PDF Download
3676281.pdf
29 December 2025
Total Citations: 2
Total Downloads: 2826

Published: 16 June 2025
Online AM: 02 July 2024
Accepted: 26 May 2024
Revised: 07 April 2024
Received: 30 January 2024

Citation in BibTeX format

RESEARCH-ARTICLE

ChatGPT on ChatGPT: An Exploratory Analysis of its Performance in the Public Sector Workplace

JIESHU WANG, Arizona State University, Tempe, AZ, United States

ELIF KIRAN, Arizona State University, Tempe, AZ, United States

S R AURORA, The University of Texas Rio Grande Valley, Brownsville, TX, United States

MICHAEL SIMEONE, Arizona State University, Tempe, AZ, United States

JOSÉ LOBO, Arizona State University, Tempe, AZ, United States

Open Access Support provided by:

The University of Texas Rio Grande Valley

Arizona State University

Digital Government: Research and Practice, Volume 6, Issue 2 (June 2025)
https://doi.org/10.1145/3676281
EISSN: 2639-0175

ChatGPT on ChatGPT: An Exploratory Analysis of its Performance in the Public Sector Workplace

JIESHU WANG, Decision Theater, Arizona State University, Tempe, United States
ELIF KIRAN, School of Applied Professional Studies, Arizona State University, Tempe, United States
S.R. AURORA, Robert C. Vackar College of Business and Entrepreneurship, The University of Texas Rio Grande Valley, Brownsville, United States
MICHAEL SIMEONE, School of Complex Adaptive Systems, Arizona State University, Tempe, United States
JOSÉ LOBO, School of Sustainability, Arizona State University, Tempe, United States

This study explores the impact of Generative Artificial Intelligence (GenAI), in particular, ChatGPT, on the public sector workforce in the United States, focusing on task replacement, assistance potential, and the evolving landscape of skills. Utilizing GPT-4 to evaluate 1,022 core tasks across 51 public sector occupations, we provide an exploratory analysis of the roles susceptible to ChatGPT automation and those in which ChatGPT can augment human efforts. Our findings reveal that while 63% of tasks are resistant to ChatGPT replacement, primarily due to their requirement for physical presence, emotional intelligence, and complex decision-making, tasks that are routine, rule-based, and involving basic content generation show a high potential for automation. The study also identifies key skills that will remain vital, those likely to become obsolete, and new skills that will emerge as essential, highlighting the need for a strategic approach to workforce development in the face of AI advancements. In particular, our findings underscore the growing importance of skills in applying AI technologies and the ability to validate and interpret AI-generated content for humans to remain competitive. We offer insights into public-sector-specific impacts and propose a methodological framework for future research, emphasizing the importance of adapting educational curricula and policies to prepare for an AI-integrated future.

CCS Concepts: • **Social and professional topics** → **Employment issues**; **Automation**; *Economic impact*; Adult education; Socio-technical systems; • **Computing methodologies** → **Natural language generation**; • **Applied computing** → *E-government;*

Additional Key Words and Phrases: Public sector, workforce, artificial intelligence, large language models, ChatGPT, future of work

ACM Reference Format:
Jieshu Wang, Elif Kiran, S.R. Aurora, Michael Simeone, and José Lobo. 2025. ChatGPT on ChatGPT: An Exploratory Analysis of its Performance in the Public Sector Workplace. *Digit. Gov. Res. Pract.* 6, 2, Article 29 (June 2025), 28 pages. https://doi.org/10.1145/3676281

Authors' Contact Information: Jieshu Wang, Decision Theater, Arizona State University, Tempe, Arizona, United States; e-mail: jwang490@asu.edu; Elif Kiran, School of Applied Professional Studies, Arizona State University, Tempe, Arizona, United States; e-mail: ekiran@asu.edu; S.R. Aurora, Robert C. Vackar College of Business and Entrepreneurship, The University of Texas Rio Grande Valley, Brownsville, Texas, United States; e-mail: s.r.aurora.2023@gmail.com; Michael Simeone, School of Complex Adaptive Systems, Arizona State University, Tempe, Arizona, United States; e-mail: Michael.Simeone@asu.edu; José Lobo, School of Sustainability, Arizona State University, Tempe, Arizona, United States; e-mail: Jose.Lobo@asu.edu.

ACM 2639-0175/2025/06-ART29
https://doi.org/10.1145/3676281

1 Introduction

The expansion of available tools that use **artificial intelligence (AI)** has created a rapidly accelerating wave of technological changes. The implications of these changes, both challenges and opportunities, have surged to the forefront of concern. Central to this conversation recently is the emergence and availability of **generative AI (GenAI)** technologies like ChatGPT, sophisticated models that not only interact with users but can also create new, contextually relevant content, raising profound questions about their role in social and economic life. As these technologies could automate, augment, or even obviate certain skills and labor activities, questions about a person's changing role in social and economic life turn to the effect of these technologies on livelihood, employment, and how jobs are carried out.

This article investigates the impact of ChatGPT on the public sector workforce. While employment in the public sector in the United States—federal, state, and local governments—represents only about 15% of all jobs, the public sector represents a valuable and important setting in which to examine the effects of generative AI on employment and the consequences of these effects. Public services directly and indirectly affect the well-being of nearly every resident of the United States, and the efficacy with which these services are provided greatly depends on public sector workers.

Many dynamics and concerns regarding the effect of technology adoption on jobs are common in the private and public sectors. However, the manner and pace of technology adoption in the public sector and the normative criteria governing the adoption differ in important and ultimately revealing ways. Our investigation began with identifying 51 occupations in the public sector. We then utilized the GPT-4 **Application Programming Interface (API)** to assess all 1,022 core tasks associated with these occupations, evaluating how they could be replaced or augmented by ChatGPT. Additionally, we explored what existing skills are essential and what new skills may be required in the future in the presence of Generative AIs.

Our findings reveal a nuanced situation in which ChatGPT has varying degrees of impact across different occupations in the public sector. While some tasks are highly susceptible to automation, others remain firmly within the human domain, necessitating skills that AI currently cannot replicate. We discovered that tasks involving repetitive, rule-based activities are more likely to be replaced by ChatGPT, whereas those requiring physical interaction, complex decision-making, or emotional intelligence are less prone to automation. Furthermore, our analysis highlights the emerging necessity for skills in efficiently utilizing AI tools and interpreting AI-generated content, underscoring the evolving nature of skill requirements in an AI-integrated workplace. This study offers a preliminary overview of AI's potential roles and limitations in transforming the public sector workforce, providing valuable insights for policymakers, educators, and professionals navigating this rapidly evolving landscape.

2 Literature Review

2.1 Automation and Future of Work

The concept of work in the 21st century is being reshaped by automation and technological advancements. Automation, encompassing AI and robotics, is altering job roles, skill requirements, and overall work dynamics in industries such as manufacturing, transportation, education, and even scientific research [1, 13, 17, 31, 45]. The International Society of Automation describes automation as the application of technology to monitor and control production and service delivery [25]. This technological advancement has led to the phenomenon of "technological unemployment," a term that encapsulates the replacement of traditional roles such as switchboard and elevator operators with automated systems [41] and of which the discussion goes back to Keynes' article at the beginning period of the Great Depression in the 1930s [27].

Keynes' foresight nearly a century ago anticipated the challenges and transformations brought about by technological advancements. Today, we grapple with the double-edged sword of technology, experiencing both its benefits and the pitfalls of technological unemployment, which exacerbates inequalities and disrupts the balance between labor and capital. Various solutions have been suggested to address this issue, from the Luddite approach of resisting innovation to progressive strategies such as welfare, public employment programs, and universal basic income. Economists also advocate for measures such as subsidies for small businesses, reduced workweeks, and public ownership of tech infrastructure [36]. On the flip side, technological progress has reduced the overall work time, enhanced the quality of work and leisure, improved working conditions, and opened up opportunities for workforce emancipation, particularly for women. It has also expanded leisure options and made travel more accessible [46].

The rise of AI and robotics, particularly evident in the current Fourth Industrial Revolution, is transforming economies. Building on digital advancements since the 1950s, the current revolution is characterized by a fusion of technologies that blur lines between physical, digital, and biological realms [31, 39]. While they promise enhanced productivity and wealth creation, they also pose risks to job security, particularly for roles susceptible to automation [4, 9, 11, 26, 29]. Frey and Osborne's prediction that almost half of U.S. jobs could be automated [21] and the World Economic Forum's projection of 85 million jobs being displaced by 2025 highlight these risks [44]. The challenge is managing inequality, as those with automation-complementing skills may thrive while others face reduced employment opportunities [6].

Nevertheless, the recent development of GenAI technologies like ChatGPT is reshaping our understanding of the future of work. Contrary to earlier beliefs that AI would mainly displace low-skilled workers, recent studies suggest that GenAIs might offer significant benefits to mid- or low-skilled workers, enabling them to produce content that meets average standards [8, 10, 22, 42].

2.2 AI and Public Sector Workforce

While discussions about automation and technological innovation often center on the private sector, the public sector is equally subject to the profound transformations driven by these advancements. The U.S. federal government, for instance, has increasingly integrated AI-powered technologies into a wide array of public services, including healthcare, education, housing, and social benefits [2, 15, 34, 38, 43, 47]. This integration is widespread, with nearly half of the U.S. federal agencies employing more than 400 people either implementing or planning to implement AI in their operations [20]. The overarching goal of these AI initiatives is to utilize the vast data collected by government agencies to enhance service quality and efficiency while increasing responsiveness to public needs [30].

However, the path of technological adoption in the public sector differs significantly from that in the private sector, not only in terms of pace but also in the nature of challenges and objectives. The public sector has certain characteristics, which also influence its interaction with automation and, specifically, AI. The adoption of new technologies is not governed mainly by considerations of efficiency or cost reduction. Whereas the management of firms might need to worry only about the concerns of proprietors or shareholders, the stakeholders with regard to the operations of the public sector are more numerous and more varied. This can have an impact on the way that automation and AI are integrated and utilized. For example, public sector organizations often have a strong focus on accountability, transparency, and equity, which can influence how AI systems are designed and implemented [12]. Additionally, the public sector may also have different data privacy and security requirements compared to the private sector, which can further impact the use of AI technologies. Besides these, different from the environments fostering new ideas and innovation, the public sector's natural tendency towards stability and predictability often hinders change [5].

Commonly cited hindrances to technological experimentation in the public sector include heavy bureaucracy, universally applied and inflexible rules and procedures discouraging innovation, intricate accountability and participation processes, funding shortages, unclear laws limiting innovation, lack of support from managers, narrow vision about the future, and risk aversion. Achieving change in the public sector can be challenging, given these factors [5]. Thus, Craig et al. [14] argued that due to the perception that new technologies pose a threat to traditional bureaucratic employment, the adoption of technology is hindered by public sector unionization [14].

Moreover, the public sector differs from the private sector in terms of certain core values, the embracing of which may also cause a challenge in adopting automation in its natural workflow. The dedication to fair employment and the avoidance of discrimination are vital aspects of this framework of ethical issues. The potential asymmetric distribution of chances for introduction and opportunities for improvement in the vast landscape of digital technologies embodies the risk of diminishing these values. For instance, if automation focuses on labor divisions already divided based on gender or race, then it is crucial to examine how the process of automation might affect the public sector's strong emphasis on fostering diversity and equality in employment opportunities [7].

Therefore, while the adoption of AI in the public sector holds immense potential for enhancing service delivery and efficiency, it must be navigated with a nuanced understanding of the sector's specific challenges and core values. Balancing technological innovation with these considerations is crucial to ensure that the public sector not only keeps pace with technological advancements but also upholds its foundational principles.

3 Research Questions

We embarked on this study with the aim to address the following research questions:

RQ1. How does ChatGPT evaluate its ability to *replace* humans in performing tasks involved in occupations within the public sector in the United States?

RQ2. How does ChatGPT evaluate its ability to *augment* humans in performing tasks involved in the public sector workplace? Specifically, how can ChatGPT assist human efforts, and how does it enhance task productivity?

RQ3. Based on ChatGPT's evaluations on tasks, how does ChatGPT influence public sector occupations? Specifically, which occupations are most susceptible to and which are least likely to be affected by ChatGPT's capabilities?

RQ4. How does ChatGPT evaluate the *skills* that humans will need to retain or develop to stay relevant, and which existing skills might become obsolete in the near future?

While **RQ1** and **RQ2** might seem alike, they examine the distinctions between AI's replacement and augmentation of human labor. ChatGPT's replacement of humans (**RQ1**) focuses on its capability to autonomously perform tasks, reducing or eliminating the need for human intervention. However, this does not imply the possibility of assisting humans or an enhancement in productivity, quality, or efficiency. Conversely, augmentation by ChatGPT (**RQ2**) emphasizes a collaboration between human and AI that may enhance task execution speed and quality. As Engelbart [1962] noted, augmenting human intellect is about "increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems" [19].

Regarding **RQ2**, it is important to distinguish between two seemingly similar yet distinct aspects of ChatGPT's augmentation of human labor: assistance and productivity enhancement. The first aspect focuses on the extent to which ChatGPT can support humans in executing tasks, while the second evaluates how ChatGPT's assistance can accelerate task completion—specifically, quantifying how much faster a task can be accomplished with ChatGPT's involvement compared to relying solely on human effort.[1] This differentiation stems from the diverse nature of occupational tasks. While it might seem intuitive that tasks significantly aided by ChatGPT would naturally see notable productivity boosts, this correlation is not always direct. Tasks receiving similar

[1] The *exposure rubric*, as defined by Eloundou et al. [2023], utilizes a binary measure to evaluate an occupation's exposure to large language models (LLMs) [18]. A task is considered exposed to LLMs if, with LLM assistance, it can be executed in less than 50% of the original time required [18]. Adopting the reasoning behind such a framework, we acknowledge the significant role that time reduction (or as conceptualized in our study, productivity enhancement) plays in demonstrating the potential of LLMs to augment human labor. However, instead of a binary measure, we propose a continuous scale—quantifying the extent of time reduction or the factor by which productivity could be multiplied—to offer a more detailed understanding of LLMs' impact.

levels of assistance from ChatGPT might not experience comparable increases in productivity due to the diverse ways in which AI support manifests.[2]

4 Methods and Data

This study uses GPT-4, a **large language model (LLM)** to gain insight into the future of public sector labor in the context of the widespread adoption of AI technologies. The procedure of "asking an AI about AI" has recently been explored by researchers [18] and is used in this study to create estimations based on a large sample of text material contained in the model training data. The estimations generated by GPT-4 are not a simulation or traditional forecasting model; they summarize the extant training data that comprise the LLM model. With this in mind, the estimations from GPT-4 are treated as concise representations of extensive material on a given subject, not the equivalent of an expert opinion. That is to say, although we validated output for quality and accuracy, experts may disagree with the substance of certain estimates or responses from the model. Overall, our evaluation found that GPT-4 was able to provide a broad and consistent set of estimates about the future of public sector labor. Key to our rigor in this research is the consistency and systematic approach to prompting.

The empirical work reported here utilizes terms that, while used prominently in casual conversations about the economy, have a precise meaning when it comes to data collection. We find it useful to distinguish between employment, jobs, occupations, and tasks using standard definitions from labor economics. An *occupation* is a craft, trade, profession, or other means of earning a living. A *task* is a specific work activity performed in an occupation. Tasks are typically conceptualized as the smallest unit of activity with a recognizable outcome. Employees who are in the same occupation perform essentially the same tasks, whether or not they work in the same industry. While occupations are defined by the set of tasks they require, different occupations can share tasks. *Skills* are the application of knowledge, basic or specialized, in a work setting. A *job* is a specific instance of employment, a position of employment to be filled at a workplace. *Employment* denotes the number of jobs in an occupation, including full-time jobs, part-time jobs, and self-employment.

4.1 Identifying Public Sector Occupations

In this study, we identified a total of 51 occupations representing the public sector workforce in the United States as of May 2022. These occupations encompass 1,022 core tasks and involve 6,679,080 workers. A list and description of these 51 occupations are available in Appendix C.

To compile this dataset, we initially sourced occupation information from the **Occupational Information Network (O*NET)**, a database developed by the **U.S. Department of Labor's Employment and Training Administration (USDOL/ETA)**.[3] This database provides extensive details about the activities, tasks, abilities, and skills associated with an occupation. Skills are divided into basic skills (such as reading, which facilitates the acquisition of new knowledge) and cross-functional skills (such as problem-solving, which extends across several domains of activities). Subsequently, we obtained national occupational employment statistics, segmented by employee type, from the **Bureau of Labor Statistics (BLS)** as of May 2022.[4] This dataset classifies employment

[2] For instance, our findings reveal instances where tasks estimated by GPT-4 to receive high assistance levels did not correspondingly show substantial productivity gains. One notable example involves the task of career or technical education teachers instructing students in specific occupational skills using a systematic plan of lectures, discussions, audio-visual presentations, and laboratory, shop, and field studies. Despite GPT-4 assigning a high assistance score (8/10), indicating substantial support from ChatGPT in enriching lesson plans and creating engaging content, the estimated productivity increase was only moderate (1.3). This discrepancy highlights ChatGPT's limitations in substituting for hands-on guidance and real-time problem-solving that are crucial in career training. Such observations underscore the connections yet differences between ChatGPT's assistance and potential productivity enhancement, a consideration that has informed how we designed our prompts (see Section 4.2.2).

[3] We utilized version 28.0 of the O*NET database, updated in August 2023. Detailed information on this and previous versions can be found in the O*NET Database Release Archive (https://www.onetcenter.org/db_releases.html)

[4] https://www.bls.gov/oes/current/999001.htm

ownership into two main categories: private sector and government, which includes federal, state, and local government entities.

We merged the occupational tasks dataset with the employment statistics dataset using the ***Standard Occupational Classification*** **(SOC)** code, the coding system used by all agencies of the U.S. federal government to classify workers into occupational categories. It is important to note that while the SOC system used by O*NET and the **occupation codes (OCC)** used by the BLS are generally aligned, there are minor discrepancies. For example, one OCC code, such as 33-3021 (Detectives and Criminal Investigators), might correspond to multiple SOC occupations such as Police Identification and Records Officers (33-3021.02) and Intelligence Analysts (33-3021.06). In such cases, we treated these related occupations as a single occupation due to their closely related job functions and the availability of employment data only at OCC levels.

In this study, an occupation was classified as public if it simultaneously met two criteria: (1) at least 80% of its nationwide employment was within federal, state, or local governments, and (2) the occupation's national employment in the public sector exceeded 2,000 individuals. This approach initially identified 56 occupations. However, due to minor data inconsistencies across different government agencies, 5 of these 56 occupations either lacked detailed task information or were not included in the O*NET database version 28.0 that we used.[5] Consequently, our final dataset comprises 51 occupations.

The 51 public sector occupations we selected are associated with a total of 1,262 tasks, comprising 1,022 core tasks and 240 supplementary tasks. According to the O*NET documentation, *core* tasks are defined as tasks central to the occupation and expected to be performed by the majority of incumbents. In contrast, *supplementary* tasks include those not central to the occupation, as well as support tasks such as supervisory and report writing activities [16]. This study focuses exclusively on the 1,022 core tasks.

These identified public sector occupations span 11 major groups, detailed in Appendix C. These groups range from business and financial operations to healthcare practitioners and technical occupations, encompassing management and legal occupations. Notably, 15 of these occupations fall under protective services, including transportation security screeners, animal control workers, transit and railroad police, and firefighters. Thirteen occupations are in the field of educational instruction and library occupations, which include kindergarten, elementary, and middle school teachers, as well as special education teachers. Additionally, the occupations cover office and administrative support roles, such as postal service mail carriers, public safety telecommunicators, eligibility interviewers for government programs, and library assistants. Other notable occupations include judges, police patrol officers, urban planners, forest technicians, and air traffic controllers.

A sample of the public sector occupational task dataset can be found in Table 1. The full dataset is openly available; please refer to the Data Availability section for more information.

4.2 Prompt Engineering

4.2.1 The Process of Iterative Prompt Engineering. ***Generative artificial intelligence*** **(generative AI or GenAI)** is a type of artificial intelligence (more precisely, a type of machine learning) that can generate new content such as text, images, or other media using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics [23]. A *prompt* refers to the input given by a user to a generative AI model. The practice of creating such prompts to guide an AI model in producing specific outcomes is known as *prompt engineering* [37]. We developed our prompts through an iterative and heuristic approach, with the objective of instructing OpenAI's GPT models to generate both quantitative estimates and qualitative explanations. Iterative refinement is widely regarded as a best practice in prompt engineering and has been integrated into various prompt engineering protocols. This approach is designed to enhance GenAI's performance through multiple rounds of validation and refinement

[5] The five occupations not fully accounted for are (1) legislators, (2) property appraisers and assessors, (3) special education teachers in kindergarten and elementary school, (4) short-term substitute teachers, and (5) teachers and instructors in various other categories.

Table 1. A Sample of the Occupational Task Dataset

Job title	Employment	Gov-owned pct	Task
Detectives and Criminal Investigators	107,310	99.9%	Look for trace evidence, such as fingerprints, hairs, fibers, or shoe impressions, using alternative light sources when necessary.
Fish and Game Wardens	6,530	100%	Inspect commercial operations relating to fish or wildlife, recreation, or protected areas.
Postal Service Mail Carriers	326,760	100%	Deliver mail to residences and business establishments along specified routes by walking or driving, using a combination of satchels, carts, cars, and small trucks.
Judges, Magistrate Judges, and Magistrates	28,230	100%	Interpret and enforce rules of procedure or establish new rules in situations where there are no procedures already established by law.
Court, Municipal, and License Clerks	156,900	98.2%	Answer inquiries from the general public regarding judicial procedures, court appearances, trial dates, adjournments, outstanding warrants, summonses, subpoenas, witness fees, or payment of fines.
Transit and Railroad Police	3,010	89.3%	Examine credentials of unauthorized persons attempting to enter secured areas.

[28]. By "heuristic," we refer to the process of assessing and validating GenAI's performance using human judgment grounded in common sense. This process was essential in addressing our research questions, and we describe the process as follows:

The iterative development involved extensive testing of various prompt variations within OpenAI's Playground interface,[6] experimenting with different combinations of model settings. To establish a baseline, we selected a range of occupational tasks from our dataset, particularly those tasks that require a high degree of physical interaction—such as driving vehicles and guarding facilities—which GPT models, especially without supplementary hardware, would presumably find challenging. Despite advancements in AI technologies like autonomous vehicles, it is evident that GPT-driven technologies still face limitations in physical task execution. The iterative process included using these physically demanding tasks as baseline queries alongside other randomly chosen tasks. We then assessed the GPT-generated responses, continuously refining our prompts to elicit more accurate and logically sound answers based on heuristic common sense. For example, we observed that the GPT-3.5 model tended to overestimate ChatGPT's capabilities in performing physical tasks. To address this, we incorporated specific instructions in our prompts, directing the model to consider whether ChatGPT can *physically* execute the tasks. This adjustment significantly enhanced the models' performance in responding to our baseline questions.

The "temperature" variable was set as zero, allowing for minimum randomness in the model's responses, thereby ensuring that the output is more deterministic and consistent.

In this study, we also experimented with different scales to assess the extent to which GPT models can substitute human labor in performing tasks. We tested both a percentage-based approach and fixed scales ranging from 1 to 5 or 1 to 10. We found that a 1 to 10 scale—where one indicates that ChatGPT is completely unable to perform any part of the task, and 10 signifies full autonomous performance by ChatGPT without human

[6] https://platform.openai.com/playground?mode=chat

intervention—allowed for more realistic evaluations by the models. Recognizing the potential of AI-powered tools to augment and assist in tasks, thereby enhancing productivity (**RQ2**), we included instructions in our prompts to estimate how ChatGPT could improve human efficiency and task productivity.

In our iterative prompt engineering, we tested the same prompts on two models representing two generations of GPT models—GPT-3.5 (gpt-3.5-turbo[7]) and GPT-4 (gpt-4-1106-preview[8]). Because we observed that GPT-4 consistently delivers more accurate, reasonable, and contextually relevant responses than GPT-3.5,[9] we ultimately chose to utilize GPT-4 instead of GPT-3.5 for this study.

4.2.2 The Prompt Design. The finalized version of our prompt, as developed through the iterative process described in Section 4.2.1, is included in Appendix A. This prompt directs GPT-4 to engage with questions related to our first, second, and fourth research questions outlined in Section 3.[10]

For **RQ1**, it requires GPT-4 to quantitatively assess, on a 1–10 scale, the extent to which ChatGPT could replace human labor in specific tasks, accompanied by an 80-word rationale for the given rating. Addressing **RQ2**'s dual focus—assistance and productivity enhancement—our prompt instructs GPT-4 to first determine, also on a 1–10 scale, how GPT models can aid human labor. Subsequently, it must estimate the factor by which ChatGPT could increase human productivity; in other words, how it could expedite tasks or reduce their completion time. The prompt further requests two explanations from GPT-4: one detailing the reasoning behind these quantitative assessments regarding task augmentation and the other outlining ChatGPT's *limitations* in task execution.

For **RQ4**, which explores the evolving skill landscape due to technologies like ChatGPT, the prompt is designed to elicit qualitative estimations and explanations on four aspects: (1) the reasons why humans may be still relevant in performing the task; (2) the existing skills humans need to retain to stay relevant; (3) the new skills humans should acquire in an era marked by GPT-like technologies; and (4) the current skills likely to become obsolete due to GPT-like technologies. GPT-4 is instructed to succinctly answer these queries within an 80-word limit for each.

4.3 Aggregating Task Ratings to Occupation Ratings

To answer **RQ3**, we aggregated three specific task ratings estimated by GPT-4—namely, the GPT-replacement scale, GPT-assistance scale, and GPT-caused productivity increase—to create occupation-level ratings. In subsequent sections, we denote the task-level GPT-replacement scale, GPT-assistance scale, and GPT-caused productivity increase as π, ρ, and τ, respectively. The corresponding aggregated scales at the occupation level are denoted as Π, P, and T.

[7] The gpt-3.5-turbo model, trained with data up to September 2021, is detailed in its documentation: https://platform.openai.com/docs/models/gpt-3-5

[8] The gpt-4-1106-preview model, trained with data up to April 2023, is detailed in its documentation: https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[9] We noted a tendency in GPT-3.5 to overestimate the capabilities of ChatGPT in performing certain tasks. For instance, in the case of administrative law judges issuing subpoenas and administering oaths for formal hearings, GPT-3.5 assigned a replacement score of 7 out of 10. It reasoned that GPTs could largely take over this role by generating legal documents, including subpoenas, based on provided information. In contrast, GPT-4's assessment was significantly lower, at 1 out of 10, suggesting that ChatGPT is incapable of substituting for any part of this task, because these are legal actions requiring human officials. Another instance where the assessment of ChatGPT's capabilities differed significantly between GPT-3.5 and GPT-4 is in the task of teachers preparing materials and classrooms for class activities. GPT-3.5 gave this task a replacement score of 7 out of 10, suggesting that ChatGPT could partially take over by offering instructions and suggestions for setting up classroom materials or equipment. In contrast, GPT-4's evaluation indicated that ChatGPT is entirely incapable of replacing humans in this aspect, assigning a score of 1 out of 10. This discrepancy highlights GPT-4's "recognition" of the inherent physical limitations of GPT models, acknowledging that despite its ability to provide digital assistance, ChatGPT cannot execute tasks requiring physical interaction, such as the actual preparation of materials or the physical setup of classrooms. Upon reviewing these responses from the two GPT models, our team members concur that GPT-4's evaluation aligns more closely with the realistic capabilities of ChatGPT. Therefore, we chose to use GPT-4 instead of GPT-3.5 for this study.

[10] Addressing **RQ3** requires aggregating insights from RQ1 and RQ2. Therefore, our prompt does not directly ask about occupation-level evaluations.

The occupation-level scales in our study are determined through a weighted aggregation approach. This means that each task within an occupation is assigned a specific weight, and these weighted values are then summed to represent the overall scale of the occupation. The rationale behind this weighted method is the recognition that not all tasks within an occupation hold equal significance. Some tasks are more important than others. For instance, in the case of firefighters, the task of rescuing victims from a burning building is far more important than cleaning fire stations. By assigning greater weight to more important tasks, their impact is more significantly reflected in the overall scale. As a result, the aggregated scale provides a comprehensive representation of the relative importance of different tasks within an occupation.

Let S_i represent any of the three aggregated ratings for occupation i, and let s_{ij} denote the rating of a core task j within occupation i. The calculation of S_i follows this equation:

$$S_i = \sum_{j=1}^{n_i} \left(s_{ij} \cdot w_{ij} \right), \tag{1}$$

where n_i represents the total number of core tasks associated with occupation i, and w_{ij} signifies the weight assigned to task j within that occupation. In our methodology, the importance of each task within an occupation dictates its weight in the overall calculation. We utilize the importance rating (IM) provided for each task in the O*NET task dataset.[11] This IM value, which varies between 1 and 5, reflects the importance of a task as determined by surveys conducted with incumbents of the respective occupations. Here, a rating of 1 denotes a task that is not important to an occupation, while a rating of 5 signifies a task that is extremely important.

The weight of a task, w_{ij}, is calculated as its proportional importance relative to the total importance of all tasks within that occupation. In essence, a task deemed more important for an occupation receives a higher weight, thereby exerting greater influence on the occupation's overall rating. w_{ij} is computed using the following equation:

$$w_{ij} = \frac{IM_{ij}}{\sum_{q=1}^{n_i} IM_{iq},} \tag{2}$$

where IM_{ij} denotes the importance rating of task j within occupation i.

For example, consider an occupation i with two tasks: task one (n) rated as 1 for importance ($IM_{in} = 1$), signifying minimal importance, and task two (m) with a substantially higher importance rating of 4 ($IM_{im} = 4$). Consequently, the total importance rating for occupation i is the sum of these ratings: $1 + 4 = 5$. The weight assigned to task one (w_{in}) is calculated by dividing its importance rating by the total importance rating, namely, $1/5 = 0.2$. Similarly, the weight for task two (w_{im}) is $4/5 = 0.8$. Suppose task one is estimated to be fully replaceable by ChatGPT (with a GPT-replacement score $\pi_n = 10$) and task two is deemed not replaceable (with a score $\pi_m = 1$). The aggregated GPT-replacement scale for occupation i (Π_i) is then calculated as $w_{in}\pi_n + w_{im}\pi_m = 10 \times 0.2 + 1 \times 0.8 = 2.8$. This aggregated score of 2.8 (on a scale from 1 to 10) suggests that occupation i is relatively unlikely to be completely replaced by ChatGPT. This inference is drawn from the fact that its more important task (m) has a low replaceability score despite the high replaceability of the less-important task (n).

5 Results

Figure 1 displays the distribution of both estimated and aggregated variables. Table 2 provides the descriptive statistics for these variables, broken down at both the task and occupation levels. Correlation tables of variables can be found in Appendix B.

[11] The definition and explanation of IM values can be found on O*NET website at https://www.onetonline.org/help/online/scales

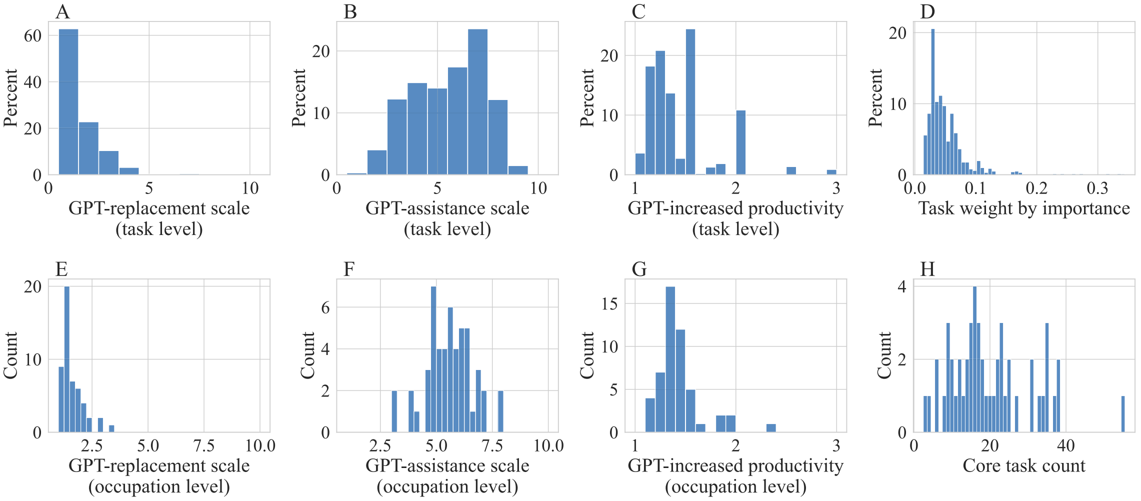


Fig. 1. Distributions of (A) GPT-estimated task-level ChatGPT-replacement scale, (B) GPT-estimated task-level ChatGPT-assistance scale, (C) GPT-estimated task-level ChatGPT-caused productivity increase, (D) task weight by importance, (E) aggregated occupation-level ChatGPT-replacement scale, (F) aggregated occupation-level ChatGPT-assistance scale, (G) aggregated occupation-level ChatGPT-caused productivity increase, and (H) core task count.

Table 2. Descriptive Statistics

Level	**Variable**	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Task	GPT-replacement scale (π) (1–10)	1,022	1.570	0.910	1.000	1.000	1.000	2.000	7.000
	GPT-assistance scale (ρ) (1–10)	1,022	5.540	1.790	1.000	4.000	6.000	7.000	9.000
	GPT-increased productivity (τ)	1,022	1.410	0.340	1.000	1.200	1.300	1.500	3.000
	Task importance (IM) (1–5)	1,022	4.170	0.400	3.000	3.910	4.230	4.460	4.960
	Task weight by importance (w)	1,022	0.050	0.030	0.010	0.030	0.040	0.060	0.340
Occupation	Gov employment	51	130,962	235,323	2,120	12,395	52,770	124,865	1,246,620
	Core task count	51	20.039	10.718	3.000	12.500	17.000	25.000	55.000
	GOV-owned percentage	51	93.40%	5.57%	82.59%	89.08%	93.69%	99.46%	100.00%
	GPT-replacement scale (Π) (1–10)	51	1.610	0.475	1.000	1.314	1.456	1.817	3.321
	GPT-assistance scale (P) (1–10)	51	5.592	1.035	3.012	4.908	5.701	6.212	7.848
	GPT-increased productivity (T)	51	1.434	0.223	1.122	1.313	1.374	1.461	2.391

We observed that the distribution of the GPT-replacement scale, both at the task and occupation levels, is right-skewed, as depicted in Figures 1(A) and 1(E). This skewness suggests a power-law distribution, indicating that while a majority of tasks and occupations are estimated to have low GPT-replacement scores, a small number are evaluated with high scores. In contrast, the GPT-assistance scales, shown in Figures 1(B) and 1(F), exhibit a more normal distribution. The mean values of these scales are centered, suggesting that a significant proportion of tasks and occupations in the public sector have a moderate potential for enhancement through the utilization of ChatGPT. Regarding the potential increase in productivity facilitated by ChatGPT, depicted in Figures 1(C) and 1(G), the distributions are slightly right-skewed. The average increase in productivity is estimated to be around 1.4 times, indicating a general tendency towards moderate improvements in productivity across various tasks and occupations.

5.1 Task Replacement Estimation

As shown in Figure 1(A), GPT-4's estimations suggest that approximately 63% (642 out of 1,022) of the core tasks in public sector occupations are extremely unlikely to be replaced by ChatGPT at all ($\pi = 1$).[12] These tasks typically involve complex real-world problem-solving, social interpretation, and, most crucially, physical interaction with people and the world—an area currently beyond the capabilities of GPT models. For instance, tasks like a subway operator's role in driving subways to transport passengers are beyond ChatGPT's purview, as operating machinery and vehicles are outside its functional scope. Similarly, tasks requiring physical presence, such as a teacher's responsibility to enforce behavior rules and maintain order in classrooms, are unfeasible for ChatGPT-like technologies, given their inability to physically exist in such environments.

Additionally, tasks necessitating human accountability are also beyond the realm of ChatGPT. For example, a detective's duty to testify in court and present evidence cannot be undertaken by ChatGPT, as they are not recognized legal entities and are ineligible to perform such legally binding activities. Another notable instance is a judge's responsibility to conduct hearings and make decisions on issues such as social program eligibility, environmental protection, or health and safety regulation enforcement. Such tasks demand human judgment and legal authority, which ChatGPT cannot provide due to their lack of legal status and human discernment.

According to GPT-4's estimation, only nine tasks across various occupations have a GPT-replacement scale (π) value of 5 or higher, indicating a significant likelihood of being automated by ChatGPT in the foreseeable future. Notably, this group includes four tasks typically performed by court, municipal, and license clerks, including assisting in preparing budget and reviewing expenditure, and two tasks assigned to library assistants, such as identifying overdue materials by reviewing microfilm or issue cards. Additionally, tasks such as maintaining records and logbooks for forest fire inspectors, coding license application information for computer entry by clerks, and maintaining operational records by patrol officers are identified as areas where ChatGPT could be highly effective. GPT-4 suggests that ChatGPT can "streamline data entry, automate and optimize cataloging and retrieval process, speeding up the process, improving record accuracy, and reducing manual coding errors." Interestingly, despite the majority (26 out of 27) of firefighters' tasks being deemed challenging for ChatGPT, there is an exception: preparing written reports on fire incidents, where ChatGPT could significantly contribute or even potentially replace human labor.

5.2 Task Augmentation Estimation

RQ2 explores the estimation by GPT-4 of ChatGPT's role in augmenting human task performance within the public sector. This inquiry has two dimensions: (1) the GPT assistance scale, which quantifies the extent to which ChatGPT can aid in executing each task, and (2) the productivity increase, which measures the potential enhancement in task productivity attributable to ChatGPT's assistance. In other words, it assesses how much ChatGPT can reduce the task completion time. Results of the two dimensions are discussed in Sections 5.2.1 and 5.2.2, respectively.

5.2.1 Task Assistance. In terms of assisting human labors, GPT-4 estimates that ChatGPT in general can substantially aid in numerous tasks. Nevertheless, the extent and nature of this assistance vary by task. According to the data presented in Figure 1(B), approximately 55% of tasks receive a GPT-assistance scale (ρ) rating above 5, on a scale from 1 to 10, indicating a moderate to high level of potential assistance. Notably, the mode of the GPT-assistance scale is 7, representing 23.5% of all evaluated tasks. Tasks that typically benefit from the highest levels of GPT assistance include those related to record-keeping, report preparation, routine communication such as answering queries, and performing basic research.

[12] A sample of the task-level results obtained from GPT-4 queries is presented in Appendix D. The complete datasets are openly accessible on Harvard Dataverse [40]. Please refer to the Data Availability section for further details.

Generally, there exists a positive correlation between the scales of task replacement and assistance.[13] Yet, we have identified 189 tasks with a low replacement risk but high augmentation potential, suggesting they could benefit from ChatGPT's capabilities to augment task completion without completely replacing human labor. Such tasks typically demand greater human engagement, emphasizing accountability and the ability to interpret real-world scenarios and social cues. For example, ChatGPT could assist healthcare practitioners in clinical research by supporting literature review, data analysis, and drafting research documents despite not being able to conduct research directly or interact with patients. Similarly, for special education teachers, while ChatGPT cannot replace the crucial human elements of counseling, such as personal experience and empathy, it can provide valuable support in information gathering related to academic programs and career options. Air traffic controllers' task of compiling information about flights presents another example; while ChatGPT can help streamline data compilation, it cannot fully understand the context or ensure the accuracy of compiled data. Thus, despite ChatGPT's high assistance rating ($\rho = 9$) for this task, it cannot eliminate the need for human verification and interpretation, reflected in a lower replacement scale ($\pi = 3$).

Among the tasks analyzed, only 44 (4.3%) received substantially low assistance scales ($\rho \leq 2$) from GPT-4, indicating minimal potential for ChatGPT's support. Notably, three tasks were deemed extremely unlikely to benefit from ChatGPT's assistance at all ($\rho = 1$). These tasks encompass first-line supervisors overseeing the transfer of offenders or driving crew carriers to transport firefighters to fire sites, as well as library assistants delivering items by hand or using push carts. The common denominator among these tasks is the essential requirement for physical presence, a domain where ChatGPT and similar technologies fall short, unable to offer assistance or serve as a replacement for human labor.

5.2.2 Productivity Increase. The augmentation of human labor by ChatGPT not only encompasses task assistance but also the potential to increase productivity. We found that GPT-4's predictions regarding productivity increase are cautiously modest. According to its analysis, 56.3% of public sector tasks could see a speed increase of no more than 1.3 times with ChatGPT's assistance, with 15 tasks showing no productivity increase at all ($\tau = 1$).

Only 9 out of the 1,022 tasks are estimated by GPT-4 to achieve a maximum productivity boost of threefold. These include forest fire inspectors' tasks of maintaining logbooks and farm management educators researching information requested by farmers. Tasks with significant productivity increase ($\tau \geq 2$) constitute 13% of all tasks.

Particularly, tasks that stand to gain significant speed improvements by integrating tools like ChatGPT primarily involve repetitive information gathering, data entry, and calculations, which do not require extensive contextual understanding or social interactions. For instance, eligibility interviewers for government programs, responsible for computing and authorizing assistance amounts for various programs such as grants and food stamps, can benefit from ChatGPT-supported technologies, which can assist with calculations and offer suggestions, thereby accelerating the process. However, it is important to note that human judgment remains essential in complex cases. Another task where GPT-powered tools can significantly contribute is the work of tax examiners, particularly in notifying taxpayers of overpayments or underpayments. The integration of ChatGPT can expedite the notification process by automating content creation for notices and tracking delinquencies, thus improving efficiency.

There is a notable correlation between ChatGPT's assistance scale and the potential productivity increase.[14] We observe that higher assistance scores are often associated with high productivity boosts, exemplified by tasks like teachers maintaining student attendance records ($\rho = 9$) potentially tripling in speed. Tasks estimated to have minimal productivity increases also show lower assistance scales. Nonetheless, exceptions exist where

[13] As illustrated in Figure B.1 within Appendix B, the Pearson correlation coefficient indicative of the relationship between the GPT-replacement scale and the GPT-assistance scale stands at 0.64.

[14] The Pearson correlation coefficient between the GPT-assistance scale and productivity increase, shown in Figure B.1 within Appendix B, is 0.8.

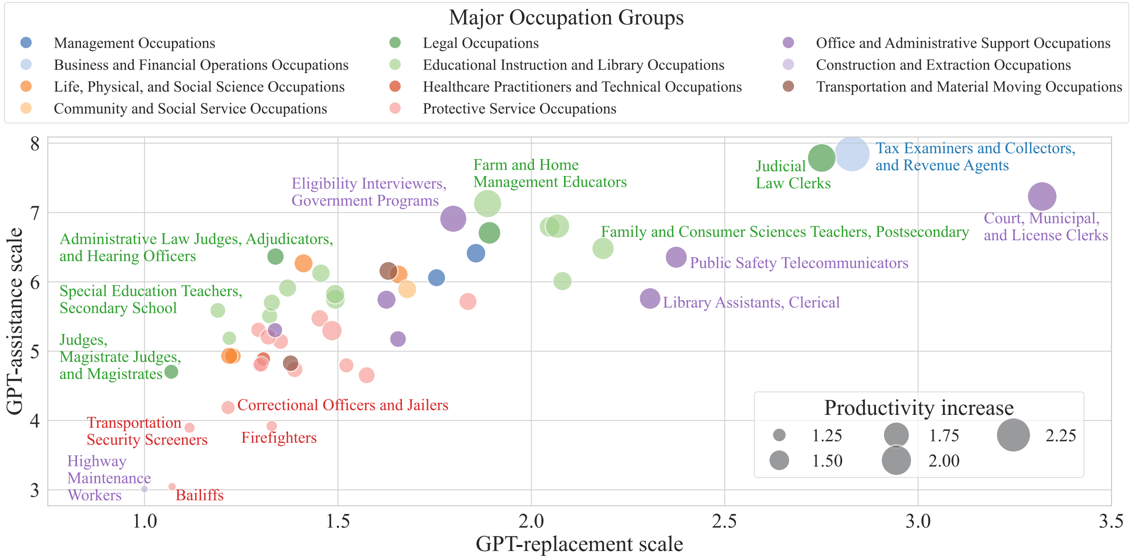


Fig. 2. A scatter plot presenting the aggregated, occupation-level scales of GPT replacement and GPT assistance, with marker size based on aggregated productivity increase. Several occupations are highlighted in the plot. The color coding of the dots indicates the major group to which each occupation belongs, according to the Standard Occupational Classification (SOC) system defined by the Bureau of Labor Statistics (BLS). For example, firefighters and bailiffs both belong to the “Protective Service Occupations” major group, so their dots, appearing in the lower left corner, have the same color (light pink).

tasks with high assistance scores see limited productivity gains. For instance, while ChatGPT can assist teachers in delivering instructions (see Footnote 2) or healthcare practitioners in organizing study materials—each task benefiting from substantial AI assistance —its inability to interact with the environment and people (such as hands-on guidance to students or practice medicine) limits its productivity impact to a small increase.

It is important to acknowledge that these estimations of productivity increase, while insightful, are not precise measurements and only reflect GPT-4’s assessments based on its training data and analytical capabilities.

5.3 Occupation-level Results

Aggregating GPT-4’s task-level estimations to occupation-level reveals a correlation between the scales of replacement, assistance, and productivity increase, as illustrated in Figure 2. The data suggest that occupations highly replaceable by ChatGPT also tend to benefit significantly from ChatGPT’s assistance, leading to substantial productivity increases, with a few exceptions.

For instance, court, municipal, and license clerks who handle clerical tasks in courts, municipalities, and governmental licensing agencies are identified as having the highest likelihood of being replaced by language models like ChatGPT, receiving a score of 3.32 on a scale of 1 to 10 (note that 3.32 is not a significantly high score). ChatGPT is also projected to greatly assist their work, potentially doubling their productivity. Analyzing their tasks, such as entering information, recording court orders, checking records, preparing meeting agendas, editing meeting minutes, and responding to queries, reveals that these activities are relatively repetitive, routine, and largely governed by established protocols. Properly integrating ChatGPT could markedly enhance productivity in these areas. For instance, ChatGPT can help clerks examine legal documents submitted to the courts, particularly by suggesting potential issues in the document based on legal rules.

Tax examiners, collectors, and revenue agents responsible for tax assessment and collection are ranked second highest in terms of replacement potential, scoring 2.83 on the replacement scale. This occupation could see as high as a 2.4-fold increase in productivity with ChatGPT's assistance. ChatGPT can facilitate tasks such as answering common tax-related questions, organizing and retrieving information, and automating the generation of notifications. However, it is important to note that ChatGPT is not capable of engaging in verbal communication with taxpayers to discuss specific issues or explain complex policies and procedures, tasks that require advanced interpersonal communication skills and the ability to address complex, individualized legal matters.

The occupations that demonstrate the lowest potential for replacement by ChatGPT also exhibit minimal prospects for augmentation and productivity enhancement. These occupations include roles such as highway maintenance workers, judges, bailiffs, transportation security screeners, correctional officers and jailers, forensic science technicians, kindergarten teachers, and forest and conservation technicians. Occupations in construction and extraction, like highway maintenance workers, as well as those in protective services, such as bailiffs and firefighters, are generally less susceptible to replacement by ChatGPT. This observation is evident in Figure 2, where light purple and pink dots predominantly cluster in the plot's left half. These occupations typically require direct physical interaction with people and environments as well as manual operation of equipment, which are tasks currently outside the scope of ChatGPT's capabilities.

As we have presented, we observed a broad correlation across occupations between the ChatGPT replacement scale, the ChatGPT assistance scale, and the productivity enhancement attributable to ChatGPT, as depicted in Figure 2 and Figure B.2 in Appendix B. Here in Figure 2, most of the data points align along an upward-trending trajectory, with dots positioned further to the right (indicating higher assistance levels) also displaying larger radii, signifying greater productivity increases. However, there are notable exceptions among the occupations. For instance, eligibility interviewers for government programs, represented by a purple dot situated in the upper middle section of Figure 2, exhibit a significantly higher productivity increase (1.8) compared to surrounding points. This suggests that the inclusion of ChatGPT-like technologies into their workflow could substantially benefit and augment their operations, without the imminent threat of replacement. A closer look at the task-level results for eligibility interviewers reveals the underlying reason: Their responsibilities heavily involve routine information collection and processing, calculations, and report drafting—tasks that ChatGPT can enhance markedly. At the same time, their role necessitates direct interaction with people, interpreting individual needs, and applying empathy and ethical judgment—elements of the job where ChatGPT falls short of replacing human capabilities.

5.4 ChatGPT Reshaping Skill Landscape

5.4.1 Crucial Skills Humans Must Preserve.

GPT-4's analysis indicates that tasks less vulnerable to replacement by ChatGPT often require skills involving physical manipulations, managing complex scenarios, or interacting with people. These skills include critical thinking, active listening, negotiation, persuasion, time management, coordination, public speaking, and troubleshooting. Notably, these are also among the skills GPT-4 identifies as crucial for humans to preserve to remain relevant, addressing the first part of **RQ4**. For instance, consider the task of technical teachers at middle schools who meet with parents to discuss their children's progress. This task demands a multifaceted skill set, including critically assessing children's progress, understanding parents' responses and emotions, and persuading them to adhere to certain guidelines. It also involves complex problem-solving and decision-making skills to prioritize children's needs and effectively communicate these to the parents. Another example is the task of parking-enforcement workers who appear in court for hearings on contested traffic citations. GPT-4 categorizes this task as non-replaceable due to the need for active listening, writing court documentation, communicating observations and judgments, as well as skills in persuasion and service orientation. Similarly, the task of subway and streetcar operators in directing emergency evacuation procedures ($\pi = 1$) requires a complex array of skills for interacting with people and the environment in challenging situations and responding promptly to crises, including monitoring, persuasion, coordination, instructing, service orientation, time management, and managing human resources.

However, skills related to selecting, operating, maintaining, and repairing equipment are among those least susceptible to AI replacement. For instance, tasks such as preventative maintenance performed by highway maintenance workers on vehicles and equipment or firefighters cleaning firefighting gear require substantial human involvement. While GPT-4 acknowledges that ChatGPT can assist in predicting maintenance needs, diagnosing mechanical issues, managing inventory, and establishing schedules, it lacks the capability to interact with equipment physically.

In its estimation regarding the existing skills humans need to maintain, GPT-4 also underscores tacit, task- or job-specific skills that require a deep understanding of the job context, often rooted in common sense and experience. For instance, in the role of library assistants organizing records and sorting books, GPT-4 highlights the necessity of retaining comprehensive knowledge of library systems and classification schemes. Additionally, for tasks such as identifying overdue materials and delinquent borrowers, GPT-4 points out the importance of understanding library policies. Furthermore, GPT-4 suggests that animal-control workers involved in prosecutions related to animal treatment must preserve their legal knowledge, public speaking skills, and the ability to provide court testimony. These identified skills and knowledge bases are intrinsically linked to the hands-on experience acquired within their respective fields, characterizing many as tacit skills and knowledge that are challenging to codify for AI. In protective service occupations, such as those of police officers and firefighters, physical fitness and rescue skills are also emphasized as crucial.

Furthermore, GPT-4 identifies a critical set of skills centered around human interaction and understanding, such as ethical judgment, decision-making, and empathy. For instance, eligibility interviewers for government programs assess personal and financial data to determine eligibility. This process demands ethical judgment and decision-making grounded in common sense and empathy, such as prioritizing applicants based on need. While human judgment is not infallible, it incorporates a nuanced interpretation of interpersonal dynamics that ChatGPT-like technologies may not fully comprehend. Similarly, elementary school teachers designing and leading activities must manage classrooms, empathize with students, and inspire motivation. These skills transcend mere experience, requiring a deep empathy to appreciate the diverse behaviors and needs of individual students. Such nuanced understanding underscores the indispensable role of human sensitivity in tasks where judging, interpreting, and reacting to human characteristics and behaviors are paramount.

5.4.2 Skills That May Become Obsolete. Skills that GPT-4 predicts will become obsolete include memorizing information, manual record-keeping, preparation of training materials, document drafting, scheduling routine events, routine information gathering, basic research activities (such as preliminary legal research and routine case law retrieval), some aspects of procedural verification, simple calculations, and basic fact-finding. This aligns with the conclusions of Frey and Osborne [21], which state that skills required for routine, repetitive tasks that lack creativity and real-world interaction are likely to become obsolete. Furthermore, due to the capacity of ChatGPT to produce new content, GPT-4 anticipates that the skills associated with basic content creation could also soon become obsolete and diminish in demand. By basic content creation, we refer to the development of new content derived from existing materials (through a combination of learned patterns and probabilistic modeling), necessitating a minimal degree of manipulation and lower levels of creativity. This encompasses the generation of written materials for training, marketing promotions, planning, public education, and classes, as well as the compilation, editing, and reporting of information or data analysis based on pre-existing content.

However, GPT-4 highlights that not every aspect of content creation is at risk of becoming obsolete. Skills that leverage AI tools for content creation, especially in producing customized content that demands an understanding of complex real-world contexts, are expected to be increasingly sought after. Additionally, the competency to effectively present and communicate this content remains crucial. For instance, police officers involved in drug-related community programs might find basic content creation skills less necessary in the near future. Yet, they could benefit from using AI for program customization. Skills in public speaking, community engagement, and developing programs tailored to specific community and individual needs will continue to be vital for the success of such initiatives.

Teachers face a similar transition. While basic content creation for course materials may become obsolete in the near future because of ChatGPT, their roles could be enhanced by integrating AI into content creation and adopting digital teaching tools. Skills in curriculum development, pedagogical strategies, and adapting and customizing materials will stay indispensable. The ability for teachers to contextualize educational content and respond to student feedback will maintain its significance in the educational landscape.

5.4.3 Essential New Skills for the Future. In our prompt, we directed GPT-4 to assess what new skills might become important in the age of generative AI. Notably, GPT-4 strongly suggests that proficiency in utilizing AI tools across a range of tasks will become increasingly essential. This proficiency is encapsulated by the concept of "AI literacy," which encompasses understanding, applying, evaluating AI tools, and ethical considerations [35]. However, GPT-4's analysis primarily highlighted the skills of applying and evaluating AI as essential.

For application skills, GPT-4 illustrated how various public sector occupations could learn the skills of utilizing AI for both analytical tasks and content generation. For example, detectives might acquire skills of using AI for predictive analytics and network analysis in criminal investigations, while firefighters could learn to use AI to analyze fire patterns and predict hot spots. Highway maintenance workers can adopt AI for predictive maintenance and traffic management. Additionally, GPT-4 emphasizes the significant potential for various public sector occupations to enhance their functions through the integration of varied sources of information assisted by AI capabilities. This includes the application of AI in public education for initiatives such as fire prevention, the creation of reports for administrative purposes, the development of training scenarios, and the formulation of fitness and wellness plans. Specifically, GPT-4 points out that correctional officers could improve conflict-resolution strategies with AI-generated guidelines. Teachers have the opportunity to revolutionize curriculum development, test creation, and grading processes through AI tools. Similarly, healthcare practitioners could significantly augment patient education by incorporating AI-generated insights.

Regarding the skills of evaluating AI, GPT-4 underscores the emerging necessity for public sector professionals to learn to interpret, analyze, and validate AI-generated content. Take, for instance, the role of medical personnel in referring patients to surgeons or other physicians. They can leverage AI tools to organize referral options based on symptoms and generate potential referral recommendations, enhancing their productivity. However, they must then interpret these AI suggestions and make clinical decisions based on their medical knowledge, human judgment, and professional networks. This process involves understanding healthcare pathways and even insurance policies, highlighting the need for a nuanced blend of AI utilization and human expertise. Similarly, in education, while AI can aid in program development, the capacity to evaluate AI-driven data for pedagogical enhancement is vital. Firefighters, too, can harness AI for insights on fire behavior and building safety, yet the imperative lies in their ability to critically evaluate such information, determining whether and how to integrate such AI-generated information into decision-making processes.

We observe that GPT-4's analysis overlooks the first and fourth dimensions of AI literacy as defined by Ng et al. [2021], specifically the technical understanding of AI and the ethical considerations associated with its use. While a deep technical comprehension of AI's workings is not strictly necessary for its application and assessment, such knowledge can enhance the effectiveness of these processes. Realistically, expecting every public sector worker to grasp the intricacies of AI algorithms and coding is impractical.

However, we found that GPT-4's evaluation misses a critical component: the skills to discern the ethical implications of employing AI tools. GPT-4's focus on skills interpreting AI-generated content primarily concerns evaluating its accuracy, not its ethical or moral ramifications. Accuracy, while important, carries inherent moral weight—for instance, decisions affecting the care hours for individuals with disabilities have profound ethical consequences. Ethical considerations often navigate complex situations involving privacy, security, happiness, trust, equity, and equality, influencing whether to employ AI technology.

Thus, despite GPT-4's omission, we argue that the skills to navigate and evaluate the ethical implications of AI use emerges as an indispensable skill in the future.

6 Discussion and Conclusions

There has been considerable debate surrounding the impact of LLM-powered technologies, such as ChatGPT, on occupations and employment. This debate is part of a wider discussion about the consequences of another wave of automation (the application of technology, robotics, or processes to achieve outcomes with minimal human input). If this discussion is to be anything more than airing concerns or aspirations, then how LLMs can replace or enhance the skills that constitute occupations needs to be identified and quantified.

The results presented here demonstrate one possible approach in the context of public sector jobs in the United States. In this study, we leverage the "reasoning" capabilities of the GPT-4 LLM model aiming to shed light on the transformative potential of ChatGPT in reshaping occupational tasks and skill requirements typical of public sector jobs. The findings indicate that while many tasks are not yet susceptible to AI replacement, primarily due to the need for physical interaction and complex human judgment, there are combinations of skill requirements and task executions that can be performed or augmented by generative AI.

In addressing our **RQ1**, which explores how ChatGPT can *replace* humans in performing occupational tasks within the public sector, our findings provide a significant insight. According to GPT-4's estimations, approximately 63% of tasks in the public sector cannot be replaced by GPT-powered technologies, underscoring the importance of human cognition and abilities in tasks requiring physical presence, emotional intelligence, accountability, and complex decision-making. However, the fact that nine tasks have a high potential for replacement by ChatGPT does signal a potential shift towards automation in certain clerical and administrative functions.

In response to **RQ2**, which examines ChatGPT's capacity to *augment* human labors in the public sector, GPT-4's estimations indicate that over half of the tasks in the public sector could significantly benefit from integrating ChatGPT. This finding underscores a substantial opportunity for enhancing public sector operations. The areas where ChatGPT's assistance is particularly notable include tasks that involve record-keeping, basic information gathering, and routine communication. However, ChatGPT is less likely to be effective in tasks demanding direct physical interaction with people or the environment. Additionally, GPT-4's estimation for productivity improvements with ChatGPT's assistance tends to be conservative, projecting that most tasks would see less than a 1.3-fold increase in completion speed. Despite this, a noticeable relationship exists between the degree of ChatGPT's assistance and the anticipated productivity gains, albeit with certain outliers.

By aggregating task scales estimated by GPT-4 to the occupation level, we analyzed ChatGPT's impact on 51 occupations in the public sector (**RQ3**). Consistent with task-level results, we found that occupations least replaceable by ChatGPT typically involve significant physical labor or require empathy and in-depth human interaction, such as occupations in healthcare, education, and emergency services (for example, firefighters, bailiffs, and special education teachers). Conversely, occupations with a high degree of routine, rule-based tasks, particularly in administrative and clerical occupations, show a higher potential for replacement by ChatGPT, for example, court clerks, tax examiners, and library assistants.

Addressing **RQ4**, our investigation with GPT-4's estimations sheds light on the evolving landscape of skills in the public sector in the age of GenAIs, with a focus on three main areas: skills to retain, skills becoming obsolete, and essential new skills to acquire. GPT-4 highlights the importance of maintaining critical thinking, interpersonal communication, ethical reasoning, complex problem-solving, decision-making, and physical skills (such as fitness and rescue skills) to stay relevant. It also stresses the value of tacit, task- or occupation-specific skills, emphasizing the need for a deep understanding of particular work contexts and the accumulation of experience.

On the flip side, skills linked to routine, repetitive tasks, basic data processing, and content creation are flagged as becoming less essential, potentially obsolete, due to AI's capacity to automate these functions. Regarding new skills, GPT-4 points out the growing necessity for AI literacy, especially in applying AI tools and interpreting and evaluating AI-generated content, to remain relevant and competitive in the public sector.

Our findings specifically underscore the importance of incorporating AI tools into existing workflows, such as healthcare professionals utilizing AI for referral recommendations or teachers using ChatGPT for efficient lecture material preparation. Furthermore, the ability to critically assess AI-generated content is identified as crucial. While GPT-4 focuses on the accuracy and fact-checking aspects, we argue that evaluating the ethical implications and moral consequences of AI-generated content is equally vital and warrants attention.

This research, while providing useful insights, has certain limitations primarily stemming from its reliance on ChatGPT's analysis. It is important to recognize that while GPT models provided by OpenAI offer an extensive summary of information in their training data, they are not without shortcomings. Issues such as the absence of the most recent updates and biases inherent in the training data can affect the accuracy and objectivity of the estimations [3, 33]. Despite these limitations, our results demonstrate both face validity, as they appear plausible and credible to the authors, and internal validity, demonstrating consistency and reliability within the scope of the study.

To further strengthen these findings, and extend their usefulness when considering how to manage and respond to AI-powered automation, future research could focus on external validity by involving human experts such as human-resource professionals, hiring managers, and public sector workers. Their perspectives, gathered through interviews and focus groups, would offer a practical and experiential dimension to validate our results. Acknowledging these prospective improvements, we are currently developing a human-GenAI integrative framework, building upon the methodologies used in this study. Our goal is to conduct comprehensive research that not only reaffirms our current findings but also explores the implications of generative AI across a broader spectrum of occupations, extending beyond the public sector.

Additionally, a comparative study between sectors could offer valuable insights into how different types of occupations are uniquely affected by AI, highlighting sector-specific challenges and opportunities. This could involve not only quantitative analysis similar to what was conducted in this project but also qualitative research to understand the subjective experiences and perceptions of workers in these sectors. Tapping into the detailed descriptions of occupations, tasks, and skills produced and maintained by the U.S. Department of Labor can provide a rich basis for identifying and comparing the possible effects of AI automation on the workforce of the U.S. Incorporating a geographic dimension into this analysis could be highly informative as well. By merging the data generated from this research with labor statistics, it would be possible to assess how different states, cities, and demographics are differently impacted by generative AI. Moreover, examining the long-term effects of AI integration on workforce skills and employment patterns would provide a more comprehensive understanding of AI's impact. Investigating the educational and policy implications of these changes would be crucial in guiding effective workforce development and adapting educational curricula to meet future demands.

In addition to exploring the impact of ChatGPT-like technologies on public sector workplaces, it is crucial to consider the associated risks, particularly given the unique context of the public sector. For example, as discussed in Section 5.2, our findings indicate that technologies like ChatGPT can enhance productivity for eligibility interviewers in government programs by aiding in calculations and offering suggestions for assistance amounts. However, real-world instances reveal the potential pitfalls of automation in this domain. For example, automated tools have previously failed to account for the specific needs of individuals with cerebral palsy or diabetes, leading to unjust reductions in care hours, with a disproportionately negative impact on low-income populations with disabilities across various states [2, 32]. In certain states, automated systems deployed for detecting fraudulent unemployment claims has led to incorrect fraud allegations against more than 34,000 individuals seeking unemployment benefits [24]. Such cases illustrate not only the limitations of AI in understanding complex human needs but also the dangers of over-reliance on AI-generated results. Even with human oversight, there can be a misplaced trust in AI's accuracy, resulting in decisions that lack empathy. Errors such as misclassifying an individual with double amputations as not having a mobility problem because they use a wheelchair only worsen the outcomes [2, 32]. These scenarios highlight how AI failures, combined with human errors and poor judgment, can cause significant harm, especially to underrepresented groups. Therefore, conducting further research into the risks of replacing or augmenting human workers with technologies like ChatGPT in the public sector is not only crucial but urgent. Such investigations are essential for ensuring these technologies serve to enhance public services without compromising the welfare of vulnerable populations.

In conclusion, our research highlights a dual trend: the persistent relevance of human-centric skills in the public sector and the rising importance of AI literacy and skills to interpret AI-generated content. As AI continues to evolve, public sector workforce development must focus on nurturing these complementary skills. Embracing this human-AI partnership can lead to a more efficient, innovative, and responsive public sector that is well equipped to meet the challenges of the 21st century.

Appendices

A Final Prompt

I will give you a list of json about occupation information of workers employed by the U.S. governments. I want to know how OpenAI's GPT models affects those occupations. For each json, review the Job title, Job_title_description, Activity and Task, and then, assess how ChatGPT can replace and/or augment human in performing this "Task."

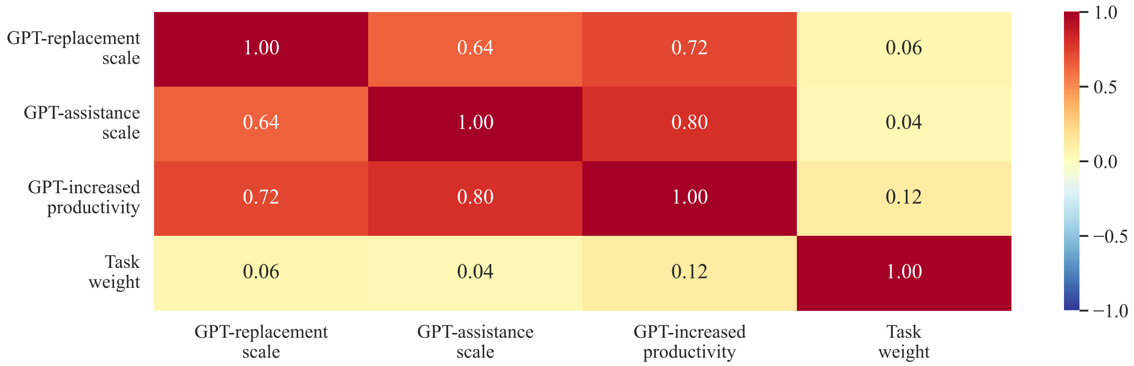
Next, give your answer in a list of json. Each json will have the following fields:

- *index (integer, identical to the index from original json),*
- *GPT_replacing_scale (integer, on a scale of 1 to 10, estimate, if GPT models completely replace humans, how well GPTs perform in physically completing the "Task," with 1 representing GPTs unable to physically perform any part of this "Task," while 10 representing GPTs able to physically perform this task completely without human intervention),*
- *GPT_replacing_explanation (maximum 80 words, estimate to what extent GPTs can replace humans in performing the "Task," and explain why),*
- *GPT_assistance_scale (integer, on a scale of 1 to 10, estimate how well GPT models can assist (instead of replace) humans performing the "Task," 1 representing GPTs unable to assist humans in any part of the "Task" while 10 representing GPTs extremely capable in assisting humans and increase productivity significantly),*
- *productivity_increase (float, estimate about how many times the productivity of this "Task" can be increased when augmented by GPTs),*
- *GPT_assistance_explanation (maximum 80 words, explain to what extent GPTs can assist humans in performing the "Task" to increase the productivity, quality, and efficiency),*
- *GPT_limitation (maximum 80 words, describe what are GPTs' current limitations in performing this "Task"),*
- *human_relevance (maximum 80 words, explain why humans can still be relevant in performing this "Task" in the advent of GPTs),*
- *skills_to_maintain (maximum 80 words, predict what existing skills human workers need to maintain in order to stay relevant in performing this "Task" in a future with GPTs),*
- *skills_to_acquire (maximum 80 words, predict what new skills human workers need to acquire in order to stay relevant in performing this "Task" in a future with GPTs),*
- *obsolete_skills (maximum 80 words, estimate what skills that are currently important in performing the "Task" may become obsolete because of the advent of GPTs).*

When giving your answers, be as specific and accurate as possible. Please consider the context: those are workers in the public sector.

B Correlation Tables

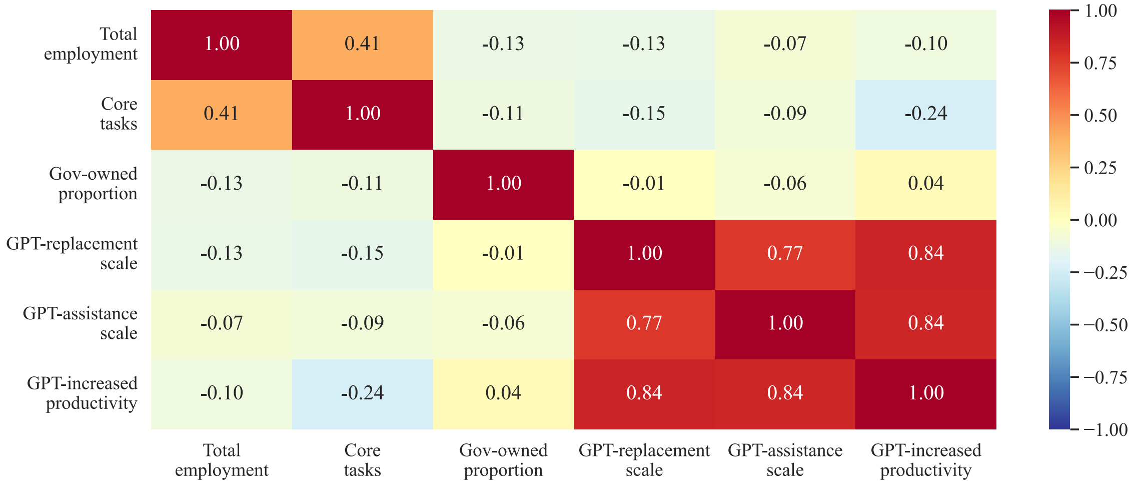
The following plot shows the correlation table of the task-level variables.



	GPT-replacement scale	GPT-assistance scale	GPT-increased productivity	Task weight
GPT-replacement scale	1.00	0.64	0.72	0.06
GPT-assistance scale	0.64	1.00	0.80	0.04
GPT-increased productivity	0.72	0.80	1.00	0.12
Task weight	0.06	0.04	0.12	1.00

Fig. B.1. Correlation table of task-level variables.

The following plot shows the correlation table of the occupation-level variables.



	Total employment	Core tasks	Gov-owned proportion	GPT-replacement scale	GPT-assistance scale	GPT-increased productivity
Total employment	1.00	0.41	-0.13	-0.13	-0.07	-0.10
Core tasks	0.41	1.00	-0.11	-0.15	-0.09	-0.24
Gov-owned proportion	-0.13	-0.11	1.00	-0.01	-0.06	0.04
GPT-replacement scale	-0.13	-0.15	-0.01	1.00	0.77	0.84
GPT-assistance scale	-0.07	-0.09	-0.06	0.77	1.00	0.84
GPT-increased productivity	-0.10	-0.24	0.04	0.84	0.84	1.00

Fig. B.2. Correlation table of occupation-level variables.

C Public Sector Occupations and ChatGPT's Impact

Job title	National gov employment	Gov-owned pct	GPT-repl. scale (aggregated)	GPT-assis. scale (aggregated)	GPT-increased productivity (aggregated)	Major category
Tax Examiners and Collectors, and Revenue Agents	50,610	1.000	2.829	7.848	2.391	Business and Financial Operations Occupations
Probation Officers and Correctional Treatment Specialists	87,430	0.972	1.679	5.894	1.430	Community and Social Service Occupations
Highway Maintenance Workers	135,220	0.943	1.000	3.012	1.122	Construction and Extraction Occupations
Agricultural Sciences Teachers, Postsecondary	7,720	0.937	2.047	6.794	1.512	Educational Instruction and Library Occupations
Special Education Teachers, Secondary School	139,600	0.915	1.190	5.588	1.319	Educational Instruction and Library Occupations
Special Education Teachers, Middle School	78,440	0.945	1.329	5.701	1.361	Educational Instruction and Library Occupations
Career/Technical Education Teachers, Secondary School	85,150	0.965	1.456	6.125	1.409	Educational Instruction and Library Occupations
Secondary School Teachers, Except Special and Career/Technical Education	896,190	0.860	1.493	5.823	1.445	Educational Instruction and Library Occupations
Career/Technical Education Teachers, Middle School	10,520	0.947	1.493	5.748	1.467	Educational Instruction and Library Occupations
Middle School Teachers, Except Special and Career/Technical Education	541,590	0.886	1.370	5.911	1.401	Educational Instruction and Library Occupations
Elementary School Teachers, Except Special Education	1,246,620	0.894	1.324	5.509	1.331	Educational Instruction and Library Occupations
Kindergarten Teachers, Except Special Education	102,140	0.856	1.219	5.187	1.277	Educational Instruction and Library Occupations
Family and Consumer Sciences Teachers, Postsecondary	2,120	0.876	2.185	6.482	1.574	Educational Instruction and Library Occupations
Farm and Home Management Educators	7,400	0.900	1.887	7.129	1.877	Educational Instruction and Library Occupations
Library Technicians	63,170	0.861	2.068	6.804	1.670	Educational Instruction and Library Occupations
Geography Teachers, Postsecondary	2,840	0.850	2.081	6.011	1.442	Educational Instruction and Library Occupations
Healthcare Diagnosing or Treating Practitioners, All Other	19,700	1.000	1.307	4.887	1.276	Healthcare Practitioners and Technical Occupations
Judges, Magistrate Judges, and Magistrates	28,230	1.000	1.069	4.703	1.291	Legal Occupations
Arbitrators, Mediators, and Conciliators	2,710	1.000	1.892	6.707	1.587	Legal Occupations

Administrative Law Judges, Adjudicators, and Hearing Officers	12,490	1.000	1.339	6.365	1.371	Legal Occupations
Judicial Law Clerks	15,480	1.000	2.751	7.787	1.903	Legal Occupations
Forensic Science Technicians	15,430	0.877	1.219	4.933	1.361	Life, Physical, and Social Science Occupations
Forest and Conservation Technicians	26,480	0.909	1.229	4.929	1.341	Life, Physical, and Social Science Occupations
Urban and Regional Planners	33,510	0.840	1.411	6.267	1.431	Life, Physical, and Social Science Occupations
School Psychologists	52,770	0.876	1.657	6.107	1.424	Life, Physical, and Social Science Occupations
Education Administrators, Kindergarten through Secondary	236,130	0.826	1.857	6.413	1.455	Management Occupations
Postmasters and Mail Superintendents	13,460	1.000	1.755	6.060	1.392	Management Occupations
Postal Service Mail Sorters, Processors, and Processing Machine Operators	119,200	1.000	1.337	5.304	1.299	Office and Administrative Support Occupations
Postal Service Mail Carriers	326,760	1.000	1.626	5.743	1.433	Office and Administrative Support Occupations
Postal Service Clerks	77,690	1.000	1.655	5.176	1.348	Office and Administrative Support Occupations
Public Safety Telecommunicators	86,130	0.900	2.375	6.356	1.561	Office and Administrative Support Occupations
Library Assistants, Clerical	65,070	0.838	2.307	5.760	1.539	Office and Administrative Support Occupations
Eligibility Interviewers, Government Programs	137,880	0.921	1.798	6.911	1.811	Office and Administrative Support Occupations
Court, Municipal, and License Clerks	156,900	0.982	3.321	7.232	1.955	Office and Administrative Support Occupations
First-line Supervisors of Correctional Officers	54,300	0.972	1.297	4.801	1.277	Protective Service Occupations
Transportation Security Screeners	42,240	0.888	1.117	3.896	1.187	Protective Service Occupations
Animal Control Workers	10,390	0.904	1.574	4.653	1.355	Protective Service Occupations
Transit and Railroad Police	3,010	0.893	1.522	4.796	1.297	Protective Service Occupations
Parking Enforcement Workers	7,570	0.929	1.302	4.812	1.308	Protective Service Occupations
Fish and Game Wardens	6,530	1.000	1.320	5.205	1.333	Protective Service Occupations
Detectives and Criminal Investigators	107,310	0.999	1.836	5.717	1.395	Protective Service Occupations
Correctional Officers and Jailers	346,940	0.955	1.216	4.186	1.267	Protective Service Occupations

Bailiffs	16,260	1.000	1.071	3.047	1.131	Protective Service Occupations
Forest Fire Inspectors and Prevention Specialists	2,200	0.965	1.485	5.296	1.489	Protective Service Occupations
Fire Inspectors and Investigators	12,300	0.848	1.453	5.474	1.374	Protective Service Occupations
Firefighters	298,870	0.930	1.329	3.920	1.195	Protective Service Occupations
First-line Supervisors of Firefighting and Prevention Workers	80,680	0.960	1.352	5.141	1.322	Protective Service Occupations
First-line Supervisors of Police and Detectives	130,530	0.990	1.294	5.308	1.303	Protective Service Occupations
Police and Sheriff\'s Patrol Officers	649,400	0.990	1.388	4.738	1.341	Protective Service Occupations
Air Traffic Controllers	19,400	0.913	1.630	6.158	1.425	Transportation and Material Moving Occupations
Subway and Streetcar Operators	8,370	0.918	1.378	4.827	1.341	Transportation and Material Moving Occupations

D Sample Task Estimations

Job title	Task	IM	GPT rep. scale	GPT rep. expl.	GPT ast. scale	GPT ast. expl.	Prod. inc.	GPT limit.	Human relev.	Skills to maintain	Skills to acquire	Obsolete skills
Administrative Law Judges, Adjudicators, and Hearing Officers	Conduct hearings to review and decide claims regarding issues, such as social program eligibility, environmental protection, or enforcement of health and safety regulations.	4.59	1	GPT cannot physically conduct hearings or make legal decisions, as these require human judgment and presence.	7	GPT can assist by providing legal research, drafting documents, and summarizing case law to support decision-making.	1.5	GPT lacks the ability to interpret nuances in testimony, make judgments, and physically preside over hearings.	Human judges are essential for interpreting law with understanding of social context and exercising discretion.	Judges should maintain skills in legal reasoning, judgment, and decision-making.	Judges may need to learn how to integrate AI tools into their workflow for efficiency.	Routine legal research and some documentation tasks may become less critical with AI assistance.
Detectives and Criminal Investigators	Prepare reports that detail investigation findings.	4.53	3	While GPT can generate reports, it cannot conduct physical investigations or verify findings.	8	GPT can significantly speed up report writing and ensure thoroughness and accuracy of documentation.	2	GPT cannot replace the human investigative process or the nuanced understanding of case details.	Investigators provide the critical insights and judgments that form the basis of reports.	Investigative skills and the ability to synthesize information are crucial.	Learning to use AI for data analysis and report generation is advantageous.	Basic report drafting may become less important with AI assistance.
First-line Supervisors of Correctional Officers	Restrain, secure, or control offenders, using chemical agents, firearms, or other weapons of force as necessary.	4.49	1	GPT cannot engage in physical restraint or control of offenders, which is inherently a physical task.	2	GPT can offer theoretical training on control techniques but cannot assist in physical application.	1.1	GPT lacks physical capabilities and cannot respond to real-time security threats.	Physical intervention and real-time decision-making are necessary and cannot be replaced by GPT.	Physical restraint techniques and situational awareness are vital skills to maintain.	Understanding of AI-assisted security systems could be beneficial.	Manual reporting post-incident may be reduced with automated systems.
Fish and Game Wardens	Patrol assigned areas by car, boat, airplane, horse, or on foot to enforce game, fish, or boating laws or to manage wildlife programs, lakes, or land.	4.75	1	GPT cannot patrol areas or enforce laws, as it lacks physical capabilities and authority.	2	GPT can assist by providing information or communication support during patrols.	1.1	GPT cannot physically patrol or directly enforce regulations.	Human judgment and physical presence are essential for law enforcement.	Patrolling skills, law enforcement, and environmental knowledge.	Use of AI for data analysis and communication enhancement.	Routine data entry and analysis may be reduced.
Subway and Streetcar Operators	Operate controls to open and close transit vehicle doors.	4.83	1	GPT cannot physically operate doors; this requires manual or automated systems.	2	GPT could provide verbal commands or reminders, but minimal impact on productivity.	1	GPT lacks physical presence and cannot interact with control mechanisms.	Humans are needed for manual control and to handle unexpected situations.	Operators should maintain manual dexterity and situational awareness.	Learning to work with AI for operational efficiency could be beneficial.	Simple repetitive tasks may be automated, reducing the need for manual operation.
Arbitrators, Mediators, and Conciliators	Prepare written opinions or decisions regarding cases.	4.89	4	GPT can help draft written opinions but cannot replace the legal reasoning and judgment required for official decisions.	8	GPT can assist by providing legal research and drafting support, improving the efficiency of decision writing.	2	GPT cannot interpret legal nuances or apply judicial discretion as a human arbitrator would in decision-making.	Human judgment is critical for weighing evidence, legal arguments, and making fair decisions in legal proceedings.	Arbitrators must maintain their legal reasoning, judgment, and decision-making abilities.	Arbitrators should learn to use AI for research and initial drafting to enhance their decision-making process.	Basic legal research and initial drafting may be streamlined with GPT\’s capabilities.
Firefighters	Create openings in buildings for ventilation or entrance, using axes, chisels, crowbars, electric saws, or core cutters.	4.55	1	GPT cannot physically handle tools or perform tasks requiring manual dexterity and strength.	2	GPT could provide guidance on the use of tools based on building materials and design.	1.05	GPT lacks the ability to manipulate physical objects and respond to real-time physical scenarios.	The physical execution of creating openings in structures requires human intervention.	Manual dexterity, use of firefighting tools, and physical strength are essential.	Understanding AI-generated structural analysis could be beneficial.	Manual research on building construction for ventilation may be reduced.
Family and Consumer Sciences Teachers, Postsecondary	Supervise undergraduate or graduate teaching, internship, and research work.	4.12	1	GPT cannot supervise hands-on research or internships, as it lacks physical presence and nuanced judgment.	5	GPT can assist in providing research guidance, resources, and preliminary feedback on student work.	1.2	GPT cannot evaluate practical skills, provide in-person mentorship, or make complex judgments.	Teachers offer expertise, mentorship, and personalized feedback that GPT cannot replicate.	Mentorship, expertise in the field, and the ability to provide personalized feedback are crucial.	Teachers should learn to use AI for administrative tasks and initial student assessments.	Basic information retrieval and initial draft feedback may be less needed from humans.

(Continued)

Continued

Library Assistants, Clerical	Open and close library during specified hours and secure library equipment, such as computers and audio-visual equipment.	4.31	1	GPT cannot physically open/close libraries or secure equipment.	2	GPT can help organize schedules for opening/closing but cannot perform physical tasks.	1.1	GPT lacks physical presence and cannot interact with real-world objects.	Physical presence is required to manage library security and equipment.	Physical security management and equipment handling remain crucial.	Learning to integrate AI for scheduling and security system management could be beneficial.	Manual scheduling and some aspects of security monitoring may become less needed.
Eligibility Interviewers, Government Programs	Compile, record, and evaluate personal and financial data to verify completeness and accuracy, and to determine eligibility status.	4.55	2	GPT can\not physically compile or record data, but can process and evaluate data if digitized.	8	GPT can streamline data evaluation, flag inconsistencies, and suggest eligibility outcomes, increasing efficiency.	2.5	GPT cannot interact with physical documents or verify their authenticity, and lacks human judgment.	Humans are needed for nuanced decisions, empathy in interviews, and handling sensitive information.	Interviewing skills, ethical judgment, and decision-making remain crucial for eligibility workers.	Workers should learn to integrate AI insights into their workflow and data analysis.	Manual data compilation may decline as digitization and AI processing become prevalent.
Tax Examiners and Collectors, and Revenue Agents	Send notices to taxpayers when accounts are delinquent.	4.59	3	GPT can automate delinquency notices but cannot address individual circumstances without human oversight.	9	GPT can streamline the process of sending notices by automating content creation and tracking delinquencies.	3	GPT cannot manage the personalization required for specific taxpayer situations.	Humans are needed to handle complex cases and to provide a personal touch in sensitive situations.	Empathy, discretion, and complex problem-solving skills are essential.	Learning to manage and interpret AI-generated reports and data analytics is important.	Routine notification tasks may become automated, reducing the need for manual processing.
Forensic Science Technicians	Examine footwear, tire tracks, or other types of impressions.	3.28	1	GPT cannot physically examine or identify impressions as it is not capable of interacting with the physical world.	3	GPT can assist by providing information on analysis techniques or helping to draft preliminary findings.	1.1	GPT cannot handle physical evidence or perform the tactile and visual tasks required.	Forensic analysis requires human expertise for accurate evidence interpretation and handling.	Skills in evidence handling, pattern recognition, and forensic analysis must be maintained.	Technicians should acquire skills in digital evidence management and AI-assisted analysis.	Some aspects of evidence cataloging may become less needed with AI assistance.
Probation Officers and Correctional Treatment Specialists	Prepare and maintain case folder for each assigned inmate or offender.	4.46	2	GPT cannot manage physical case folders but can help organize and retrieve digital case information.	7	GPT can assist in structuring case notes, automating data entry, and providing templates for case management.	1.5	GPT cannot interact with inmates or offenders and lacks the ability to make nuanced judgments.	Human empathy and judgment are crucial for managing offender rehabilitation and interpreting behavior.	Probation officers should maintain interpersonal skills, judgment, and case management expertise.	Officers should learn to use AI for administrative tasks and data analysis.	Manual data entry and some aspects of record-keeping may become less critical.
Highway Maintenance Workers	Set out signs and cones around work areas to divert traffic.	4.5	1	GPT cannot set out physical signs or cones on roads.	4	GPT could help optimize traffic diversion plans or train workers on safety protocols.	1.2	GPT cannot perform physical tasks or adapt to on-site conditions.	Human presence is required for setting up and adjusting traffic control measures.	Physical setup of signs and cones and on-site adaptability are crucial skills.	Learning to integrate AI tools for traffic management could be beneficial.	Manual traffic diversion planning might be streamlined with AI assistance.

Data Availability

The datasets generated and used to support the findings of this study are openly available in Harvard Dataverse at: Jieshu Wang, Elif Kiran, S. R. Aurora (also known as Mai P. Trinh), Michael Simeone, and José Lobo. 2024. Replication Data for: ChatGPT on ChatGPT: An Exploratory Analysis of its Performance in the Public Sector Workforce (Version 1) [dataset]. *Harvard Dataverse* (2024). https://doi.org/10.7910/DVN/P3CDHS.

Acknowledgments

We thank our colleagues at the Decision Theater at Arizona State University (ASU) for their invaluable contributions to our discussions, especially the team at Positive Action to Knowledge (PAK) for their active engagement and insights, which have not only made this collaboration possible but also profoundly enhanced the quality of our research. Additionally, we acknowledge the AI Innovation Challenge at ASU, in collaboration with OpenAI, which provided us the opportunity to experiment with advanced generative AI technologies. We also thank the OpenAI Researcher Access Program for granting us access to their cutting-edge API tools, which were instrumental in conducting our study. Furthermore, we extend our gratitude to the editor and anonymous reviewer for their insightful comments and constructive feedback, which have significantly improved the quality and clarity of this article.

References

[1] Ajay Agrawal, John McHale, and Alex Oettl. 2018. *Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth*. Working paper w24541. National Bureau of Economic Research, Cambridge, MA. DOI: https://doi.org/10.3386/w24541

[2] AI Now Institute. 2018. *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*. Technical Report. AI Now Institute. Retrieved from http://www.law.nyu.edu/sites/default/files/litigatingalgorithms_0.pdf

[3] Hussam Alkaissi and Samy I. McFarlane. 2023. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* (Feb. 2023). DOI: https://doi.org/10.7759/cureus.35179

[4] David H. Autor. 2015. The paradox of abundance: Automation anxiety returns. In *Performance and Progress: Essays on Capitalism, Business, and Society*, Subramanian Rangan (Ed.). Oxford University Press.

[5] Evan M. Berman and Imane Hijal-Moghrabi. 2022. *Performance and Innovation in the Public Sector: Managing for Results* (3rd ed.). Routledge, New York. DOI: https://doi.org/10.4324/9781003304753

[6] Joel Blit, Samantha St. Amand, and Joanna Wajda. 2018. *Automation and the Future of Work: Scenarios and Policy Options*. Technical Report. Centre for International Governance Innovation, Waterloo, Ontario. Retrieved from http://www.zbw.eu/econis-archiv/handle/11159/2032

[7] Erin L. Borry and Heather Getha-Taylor. 2019. Automation in the public sector: Efficiency at the expense of equity? *Pub. Integ.* 21, 1 (Jan. 2019), 6–21. DOI: https://doi.org/10.1080/10999922.2018.1455488

[8] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. Generative AI at Work. DOI: https://doi.org/10.3386/w31161

[9] Erik Brynjolfsson and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (1st ed.). W. W. Norton & Company, New York.

[10] Erik Brynjolfsson and Tom Mitchell. 2017. What can machine learning do? Workforce implications. *Science* 358, 6370 (Dec. 2017), 1530–1534. DOI: https://doi.org/10.1126/science.aap8062

[11] Nicholas G. Carr. 2014. *The Glass Cage: Automation and Us* (1st ed.). W. W. Norton & Company, New York.

[12] Tom Christensen, Per Lægreid, and Kjell Arne Røvik. 2020. *Organization Theory and the Public Sector: Instrument, Culture and Myth* (2nd ed.). Routledge. Retrieved from https://www.routledge.com/Organization-Theory-and-the-Public-Sector-Instrument-Culture-and-Myth/Christensen-Laegreid-Rovik/p/book/9780367428914

[13] Iain M. Cockburn, Rebecca Henderson, and Scott Stern. 2019. The impact of artificial intelligence on innovation: An exploratory analysis. In *The Economics of Artificial Intelligence: An Agenda*, Ajay Agrawal, Joshua Gans, and Avi Goldfarb (Eds.). University of Chicago Press. Retrieved from http://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda

[14] Steven G. Craig, Edward C. Hoang, and Janet E. Kohlhase. 2023. Adoption of technological change in the public sector: Evidence from US states. *Int. Region. Sci. Rev.* 46, 3 (May 2023), 299–327. DOI: https://doi.org/10.1177/01600176221125692

[15] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. 2019. *AI Now 2019 Report*. Technical Report. AI Now Institute, New York. Retrieved from https://ainowinstitute.org/AI_Now_2019_Report.html

[16] Erich C. Dierdorff and Jennifer J. Norton. 2011. *Summary of Procedures for O*NET Task Updating and New Task Generation*. Technical Report. National Center for O*NET Development. Retrieved from https://www.onetcenter.org/dl_files/TaskUpdating.pdf

[17] Paul R. Donnellan. 2018. The future of mobility-electric, autonomous, and shared vehicles. *IEEE Eng. Manag. Rev.* 46, 4 (2018), 16–18. DOI: https://doi.org/10.1109/EMR.2018.2880987

[18] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. DOI: https://doi.org/10.48550/arXiv.2303.10130

[19] D. C. Engelbart. 1962. *Augmenting Human Intellect: A Conceptual Framework.* Technical Report 3223. Stanford Research Institute, Washington, D.C. Retrieved from https://csis.pace.edu/~marchese/CS835/Lec3/DougEnglebart.pdf

[20] David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. DOI: https://doi.org/10.2139/ssrn.3551505

[21] Carl Frey and Michael Osborne. 2013. The Future of Employment: How Susceptible Are Jobs to Computerization? Retrieved from http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

[22] Carl Frey and Michael Osborne. 2023. Carl Benedikt Frey and Michael Osborne on how AI benefits lower-skilled workers. *The Economist* (Sept. 2023). Retrieved from https://www.economist.com/by-invitation/2023/09/18/carl-benedikt-frey-and-michael-osborne-on-how-ai-benefits-lower-skilled-workers

[23] Francisco Jose Garcia-Penalvo and Andrea Vazquez-Ingelmo. 2023. What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *Int. J. Interact. Multim. Artif. Intell.* 8, 4 (2023). DOI: https://doi.org/10.9781/ijimai.2023.07.006

[24] Sarah N. Giest and Bram Klievink. 2024. More than a digital system: How AI is changing the role of bureaucrats in different organizational contexts. *Public Management Review* 26, 2 (2024), 379–398. https://doi.org/10.1080/14719037.2022.2095001

[25] ISA. [n. d.]. What Is Automation? Retrieved from https://www.isa.org/about-isa/what-is-automation

[26] Nir Jaimovich and Henry E. Siu. 2012. *The Trend Is the Cycle: Job Polarization and Jobless Recoveries.* Working Paper 18334. National Bureau of Economic Research. DOI: https://doi.org/10.3386/w18334

[27] John Maynard Keynes. 1931. Economic possibilities for our grandchildren. *Essays in Persuasion.* MacMillan. Retrieved from http://gutenberg.ca/ebooks/keynes-essaysinpersuasion/keynes-essaysinpersuasion-00-h.html#Economic_Possibilities

[28] Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2023. Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Educ. Inf. Technol.* (Oct. 2023). DOI: https://doi.org/10.1007/s10639-023-12249-8

[29] Wassily Leontief. 1983. National perspectives: The definition of problems and opportunities. In *The Long-term Impact of Technology on Employment and Unemployment: A National Academy of Engineering Symposium, June 30, 1983.* National Academies. Google-Books-ID: hS0rAAAAYAAJ.

[30] Helen Margetts and Cosmina Dorobantu. 2019. Rethink government with AI. *Nature (London)* 568, 7751 (2019), 163–165. DOI: https://doi.org/10.1038/d41586-019-01099-5

[31] Andrew Maynard. 2015. Navigating the fourth industrial revolution. *Nat. Nanotechnol.* 10, 12 (2015), 1005–1006. DOI: https://doi.org/10.1038/nnano.2015.286

[32] Erin McCormick. 2021. What happened when a “wildly irrational” algorithm made crucial healthcare decisions. *The Guardian* (July 2021). Retrieved from https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions

[33] Konradin Metze, Rosana C. Morandin-Reis, Irene Lorand-Metze, and João B. Florindo. 2024. Bibliographic research with ChatGPT may be misleading: The problem of hallucination. *J. Pediat. Surg.* 59, 1 (Jan. 2024), 158. DOI: https://doi.org/10.1016/j.jpedsurg.2023.08.018

[34] Lucia Nalbandian. 2022. An eye for an “I”: A critical assessment of artificial intelligence tools in migration and asylum management. *Compar. Migrat. Stud.* 10, 1 (2022), 32–32. DOI: https://doi.org/10.1186/s40878-022-00305-0

[35] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Comput. Educ.: Artif. Intell.* 2 (Jan. 2021), 100041. DOI: https://doi.org/10.1016/j.caeai.2021.100041

[36] Michael A. Peters. 2019. Technological unemployment: Educating for the Fourth Industrial Revolution. In *The Chinese Dream: Educating the Future.* Routledge, 99–107.

[37] James Phoenix and Mike Taylor. 2024. *Prompt Engineering for Generative AI* (early release ed.). O’Reilly Media, Inc. Retrieved from https://www.oreilly.com/library/view/prompt-engineering-for/9781098153427/

[38] Daniel S. Schiff, Kaylyn Jackson Schiff, and Patrick Pierson. 2022. Assessing public value failure in government adoption of artificial intelligence. *Pub. Admin. (London)* 100, 3 (2022), 653–673. DOI: https://doi.org/10.1111/padm.12742

[39] Klaus Schwab. 2016. The Fourth Industrial Revolution: What it means and how to respond. Retrieved from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

[40] Jieshu Wang, Elif Kiran, S. R. Aurora (also known as Mai P. Trinh), Michael Simeone, and José Lobo. 2024. Replication data for: ChatGPT on ChatGPT: An exploratory analysis of its performance in the public sector workforce (version 1) [dataset]. *Harvard Dataverse.* DOI: https://doi.org/10.7910/DVN/P3CDHS

[41] Weiyu Wang and Keng Siau. 2019. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *J. Datab. Manag.* 30, 1 (Jan. 2019), 61–79. DOI: https://doi.org/10.4018/JDM.2019010104

[42] Michael Webb. 2019. The Impact of Artificial Intelligence on the Labor Market. DOI: https://doi.org/10.2139/ssrn.3482150

[43] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer. 2019. Artificial intelligence and the public sector—Applications and challenges. *Int. J. Pub. Admin.* 42, 7 (May 2019), 596–615. DOI: https://doi.org/10.1080/01900692.2018.1498103

[44] World Economic Forum. 2020. *The Future of Jobs Report 2020.* Technical Report. Geneva Switzerland. Retrieved from https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

[45] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 1 (2019), 1–27. DOI: https://doi.org/10.1186/s41239-019-0171-0

[46] Fabrizio Zilibotti. 2008. Economic possibilities for our grandchildren 75 years after: A global perspective. In *Revisiting Keynes: Economic Possibilities for Our Grandchildren*, Lorenzo Pecchi and Gustavo Piga (Eds.). The MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262162494.003.0003

[47] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. 2021. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Gov. Inf. Quart.* 38, 3 (2021). DOI: https://doi.org/10.1016/j.giq.2021.101577

Received 30 January 2024; revised 7 April 2024; accepted 26 May 2024