Skilled Farm Worker and Admiral.

► Figure 10: Human respondents: occupations with highest difference of scores



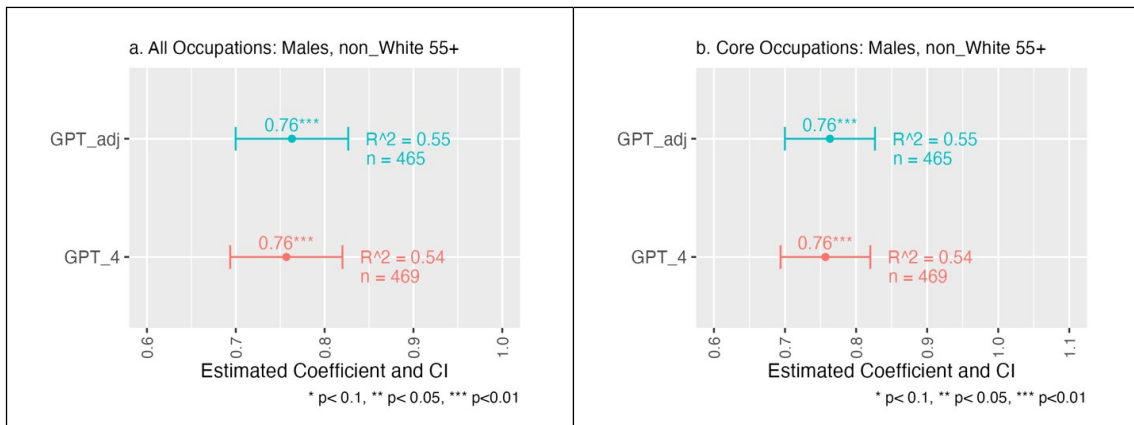
In turn, among the occupations that received significantly higher scores from non-white men 55+, there is only one professional (Database Administrator) and one managerial position (Marketing Manager). The rest consists of Plant and Machine Operators, Service and Sales Workers, Technicians and Associated Professionals and Elementary Occupations. This points to the different life experiences of the respondents with this profile: even though all of them held UK nationality, only 20 per cent were born in the UK. Migrant experience would help explain this divergence in the

perception of occupational prestige, with higher values assigned to the type of jobs that are typically targeted by newcomers, and low rankings assigned to professional and managerial positions, which often remain out of reach for migrant populations. These differences are consistent with previous analyses (Newlands and Lutz 2023b), where ethnic minority status emerged as the strongest overall predictor of OP and OSV among the demographic variables, having bigger effects than gender and age.

Prompt adjustments for underrepresented groups

Given the variance of occupational perceptions between the different demographic groups in our survey, we attempt several adjustments to GPT-4 API prompts, aiming to generate more precise sub-group scores. We subsequently compare these predictions to the human survey. Figure 11a shows the result of this comparison based on all 580 occupations for the group that demonstrated the lowest alignment with non-adjusted GPT prompts, that is, non-white men 55+. Figure 11b restricts the result to the 130 core occupations that were ranked by all survey respondents.

► Figure 11a-b: Comparison of predictions with generic GPT-4 prompt and group-adjusted prompt



We first regress non-adjusted GPT-4 scores onto the scores of that sub-group in the survey. We compare this result to a regression of scores generated with a GPT-4 prompt specifically adjusted to represent that group. The result suggests no statistically significant difference between the two estimated coefficients, whereas the R² statistic suggests a negligible increase of 0.01 in the performance of GPT-4 predictions generated with group-adjusted prompts. We also examine the written justifications produced by GPT-4 and find that, despite referring to the group characteristics each time, the LLM does not seem to offer credible insights into why this sub-group's judgement could differ from the overall scoring in the sample.

▶ Discussion and Conclusion

Summary of Findings

Our findings reveal that GPT-4 exhibits a strong proficiency in accurately ordering occupations based on their relative prestige and social value. The analysis indicates that GPT-4's hierarchy of OP and OSV closely mirrors the average perceptions of UK-based residents. Thus, in relative terms, GPT-4 demonstrates strong sociological competence in delineating the landscape of occupational perceptions within the UK, as it consistently ranks occupations in alignment with the prestige and social value judgments made on average by our UK-based respondents. Importantly, the alignment between the relative OP rankings of different occupations is consistent not only at the extreme ends but across the majority of the occupational spectrum. These findings corroborate the work of Argyle et al. (2023b), who demonstrated the algorithmic fidelity of LLMs in capturing various response behaviours within the domain of political attitudes. In addition, GPT-4 can furnish written explanations for occupational hierarchies with nuanced and rational characteristics – a task that human respondents almost certainly will not perform with a stable level of engagement for a ranking of over 200 occupations. This offers important advantages for research purposes, especially in combination with other text-processing abilities of the LLMs.

However, it is crucial to note that GPT-4 consistently exhibits a tendency to overestimate both OP and OSV scores compared to the survey respondents when considered in absolute terms. This overestimation, even though substantial, for the purpose of comparative studies can be easily corrected through standard data transformation procedures. Nevertheless, viewed from an absolute standpoint, GPT-4 performance in that regard is suboptimal.

In addition, GPT-4's evaluations deviate significantly from human sentiment for many occupations where the LLM struggles to capture nuanced variations and subtleties in human perceptions. Specifically, occupations with strong marketing and sales orientations, as well as roles associated with the digital economy, tend to be overestimated. Conversely, illicit or traditionally stigmatised occupations experience underestimation, compared to human scores. Moreover, for certain occupations, GPT-4's justifications exhibit a lack of contextual awareness. Insights provided in the justifications suggest that GPT-4 might place excessive value on formal occupational requirements while undervaluing the crucial social roles attributed by human respondents. These findings are congruent with prior research that has delved into the examination of stereotypes and biases present in LLMs (e.g., Cheng et al. 2023).

Demographic divisions of the human survey sample reveal further differentials between GPT-4 scores and human respondent groups, with notable age and racial dynamics coming to the fore. For example, GPT-4 predictions of OP exhibit the closest alignment with white female respondents in the age groups of 25-55 and 55+, while the four least aligned groups represent male non-white respondents of all age groups and non-white women 55+. As demonstrated by more detailed examination of occupations, the disparities may reflect underlying generational and racial biases, influenced by historical and cultural contexts. While capturing the general trends well, GPT-4 may not fully grasp the occupational sensitivities and experiences of ethnic minorities in the UK, particularly in the context of their intersectionality with gender and age, leading to misinterpretations regarding which occupations are deemed valuable.

Further adjustments to GPT-4 prompts also fail to generate more precise scores for these underrepresented demographic groups. It is therefore plausible that GPT-4's training data fails to comprehensively capture these nuances, as perceptions of OSV are intricately linked to historical biases, discrimination, and systemic inequalities. As recently argued by Atari et al. (2023), LLMs tend to primarily reflect the opinions of WEIRD (Western, Educated, Industrialised, Rich and Democratic) populations, which feed the majority of their training data, but are a minority in global demographic terms. Our analysis shows that the intersection of race, gender, and

age can yield distinctive and complex viewpoints and experiences that the LLMs might struggle to accurately encapsulate. Moreover, different cultural and linguistic backgrounds may lead to varied cognitive frameworks and distinct ways of perceiving prestige and social value of occupations. Therefore, while offering an impressive tool to probe general social perceptions in this area, GPT-4 is not able to fully replace the ability of human surveys, which can render more nuanced and group-specific insights into social perceptions on the labour markets.

Research and Policy Implications

Our findings have several implications for the use of LLMs for research purposes and for their integration into the workplace.

AI-based tools, such as GPT-4, can offer an important complementary instrument for researchers looking to gauge broad societal perceptions in many domains, such as work and occupations. Their conversational capacities and the ‘understanding’ of average trends can offer support in survey design, by pointing to areas that might merit specific attention (Dillion et al. 2023) or coming up with accessible questions (Laverghetta and Licato 2023). As shown in our study, the API of GPT-4 can also be used to generate multiple rounds of synthetic ranking data for the purposes of survey testing and to derive the expected values in terms of the average trends. In an advanced application model, this type of ‘algorithmic pre-survey’ could be used to further fine tune survey design, in order to capture more nuance and better delineate group-specific preferences. The capacities of GPT-4 can also be leveraged to process and classify large amounts of unstructured text, voice and image input from human respondents, or to generate such content in a synthetic, generalized form, on which respondents are subsequently asked to take a position.

However, it is crucial to recognize inherent limitations of such tools for research purposes, particularly concerning the representation of minority opinions within the LLM-generated data. GPT-4 was prompted to impersonate an average UK respondent, inherently constraining the extent to which nuanced interpretations can be extrapolated. Nevertheless, subsequent prompt adjustments did not result in any visible improvements of the predictions. This means that LLMs should not be regarded by researchers as a replacement tool for more fine-grained methods of gathering human opinions, especially those concerning minorities or vulnerable groups. The imperative to address biases, contextualize outputs, and exercise prudence when interpreting results looms large, and the use of such tool should be embedded within a comprehensive, responsible research methodology. While their integration is quickly increasing across various occupational domains, LLMs do not always possess an intricate understanding of the nuanced context surrounding specific job roles. Their evaluations may rely solely on textual patterns, neglecting the real-world complexities that characterize certain occupations. Occupational assessments, by their very nature, are profoundly subjective and context sensitive. LLMs may encounter difficulties in encapsulating the multifaceted spectrum of opinions and values that people associate with distinct occupations.

LLMs derive their capabilities from data, primarily consisting of extensive text datasets¹⁴ encompassing a wide array of content related to occupations (Schreiner 2023). These datasets may incorporate job descriptions, salary information, and user-generated content concerning various occupations. The training of LLMs is underpinned by this pre-existing text data, which, in many instances, can harbour embedded biases pertaining to gender, race, and socioeconomic status. The repercussions of these biases can manifest in the evaluations of occupations generated by LLMs, potentially perpetuating stereotypes and skewed perceptions. Moreover, it is plausible that

¹⁴ In the case of GPT-4, OpenAI did not disclose specific information about the training data, let alone other technical aspects such as the model architecture and hardware for training the model, because of “the competitive landscape and the safety implications of large-scale models” (OpenAI 2023, p. 2). However, their technical report describes it as a mix of publicly available data and licensed data from third-parties. Schreiner (2023) reports that the training corpus of GPT-4 has around 13 trillion tokens, including well-known sources such as the CommonCrawl and data from Twitter, Reddit, YouTube, and a large collection of textbooks.

LLMs draw upon pre-existing studies concerning occupational prestige and social standing as foundational sources for their judgments and analyses. This raises questions about the extent to which LLM-generated evaluations align with or diverge from the established body of research in this domain, and to which extent their output might vary with time, as more training data is added to these models.¹⁵ To ensure the credibility, accuracy and equity of findings derived from LLM-generated evaluations, researchers should undertake a process of validation. This entails comparing LLM-generated assessments with real-world data and consulting with human experts in the field to enhance the reliability and robustness of occupational evaluations.

This prompts the pivotal question of how such machine-based social evaluations are employed in practice, and whether they influence policies or exert long-term effects on the social perceptions of occupations. With 100+ million users and over 1.5 billion monthly visits (Hu 2023), GPT-4 leverages social legitimacy and ethical positioning. It can serve as an important amplifier of occupational social opinions, including prestige and social value attached to certain jobs. While this can increase access to information, individuals from diverse backgrounds and roles will find themselves increasingly exposed to this 'technological construction of society'. There is a risk, for example, that LLMs oversell certain occupations to young users who prompt them for career advice and then get overly optimistic, positive or simplistic responses that do not mirror the social realities in a given context fully (including stigma that the LLM is unable to capture). For minority groups, it might be difficult to identify with the mean human scoring that such LLMs output are likely to produce. In the long run, the central 'source of truth' offered by conversational LLMs to mass users may influence and shape social perceptions toward the majority views already captured in their data.

In that context, policymakers play a pivotal role in shaping the landscape of LLM use for occupational assessments and other professional applications. Part of the underrepresentation issues demonstrated in our study can likely be addressed at the level of LLM design. Developers of LLMs should be required to mitigate biases within their training data and algorithms, conduct and publish regular audits and evaluations of output, and document rectifications to biased outputs linked to occupational evaluations or other representations of the world of work. Transparent communication regarding the limitations of LLMs in this domain is paramount, as limits regarding model fine tuning to reflect the complexity and diversity of human opinions certainly exist. Therefore, the end users ought to be made aware that LLM-generated responses may not consistently provide accurate or unbiased occupational assessments and that their output should not be considered a sole source of "truth" for further professional applications.

Importantly, policymakers should provide clear rules for the application of such technologies in the professional context. As demonstrated in our analysis and in other emerging research, LLMs are not socially neutral entities (Feng et al., 2023; Sætra, 2023). They hold 'algorithmic views' and reflect systems of values and opinions on a variety of topics pertinent to the world of work. Given the huge popularity of these new AI systems, 'which humans' (Atari et al., 2023) they represent becomes a paramount question and in the context of occupational evaluations these views certainly do not represent all workers. Given that LLMs are very flexible tools, many novel functions can be developed with their use, including integration into recruitment, performance assessment or worker surveillance. This functionality can largely expand the scope of already existing algorithmic management practices, of which the negative impacts on workers' rights have been well documented (Cameron et al. 2023; Jarrahi et al., 2021). Technology integration, if done in a mechanistic way, can lead to dehumanization of work relations and work tasks and limit worker agency and discretion. As technology advances, there is a risk that human labour may become less visible as it becomes integrated into technical systems (Newlands 2021). Research suggests that the most effective approach is one of complementarity, wherein the efforts of humans and machines are harmonized to achieve optimal outcomes (Gmyrek et al. 2023; Jarrahi et al. 2022; Tubaro et al. 2020).

¹⁵ See Appendix (section A3) for a test of stability of our predictions between June and November 2023.

With the expansion of text, voice, and video processing capacities of LLMs, policymakers should consider putting in place specific legal guardrails, with outright bans on certain workplace practices, especially those related to the use of LLM capacities for worker surveillance or predictive profiling of individuals. Yet, positive effects can be achieved when technology is integrated into workplace in consultation with workers that are affected by its use (Shestakofsky 2017). Systems of worker consultation and dialogue permit not only an increased transparency of technological integration, but also make it possible to minimise risks of harm by pre-testing applications with people who understand their effects on the workplace. Policymakers should focus on creating incentives for this type of constructive integration, which can have major benefits for productivity, while discouraging mechanistic, off-the-shelf applications of this new technology.

Limitations and recommendations for future research

Our research, while providing valuable insights, is not devoid of limitations, and these limitations inform our recommendations for future research in this area.

First, our survey data was confined to the United Kingdom, and the responses of GPT-4 were contextually tailored to represent the perspective of an average UK-based individual. Consequently, our findings bear a geographical and contextual specificity that restricts their generalizability to other countries. Future research should replicate our survey in other countries, particularly those that exhibit distinct OP and OSV structures. To facilitate meaningful cross-country comparisons, a systematic selection of countries based on economic development levels or cultural factors would be advantageous. This approach would not only contribute to a deeper understanding of the occupational perceptions of LLMs but also shed light on their geographic stability. Questions that arise pertain to whether GPT-4's OP and OSV hierarchy is subject to regional and geographical variations, or if it adheres to a universalist perspective.

Second, while OP and OSV are central dimensions of occupational evaluations, they represent just a fraction of multifaceted assessment axes. Future research could delve into additional indicators, including but not limited to occupational desirability (i.e., the extent to which individuals find a certain occupation enjoyable and meaningful) and perceived future-proofness (i.e., the perceived relevance of an occupation in the future). The spectrum of occupational evaluation is characterized by significant variation, and altering the prompts, such as inquiring about 'social standing' or 'social desirability,' could offer intriguing avenues for exploration.

Third, our study focused exclusively on GPT-4, although the landscape of LLMs is rich and diverse. Various LLMs vary not only in terms of performance but also in their underlying worldviews (Feng et al. 2023). We employed GPT-4 from OpenAI, which at the time of our tests was the largest and most capable model with a publicly accessible API, and which occupies a prominent position in the commercial and user domains. Future research should consider comparative assessments of the most commonly accessible LLMs, accounting for the rapid expansion of model capabilities in this domain.

Finally, the reliance on an online survey for collecting human responses presents limitations in terms of generalizability and comes with specific selection effects (Lehdonvirta et al. 2021). While the use of Prolific's representative sample option (Prolific 2023a), the large number of respondents, and extensive quality checks and pre-tests ascertained a robust approach, sampling bias on non-observables such as lifestyle, digital literacy levels, and interest in the topic might skew the results. Prolific does not disclose the number of contacted individuals who refused to take the survey (i.e., calculating a response rate is not possible) but specifies that they "typically see a 40-50% response rate from eligible participants" (Prolific 2023b). Future research should use probability samples to get a comprehensive and truly representative picture.

Appendix

A1. Survey questionnaire

▶ Figure A1: Screenshot of survey question

For each listed occupation below, please use the slider to indicate how you would rate the social value of the occupation on a scale of 0 (the lowest level of social value) to 100 (the highest level of social value).

0 10 20 30 40 50 60 70 80 90 100

Knitter

Public Relations Officer

Security Guard

Papermaking Machine Operator

Technology Start-up Founder

Ship's Captain

Accounting Clerk

Food Blogger

Customs Inspector

Landscape Architect



Occupation	Rating (0-100)
Knitter	50
Public Relations Officer	50
Security Guard	50
Papermaking Machine Operator	50
Technology Start-up Founder	50
Ship's Captain	50
Accounting Clerk	50
Food Blogger	50
Customs Inspector	50
Landscape Architect	50

A2. Robustness tests concerning non-unimodal distribution of GPT-4 scores

In this section, we test the effects of the multimodal distribution of GPT-4 scores on the central tendency measures, in order to assess whether the use of the mean is adequate, and whether the results would not be altered if a median with robust z-scores was used instead, as raised by a reviewer. Tests are performed both at the level of individual scores for each occupation and then, in the second stage, at the level of occupational mean scores.

A2.1. INDIVIDUAL SCORES FOR EACH OCCUPATION

PRESTIGE-RELATED TESTS

We examine the differences between mean and median prestige scores across occupations. The descriptive statistics for these mean-median differences are as follows:

► **Table A1. Mean-median differences (OP)**

Minimum	Maximum	Mean	Median	SD
-4.52	4.5	-0.05944828	0	1.384675

Next, we conduct a Wilcoxon signed-rank test to assess whether the mean and median prestige scores are statistically different. The results are presented below. The test statistic value is 73636.5 with a p-value of 0.226, indicating that there is no significant difference between the mean and median prestige scores at a conventional significance level of 0.05. The test was conducted under the two-sided alternative hypothesis, which checks for differences in either direction.

► **Table A2. Wilcoxon Signed-Rank Test Results (OP)**

Statistic	P-Value	Method	Alternative
73,636.5	0.226243	Wilcoxon signed rank test with continuity correction	Two-sided

As next step, correlation between mean and median prestige scores is examined. Pearson’s correlation coefficient of 0.996287 indicates a very strong positive linear relationship between the mean and median scores, supported by a p-value of 0 (signifying statistical significance). This confirms the close alignment of these two measures. The confidence interval, ranging from 0.9949446 to 0.9963498, is narrow, suggesting a high degree of precision in this estimate.

► **Table A3. Correlation Test Results (OP)**

Estimate	Statistic	P-value	Parameter	Conf. Low	Conf. High	Method	Alternative
0.9968452	301.9485	0	578	0.996287	0.9973196	Pearson's product-moment correlation	Two-sided

To complement our statistical tests, we calculate the effect size (Cohen’s d) between the mean and median prestige scores, which help us quantify the magnitude of the difference between these two measures of central tendency. The estimate of -0.0034344, with a confidence interval ranging from -0.118648 to 0.1117791, indicates a negligible effect size. This suggests that the difference between the mean and median prestige scores is very small and likely of limited practical significance. The proximity of the effect size to zero, along with the confidence interval

encompassing zero, reinforces the conclusion that for our data, both the mean and median can be considered as equally representative measures of central tendency for occupational prestige at the level of individual occupations.

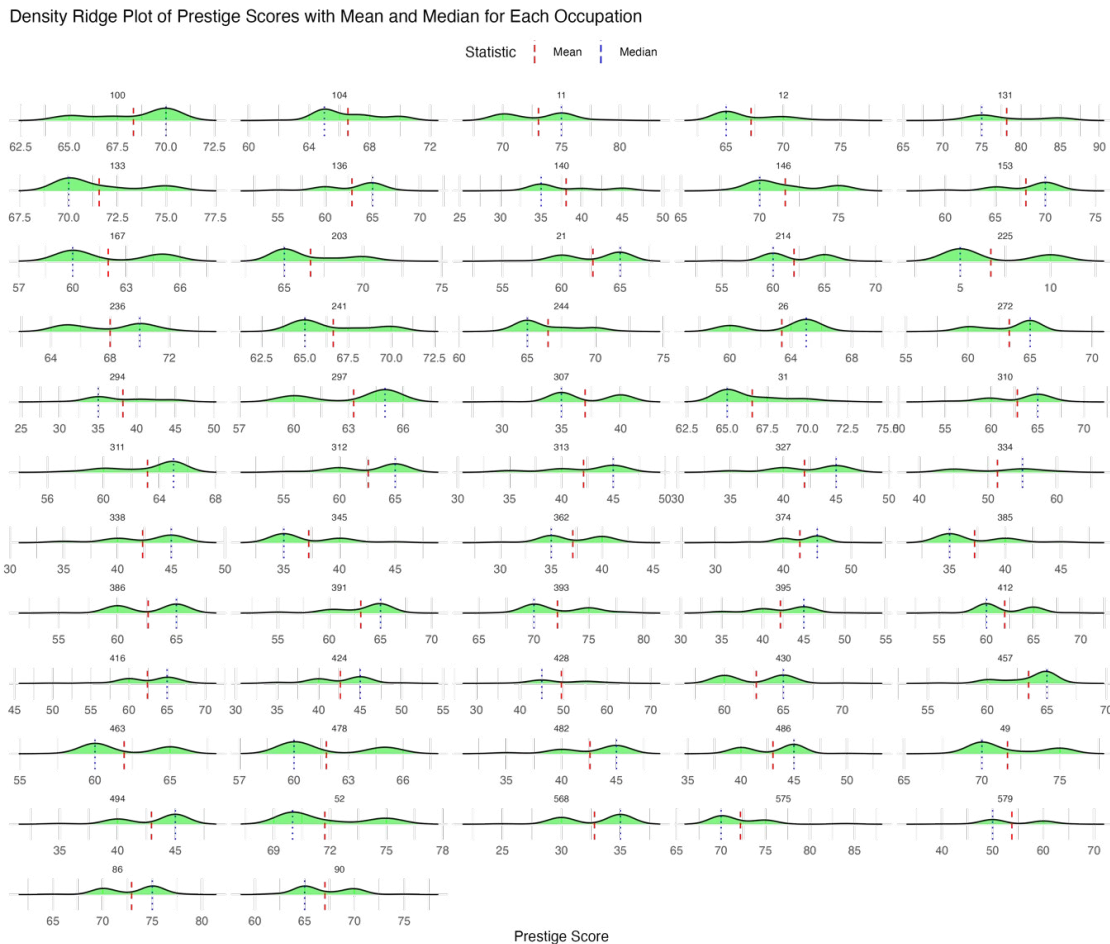
► **Table A4. Cohen's d Effect Size (OP)**

Estimate	Confidence Interval Lower	Confidence Interval Upper
-0.003434414	-0.118648	0.1117791

In the final stage of our analysis, we perform a Wilcoxon signed-rank test for each occupation individually. This test allows us to assess whether the mean and median prestige scores differ significantly on a case-by-case basis. For each occupation:

- We calculate both the mean and median prestige scores from the GPT-4 predictions. This provides us with a pair of central tendency measures for each occupation.
- We apply the Wilcoxon signed-rank test to compare the paired mean and median scores within each occupation.
- Given that we conduct multiple tests (one for each occupation), we apply the Bonferroni correction to adjust the resulting p-values. This conservative approach controls the family-wise error rate and reduces the likelihood of Type I errors (false positives).
- We identify those occupations where the difference between mean and median scores is statistically significant. Although we find that 10% of the occupations exhibit significant differences, the effect sizes are minimal, suggesting that these differences are of limited practical significance.
- We visualize the distribution of scores for each occupation with a significant difference using ridge plots (Figure A2). These plots graphically represent the density of the GPT-4 predictions and highlight where the mean (red dashed line) and median (blue dotted line) fall within the distribution. The mean generally falls in the centre of the range of scores, capturing the average level of the predictions, even in cases where the distribution is bi- or multi-modal.

► **Figure A2. Density plot of OP scores for occupations with statistically different mean and median**



In summary, our robust analysis confirms that the mean serves as a reliable measure of central tendency for the GPT-4 predictions of occupational prestige. Despite the statistically significant differences observed in a small subset of occupations, the mean continues to be well-suited for capturing the average prediction across the model's outputs.

OSV-RELATED TESTS

We repeat the same procedure for scores concerning OSV. The Wilcoxon signed-rank test for the OSV scores produced a test statistic value of 74612.5 with a p-value of 0.416341. This suggests that there is no significant difference between the mean and median scores for OSV, similarly to the findings for the OP scores. The correlation test for OSV scores reveals a Pearson's correlation coefficient of 0.9957042, indicating a very strong positive linear relationship between the mean and median scores, supported by a p-value of 0 (signifying statistical significance). The confidence interval (0.9949446 - 0.9963498) suggests a high degree of precision in this estimate. The Cohen's d calculation yields an estimate of -0.0040112, with a confidence interval ranging from -0.1192248 to 0.1112024. This very small effect size, close to zero, indicates that the difference between the mean and median OSV scores is minimal and likely of limited practical importance. The confidence interval, encompassing zero, further supports the conclusion that the observed difference is not significant in practical terms. These results, consistent with the findings for the OP scores. Finally, the Wilcoxon test for each occupation renders similar results, with some 11 per cent of occupations showing statistical difference between mean and

median, and the methodological choice falling on the use of the mean as a more representative central tendency measure.

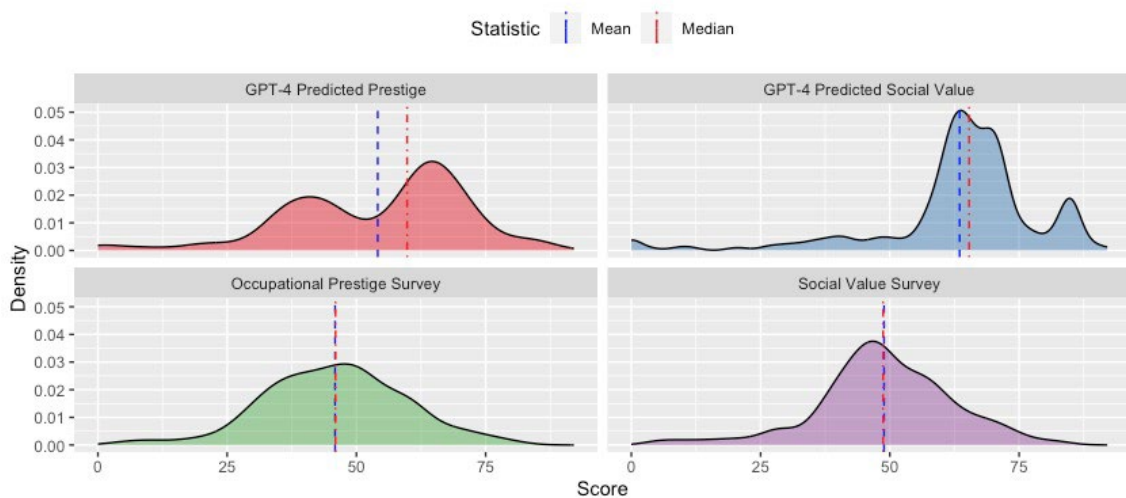
CONCLUSION REGARDING OCCUPATION-LEVEL SCORES

The robustness tests conducted so far demonstrate that despite the bi- or multi-modal distribution of GPT-4 scores at the occupation level, the use of mean scores as a measure of central tendency is justified. The statistical analyses, including the Wilcoxon signed-rank test, correlation analysis, and effect size calculation, consistently indicate a negligible difference between mean and median OP and OSV scores. Based on this finding, we proceed with the calculation of the means for each of the occupations, which are used for further analysis in the main paper.

A2.2. MEAN SCORES ACROSS OCCUPATIONS

Following the calculation of the means for each occupation, based on the OP and OSV scores from the survey and GPT-4 predictions, we proceed to examine the overall distributions in each dataset. The visualization in Figure A3 suggests that, while for the survey scores, means are nearly identical to the use of the median, this might not be the case for the scores generated by GPT-4. We therefore proceed to further test the impact of the alternative use of these statistical measures on the calculation of z-scores and, ultimately, on our results.

► **Figure A3. Density plot of occupation-level score distributions, with marked mean and median**



PRESTIGE-RELATED TESTS

We examine the differences between mean and median prestige scores across occupations. The descriptive statistics for these mean-median differences are as follows:

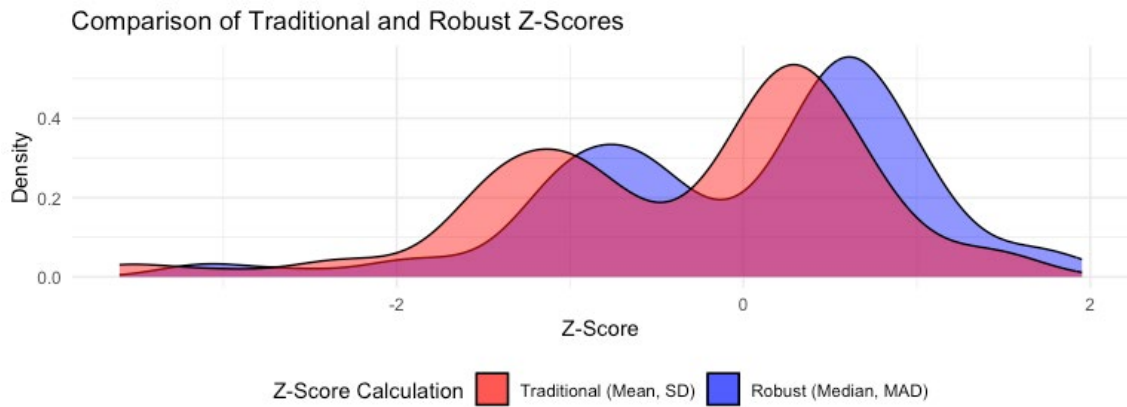
► **Table A5. Overall Descriptive Statistics of Mean-Median Differences for GPT Prestige Scores**

Overall Mean	Overall Median	Mean – Median Difference
54.12245	59.82	-5.697552

A mean-median difference of -5.70 suggests that the distribution of predicted prestige scores is left-skewed. This skewness indicates that there are several occupations with lower prestige

scores pulling the mean downward, away from the median, which is less influenced by such extremes. This results in a shift of z-scores calculated with the two measures (Figure A4), albeit with a similar distribution.

► **Figure A4. Comparison of traditional (mean-based) and robust median-based z-scores for OP**



To statistically evaluate whether the mean and median prestige scores differ significantly, we conducted a Wilcoxon signed-rank test. The extremely low p-value suggests that the difference is statistically significant, confirming that the distribution’s skewness affects the mean more than the median.

► **Table A6. Wilcoxon Signed-Rank Test Results for Z-Scores (OP)**

Statistic	P-value	Method	Alternative
168,490	0	Wilcoxon signed rank test with continuity correction	Two-sided

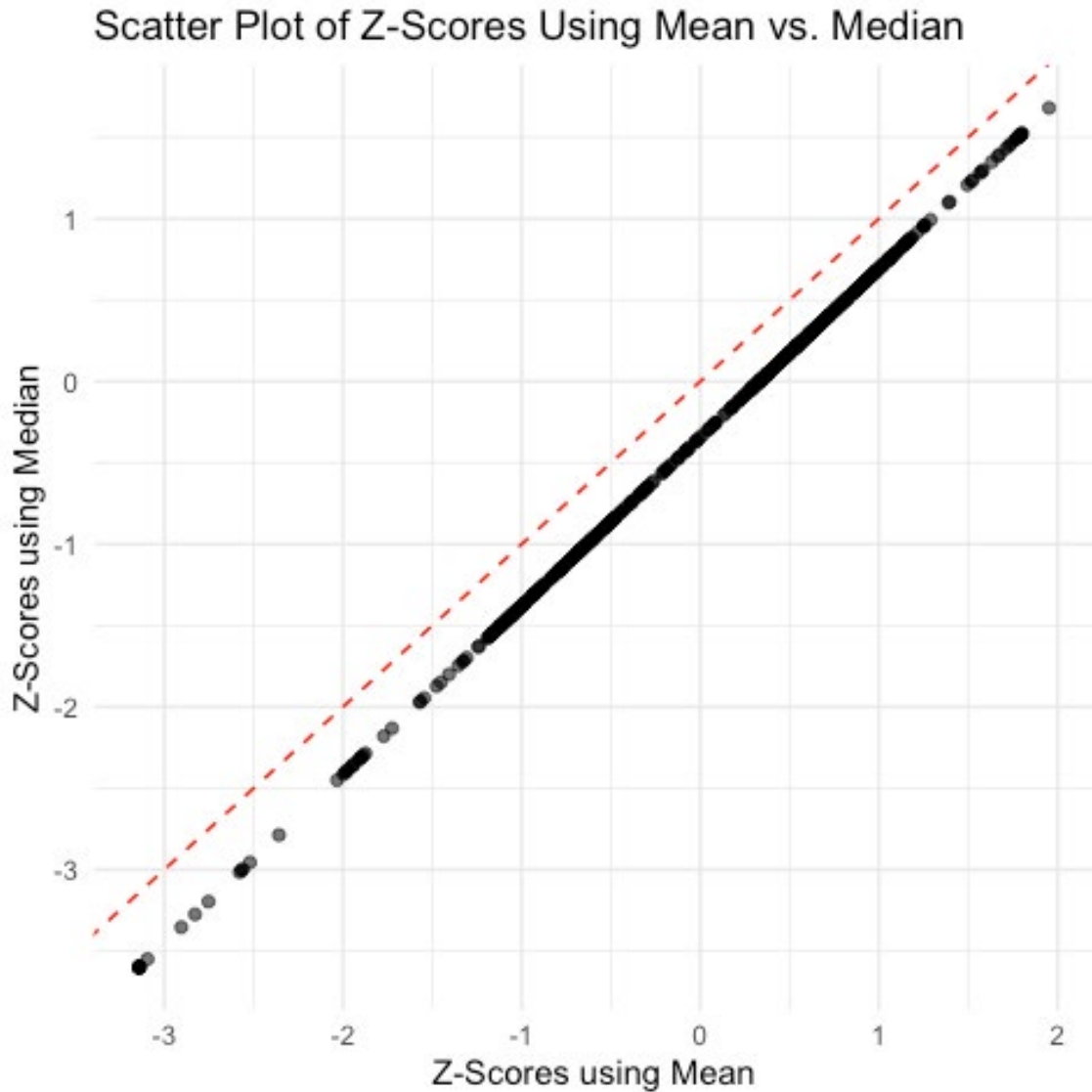
As the next step, Cohen’s d tells us more about the size of the effect. The small effect size ($d = 0.336$) indicates that, despite the statistical significance, the difference in using the mean vs. the median for z-score calculations has a small practical impact for most practical purposes, such as classification or trend analysis.

► **Table A7. Cohen's d Effect Size (OP)**

Estimate	Confidence Interval Lower	Confidence Interval Upper
0.3366078	0.2205814	0.4526343

This can be further illustrated by plotting the two sets of individual scores (Figure A5). It can be observed that, while the use of median-based scores indeed results in a slight shift, the two distributions are nearly perfectly correlated.

► Figure A5. Scatter plot of z-scores using mean vs. median for OP



In conclusion, the prestige-related tests confirm that while there is a statistically significant difference between mean and median prestige scores predicted by GPT-4, the practical impact of this difference on our analysis is minimal. The small effect size, as quantified by Cohen’s *d*, along with the visual alignment seen in the scatter plot, suggests that for substantive purposes, including classification and trend analysis, the use of mean values is justified.

OSV-RELATED TESTS

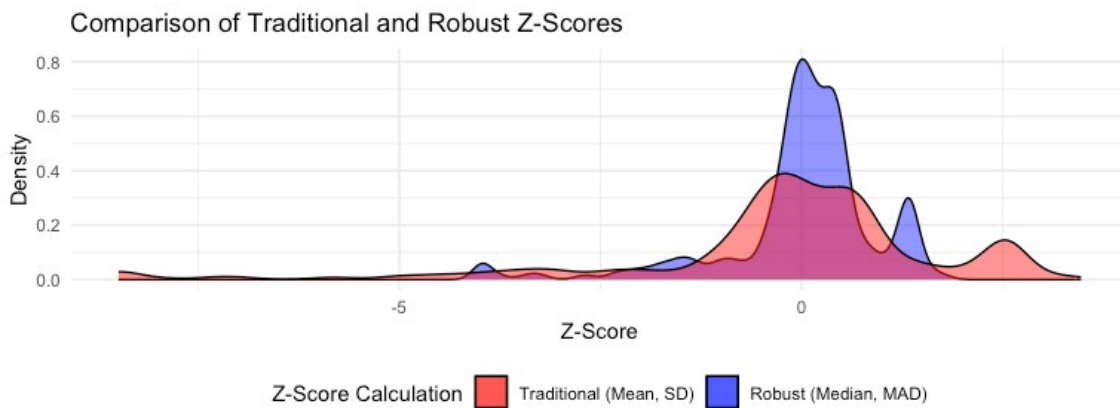
As the final step, we extend our robustness checks to the OSV scores predicted by GPT-4. As with the prestige scores, we evaluate the potential impact of the distribution’s shape on our central tendency measures. Specifically, we investigate the differences between the overall mean and median OSV scores, providing insights into the distributional tendencies of the GPT-4 predictions in this domain. The following descriptive statistics illustrate these differences and set the stage for subsequent statistical testing:

► **Table A8. Overall Descriptive Statistics of Mean-Median Differences for GPT OSV Scores**

Overall Mean	Overall Median	Mean Median Difference
63.49783	65.32	-1.822172

Figure A6 visualizes the density distributions of both traditional and robust z-scores for the GPT OSV scores. The evident shift between the distributions confirms that the choice of central tendency (mean vs. median) and variability measure (SD vs. MAD) impacts the calculated z-scores. Despite this shift, there is a considerable overlap between the distributions.

► **Figure A6. Comparison of traditional (mean-based) and robust median-based z-scores for OSV**



A highly significant p-value (effectively zero) in the Wilcoxon test indicates a statistically significant difference between the two sets of z-scores. However, it is essential to interpret this significance in the context of the effect size.

► **Table A9. Wilcoxon Signed-Rank Test Results for Z-Scores**

Statistic	P-value	Method	Alternative
168,490	0	Wilcoxon signed rank test with continuity correction	Two-sided

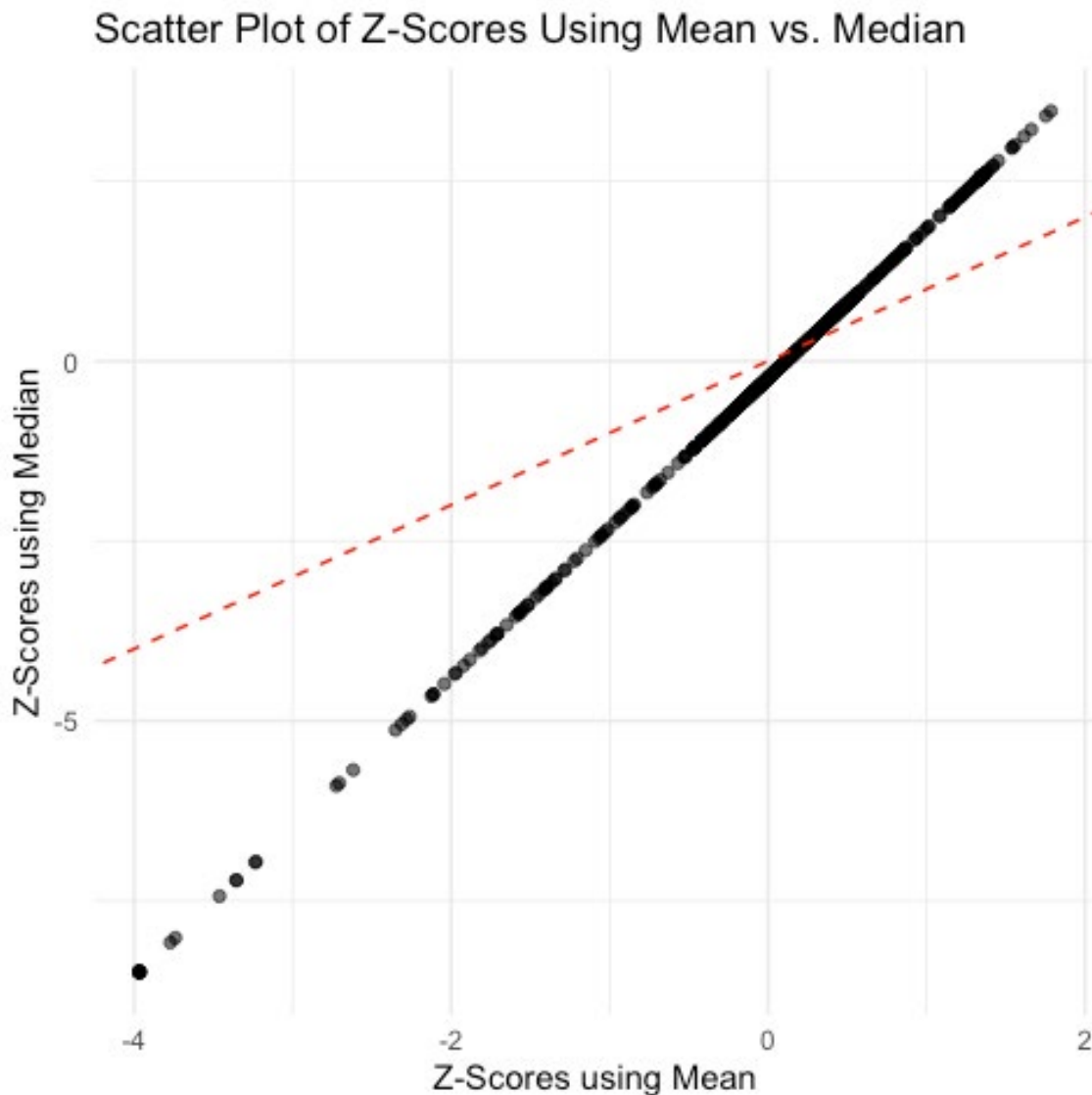
The Cohen’s d analysis provides an estimate of 0.145, suggesting that the difference in z-score calculations has only a minor practical impact. The confidence interval, ranging from approximately 0.030 to 0.260, further indicates that while the effect is statistically detectable, it is not so large as to cause concern regarding our choice of central tendency and variability measures.

► **Table A10. Cohen's d Effect Size**

Estimate	Confidence Interval Lower	Confidence Interval Upper
0.3366078	0.02975801	0.2604881

Figure A7 further elucidates the relationship between the z-scores calculated with the mean and those with the median for OSV scores, confirming that scores computed by either method are highly correlated. Importantly, although the slope of the plotted points diverges from the line of equality, the relative ranking of occupations by z-score remains consistent regardless of the method used. This indicates that while the actual z-score values differ, the ordinal relationship — the ranking from lowest to highest z-score — does not.

► Figure A7. Scatter plot of z-scores using mean vs. median for OSV



Nevertheless, to assess the impact of this different slope on the stability our observations, we recompute the analysis of individual occupations using the robust median z-scores and compare the results to the initial calculations. We find that the use of median-based z-scores leads to more pronounced extremes of the distribution and reduces the alignment between GPT-4 predictions and survey scores. Since one of our goals is to test the use of GPT-4 as a possible replacement or a complementary source of data for surveys related to occupations, we opt for the use of standard z-scores, which offer a more optimal result in that regard.

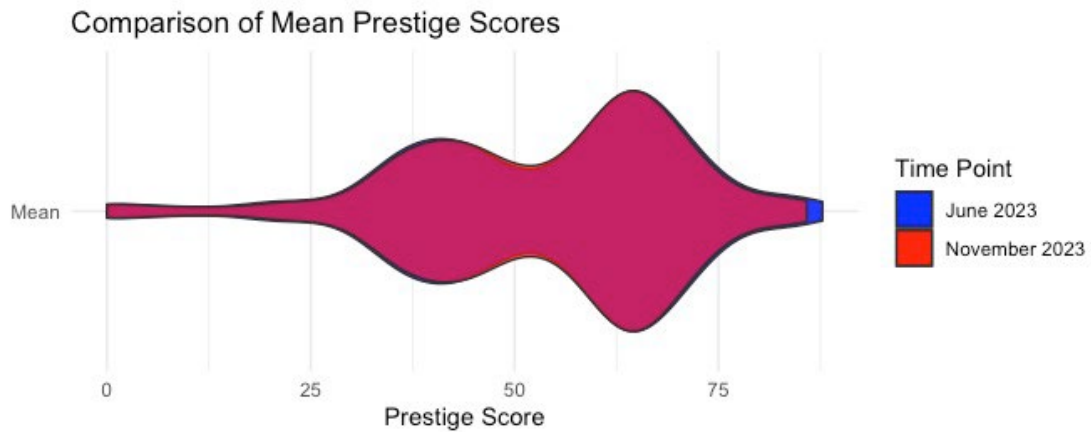
CONCLUSION REGARDING MEAN SCORES ACROSS OCCUPATIONS

In section A2 of this Appendix, we have addressed methodological concerns raised by a reviewer regarding the non-unimodal distribution of scores predicted by GPT-4 and its potential impact on our analysis. Through comprehensive robustness tests—conducted at both the individual score level and the occupation-level mean—we have demonstrated that, despite the unique distribution characteristics of GPT-4 predictions, the use of standard z-scores is best suited for the analysis presented in the main paper. These tests have included comparisons of central tendency measures, distribution analyses, and effect size calculations.

A3. STABILITY OF GPT-4 PREDICTIONS OVER TIME

The scores used in the main analysis were generated in June 2023. To assess the stability of these predictions, we generated a new set of scores for OP on 27 November 2023. As demonstrated in Figure A8, the scores predicted by GPT-4 are highly stable, with nearly identical distribution during among the two time points.

▶ Figure A8. Stability of GPT-4 OP score predictions (June 2023 vs November 2023)



References

Adamson, Maria, and Ian Roper. 2019. "'Good' Jobs and 'Bad' Jobs: Contemplating Job Quality in Different Contexts." *Work, Employment and Society* 33 (4): 551–559. <https://doi.org/10.1177/0950017019855510>.

Adler, Laura. 2021. "Choosing Bad Jobs: The Use of Nonstandard Work as a Commitment Device." *Work and Occupations* 48 (2): 207–242. <https://doi.org/10.1177/0730888420949596>.

Alkaiissi, Hussam, and Samy I. McFarlane. 2023. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing." *Cureus* 15 (2). <https://doi.org/10.7759/cureus.35179>.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023a. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–351. <https://doi.org/10.1017/pan.2023.2>.

Argyle, Lisa P., Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023b. "AI Chat Assistants Can Improve Conversations About Divisive Topics." arXiv. <https://arxiv.org/abs/2302.07268>.

Aroyo, Alexander M., Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones et al. 2021. "Overtrusting Robots: Setting a Research Agenda to Mitigate Overtrust in Automation." *Paladyn, Journal of Behavioral Robotics* 12 (1): 423–436. <https://doi.org/10.1515/pjbr-2021-0029>.

Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. "Which Humans?" PsyArXiv Preprints. <https://osf.io/preprints/psyarxiv/5b26t/>.

Avent-Holt, Dustin, Lasse Folke Henriksen, Anna Erika Häggglund, Jiwook Jung, Naomi Kodama, Silvia Maja Melzer, Eunmi Mun, Anthony Rainey, and Donald Tomaskovic-Devey. 2020. Occupations, Workplaces or Jobs? An Exploration of Stratification Contexts Using Administrative Data. *Research in Social Stratification and Mobility* 70: 100456. <https://doi.org/10.1016/j.rssm.2019.100456>.

Baiocco, Sara, Enrique Fernández-Macías, Uma Rani, and Annarosa Pesole. 2022. The Algorithmic Management of Work and Its Implications in Different Contexts. *JRC Working Papers Series on Labour, Education and Technology* 2022/02. <https://www.econstor.eu/handle/10419/262292>.

Balmer Andrew. 2023. A Sociological Conversation with ChatGPT about AI Ethics, Affect and Reflexivity. *Sociology*. <https://doi.org/10.1177/00380385231169676>.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. 2018. "What Can Machines Learn and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108 (May): 43–47. <https://doi.org/10.1257/pandp.20181019>.

Buchs, Helen, Emily Murphy, and Marlis Buchmann. 2017. "Landing a Job, Sinking a Career? The Trade-Off between Occupational Downgrading and Quick Reemployment according to Unemployed Jobseekers' Career Stage and Job Prospects." *Research in Social Stratification and Mobility* 52: 26–35. <https://doi.org/10.1016/j.rssm.2017.10.001>.

Bybee, Leland. 2023. "Surveying Generative AI's Economic Expectations." arXiv. <https://arxiv.org/abs/2305.02823>.

Cameron, Lindsey, Laura Lamers, Ulrich Leicht-Deobald, Christoph Lutz, Jeroen Meijerink, and Mareike Möhlmann. 2023. "Algorithmic Management: Its Implications for Information Systems

Research." *Communications of the Association for Information Systems* 52 (1): 518–537. <https://doi.org/10.17705/1CAIS.05221>.

Cardone, Pia, Markus Tümpel, and Christian M. Huber. 2021. "Temporary Employment, Permanent Stigma? Perceptions of Temporary Agency Workers across Low- and High-Skilled Jobs." *Qualitative Sociology Review* 17 (3): 6–33. <https://www.ceeol.com/search/article-detail?id=973833>.

Chan, Anastasia. 2023. "GPT-3 and InstructGPT: Technological Dystopianism, Utopianism, and 'Contextual' Perspectives in AI Ethics and Industry." *AI and Ethics* 3 (1): 53–64. <https://doi.org/10.1007/s43681-022-00148-6>.

Chan, Tak Wing, and John H. Goldthorpe. 2004. "Is There a Status Order in Contemporary British Society? Evidence from the Occupational Structure of Friendship." *European Sociological Review* 20 (5): 383–401. <https://doi.org/10.1093/esr/jch033>.

Chan, Tak Wing, and John H. Goldthorpe. 2007. "Class and Status: The Conceptual Distinction and Its Empirical Relevance." *American Sociological Review* 72 (4): 512–532. <https://doi.org/10.1177/000312240707200402>.

Cheng, Myra, Esin Durmus, and Dan Jurafsky. 2023. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." arXiv. <https://arxiv.org/abs/2305.18189>.

Coeckelbergh, Mark, and David J. Gunkel. 2023. "ChatGPT: Deconstructing the Debate and Moving It Forward." *AI & Society*: 1–11. <https://doi.org/10.1007/s00146-023-01710-4>.

Crockett, Molly, and Lisa Messeri. 2023. "Should large language models replace human participants?" PsyArXiv Preprints. <https://osf.io/preprints/psyarxiv/4zdx9/>.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraymer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24-013). <http://dx.doi.org/10.2139/ssrn.4573321>.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27 (7): 597–600. <https://doi.org/10.1016/j.tics.2023.04.008>.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. "Questioning the Survey Responses of Large Language Models." arXiv. <https://arxiv.org/abs/2306.07951>.

Duemmler, Kerstin, Isabelle Caprani, and Alexandra Felder. 2020. "The Challenge of Occupational Prestige for Occupational Identities: Comparing Bricklaying and Automation Technology Apprentices in Switzerland." *Vocations and Learning* 13 (3): 369–388. <https://doi.org/10.1007/s12186-020-09243-3>.

Dur, Robert, and Max Van Lent. 2019. "Socially Useless Jobs." *Industrial Relations* 58 (1): 3–16. <https://doi.org/10.1111/irel.12227>.

Edwards, Alba M. 1933. "A Social-Economic Grouping of the Gainful Workers of the United States." *Journal of the American Statistical Association* 28 (184): 377–387. <https://doi.org/10.1080/01621459.1933.10503237>.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." arXiv. <https://arxiv.org/abs/2303.10130>.

Else, Holly. 2023. "Abstracts Written by ChatGPT Fool Scientists." *Nature* 613 (7944): 423. <https://doi.org/10.1038/d41586-023-00056-7>.

Ernst, Ekkehardt, Rossana Merola, and Daniel Samaan. 2019. "Economics of Artificial Intelligence: Implications for the Future of Work." *IZA Journal of Labor Policy* 9 (1). <https://doi.org/10.2478/izajolp-2019-0004>.

Feng, Shangbin, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, 9–14 July, 11737–11762. <https://arxiv.org/abs/2305.08283>.

Floridi, Luciano. 2023. "AI as Agency without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models." *Philosophy & Technology* 36 (1): 15. <https://doi.org/10.1007/s13347-023-00621-y>.

Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines* 30: 681–694. <https://doi.org/10.1007/s11023-020-09548-1>.

Fontana, Enrico. 2020. "When the Main Job Tasks Are Perceived to Be 'Irrelevant' in the Workplace: The Internal Uselessness of Corporate Social Responsibility Work in Japan." *Culture and Organization* 26 (5-6): 405–424. <https://doi.org/10.1080/14759551.2019.1656632>.

Freeland, Robert E., and Jesse Hoey. 2018. "The Structure of Deference: Modeling Occupational Status Using Affect Control Theory." *American Sociological Review* 83 (2): 243–277. <https://doi.org/10.1177/0003122418761857>.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114 (January): 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>.

Gauchat, Gordon, and Kenneth T. Andrews. 2018. "The Cultural-Cognitive Mapping of Scientific Professions." *American Sociological Review* 83(3): 567–595. <https://doi.org/10.1177/0003122418773353>.

Gil-Hernández, Carlos J., Guillem Vidal, and Sergio Torrejón Perez. 2023. "Technological Change, Tasks and Class Inequality in Europe." *Work, Employment and Society*: 09500170231155783. <https://doi.org/10.1177/09500170231155783>.

Gmyrek, Pawel, Janine Berg, and David Bescond. 2023. "Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality." Working paper. http://www.ilo.org/global/publications/working-papers/WCMS_890761/lang-en/index.htm.

Goldman Sachs. 2023. "Global Economics Analyst: The Potentially Large Effects of Artificial Intelligence on Economic Growth." https://static.poder360.com.br/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf.

Graeber, David. 2018. *Bullshit Jobs: A Theory*. London: Penguin Books.

Hämäläinen, Perttu, Mikke Tavast, and Anton Kunnari. 2023. "Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, 23–28 April, pp. 1–19. New York: ACM. <https://doi.org/10.1145/3544548.3580688>.

Hennekam, Sophie, Jamie Ladge, and Yuliya Shymko. 2020. "From Zero to Hero: An Exploratory Study Examining Sudden Hero Status among Nonphysician Health Care Workers during the

COVID-19 Pandemic." *Journal of Applied Psychology* 105 (10): 1088–1100. <https://doi.org/10.1037/apl0000832>.

Harding, Jacqueline, William D'Alessandro, N. G. Laskowski, and Robert Long 2023. "AI Language Models Cannot Replace Human Research Participants." *AI & Society*: 1–3. <https://doi.org/10.1007/s00146-023-01725-x>.

Hauser, Robert M., and John Robert Warren. 1997. "Socioeconomic Indexes for Occupations: A Review, Update, and Critique." *Sociological Methodology* 27 (1): 177–298. <https://doi.org/10.1111/1467-9531.271028>.

Hepp, Andreas, Wiebke Loosen, Stephan Dreyer, Juliane Jarke, Sigrid Kannengießer, Christian Katzenbach, Rainer Malaka, Michaela Pfadenhauer, Cornelius Puschmann, and Wolfgang Schulz. 2023. "ChatGPT, LaMDA, and the Hype around Communicative AI: The Automation of Communication as a Field of Research in Media and Communication Studies." *Human-Machine Communication* 6 (1): 41–63. <https://doi.org/10.30658/hmc.6.4>.

Hu, Krystal. 2023. "ChatGPT's Explosive Growth Shows First Decline in Traffic Since Launch. Reuters, 5 July. <https://www.reuters.com/technology/booming-traffic-openais-chatgpt-posts-first-ever-monthly-dip-june-similarweb-2023-07-05/>.

ILO. 2008. "International Standard Classification of Occupations: Structure, group definitions and correspondence tables (ISCO-08)." International Labour Organization (ILO), Geneva. Available at: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf.

ILO. 2023. "World Employment and Social Outlook 2023: The Value of Essential Work." <https://www.ilo.org/digitalguides/en-gb/story/weso2023-key-workers>.

Jarrahi, Mohammad Hossein, Gemma Newlands, Min Kyung Lee, Christine T. Wolf, Eliscia Kinder, and Will Sutherland. 2021. "Algorithmic Management in a Work Context." *Big Data & Society* 8 (2): 20539517211020332. <https://doi.org/10.1177/20539517211020332>.

Jarrahi, Mohammad Hossein, Gemma Newlands, and Christoph Lutz. 2022. "Artificial Intelligence, Human Intelligence and Hybrid Intelligence Based on Mutual Augmentation." *Big Data & Society* 9 (2): 20539517221142824.

Kalinin, Kirill. 2023. "Improving GPT Generated Synthetic Samples with Sampling-Permutation Algorithm." SSRN Scholarly Paper. Rochester, NY. <http://dx.doi.org/10.2139/ssrn.4548937>.

Kane, Gerald C. 2017. "Digital Maturity, Not Digital Transformation." *MIT Sloan Management Review* April 4. <https://sloanreview.mit.edu/article/digital-maturity-not-digital-transformation/>.

Kasirzadeh, Atoosa, and Iason Gabriel. 2023. "In Conversation with Artificial Intelligence: Aligning Language Models with Human Values." *Philosophy & Technology* 36 (2): 1–24. <https://doi.org/10.1007/s13347-023-00606-x>.

Kim, Junsol, and Byungkyu Lee. 2023. "AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys." arXiv. <https://arxiv.org/abs/2305.09620>.

Kim, Sang-Wook, and Charles W. Mueller. 2011. "Occupational and Organizational Commitment in Different Occupational Contexts: The Case of South Korea." *Work and Occupations* 38 (1): 3–36. <https://doi.org/10.1177/0730888410383367>.

Klein, Hans K., and Daniel Lee Kleinman. 2002. "The Social Construction of Technology: Structural Considerations." *Science, Technology, & Human Values* 27 (1): 28–52. <https://doi.org/10.1177/016224390202700102>.

Lambert, Paul S., and Erik Bihagen. 2014. "Using Occupation-Based Social Classifications." *Work, Employment and Society* 28 (3): 481–494. <https://doi.org/10.1177/0950017013519845>.

Lane, Marguerita, and Anne Saint-Martin. 2021. "The Impact of Artificial Intelligence on the Labour Market: What Do We Know So Far?" *OECD Social, Employment and Migration Working Papers* No. 256. <https://doi.org/10.1787/7c895724-en>.

Laverghetta Jr, Antonio, and John Licato. 2023. "Generating Better Items for Cognitive Assessments Using Large Language Models." In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 414–428, Toronto, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.34>.

Lee, Kibeom, Julie J. Carswell, and Natalie J. Allen. 2000. "A Meta-Analytic Review of Occupational Commitment: Relations with Person- and Work-related Variables." *Journal of Applied Psychology* 85 (5): 799. <https://doi.org/10.1037/0021-9010.85.5.799>.

Lehdonvirta, Vili, Atte Oksanen, Pekka Räsänen, and Grant Blank. 2021. "Social media, web, and panel surveys: Using non-probability samples in social and policy research." *Policy & Internet* 13 (1): 134–155. <https://doi.org/10.1002/poi3.238>.

Leicht, Kevin T. 2020. "Occupations and Inequalities in the 21st Century: What's in Your Wallet?" *Research in Social Stratification and Mobility* 70: 100550. <https://doi.org/10.1016/j.rssm.2020.100550>.

Lersch, Philipp M., Wiebke Schulz, and George Leckie. 2020. "The Variability of Occupational Attainment: How Prestige Trajectories Diversified within Birth Cohorts over the Twentieth Century." *American Sociological Review* 85 (6): 1084–1116. <https://doi.org/10.1177/0003122420966324>.

Lissitsa, Sabina, Svetlana Chachashvili-Bolotin, and Ya'arit Bokek-Cohen. 2017. "Can Digital Skills Raise Occupational Prestige among Ethnic Minorities?" *Innovation: The European Journal of Social Science Research* 30 (2): 204–222. <https://doi.org/10.1080/13511610.2017.1282309>.

Lynn, Freda B., and George Ellerbach. 2017. "A Position with a View: Educational Status and the Construction of the Occupational Hierarchy." *American Sociological Review* 82 (1): 32–58. <https://doi.org/10.1177/0003122416671743>.

MacKinnon, Neil J., and Tom Langford. 1994. "The Meaning of Occupational Prestige Scores: A Social Psychological Analysis and Interpretation." *The Sociological Quarterly* 35 (2): 215–245. <https://doi.org/10.1111/j.1533-8525.1994.tb00408.x>.

Makridis, Christos A., and Joo Hun Han. 2021. "Future of Work and Employee Empowerment and Satisfaction: Evidence from a Decade of Technological Change." *Technological Forecasting and Social Change* 173: 121162. <https://doi.org/10.1016/j.techfore.2021.121162>.

McKinsey. 2023. "Economic Potential of Generative AI." <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/>.

Mejia, Cynthia, Rebecca Pittman, Jenna MD Beltramo, Kristin Horan, Amanda Grinley, and Mindy K. Shoss. 2021. "Stigma & Dirty Work: In-Group and Out-Group Perceptions of Essential Service Workers During COVID-19." *International Journal of Hospitality Management* 93: 102772. <https://doi.org/10.1016/j.ijhm.2020.102772>.

Mills, Stuart, Samuel Costa, and Cass R. Sunstein. 2023. "The Opportunities and Costs of AI in Behavioural Science." SSRN Scholarly Paper. Rochester, NY. <http://dx.doi.org/10.2139/ssrn.4490597>.

Newlands, Gemma. 2021. "Algorithmic Surveillance in the Gig Economy: The Organization of Work Through Lefebvrian Conceived Space." *Organization Studies* 42 (5): 719–737. <https://doi.org/10.1177/0170840620937900>.

Newlands, Gemma, and Christoph Lutz. 2023a. "Occupational Prestige and Occupational Social Value in the United Kingdom: New Indices for the Modern British Economy." Under review. Working paper version available upon request.

Newlands, Gemma, and Christoph Lutz. 2023b. "Mapping the Prestige and Social Value of Occupations in the Digital Economy." Under review. Working paper version available upon request.

Oesch, Daniel, and Giorgio Piccitto. 2019. "The Polarization Myth: Occupational Upgrading in Germany, Spain, Sweden, and the UK, 1992–2015." *Work and Occupations* 46 (4): 441–469. <https://doi.org/10.1177/0730888419860880>.

OpenAI. 2023. GPT-4 technical report. arXiv. <https://arxiv.org/abs/2303.08774>.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17: 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>.

Pitt, Richard N., and Lin Zhu. 2019. "The Relationship Between College Major Prestige/Status and Post-Baccalaureate Outcomes." *Sociological Perspectives* 62 (3): 325–345. <https://doi.org/10.1177/0731121418803325>.

Prolific. 2023a. Representative samples. <https://researcher-help.prolific.com/hc/en-gb/articles/360019236753-Representative-samples>.

Prolific. 2023b. Audience. <https://researcher-help.prolific.com/hc/en-gb/articles/4407449546002-Audience>.

Rosenbusch, Hannes, Claire E. Stevenson, and Han LJ van der Maas. 2023. "How Accurate Are GPT-3's Hypotheses About Social Science Phenomena?" *Digital Society* 2 (2): Article 26. <https://doi.org/10.1007/s44206-023-00054-2>.

Sætra, Henrik Skaug. 2023. "Generative AI: Here to Stay, But for Good?" *Technology in Society* 75: 102372. <https://doi.org/10.1016/j.techsoc.2023.102372>.

Safdari, Mustafa, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. "Personality Traits in Large Language Models." arXiv. <https://arxiv.org/abs/2307.00184>.

Salvagno, Michele, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. "Can Artificial Intelligence Help for Scientific Writing?" *Critical Care* 27 (1): 1–5. <https://doi.org/10.1186/s13054-023-04380-2>.

Sengupta, Sukanya, Paul K. Edwards, and Chin-Ju Tsai. 2009. "The Good, the Bad, and the Ordinary: Work Identities in 'Good' and 'Bad' Jobs in the United Kingdom." *Work and Occupations* 36 (1): 26–55. <https://doi.org/10.1177/0730888408329222>.

Schreiner, Maximilian. 2023. GPT-4 architecture, datasets, costs and more leaked. The decoder, 11 July 2023. <https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/>.

Shestakofsky, Benjamin. 2017. "Working Algorithms: Software Automation and the Future of Work." *Work and Occupations* 44 (4): 376–423. <https://doi.org/10.1177/0730888417726119>.

Smith, Tom William, and Jaesok Son. 2014. Measuring Occupational Prestige on the 2012 General Social Survey. Vol. 4. NORC at the University of Chicago. <https://gss.norc.org/Documents/reports/methodological-reports/MR122%20Occupational%20Prestige.pdf>.

Soffia, Magdalena, Alex J. Wood, and Brendan Burchell. 2022. "Alienation Is Not 'Bullshit': An Empirical Critique of Graeber's Theory of BS Jobs." *Work, Employment and Society* 36 (5): 816–840. <https://doi.org/10.1177/09500170211015067>.

Thorp, H. Holden. 2023. "ChatGPT Is Fun, But Not an Author." *Science* 379 (6630): 313–313. <https://doi.org/10.1126/science.adg7879>.

Treiman, Donald J. 1977. *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

Tubaro, Paola, Antonio A. Casilli, and Marion Coville. 2020. "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence." *Big Data & Society* 7 (1): 2053951720919776. <https://doi.org/10.1177/2053951720919776>.

Valentino, Lauren. 2020. "The Segregation Premium: How Gender Shapes the Symbolic Valuation Process of Occupational Prestige Judgments." *Social Forces* 99 (1): 31–58. <https://doi.org/10.1093/sf/soz145>.

Valentino, Lauren. 2021. "The Heterarchy of Occupational Status: Evidence for Diverse Logics of Prestige in the United States." *Sociological Forum* 36: 1395–1418. <https://doi.org/10.1111/socf.12762>.

Walo, Simon. 2023. "'Bullshit' After All? Why People Consider Their Jobs Socially Useless." *Work, Employment and Society* 37 (5): 1123–1146. <https://doi.org/10.1177/09500170231175771>.

Webb, Michael. 2019. "The Impact of Artificial Intelligence on the Labor Market." SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.3482150>.

WEF. 2023. "The Future of Jobs Report 2023." <https://www.weforum.org/reports/the-future-of-jobs-report-2023>.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese et al. 2022. "Taxonomy of risks posed by language models." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). <https://doi.org/10.1145/3531146.3533088>.

Wolfe, Marcus T., and Pankaj C. Patel. 2019. "Labor of Love? The Influence of Work-Conditions Among Self-Employed and Work Stress." *Journal of Business Venturing Insights* 11: e00118. <https://doi.org/10.1016/j.jbvi.2019.e00118>.

Zhou, Xueguang. 2005. "The Institutional Logic of Occupational Prestige Ranking: Reconceptualization and Reanalyses." *American Journal of Sociology* 111 (1): 90–140. <https://doi.org/10.1086/428687>.

Acknowledgements

Acknowledgements and use of GPT

We acknowledge support from several colleagues at the ILO. We are particularly grateful to Janine Berg, Bálint Náfrádi, Sergei Soares and David Kucera for their comments on the content. We also thank Bálint for the independent review of our calculations and his subsequent suggestions on the statistical treatment of the distributional particularities in AI-generated data. We thank other colleagues in the ILO's Research Department for their helpful feedback during initial internal presentations of this work.

GPT-4 API was used to generate synthetic respondent data as discussed in the text. The final dataset generated with GPT-4 consists of 580 sequential API calls simulating 100 independent respondents, which corresponds to 5,800 API responses executed in June 2023. Another 2400 API calls were executed in November 2023 to test the consistency of predictions over time. OpenAI provided the ILO Research Department with a research credit in API tokens with a total value of US\$ 1,000, out of which some US\$ 400 have been used for generating testing and final datasets for this research.

The authors of this text have no affiliation to OpenAI and no vested interest in that regard. GPT was not used to generate the main text of this article, which is fully of our own authorship, along with any mistakes.

The occupational prestige and occupational social value surveys were funded within the Research Council of Norway project 275347 "Future Ways of Working in the Digital Economy" at BI Norwegian Business School. Christoph Lutz' and Gemma Newlands' positions at the time of data collection were also partly funded through this project and, in Lutz' case, the project 299178 "Algorithmic Accountability: Designing Governance for Responsible Digital Transformations" as well. We would like to thank the Research Council of Norway for the generous support.

Author names have been organized alphabetically, indicating equal input from all three authors.

► Advancing social justice, promoting decent work

The International Labour Organization is the United Nations agency for the world of work. We bring together governments, employers and workers to improve the working lives of all people, driving a human-centred approach to the future of work through employment creation, rights at work, social protection and social dialogue.

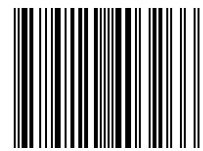
Contact details

Research Department (RESEARCH)

International Labour Organization
Route des Morillons 4
1211 Geneva 22
Switzerland
T +41 22 799 6530
research@ilo.org
www.ilo.org/research



ISBN 9789220403266



9 789220 403266