Benchmarking Large Language Models in Evaluating Workforce Risk of Robotization: Insights from Agriculture

AgriEngineering

MDPI

Article

Benchmarking Large Language Models in Evaluating Workforce Risk of Robotization: Insights from Agriculture

Lefteris Benos [1,*], Vasso Marinoudi [2,3], Patrizia Busato [4], Dimitrios Kateris [1], Simon Pearson [2] and Dionysis Bochtis [1,3]

[1] Institute for Bio-Economy and Agri-Technology (IBO), Centre of Research and Technology-Hellas (CERTH), 57001 Thessaloniki, Greece; d.kateris@certh.gr (D.K.); d.bochtis@certh.gr (D.B.)
[2] Lincoln Institute for Agri-Food Technology (LIAT), University of Lincoln, Lincoln LN6 7TS, UK; v.marinoudi@farm-b.com (V.M.); spearson@lincoln.ac.uk (S.P.)
[3] FarmB Digital Agriculture S.A., 17th Noemvriou 79, 55534 Thessaloniki, Greece
[4] Interuniversity Department of Regional and Urban Studies and Planning (DIST), Polytechnic of Turin, Viale Mattioli 39, 10125 Torino, Italy; patrizia.busato@polito.it
* Correspondence: e.benos@certh.gr

Abstract: Understanding the impact of robotization on the workforce dynamics has become increasingly urgent. While expert assessments provide valuable insights, they are often time-consuming and resource-intensive. Large language models (LLMs) offer a scalable alternative; however, their accuracy and reliability in evaluating workforce robotization potential remain uncertain. This study systematically compares general-purpose LLM-generated assessments with expert evaluations to assess their effectiveness in the agricultural sector by considering human judgments as the ground truth. Using ChatGPT, Copilot, and Gemini, the LLMs followed a three-step evaluation process focusing on (a) task importance, (b) potential for task robotization, and (c) task attribute indexing of 15 agricultural occupations, mirroring the methodology used by human assessors. The findings indicate a significant tendency for LLMs to overestimate robotization potential, with most of the errors falling within the range of 0.229 ± 0.174. This can be attributed primarily to LLM reliance on grey literature and idealized technological scenarios, as well as their limited capacity, to account for the complexities of agricultural work. Future research should focus on integrating expert knowledge into LLM training and improving bias detection and mitigation in agricultural datasets, as well as expanding the range of LLMs studied to enhance assessment reliability.

Keywords: ChatGPT; Copilot; Gemini; generative artificial intelligence; occupational information network (O*NET); human–machine interaction; task automation; expert-validated ground truth

Academic Editors: Thuseethan Selvarajah and Yakub Sebastian

Received: 18 February 2025
Revised: 17 March 2025
Accepted: 1 April 2025
Published: 3 April 2025

Citation: Benos, L.; Marinoudi, V.; Busato, P.; Kateris, D.; Pearson, S.; Bochtis, D. Benchmarking Large Language Models in Evaluating Workforce Risk of Robotization: Insights from Agriculture. *AgriEngineering* **2025**, *7*, 102. https://doi.org/10.3390/agriengineering7040102

1. Introduction

The rapid advancement of technology has significantly reshaped the workforce dynamics, leading to the partial or complete replacement of human labor by computers and robots [1,2]. Organizations are turning to automation to enhance efficiency and reduce labor costs. Routine tasks are increasingly automated, reducing human involvement [3], while artificial intelligence (AI) now enables the automation of even cognitive, non-routine work [4,5]. However, rather than solely displacing workers, automation also creates opportunities for human–machine collaboration, where machines enhance human capabilities and vice versa, enabling more innovative and flexible work environments [6,7]. To remain

competitive in this transformative landscape, employees must embrace lifelong learning and continuous skill development [8,9]. In these evolving workforce dynamics, understanding the impact of automation on workers and job content is crucial for navigating future challenges and opportunities.

In an attempt to quantify how recent technology is likely to affect the future of employment, several research works have been presented in the related literature. The effects of automation have been assessed across multiple levels, ranging from international perspectives to more specific contexts such as continents, countries, labor markets, and industries. As an illustration of the different investigating levels, Arntz [10] and Nedelkoska and Quintini [11] focused on Organization for Economic Cooperation and Development (OECD) countries. Foster-McGrego et al. [12] and Pouliakas [13] dealt with the European labor market. Furthermore, studies such as [14–16] examined the potential impact of automation on the labor markets of Brazil, China, and South Africa, respectively. Finally, works like [17] concerned manufacturing jobs, while [18] evaluated the vulnerability of agricultural jobs to the implementation of robotization. Overall, findings often vary significantly, as the influence of automation remains complex and dependent on numerous factors, including the type of technology (for example robots, AI, information, and communication technology (ICT)) and methodological approach. As noted in the review study of Filippi et al. [19], research on this subject has expanded rapidly since 2014, following the publication of Frey and Osborne's working paper [20], which focused on the U.S. labor market.

In general, when assessing automation risk through probability estimation, researchers identify occupations and tasks most susceptible to automation, calculate their likelihood of being automated, and determine the extent to which workers may be replaced. Indicatively, based on [20], the likelihood of automation for various occupations was assessed in [21] using a four-step approach and relying on a web-based platform created for the U.S. Department of Labor (O*NET) [22]. They first evaluated 70 occupations, marking them with a "1" if deemed automatable and a "0" if not on the basis of input from technology experts. Next, they identified three essential human abilities that resist automation, namely perception and manipulation, as well as social and creative intelligence. Afterwards, they used Gaussian process classifiers to calculate the automation probabilities for these roles. Finally, they used these calculated probabilities to analyze employment data towards gauging the potential impact of automation on the workforce. The above methodological approach has been followed almost exclusively by many researchers, including in [14,23–26]. Interestingly, in [11,27–29], socio-demographic (such as age, gender, and education level) and job characteristics (like working hours and wage) were also exploited.

Marinoudi et al. [18,30] introduced a new methodology to assess the impact of robotization on the agricultural labor market. Their study involved a systematic analysis of several agricultural occupations, categorizing them according to the routine versus non-routine and cognitive versus manual nature of the tasks performed. Moreover, they calculated a weighted average susceptibility rate to robotization, employing a methodological approach that included assigning importance scores to each task within these occupations. Using the O*NET database [22], they also assessed the economic impacts of replacing human labor with automation [18]. In [30], the focus was on key qualifications needed to withstand robotization. This structured framework was also partially adopted in [31] to evaluate the possible impact of large language models (LLMs) on agricultural workforce dynamics, focusing on substitution versus complementarity, job function reshaping, and skill requirements.

In summary, the common denominator of the related literature is that it relies on subjective judgements, which are often not replicable. To mitigate this drawback, the opinion of several domain experts is often averaged to minimize the impact of individual biases

and increase the reliability of the assessment. Yet, this remains a time-consuming process. Alternatively, LLMs can serve as powerful assessors of the impact of automated systems on the workforce by leveraging their capabilities in natural language understanding, knowledge extraction, data analysis, and semantic analysis [32,33]. Recently, this rationale was implemented by the International Labor Organization (ILO) in a working paper concerning the potential exposure of occupations to generative AI [34]. In particular, following a similar rubric to [35,36], ChatGPT, an AI-powered LLM developed by OpenAI [37], was utilized to assess task-level exposure to automation and aggregate the results at the occupation level via a sequence of API calls to the LLM.

While LLMs offer a scalable and efficient alternative to expert assessments, their accuracy and reliability in evaluating automation susceptibility remains uncertain, as also stressed in [34]. The lack of standardized evaluation frameworks makes it challenging to assess how well these models capture the complexities of workforce automation. Consequently, a benchmark is essential to systematically compare LLM-generated assessments against expert judgments and ensure their validity. The main objective of this study is to evaluate, for the first time, to the best of the authors' knowledge, the performance of three of the leading LLMs, namely ChatGPT, Copilot, and Gemini, in assessing occupation susceptibility to robotization. To provide a robust evaluation framework, we use human expert assessments as the ground truth [30]. By thoroughly comparing LLM-generated susceptibility scores against expert judgments, we aim to identify the strengths and limitations of these models in a workforce analysis. The methodological approach presented here is adaptable, enabling future studies to evaluate various occupations across different industries and regions. Finally, this study contributes to the development of a standardized benchmarking methodology, paving the way for future research on reliable AI-driven labor market assessments.

We focus on agriculture, as the sector presents unique challenges for robotization that distinguish it from the more standardized environments of the manufacturing sector [38,39]. While manufacturing robots operate in controlled, repetitive settings, agricultural robots must adapt to dynamic and often harsh conditions, complicating their design and deployment [40,41]. The challenges of robotization also stem from the diversity of tasks, such as planting, harvesting, weeding, and livestock care, each requiring specialized capabilities and domain-specific knowledge. Another important complication is the inherent variability of agricultural environments, including unpredictable weather and uneven terrain complicating the design and deployment of automated systems. Additionally, the sensitive nature of live produce, which is highly susceptible to environmental conditions, demands precision and adaptability from automation technologies [42].

2. Materials and Methods

2.1. Agricultural Occupations Included in the Assessment Framework

As a means of setting a common framework to compare LLM assessments against those provided by the assessors in [30], the same 15 agricultural occupations were considered as described in the open-source online tool of O*NET [22]. Necessary information for each occupation was collected statistically from a random sample of workers using standardized questionnaires that are regularly updated. Each occupation comprises carefully selected tasks, specifically described and detailed in the O*NET database, to provide a complete understanding of the associated roles and responsibilities. Table 1 summarizes the aforementioned 15 agricultural occupations and the number of tasks required for each.

These occupations cover various roles within agriculture, ranging from management and technical expertise to scientific research and manual labor, all essential for the operation and advancement of agricultural practices.

Table 1. A synopsis of the 15 reviewed agricultural occupations, including their O*NET 8-digit codes and the number of tasks required for each, as detailed in [22].

Occupation	O*NET Code	Tasks
Farmers, Ranchers, and Other Agricultural Managers	11-9013.00	30
Farm Labor Contractors	13-1074.00	8
Agricultural Engineers	17-2021.00	14
Animal Scientists	19-1011.00	9
Soil and Plant Scientists	19-1013.00	27
Agricultural Technicians	19-4012.00	26
Precision Agriculture Technicians	19-4012.01	22
Farm and Home Management Educators	25-9021.00	15
First-Line Supervisors of Farming, Fishing, and Forestry Workers	45-1011.00	30
Agricultural Inspectors	45-2011.00	22
Graders and Sorters, Agricultural Products	45-2041.00	6
Agricultural Equipment Operators	45-2091.00	17
Farmworkers and Laborers, Crop, Nursery, and Greenhouse	45-2092.00	30
Farmworkers, Farm, Ranch, and Aquacultural Animals	45-2093.00	22
Farm Equipment Mechanics and Service Technicians	49-3041.00	14

2.2. Large Language Models Used as Assessors

In this analysis, three prominent LLMs, namely ChatGPT, Copilot, and Gemini, were utilized as automated assessors. The selection of these models was driven by several key factors. First, they are freely available, which ensures broad accessibility for research purposes. Second, they have implemented measures to align with several international data regulations, like the European Union's General Data Protection Regulation (GDPR). Third, these models represent diverse approaches to language processing and understanding, offering a well-rounded basis for comparison. Lastly, these LLMs are user-friendly and do not require extensive technical knowledge to operate. In brief:

- ChatGPT, powered by OpenAI's GPT series [37], is renowned for its ability to generate coherent and contextually relevant responses across various topics. For this study, the latest freely available "ChatGPT-3.5" model was specifically utilized, given its enhanced capability to handle complex tasks, produce detailed outputs, and maintain extended dialogues.
- Copilot is an AI-powered assistant integrated into Microsoft applications [43]. In particular, the "Microsoft 365 Copilot" version, currently available for free use, exploits advanced natural language processing to provide contextually relevant and coherent responses, primarily assisting with tasks such as automating repetitive cognitive processes, making it a valuable tool for the present study.
- "Gemini 2.0 Flash", the recent version of the Gemini series developed by Google AI [44], is especially notable for its integration into Google's vast array of services and its deep learning capabilities. Its ability to handle complex reasoning tasks and provide accurate information makes it an essential resource for the current analysis.

2.3. Methodology Steps for Assessing the Susceptibility of Agricultural Occupations to Robotization and Their Cognitive/Manual Versus Routine/Non-Routine Levels

2.3.1. Data Preparation

Since ChatGPT, Copilot, and Gemini are not specifically pre-trained for the agricultural workforce landscape, domain-specific information was provided for each occupation to ensure an accurate evaluation. The O*NET database [22] served as the primary source of domain-specific data for the 15 agricultural occupations assessed in this study. It offered detailed descriptions of each occupation, including occupation-specific tasks, required technology skills, and worker requirements (such as skills, abilities, and work values). This domain-specific information was supplied to guide the models in making more precise and

contextually relevant assessments, addressing various aspects that will be detailed below. Since users generally cannot fine-tune these LLMs directly—this being a process typically carried out by the model developers (OpenAI, Microsoft, or Google, in this case)—the present study relied on prompt engineering. By thoughtfully designing input prompts, the evaluations were kept contextually relevant and tailored to the specific domain. The sequence of customized prompts employed for LLMs in the current assessment framework is described in Appendix A.

2.3.2. Sequence of Customized Prompts

To evaluate the susceptibility rate to robotization of each occupation, a three-step methodology was followed, based on a systematic approach identical to that used in [30]. For brevity, only the key aspects of the methodology are briefly mentioned in this study. Prior to the assessment procedure, a preliminary setup and contextualization were carried out to ensure consistency by using specific prompts (Appendix A). To account for potential variability in the responses, the procedure was repeated daily for ten consecutive days, from 1 to 10 February 2025. The non-deterministic output of LLMs, a consequence of their stochastic nature, explains the variability in their responses [34]. The ratings provided by each LLM were then averaged to obtain a more stable and representative assessment, reducing the impact of randomness in individual outputs. The following iterative approach helped mitigate fluctuations in the responses, ensuring that the final susceptibility scores reflected a more consistent evaluation.

Preliminary Setup and Contextualization

- LLM identification

The process begins by identifying the specific version of the LLM being used. This is important, because different versions may have variations in their training data and capabilities, which could impact the accuracy and reliability of the assessment.

- Occupation context

To establish a foundation for evaluation, the LLM is furnished with background information regarding the agricultural occupation under consideration. Crucially, this includes a list of the constituent tasks involved based on [22], enabling the LLM to effectively conduct subsequent evaluations.

- Framework initialization

The user instructs the LLM that it will be asked to assign specific scores to various aspects of the tasks associated with the occupation according to the criteria defined in [30].

Three-Step Assessment Methodology

- Step 1: Task importance assessment

In this step, the LLM evaluates the importance of each task included in an occupation by assigning an "importance weight" based on a five-point scale ranging from 1 (not important) to 5 (strongly important). This rating helped determine which tasks were central to the occupation and, therefore, should be prioritized when considering the potential to robotization.

- Step 2: Potential to robotization of each task assessment

This step concerns the assessment of the potential to robotization of each task on the technology readiness level (TRL) and the feasibility of automation. By incorporating the TRL scores, we are able to distinguish between tasks that are closer to being automated and those that may face significant practical challenges [18].

- Step 3: Task attribute indexing

The final step involved assigning an index for the nature of each task. The LLMs were instructed to quantify the contribution of four attributes to the execution of each task, namely (a) cognitive routine, (b) cognitive non-routine, (c) manual routine, and (d) manual non-routine. Specifically, each LLM was instructed to assign an index value from the set [0, 0.25, 0.5, 0.75, 1] to each task, ensuring that the sum of these values for all four attributes equaled 1 for each task.

After the LLMs rated the importance and robotization potential for each task, the results were aggregated to an overall normalized susceptibility rate to robotization, $\hat{s}_i$, for each occupation (an occupation is represented as o_i with $i \in O$ and $O = \{1, \ldots 15\}$), classifying the occupations into three zones:

- Occupations with $0 \leq \hat{s}_i < 0.33$ correspond to a low susceptibility rate to robotization (green zone);
- Occupations with $0.33 \leq \hat{s}_i < 0.66$ correspond to a moderate susceptibility rate to robotization (yellow zone);
- Occupations with $0.66 \leq \hat{s}_i \leq 1$ correspond to a high susceptibility rate to robotization (red zone).

Furthermore, the task attributes provided by LLMs were used to map the occupations into a cognitive/manual (serving as the y-coordinate) versus routine/non-routine (serving as the x-coordinate) graph.

2.3.3. Evaluation of LLM Assessments

The outputs generated by the LLMs were compared with assessments provided by human experts in the field [30] using several metrics, including (a) relative error in task importance ratings, visualized through a box plot; (b) comparison of the rated potential for robotization, based on the classification of occupations into zones (green, yellow, or red); (c) overestimation/underestimation of the LLM-assigned $\hat{s}_i$ compared to human assessments, calculated by the difference between the resulting values; (d) Pearson correlation coefficients to assess the level of agreement among the $\hat{s}_i$ values of the LLMs; (e) comparison of $\hat{s}_i$ values for occupations grouped by task nature and major group; and (f) comparison of human and LLM assessments of agricultural occupations in terms of routine/non-routine and cognitive/manual task content.

Human assessors possess in-depth, domain-specific expertise, backed by years of practical experience and a nuanced understanding of the complexities of agricultural occupations. Their judgments were not only informed by their own knowledge but also by participatory interviews with professionals actively working in agricultural occupations, ensuring that the assessments reflect real-world conditions. In contrast to LLMs, which generate responses based on patterns in large datasets, human experts are capable of accounting for the context-dependent factors that influence the labor landscape. As a consequence, the expert evaluations serve as the ground truth, establishing them as the definitive standard for comparison in this study.

To maintain objectivity and prevent potential bias, the individual scores assigned by human assessors were not disclosed to the LLMs. This ensured that LLMs generated independent evaluations based solely on the provided task descriptions and prompts rather than aligning their outputs with preexisting expert assessments.

2.3.4. Indicative Example Illustrating the Methodology Assessment

As an illustration of the methodology, described in detail in [18,30], the occupation of Animal Scientists (19-1011.00) is considered here, which consists of nine tasks [22]. Let us focus on the first task ($j = 1$), which concerns studying the nutritional requirements

of animals and the nutritive values of animal feed materials. Copilot classified it as "very important" ($\overline{w}_1$ = 4), whereas ChatGPT and Gemini rated it as "strongly important" ($\overline{w}_1$ = 5), aligning with the human assessors' ratings. A normalized weight for each task was assigned by dividing its score by the total sum of all task scores (n = 9 in this case):

$$\hat{w}_j = \frac{\overline{w}_j}{\sum_{j=1}^{n} \overline{w}_j}. \quad (1)$$

The resulting normalized weights for this task were 0.131, 0.118, and 0.125 for ChatGPT, Copilot, and Gemini, respectively, while the corresponding human-based weight was 0.141.

Regarding the task's potential for robotization, Copilot and Gemini estimated that a significant portion could be automated (score: 0.5), whereas ChatGPT—consistent with human assessors—found no reasonable indication that the task could be automated (score: 0). By multiplying these scores with the above normalized weights, the weighted potential for robotization was calculated for this task ($\overline{r_j}$): Copilot: $\overline{r_1} = 0.059$, Gemini: $\overline{r_1} = 0.063$, and ChatGPT and human assessors: $\overline{r_1} = 0$. The overall normalized susceptibility rate to robotization for the occupation under investigation, $\hat{s}$ was calculated by summing the weighted potential for robotization of all tasks:

$$\hat{s} = \sum_{j=1}^{n} \overline{r_j}. \quad (2)$$

The calculated values were as follows: ChatGPT: 0.28 (green zone), Copilot: 0.40 (yellow zone), Gemini: 0.60 (yellow zone), and human assessors: 0.17 (green zone).

In terms of cognitive/manual and routine/non-routine classification, ChatGPT categorized the task as primarily cognitive, assigning 0.75 to cognitive routine attribute and 0.25 to the cognitive non-routine one. However, Copilot and Gemini assigned a value of 0.50 to the cognitive routine attribute, while both the cognitive non-routine and manual routine attributes received 0.25. In comparison, the human assessors classified the task as entirely cognitive non-routine by assigning 1.00, with zero values for cognitive routine, manual routine, and manual non-routine attributes. The routine/non-routine balance of a task was derived by summing the non-routine scores while subtracting those of the routine ones. The x-coordinate was calculated by aggregating the routine/non-routine balances of all tasks, weighted based on their relative importance in the occupation. Similarly, the cognitive/manual balance of a given task was determined by adding the scores of the cognitive aspects while subtracting those of the manual aspects. The y-coordinate was obtained by aggregating the cognitive/manual balances of all tasks, weighted based on the relative importance of each task.

3. Results

3.1. Task Importance Assessment

As detailed in Section 2.3.2, the first scoring step involved rating the importance of each task within the occupation through a five-point scale, ranging from "not important" (Score 1) to "strongly important" (Score 5). This scoring task is crucial, as it assigns weights to individual tasks, directly impacting the subsequent analysis of occupational susceptibility to robotization by ensuring that critical tasks are given more consideration. Figure 1 presents a box plot comparing the relative error in task importance ratings assigned by the three LLMs against the assessments provided by human experts in [30], calculated through

$$Relativeerror = \frac{LLM\ score – Human\ assessors\ average\ score}{Human\ assessors\ average\ score} \cdot 100\%. \quad (3)$$

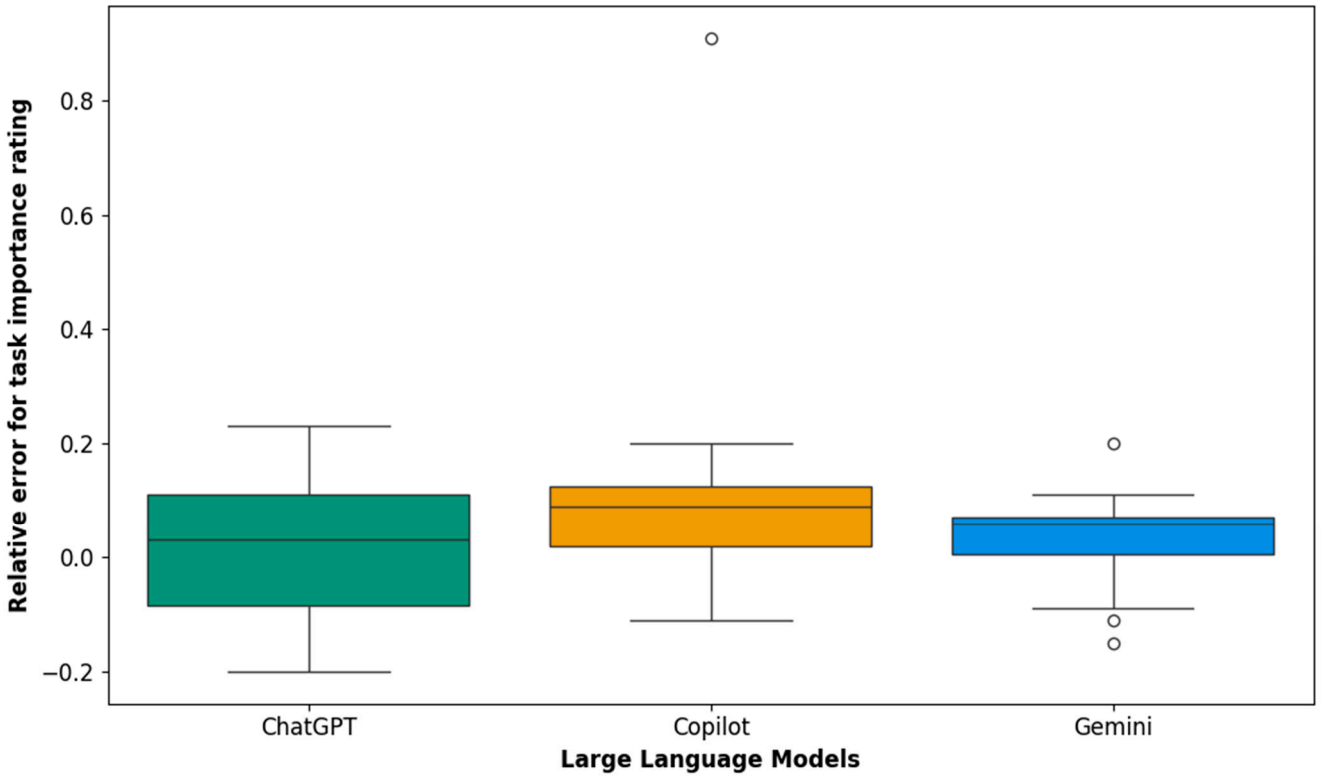


Figure 1. Box plot of the relative error in the task importance rating for the three large language models, namely ChatGPT, Copilot, and Gemini, as compared to human assessors.

Each box plot illustrates key statistical properties, including the median (represented by the central black line within the box) and the interquartile range (IQR), which encompasses the middle 50% of the data. The bottom and top edges of the box correspond to the 25th and 75th percentiles, respectively. Outliers are data points that lie an abnormal distance from other values and are depicted as individual points (circles) plotted beyond the whiskers.

Focusing on the present analysis, the median relative error for all three LLMs appears to be close to zero, signifying that, on average, their central tendency aligns reasonably well with human assessors' ratings. While the medians of all box plots fall within the IQR (boxes) of the others, suggesting no clearly statistically significant differences in the central tendency, there are notable differences in data dispersion. Specifically, ChatGPT exhibits greater dispersion than Copilot and Gemini, as illustrated by its longer box in Figure 1. Furthermore, the wider range of extreme values (whiskers) for ChatGPT indicates a broader overall distribution.

Notably, ChatGPT's box plot looks the most symmetric compared to the others, as the median line is roughly in the center of the box, and the whiskers extend a somewhat similar distance on both sides. The absence of outliers in ChatGPT suggests that its predictions, while slightly more variable overall, remained within a consistent range without extreme deviations. In contrast, both Copilot and Gemini show a possible left skew, with the latter suggesting the strongest indication. This means that the LLMs tend to overestimate the importance of tasks more often than they underestimate it. The single outlier above the upper whisker of Copilot and Gemini does suggest overestimation. However, the two outliers below the Gemini's lower whisker demonstrate instances where Gemini also underestimated the importance of tasks. With these outliers in mind, it can be inferred that Gemini might be relying on features that are not reliably indicative of task importance, leading to highly variable and inaccurate ratings in some cases.

3.2. Potential to Robotization of Each Occupation Assessment

3.2.1. Summary of the Calculated Susceptibility Rates to Robotization

Following [18,30], by using a three-tier scoring for evaluating the vulnerability to automation of each individual task and exploiting the weights to task importance mentioned above, an overall normalized susceptibility rate to robotization, $\hat{s}_i$, was calculated for each occupation. Table 2 summarizes the final estimated values of $\hat{s}_i$ for both human assessors [30] and LLMs in 15 occupations, as well as the mean LLM assessment in the final column. Each occupation is paired with its O*NET code for reference.

Table 2. Calculated susceptibility rates to robotization, $\hat{s}_i$, for humans [30], and large language models (LLMs), ChatGPT, Copilot, and Gemini, across the 15 reviewed occupations, accompanied by their O*NET 8-digit codes.

Occupation	O*NET Code	Humans	ChatGPT	Copilot	Gemini	Mean LLM Rating
Farmers, Ranchers, and Other Agricultural Managers	11-9013.00	0.21	0.43	0.54	0.50	0.49
Farm Labor Contractors	13-1074.00	0.37	0.31	0.69	0.34	0.45
Agricultural Engineers	17-2021.00	0.22	0.47	0.54	0.68	0.56
Animal Scientists	19-1011.00	0.17	0.28	0.40	0.60	0.43
Soil and Plant Scientists	19-1013.00	0.06	0.43	0.43	0.57	0.48
Agricultural Technicians	19-4012.00	0.21	0.52	0.81	0.62	0.65
Precision Agriculture Technicians	19-4012.01	0.51	0.72	0.88	0.70	0.77
Farm and Home Management Educators	25-9021.00	0.08	0.36	0.56	0.59	0.50
First-Line Supervisors of Farming, Fishing, and Forestry Workers	45-1011.00	0.15	0.39	0.56	0.56	0.50
Agricultural Inspectors	45-2011.00	0.36	0.44	0.66	0.59	0.56
Graders and Sorters, Agricultural Products	45-2041.00	0.92	0.84	0.87	0.94	0.88
Agricultural Equipment Operators	45-2091.00	0.71	0.67	0.73	0.70	0.70
Farmworkers and Laborers, Crop, Nursery, and Greenhouse	45-2092.00	0.67	0.52	0.66	0.57	0.58
Farmworkers, Farm, Ranch, and Aquacultural Animals	45-2093.00	0.46	0.73	0.60	0.53	0.62
Farm Equipment Mechanics and Service Technicians	49-3041.00	0.21	0.58	0.50	0.63	0.57

3.2.2. Classification of Occupations into Three Susceptibility to Robotization Zones

An important aspect of [30] was the classification of occupations into three zones (green, yellow, and red) based on the overall normalized susceptibility rate to robotization, $\hat{s}_i$, indicating low, moderate, or high potential, as detailed in Section 2.3.2. Figure 2 reveals that the human assessors identified eight occupations as belonging to the green zone, signifying minimal susceptibility to robotization. In contrast, most LLMs predominantly classified occupations in the yellow or red zones, suggesting moderate to high automation potential, thus showing a first overestimation. ChatGPT was the only LLM to place two occupations in the green zone: Animal Scientists (19-1011.00), aligning with the human assessments, and Farm Labor Contractors (13-1074.00). In addition, ChatGPT accurately classified Agricultural Inspectors (45-2011.00) as being in the yellow zone. All LLMs consistently identified Graders and Sorters, Agricultural Products (45-2041.00), and Agricultural Equipment Operators (45-2091.00) as high-risk occupations (red zone), in agreement with human judgments. Copilot also matched the human assessments for Farmworkers and Laborers, Crop, Nursery, and Greenhouse (45-2092.00). Gemini's performance aligned most closely with the human classifications, achieving five matches, followed by ChatGPT with four and Copilot with three. Overall, the LLMs demonstrated an inter-rater agreement of 53%.

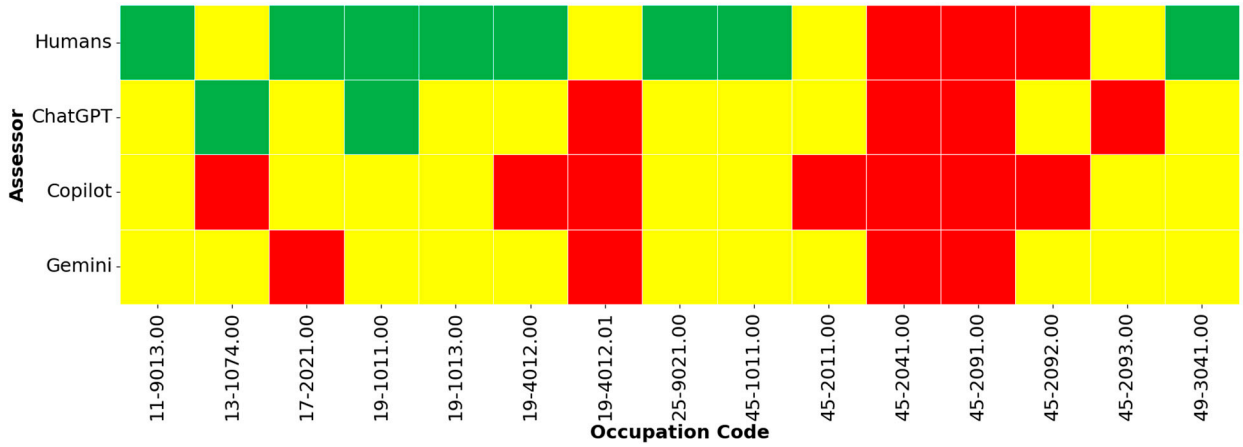


Figure 2. Heatmap illustrating the classification of the occupations into green, yellow, and red zones representing low, moderate, and high potential to robotization, respectively, according to all the assessors (Humans, ChatGPT, Copilot, and Gemini).

3.2.3. Error Analysis: LLM Versus Human Robotization Susceptibility Assessments

Classifying agricultural occupations into three zones provided a broad measure of agreement between LLMs and human assessments. To gain deeper insight, we now shift focus to analyzing the errors in robotization susceptibility assessments made by the three examined LLMs and comparing them to human evaluations for each occupation. This approach allows for a more nuanced understanding of how closely LLMs align with human judgment and where discrepancies arise.

Figure 3 presents a grouped bar chart visualizing the aforementioned error (*LLM score–Human average Score*). The x-axis displays the 15 agricultural occupations identified by their O*NET code, whereas the y-axis represents the error in the susceptibility rate, where positive values indicate overestimation and negative values indicate underestimation by the LLMs. For each occupation, the results are displayed as clustered bars, with each LLM represented by a distinct color—green for ChatGPT, orange for Copilot, and blue for Gemini—facilitating a direct comparison of their performance. The height of each bar reflects the magnitude of the error, with taller bars signifying a greater discrepancy between the LLM assessment and the human assessment. In general, the degree of error varies significantly across the 15 occupations, while no single LLM is consistently the most accurate for all job types, though ChatGPT showed a slightly better overall performance. For most occupations, all LLMs overestimate the susceptibility rate, although the extent of overestimation varies across LLMs and occupations. The exception is occupation Farmworkers and Laborers, Crop, Nursery, and Greenhouse (45-2092.00), where all LLMs underestimate the potential for robotization. Copilot exhibits the smallest underestimation, while ChatGPT exhibits the largest. However, for occupations Farm Labor Contractors (13-1074.00), Graders and Sorters, Agricultural Products (45-2041.00), and Agricultural Equipment Operators (45-2091.00), two LLMs underestimate the rate, while the third overestimates it. The overall tendency of LLMs to overestimate the susceptibility of agricultural occupations to robotization, with an average error of 0.229 ± 0.174, is likely influenced by multiple factors, which will be explored in detail in the Discussion section.

Figure 4 presents a heatmap displaying the Pearson correlation coefficients, which measure the agreement between ChatGPT, Copilot, and Gemini in predicting the susceptibility of agricultural occupations to robotization. The results indicate a moderate correlation between ChatGPT and Copilot (0.64) and between ChatGPT and Gemini (0.68), suggesting that ChatGPT's predictions align relatively well with both models. In contrast, the corre-

lation between Copilot and Gemini (0.43) is weaker, highlighting greater divergence in their assessments.

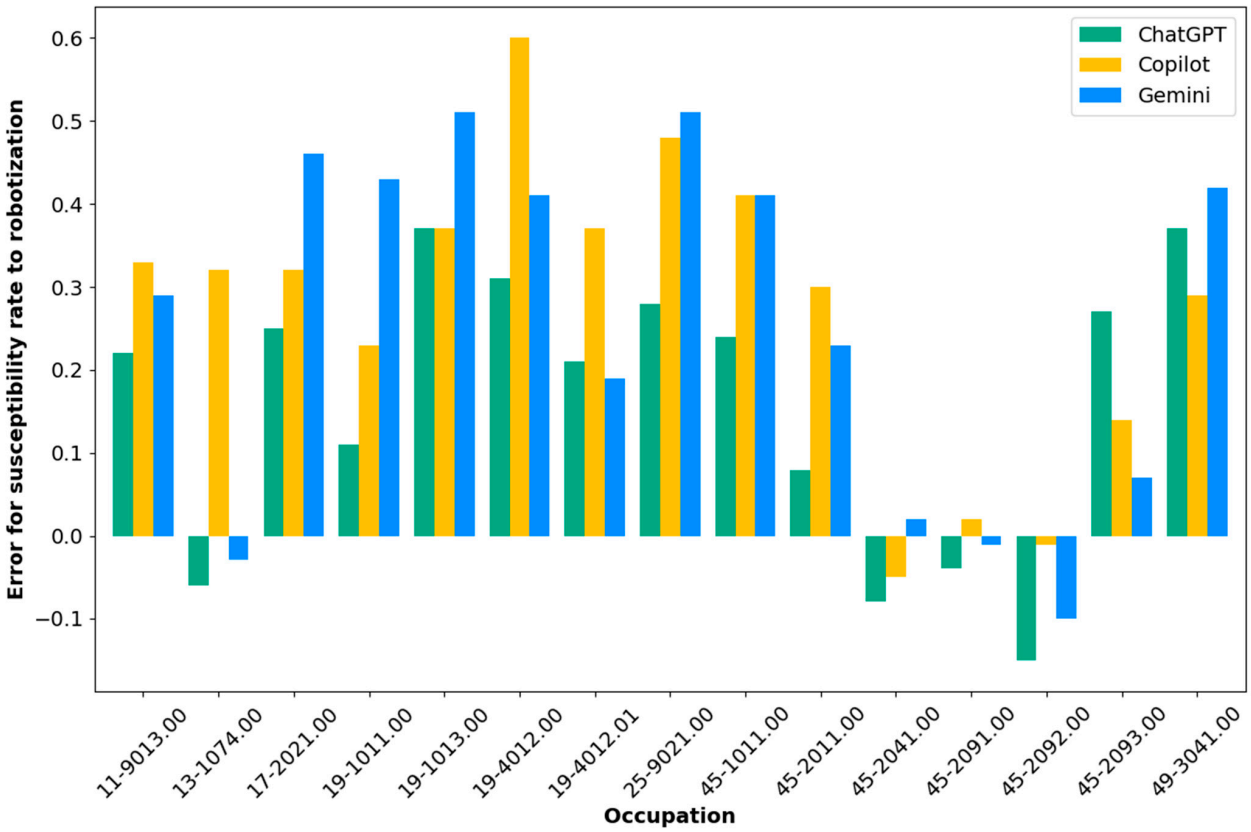


Figure 3. Overestimation/underestimation of the susceptibility rate to robotization, $\hat{s}_i$, by large language models, specifically ChatGPT, Copilot, and Gemini, compared to human assessors for each occupation.

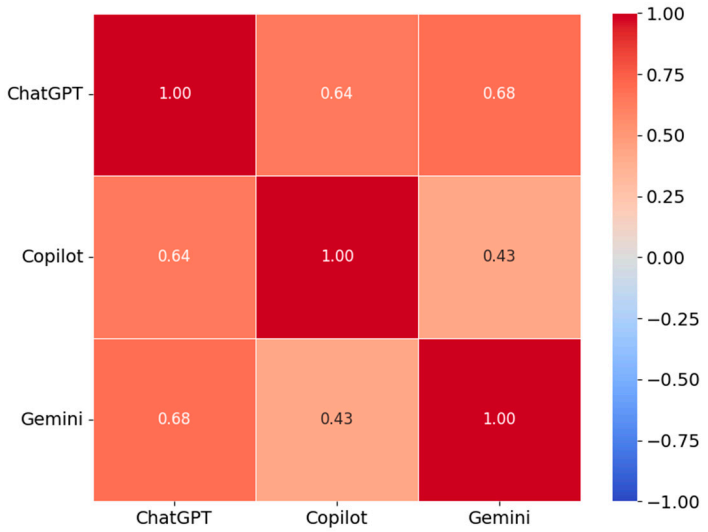
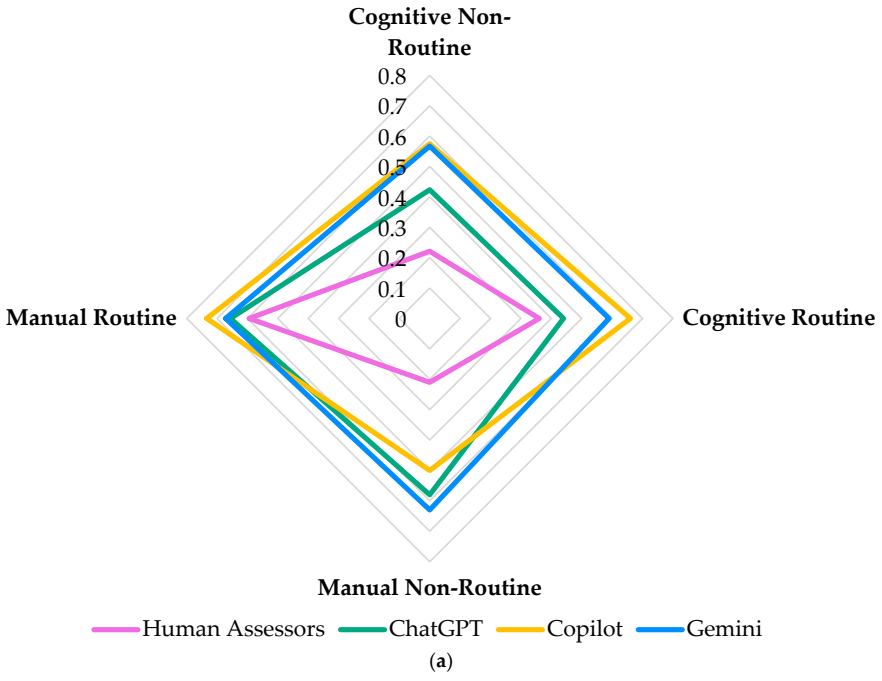


Figure 4. Pearson correlation coefficients assessing the agreement between ChatGPT, Copilot, and Gemini in predicting robotization susceptibility of the reviewed agricultural occupations.

3.2.4. Task Characteristics and Occupational Group Analysis

Following the classification of agricultural occupations based on the nature of the tasks they involve according to human evaluation [30], a spider chart is illustrated in Figure 5a. Four axes represent four distinct task categories: (a) cognitive non-routine, (b) cognitive

routine, (c) manual non-routine, and (d) manual routine, while the grey lines represent the susceptibility to robotization, $\hat{s}_i$. As a general remark, there is a strong tendency toward overestimation, as highlighted above, in all cases that can lead to misleading results, that will be elaborated next. The only case where LLMs and human assessments tend to align, to some extent, is the occupations characterized as a manual routine in [30]. Those occupations presented the highest $\hat{s}_i$, as these tasks often involve minimal cognitive decision-making and repetitive, predictable actions that can be easily automated by robotic systems. The LLMs are also high on this axis, confirming that they also recognize the high susceptibility of these tasks.



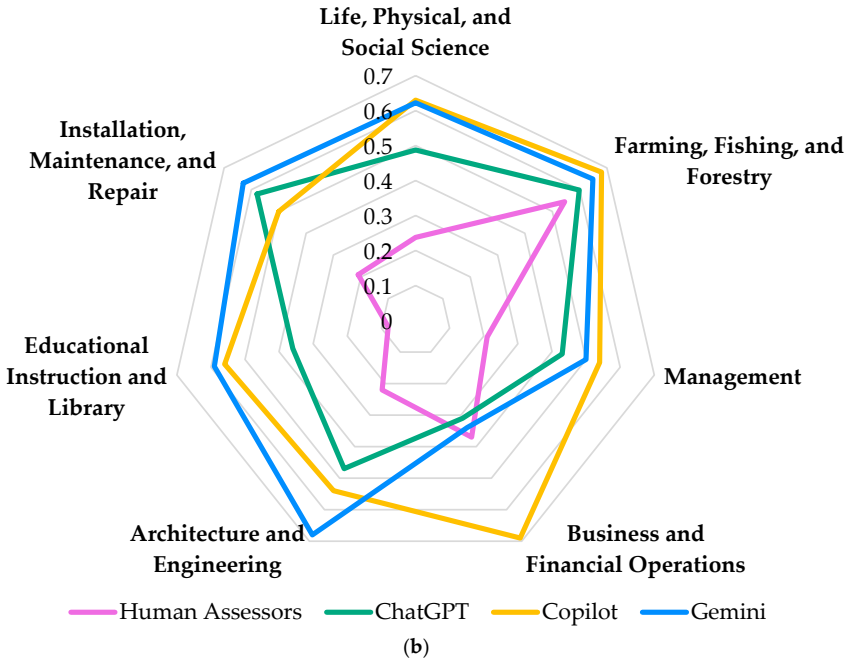


Figure 5. Spider charts illustrating the susceptibility rate to robotization based on (**a**) the nature of tasks within an occupation and (**b**) the classification of the reviewed occupations into major groups, both according to [30], as evaluated by large language models, namely ChatGPT, Copilot, and Gemini, and human assessors.

A comparable analysis was also carried out for the major occupational groups [22]. Again, an overall overestimation is observed regarding the LLM judgement for the susceptibility rate to robotization. As can be deduced from Figure 5b, the highest mean LLM errors (*Mean LLM score–Human average Score*) are for Installation, Maintenance, and Repair and Educational Instruction and Library major groups. Notably, the Farming, Fishing, and Forestry group, consisting of six agricultural occupations, demonstrates the best alignment between LLMs and human-calculated $\hat{s}_i$. The other group showing a relative agreement between humans and ChatGPT is that of Business and Financial Operations, which consists only of occupation with code 13-1074.00 (Farm Labor Contractors). The small underestimation of ChatGPT and Gemini in this specific case, combined with the overestimation of Copilot, is associated with the results depicted in Figure 3 corresponding to that occupation.

3.3. Mapping of Occupations Based on Their Cognitive/Manual and Routine/Non-Routine Characteristics

Based on the calculations of the routine/non-routine and cognitive/manual balance, as in [30], a two-dimensional occupation mapping was achieved for all LLMs. Figure 6 summarizes the resulting mapping for (a) human assessors, (b) ChatGPT, (c) Copilot, and (d) Gemini. Considering the analysis according to the average assessments of a group of domain experts (Figure 6a), a single occupation was found in the fourth quadrant, including tasks mainly of a manual non-routine nature: Farm Equipment Mechanics and Service Technicians (49-3041.00). In contrast, no occupation was placed in this quadrant from LLMs. Indicatively, this occupation was placed in the second quadrant, representing cognitive routine occupations, by all LLMs failing to a great extent to capture the nature of the majority of the tasks the occupation involves.

As far as the first quadrant is concerned regarding cognitive non-routine occupations, the experts' ratings placed eight occupations in this quadrant. Out of them, ChatGPT succeeded only in one case (Agricultural Engineers (17-2021.00)). Additionally, it incorrectly placed the occupation of Farmworkers and Laborers, Crop, Nursery, and Greenhouse (45-2092.00) in the first quadrant. This specific misclassification was also observed for the occupation mapping of Gemini. Gemini placed six occupations in the first quadrant. However, out of these placements, only three were correct (Agricultural Engineers (17-2021.00); Farm and Home Management Educators (25-9021.00); and First-Line Supervisors of Farming, Fishing, and Forestry Workers (45-1011.00)). Lastly, Copilot placed no occupation in the cognitive non-routine quadrant.

As illustrated in Figure 6, the majority of occupations were placed by LLMs in the second quadrant, representing agricultural occupations that are predominantly cognitive and routine. This contrasts with human assessments, which classified only Agricultural Inspectors (45-2011.00) in this category. ChatGPT and Copilot assigned ten occupations to this quadrant, while Gemini placed eight.

The fourth quadrant, which represents occupations with mainly routine and manual components, was the second-most populated quadrant in Figure 6a (human assessments). However, this trend was not reflected in the LLM evaluations. ChatGPT and Copilot correctly placed Graders and Sorters, Agricultural Products (45-2041.00); and Farmworkers, Farm, and Ranch (45-2093.00) in the second quadrant. Additionally, Copilot accurately assigned Farmworkers and Laborers, Crop, Nursery, and Greenhouse (45-2092.00) to this category. Notably, Gemini failed to classify any occupation as manual routine.

Overall, these findings highlight the limitations of LLMs in recognizing most of the manual aspects of occupations. Most occupations were clustered in the first and second quadrants, associated with cognitive tasks, with some positioned near the boundary between routine and non-routine work (vertical axis), demonstrating a clear deviation from the human assessments. Finally, in all cases, there were substantial devia-

tions from the human assessments regarding both the *x*- and *y*-coordinates, indicating that LLMs struggle to accurately map occupations based on their cognitive/manual and routine/non-routine characteristics.

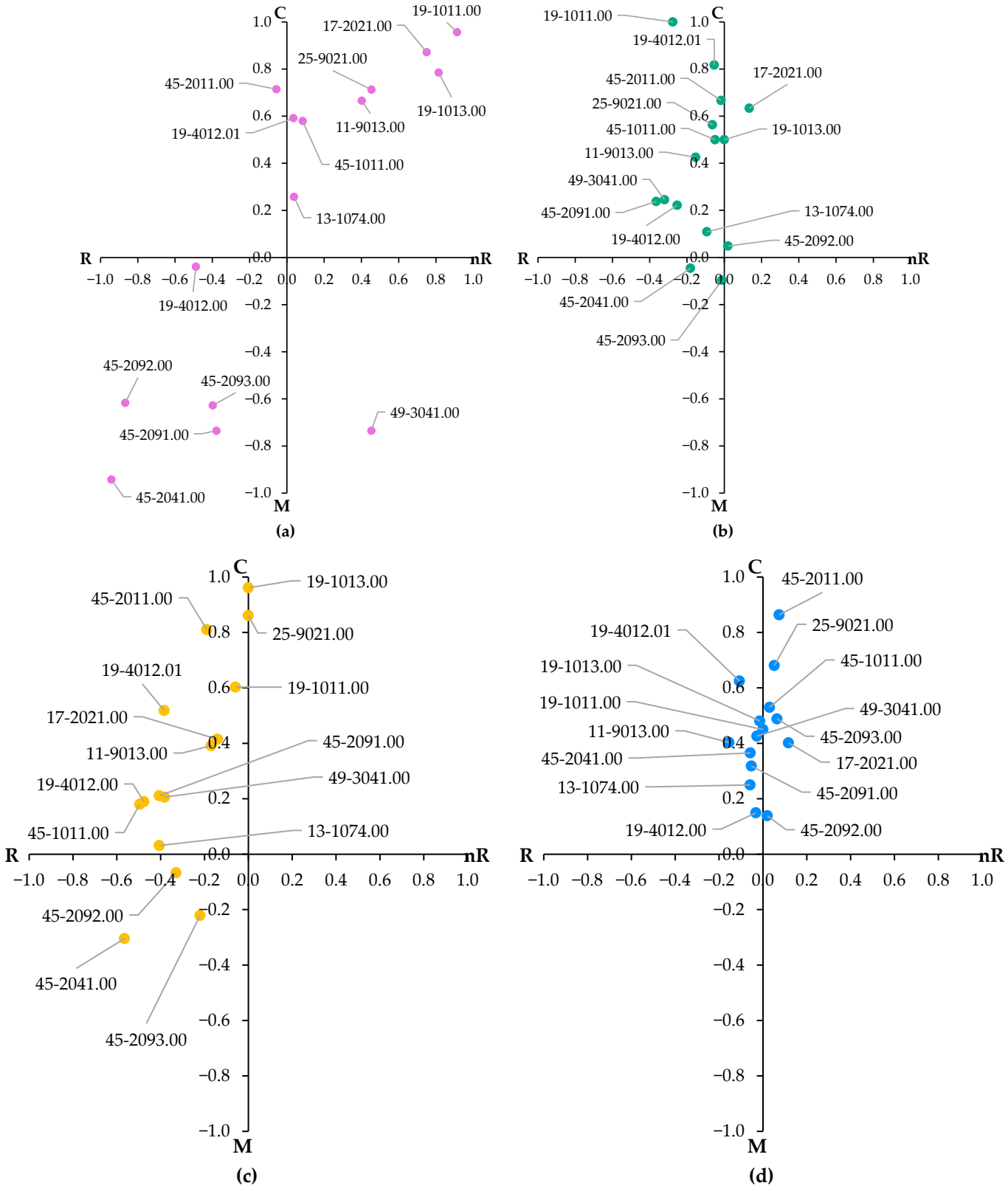


Figure 6. Comparison of human and large language model assessments of agricultural occupations in terms of the routine (R)/non-routine (nR) and cognitive (C)/manual (M) task content: (**a**) human assessors, (**b**) ChatGPT, (**c**) Copilot, and (**d**) Gemini.

4. Discussion

This study investigated the ability of LLMs, namely ChatGPT, Copilot, and Gemini, to assess the susceptibility of agricultural occupations to robotization. Each of these LLMs has its own strengths and is suited to specific use cases in assessing agricultural robotization potential. ChatGPT excels at processing and synthesizing information, making it ideal for a starting point for further expert analysis, offering general insights into the potential for agricultural robotization. On the other hand, Copilot is particularly effective in contexts where automation technologies need to be broken down into specific technical components or when analyzing the programming aspects of robotized systems. Lastly, Gemini, with its enhanced real-time data access and multimodal processing, offers context-aware assessments of agricultural robotization potential.

In this analysis, the rubric developed by Marinoudi et al. [18,30] was adopted. In essence, LLMs were instructed to evaluate the same 15 agricultural occupations as [30], with the domain experts assessments serving as the ground truth. After first being fed information from O*NET database [22], the LLMs followed a three-step evaluation procedure identical to [30] regarding (a) task importance, (b) potential to robotization of each task, and (c) task attribute indexing. The assessments produced by the LLMs were extensively compared with human evaluations analyzing their alignment with expert judgments and identifying potential discrepancies. As a general comment, a remarkable tendency of LLMs to overestimate the automation potential in agriculture was observed, which can be attributed to several factors, both intrinsic to the models' design and limitations, and the unique nature of the agricultural sector itself.

LLMs are trained on vast amounts of data from a variety of sources, including grey literature, referring to materials that are not formally published and usually lack thorough vetting [45,46]. Examples of grey literature that are often considered less reliable include online brochures released by companies that promote new automation technologies like agricultural machinery and robotics, articles that may be influenced by the interests of advertisers, personal opinions in forums, and white papers used for marketing, to mention but a few. These sources usually focus on idealized scenarios or the most advanced versions of technologies, which may not always reflect the practical challenges experienced in real-world agricultural settings. In simple terms, the training data often emphasize the potential for automation and technological progress while downplaying any limitations, risks, or uncertainties, leading to an inherent optimism bias. As a consequence, LLMs are likely to overestimate the immediate feasibility and applicability of robotic and AI technologies, leading them to classify agricultural occupations as more susceptible to automation than they actually are.

Agriculture is a particularly complex and diverse sector with tasks that vary widely in terms of their cognitive/manual and routine/non-routine aspects. Furthermore, many agricultural tasks are influenced by external factors such as weather and soil conditions, making them less predictable and harder to automate in a generalized manner [47,48]. LLMs, however, are often not equipped to recognize these subtle variations or the importance of localized knowledge and experience, which means they might overestimate the general applicability of automation technologies across all agricultural tasks. In addition, LLMs seem to struggle to consider the complexities of field deployment of robotic systems, especially when it comes to critical concerns like safety, labor, ethics, and regulatory factors [49]. For instance, safety concerns regarding the operation of autonomous machinery in unpredictable environments are difficult to be quantified [50] and might not be adequately captured in the data on which LLMs are trained. Labor issues, such as the displacement of workers, and the potential social impacts of automation cannot also easily be addressed by LLMs, which may tend to overlook the broader socio-economic implications [51].

LLMs also fail to fully account for various barriers to the adoption of automation technologies in agriculture, such as the need for a specialized infrastructure, the high costs associated with maintaining robotic systems, and regulatory constraints that hinder the widespread implementation of these technologies. Additionally, resistance to change within agricultural communities, where many may view these technologies as either threatening or impractical, further complicates the transition to automation [52,53]. These factors, which are deeply ingrained in the agricultural sector, can be overlooked by LLMs, leading to an incomplete assessment of the real challenges faced in adopting automation solutions. These models may also fail to adequately consider the potential shifts in employment patterns, such as the creation of new job roles or the need for reskilling workers [30], which are essential factors in the transition to automation.

In summary, human experts bring a level of practical, hands-on experience to their evaluations that LLMs lack. While LLMs can aggregate data and produce seemingly insightful conclusions, they often miss the qualitative aspects of work that human assessors rely on, especially in a highly specialized field like agriculture. Humans assess the complexity and variability of agricultural tasks with a deeper understanding of the context, while LLMs might overly generalize based on their training data characterized by optimism bias, leading to a failure in recognizing specific aspects in agricultural work. While these factors can significantly impact the feasibility and desirability of automating particular agricultural tasks, they are difficult for LLMs to accurately predict.

The overestimation of robotization potential observed in our study resonates with the trends identified in recent research, such as [54–56], which highlight how LLMs sometimes mirror cognitive biases like the Dunning–Kruger effect [57,58]. The Dunning–Kruger effect occurs when individuals with limited knowledge or expertise in a particular domain overestimate their own abilities or understanding. In our study, this was seen in how LLMs, despite lacking a deep understanding of the agricultural sector, tended to significantly overestimate the susceptibility of agricultural occupations to AI-supported automation. For instance, LLMs frequently placed many agricultural occupations in the "yellow" or "red" zones, indicating moderate to high potential for robotization, even though human experts considered these occupations to have lower susceptibility. This suggests that LLMs still exhibit overconfidence in their assessments, reflecting the Dunning–Kruger effect.

Based on the previously discussed points, a direction for future research involves developing methods to incorporate expert knowledge into the training and evaluation processes of LLMs. By collaborating with agricultural experts and farmers, researchers can curate datasets that reflect the practical realities of farming, including regional variations and crop-specific practices. This would enable LLMs to better understand the nuances of agricultural work and, thus, provide more accurate assessments of robotization potential. Moreover, the integration of detailed, reliable, sector-specific data could help LLMs comprehend the socio-economic, ecological, and safety aspects that influence agricultural practices and the adoption of new technologies. Future research could focus on developing LLMs that are better equipped to balance short-term enthusiasm with long-term sustainability. Research should also prioritize developing methods to detect and mitigate biases in agricultural datasets [59], as well as explainable AI techniques [60,61], to clarify why LLMs are making certain predictions and indicate the factors that are most influential in LLM decision-making.

Future research should also consider incorporating a broader range of LLMs to investigate the consistency of observed biases and limitations. By benchmarking multiple models, researchers can determine whether the identified tendencies (e.g., overestimation of robotization potential) are universal or specific to certain LLMs. This would provide valuable insights into the strengths and weaknesses of different models and inform the development

of improved and validated assessment tools. For example, comparing open-source and proprietary LLM performances can reveal differences in training data and methodologies, directly informing targeted improvements. In addition, ethical considerations, such as the displacement of workers and the equitable distribution of benefits, should be integrated into LLM assessments to ensure that automation technologies contribute to inclusive agricultural development.

Finally, rather than viewing LLMs as standalone tools, future research should explore frameworks for human–AI collaboration in workforce risk assessment. This could involve developing interactive systems where LLMs provide initial assessments, which are then refined and validated by human experts. Such collaborative approaches would combine the data processing and scale of LLMs with the practical expertise and contextual understanding of human assessors, leading to more accurate and actionable insights.

5. Conclusions

This study reveals a significant and consistent overestimation of the potential for robotization in agricultural occupations by the examined LLMs, namely ChatGPT, Copilot, and Gemini. This tendency is largely influenced by optimism bias in their training data, which often prioritize technological advancements over the practical realities of agricultural work. A key limitation identified in this study is that LLMs struggle to account for the unique complexities of agriculture, such as seasonal variability, unpredictable environmental factors, and the importance of hands-on expertise. Many agricultural tasks require adaptability and decision-making based on real-time conditions, which are difficult to model using static datasets. Furthermore, the feasibility of automation in agriculture is constrained by factors such as high equipment costs, the need for specialized infrastructure, and regulatory hurdles, all of which are often overlooked by LLM-generated assessments. Another critical shortcoming of LLMs is their inability to fully consider socio-economic implications, such as the displacement of farm workers and the evolving nature of agricultural labor. As a consequence, the current LLMs should be positioned as supportive tools, providing initial assessments and data analysis, rather than standalone decision-makers in agricultural robotization evaluations. Human experts deliver crucial insights drawn from their practical, hands-on experience and understanding of the sector's intricacies, which LLMs currently cannot replicate.

Moving forward, improving the integration of expert knowledge, sector-specific data, and bias mitigation techniques at several stages of LLM development will be crucial for enhancing the accuracy and reliability of LLMs in the context of automation assessments. A human–AI collaborative framework, where domain experts refine and validate AI-generated insights, will ensure that automation strategies align with the practical realities of modern farming and will ensure responsible progress in the field.

Author Contributions: Conceptualization, L.B., V.M., and D.B.; methodology, L.B. and V.M.; investigation, V.M., P.B., and D.K.; writing—original draft preparation, L.B. and V.M.; writing—review and editing, L.B., V.M., P.B., D.K., S.P., and D.B.; visualization, L.B., V.M., and S.P.; supervision, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed at the corresponding author.

Acknowledgments: The authors acknowledge the use of ChatGPT, Copilot, and Gemini in this study. These large language models were employed to follow a three-step evaluation process assessing (a) task importance, (b) potential for task robotization, and (c) task attribute indexing of 15 agriculture-related occupations. Their assessments were systematically compared with expert evaluations to analyze alignment and discrepancies, as part of the methodology.

Conflicts of Interest: Authors Vasso Marinoudi and Dionysis Bochtis were employed by the company farmB Digital Agriculture S.A. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interests.

Appendix A

The study utilized prompt engineering as a core methodological approach. By carefully crafting input prompts, the assessments were made both contextually appropriate and specifically aligned with the domain in question. A structured series of tailored prompts was applied to guide the evaluation framework for large language models:

1. *"What specific LLM version I am using now?";*
2. *"I will provide information about an agriculture-related occupation along with a list of its tasks…".* The information is sourced from the open-source online tool O*NET [22];
3. *"Based on this information, I am going to ask you to assign specific scores for several aspects of these tasks.";*
4. *"1st step: I want you to rate the importance weight of each individual task for this occupation using the following scale: (a) Not important: Score 1; (b) Slightly important: Score 2; (c) Important: Score 3; (d) Very important: Score 4; (e) Strongly important: Score 5.";*
5. *"2nd step: For each task assign an index for the potential to robotization. The rating refers to three scores, namely: (a) Score 0: there is no technology at technology readiness level (TRL) 3 or higher demonstrated, or there is no reasonable indication that the task can be computerized or robotized in the short- or mid-term future; (b) Score 0.5: a significant part (or parts) of the task can be computerized or robotized; and (c) Score 1: there is an existing technology or a technology under development at least at TRL 3 that can be implemented for the execution of the task.";*
6. *"3rd step: Assign an index for the nature of each task from the set [0, 0.25, 0.5, 0.75, 1] to each task to quantify the contribution of (a) Cognitive routine; (b) Cognitive non-routine; (c) Manual routine; and (d) Manual non-routine attributes to the execution of the task. These values must sum to 1 for each task."*

Points 4 to 6 are identical to those followed by human assessors in [30] and served as the ground truth for the present analysis.

References

1. Didier, N. Turning fragments into a lens: Technological change, industrial revolutions, and labor. *Technol. Soc.* **2024**, *77*, 102497. [CrossRef]
2. Qu, Y.; Fan, S. Is there a "Machine Substitution"? How does the digital economy reshape the employment structure in emerging market countries. *Econ. Syst.* **2024**, *48*, 101237. [CrossRef]
3. Marinoudi, V.; Sørensen, C.G.; Pearson, S.; Bochtis, D. Robotics and labour in agriculture. A context consideration. *Biosyst. Eng.* **2019**, *184*, 111–121. [CrossRef]
4. Upreti, A.; Sridhar, V. Effect of automation of routine and non-routine tasks on labour demand and wages. *IIMB Manag. Rev.* **2024**, *36*, 289–308. [CrossRef]
5. Zeyer-Gliozzo, B. Returns to formal, non-formal, and informal further training for workers at risk of automation. *J. Educ. Work* **2024**, *37*, 382–402. [CrossRef]
6. Leng, J.; Zhu, X.; Huang, Z.; Li, X.; Zheng, P.; Zhou, X.; Mourtzis, D.; Wang, B.; Qi, Q.; Shao, H.; et al. Unlocking the power of industrial artificial intelligence towards Industry 5.0: Insights, pathways, and challenges. *J. Manuf. Syst.* **2024**, *73*, 349–363. [CrossRef]

7. Gardezi, M.; Joshi, B.; Rizzo, D.M.; Ryan, M.; Prutzer, E.; Brugler, S.; Dadkhah, A. Artificial intelligence in farming: Challenges and opportunities for building trust. *Agron. J.* **2024**, *116*, 1217–1228. [CrossRef]
8. Bayly-Castaneda, K.; Ramirez-Montoya, M.-S.; Morita-Alexander, A. Crafting personalized learning paths with AI for lifelong learning: A systematic literature review. *Front. Educ.* **2024**, *9*, 1424386. [CrossRef]
9. Patino, A.; Naffi, N. Lifelong training approaches for the post-pandemic workforces: A systematic review. *Int. J. Lifelong Educ.* **2023**, *42*, 249–269. [CrossRef]
10. Arntz, M. The risk of automation for jobs in OECD countries: A comparative analysis. In *OECD Social, Employment and Migration Working Papers, No. 189*; OECD Publishing: Paris, France, 2016.
11. Nedelkoska, L.; Quintini, G. Automation, skills use and training. In *OECD Social, Employment and Migration Working Papers, No. 202*; OECD Publishing: Paris, France, 2018.
12. Foster-McGregor, N.; Nomaler, Ö.; Verspagen, B. Job Automation Risk, Economic Structure and Trade: A European Perspective. *Res. Policy* **2021**, *50*, 104269. [CrossRef]
13. Pouliakas, K. Determinants of automation risk in the EU labour market: A skills-needs approach. In *IZA Discussion Papers No. 11829*; Institute of Labor Economics (IZA): Bonn, Germany, 2018.
14. Albuquerque, P.H.M.; Saavedra, C.A.P.B.; de Morais, R.L.; Peng, Y. The Robot from Ipanema goes Working: Estimating the Probability of Jobs Automation in Brazil. *Lat. Am. Bus. Rev.* **2019**, *20*, 227–248. [CrossRef]
15. Zhou, G.; Chu, G.; Li, L.; Meng, L. The effect of artificial intelligence on China's labor market. *China Econ. J.* **2020**, *13*, 24–41. [CrossRef]
16. Le Roux, D.B. Automation and employment: The case of South Africa. *Afr. J. Sci. Technol. Innov. Dev.* **2018**, *10*, 507–517.
17. Parschau, C.; Hauge, J. Is automation stealing manufacturing jobs? Evidence from South Africa's apparel industry. *Geoforum* **2020**, *115*, 120–131. [CrossRef]
18. Marinoudi, V.; Lampridi, M.; Kateris, D.; Pearson, S.; Sørensen, C.G.; Bochtis, D. The future of agricultural jobs in view of robotization. *Sustainability* **2021**, *13*, 12109. [CrossRef]
19. Filippi, E.; Bannò, M.; Trento, S. Automation technologies and their impact on employment: A review, synthesis and future research agenda. *Technol. Forecast. Soc. Change* **2023**, *191*, 122448. [CrossRef]
20. Petrich, L.; Lohrmann, G.; Neumann, M.; Martin, F.; Frey, A.; Stoll, A.; Schmidt, V. Detection of Colchicum autumnale in drone images, using a machine-learning approach. *Precis. Agric.* **2020**, *21*, 1291–1303. [CrossRef]
21. Frey, C.B.; Osborne, M.A. The future of employment: How susceptible are jobs to computerisation? *Technol. Forecast. Soc. Change* **2017**, *114*, 254–280. [CrossRef]
22. O*NET OnLine. Available online: https://www.onetonline.org/ (accessed on 6 February 2024).
23. Crowley, F.; Doran, J.; McCann, P. The vulnerability of European regional labour markets to job automation: The role of agglomeration externalities. *Reg. Stud.* **2021**, *55*, 1711–1723. [CrossRef]
24. David, B. Computer technology and probable job destructions in Japan: An evaluation. *J. Jpn. Int. Econ.* **2017**, *43*, 77–87. [CrossRef]
25. Haiss, P.; Mahlberg, B.; Michlits, D. Industry 4.0—The future of Austrian jobs. *Empirica* **2021**, *48*, 5–36. [CrossRef]
26. Zemtsov, S. Robots and potential technological unemployment in the Russian regions: Review and preliminary results. *Vopr. Ekon.* **2017**, *7*, 1–16. [CrossRef]
27. Blanas, S.; Gancia, G.; Lee, S.Y. Who is afraid of machines? *Econ. Policy* **2019**, *34*, 627–690. [CrossRef]
28. Borjas, G.J.; Freeman, R.B. From Immigrants to Robots: The Changing Locus of Substitutes for Workers. *RSF Russell Sage Found. J. Soc. Sci.* **2019**, *5*, 22–42. [CrossRef]
29. Jung, J.H.; Lim, D.-G. Industrial robots, employment growth, and labor cost: A simultaneous equation analysis. *Technol. Forecast. Soc. Change* **2020**, *159*, 120202. [CrossRef]
30. Marinoudi, V.; Benos, L.; Villa, C.C.; Lampridi, M.; Kateris, D.; Berruto, R.; Pearson, S.; Sørensen, C.G.; Bochtis, D. Adapting to the Agricultural Labor Market Shaped by Robotization. *Sustainability* **2024**, *16*, 7061. [CrossRef]
31. Marinoudi, V.; Benos, L.; Villa, C.C.; Kateris, D.; Berruto, R.; Pearson, S.; Sørensen, C.G.; Bochtis, D. Large language models impact on agricultural workforce dynamics: Opportunity or risk? *Smart Agric. Technol.* **2024**, *9*, 100677. [CrossRef]
32. Sufi, F. Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. *Information* **2024**, *15*, 99. [CrossRef]
33. Patil, R.; Gudivada, V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl. Sci.* **2024**, *14*, 2074. [CrossRef]
34. Gmyrek, P.; Berg, J.; Bescond, D. *Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality*; ILO Working paper 96; International Labour Office: Geneva, Switzerland, 2023.
35. Eloundou, T.; Manning, S.; Mishkin, P.; Rock, D. *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*; University of Pennsylvania: Philadelphia, PA, USA, 2023.
36. Eisfeldt, A.L.; Schubert, G.; Zhang, M.B.; Taska, B. *Generative AI and Firm Values*; National Bureau of Economic Research: Cambridge, MA, USA, 2023. [CrossRef]
37. OpenAI Platform. Models Overview. Available online: https://platform.openai.com/docs/models (accessed on 6 February 2025).

38. Bechar, A. Agricultural Robotics for Precision Agriculture Tasks: Concepts and Principles. In *Innovation in Agricultural Robotics for Precision Agriculture: A Roadmap for Integrating Robots in Precision Agriculture*; Bechar, A., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 17–30. ISBN 978-3-030-77036-5.
39. Bazargani, K.; Deemyad, T. Automation's Impact on Agriculture: Opportunities, Challenges, and Economic Effects. *Robotics* **2024**, *13*, 33. [CrossRef]
40. Vasconez, J.P.; Kantor, G.A.; Cheein, F.A.A. Human–robot interaction in agriculture: A survey and current challenges. *Biosyst. Eng.* **2019**, *179*, 35–48.
41. Oliveira, L.F.P.; Moreira, A.P.; Silva, M.F. Advances in Agriculture Robotics: A State-of-the-Art Review and Challenges Ahead. *Robotics* **2021**, *10*, 52. [CrossRef]
42. Navas, E.; Fernández, R.; Sepúlveda, D.; Armada, M.; Gonzalez-de-Santos, P. Soft Grippers for Automatic Crop Harvesting: A Review. *Sensors* **2021**, *21*, 2689. [CrossRef]
43. Microsoft Microsoft 365. Copilot. Available online: https://www.microsoft.com/en-us/microsoft-365/copilot (accessed on 6 February 2025).
44. Google DeepMind. Gemini 2.0. Available online: https://deepmind.google/technologies/gemini/ (accessed on 6 February 2025).
45. Yogarajan, V.; Dobbie, G.; Keegan, T.T. Debiasing large language models: Research opportunities. *J. R. Soc. New Zeal.* **2025**, *55*, 372–395. [CrossRef]
46. Zhang, R.; Li, H.-W.; Qian, X.-Y.; Jiang, W.-B.; Chen, H.-X. On large language models safety, security, and privacy: A survey. *J. Electron. Sci. Technol.* **2025**, *23*, 100301. [CrossRef]
47. Grieve, B.D.; Duckett, T.; Collison, M.; Boyd, L.; West, J.; Yin, H.; Arvin, F.; Pearson, S. The challenges posed by global broadacre crops in delivering smart agri-robotic solutions: A fundamental rethink is required. *Glob. Food Secur.* **2019**, *23*, 116–124. [CrossRef]
48. Ukhurebor, K.E.; Adetunji, C.O.; Olugbemi, O.T.; Nwankwo, W.; Olayinka, A.S.; Umezuruike, C.; Hefft, D.I. Chapter 6—Precision agriculture: Weather forecasting for future farming. In *AI, Edge and IoT-Based Smart Agriculture*; Abraham, A., Dash, S., Rodrigues, J.J.P.C., Acharya, B., Pani, S.K., Eds.; Intelligent Data-Centric Systems; Academic Press: Cambridge, MA, USA, 2022; pp. 101–121. ISBN 978-0-12-823694-9.
49. Benos, L.; Sørensen, C.G.; Bochtis, D. Field Deployment of Robotic Systems for Agriculture in Light of Key Safety, Labor, Ethics and Legislation Issues. *Curr. Robot. Rep.* **2022**, *3*, 49–56. [CrossRef]
50. Benos, L.; Bechar, A.; Bochtis, D. Safety and ergonomics in human-robot interactive agricultural operations. *Biosyst. Eng.* **2020**, *200*, 55–72. [CrossRef]
51. Ashqar, H.I. Benchmarking LLMs for Real-World Applications: From Numerical Metrics to Contextual and Qualitative Evaluation. *TechRxiv.* **2025**. [CrossRef]
52. da Silveira, F.; da Silva, S.L.C.; Machado, F.M.; Barbedo, J.G.A.; Amaral, F.G. Farmers' perception of the barriers that hinder the implementation of agriculture 4.0. *Agric. Syst.* **2023**, *208*, 103656. [CrossRef]
53. Khanna, M.; Atallah, S.S.; Kar, S.; Sharma, B.; Wu, L.; Yu, C.; Chowdhary, G.; Soman, C.; Guan, K. Digital transformation for a sustainable agriculture in the United States: Opportunities and challenges. *Agric. Econ.* **2022**, *53*, 924–937. [CrossRef]
54. Singh, A.K.; Lamichhane, B.; Devkota, S.; Dhakal, U.; Dhakal, C. Do Large Language Models Show Human-like Biases? Exploring Confidence—Competence Gap in AI. *Information* **2024**, *15*, 92. [CrossRef]
55. Dorner, F.E.; Nastl, V.Y.; Hardt, M. Limits to scalable evaluation at the frontier: LLM as Judge won't beat twice the data. *arXiv* **2024**, arXiv:2410.13341.
56. Malberg, S.; Poletukhin, R.; Schuster, C.M.; Groh, G. A Comprehensive Evaluation of Cognitive Biases in LLMs. *arXiv* **2024**, arXiv:2410.15413.
57. Kruger, J.; Dunning, D. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Pers. Soc. Psychol.* **1999**, *77*, 1121.
58. Dunning, D. Chapter five—The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. In *Advances in Experimental Social Psychology*; Olson, J.M., Zanna, M.P., Eds.; Academic Press: Cambridge, MA, USA, 2011; Volume 44, pp. 247–296.
59. Wei, X.; Kumar, N.; Zhang, H. Addressing bias in generative AI: Challenges and research opportunities in information management. *Inf. Manag.* **2025**, *62*, 104103. [CrossRef]
60. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
61. Benos, L.; Tsaopoulos, D.; Tagarakis, A.C.; Kateris, D.; Busato, P.; Bochtis, D. Explainable AI-Enhanced Human Activity Recognition for Human–Robot Collaboration in Agriculture. *Appl. Sci.* **2025**, *15*, 650. [CrossRef]