Developing a Routine Task Index Using Large Language Models [*]

Deokjae Jeong,[†] Tae Lee [‡§]

August 16, 2024

Abstract

This study proposes a novel method for measuring Routine Task Intensity (RTI) and task cognitivity using Large Language Models (LLMs) to analyze O*NET task descriptions. We introduce AI-decided Routine Task Intensity (AIRTI) and AI-decided Cognitive Task Intensity (AICTI) as complements to existing measurements, addressing limitations in previous approaches such as outdated data sources and insufficient capture of task repetitiveness and cognitive demands. Our method employs LLMs to assess the routineness and cognitive intensity of occupational tasks on a scale from 0 to 1, averaging these values within six-digit SOC occupational categories. To mitigate potential reliability issues, we utilize two LLMs: OpenAI's GPT-4 and Anthropic's OPUS-3.

Keywords: Routine Task Intensity, Cognitive Task Intensity, Large Language Models

JEL Codes: J21, O15

[*]We extend our heartfelt thanks to

[*]Replication data and code and the most recent version of paper: https://github.com

[†]SSK Inclusive Economic Policy Research Team, South Korea; ubuzuz@gmail.com; jayjeo.com

[‡]Gyeongsang National University, South Korea; mizzoutai@gnu.ac.kr

[§]Corresponding author

1 Introduction

Many studies explain the hollowing out of middle-skilled employment by employing the concept of Routine-Biased Technological Change (RBTC), although this is not universally agreed upon (Autor et al., 2006; Goos et al., 2014, 2009). The key point of this discussion is how to define and measure Routine Task Intensity (RTI). Walo (2023) finds that there are roughly six methods of measuring RTI until now. The most popular one is from Autor et al. (2003), utilizing the Dictionary of Occupational Titles (DOT) from 1977. However, there are other methods utilizing O*NET, BERUFENET, or PIAAC. While detailed comparisons are provided by Walo (2023), they contend that "all RTI measures have conceptual strengths and weaknesses, ... , and that some measures are better predictors of occupational change than others."

This paper proposes the seventh method of RTI measurement by utilizing O*NET task descriptions. This is a novel approach as it utilizes Large Language Models (LLMs) for the first time in the literature. We claim that this method can serve as a supplement to the other RTI measurements, as they have their own weaknesses. For instance, Autor and Dorn (2013)'s RTI uses only three variables from DOT: routine, manual, and abstract, as shown in this equation: $RTI = \ln(\text{Routine}) - \ln(\text{Manual}) - \ln(\text{Abstract})$. Meanwhile, Autor et al. (2003) uses five variables from DOT: MATH, DCP, STS, FINGDEX, and EYEHAND.[1] What is lacking here is the capture of 'repetitiveness.' Haslberger (2022) asserts that "Autor et al. (2003) and Autor and Dorn (2013) completely fail to capture key aspects of the notion of routine ... most importantly, repetitiveness." Moreover, although using DOT is better for studying the economics during the 1970s to 1990s, it is outdated for recent studies. To overcome this, Goos et al. (2009) use O*NET (2006) data. The issue here is that they use 96 variables that vary from 0 to 7. Determining which variables are more related to routineness and in what direction is merely an author's manual decision. Finally, the shortcomings of RTI measure-

[1] Respectively, GED Math; Direction, Control, Planning; Set Limits, Tolerances, or Standards; Finger Dexterity; and Eye-Hand-Foot Coordination.

ment using PIAAC's survey (Marcolin et al., 2016) are explained in detail by Walo (2023), who note its limitations in explaining job polarization.

Our measurement for RTI is based on task descriptions for each task. To avoid confusion, we will refer to our measurement of RTI as AIRTI (AI-decided Routine Task Intensity). There are several tasks in each occupation. The task descriptions are verbal sentences, and an LLM understands and can even think for itself to determine the routineness versus non-routineness of each task. After giving values from 0 (Non-routine) to 1 (Routine) for each task, the task values are averaged within each six-digit SOC occupational taxonomy.

Our novel methodology, while promising, is not without limitations. The primary concern lies in the reliability of the Large Language Model's (LLM) output. Although this issue is likely to diminish in significance given the rapid advancements in LLM technology, it is important to note that the decision-making process is entirely delegated to the LLM. A secondary concern pertains to replicability. Due to the inherent variability in LLM outputs across iterations, the generated values, while similar, are not precisely identical. This variability poses challenges for exact reproduction of results.

To address the aforementioned reliability concerns, we employ a dual-model approach, utilizing two state-of-the-art Large Language Models (LLMs): OpenAI's GPT-4 and Anthropic's Opus-3. While other advanced models exist, such as Anthropic's Sonnet-3.5, Meta's Llama-3, and Google's Gemini, current consensus in the field suggests that Claude 3 Opus and GPT-4 represent the pinnacle of LLM capabilities. This approach leverages the strengths of both models to enhance the robustness of our methodology.

2 Definitions and Commands

2.1 Definitions

In the literature, universal definitions for *routineness* and *cognitiveness* do not exist. Each study uses distinct definitions. Therefore, we define these terms by referring to many existing studies. The definitions that we incorporated into the Python code are as follows:

A routine task involves activities that are predictable and can be automated, such as those performed by industrial robots on assembly lines or through computerization. This typically involves substituting human labor for routine information processing or repetitive tasks. **A non-routine task** requires handling unpredictable situations or resolving exceptions that automated systems and programs cannot adequately address. Examples include caregiving, creative writing, or artistic activities that demand human intuition and creativity.

A cognitive task involves activities that require mental processes, skills, and abilities. These include perception, thinking, reasoning, memory, learning, decision making, and other aspects of information processing. Examples of cognitive tasks are problem-solving, language comprehension, attention, and pattern recognition. **A manual task** involves physical processes, activities, and skills. These require the use of hands, the body, and sensory-motor coordination, including dexterity, precision, physical effort, and manipulation of tools or objects. Examples of manual tasks include handwriting, using tools, playing an instrument, and assembly work.

2.2 Commands

The commands below are the actual instructions that we used in the Python code to instruct the LLMs.

2.2.1 Routiness

Extremely Non-Routine Task (Score: 0): Assign a value of 0 exclusively to tasks that necessitate human creativity, intuition, or involve complex, unpredictable problem solving that cannot be replicated by current automation technologies at all. **Extremely Routine Task (Score: 1):** Assign a value of 1 only to tasks that are fully automatable with absolutely no need for human discretion or unpredictable judgment. This should be strictly limited to tasks where current technology can perform the task without any human oversight. **Moderately Routine or Non-Routine Tasks (Score range: 0.3 to 0.7):** Assign values within this range to tasks that blend elements of both routine and non-routine characteristics, or when the classification into extreme categories is not clear. The middle range should be expanded slightly to encourage less extreme scoring, using 0.5 as a central point for truly ambiguous tasks. **Uncertainty Principle:** If there is any uncertainty in classifying the task, default to a score closer to 0.5. Use the wider range of 0.3 to 0.7 to adjust the score slightly if there is a mild inclination towards routine or non-routine characteristics. This approach should ensure that only tasks with clear and definitive characteristics receive scores at the extremes. **Output Format:** Begin your response with the score, followed by a colon and a detailed explanation of your reasoning. The explanation should thoroughly consider the task's characteristics, including the potential for automation and the level of human input or creativity required.

2.2.2 Cognitivity

Extremely Manual Task (Score: 0): Assign a value of 0 only to tasks that are fully manual task. **Extremely Cognitive Task (Score: 1):** Assign a value of 1 exclusively to tasks are full cognitive task. **Moderately Manual or Cognitive Tasks (Score range: 0.3 to 0.7):** Assign values within this range to tasks that blend elements of both manual and cognitive characteristics, or when the classification into extreme categories is not clear. The middle range should be expanded slightly to encourage

less extreme scoring, using 0.5 as a central point for truly ambiguous tasks. **Uncertainty Principle:** If there is any uncertainty in classifying the task, default to a score closer to 0.5. Use the wider range of 0.3 to 0.7 to adjust the score slightly if there is a mild inclination towards manual or cognitive characteristics. This approach should ensure that only tasks with clear and definitive characteristics receive scores at the extremes. **Output Format:** Begin your response with the score, followed by a colon and a detailed explanation of your reasoning. The explanation should thoroughly consider the task's characteristics.

3 Scoring Results

We provide a full set of results at the link provided in the footnote.[2] This file includes task code, task description, routine score, cognitive score, and their corresponding reasonings. The dataset contains 20,508 observations (tasks). Using the task code, researchers can merge this information with any O*NET data. Examples are shown as follows, but we provide only two examples to conserve space. While we employed two LLMs (GPT-4 and OPUS-3) for measuring routineness, we used only GPT-4 for cognitivity to save time and money. We will include the OPUS-3 version of the cognitivity measure as we proceed with this paper.

3.1 Example 1

Task description: Resolve customer complaints regarding sales and service. **OPUS Routine score:** 0.4 **OPUS Routine reasoning:** Resolving customer complaints regarding sales and service involves a blend of routine and non-routine elements. While some aspects of the task, such as documenting complaints or following standardized resolution procedures, can be considered routine and potentially automatable, the core of the task requires human interaction, empathy, and problem-solving skills. Customer complaints often involve unique situations that

[2] https://tinyurl.com/yc8nu4t7

demand active listening, understanding the customer's perspective, and finding creative solutions to address their concerns. Each complaint may present a different challenge that requires adapting to the specific context and using human judgment to determine the most appropriate course of action. However, the task is not entirely non-routine, as there are likely established guidelines. **GPT Routine score:** 0.3 **GPT Routine reasoning:** Resolving customer complaints regarding sales and service is predominantly a non-routine task. While certain aspects of customer service can be automated, such as providing information about services or products, handling complaints often involves unpredictable elements and requires human intuition and creativity. Each customer complaint is unique and may require a different approach or solution, which is challenging to automate. The task involves understanding the customer's emotions, empathizing with them, and devising solutions to their problems, which are skills that are currently beyond the capabilities of automation technologies. However, since some aspects of the task, such as gathering initial information about the complaint, can be automated, it does not fully qualify as an extremely non-routine task. **GPT Cognitive score:** 0.8 **GPT Cognitive reasoning:** The task of resolving customer complaints regarding sales and service is primarily a cognitive task. It involves understanding the customer's problem (language comprehension), thinking of the best way to resolve the issue (problem-solving, decision making), and communicating that solution effectively (language comprehension again). It may also involve recalling relevant information about the company's policies or the customer's past transactions (memory). However, there might be some manual aspects involved such as typing responses or navigating through the customer service software, which is why the task doesn't score a full 1. The cognitive aspects of this task clearly outweigh the manual ones, hence the score leans more towards the cognitive end of the scale.

3.2 Example 2

Task description: Load or unload materials into containers or onto trucks, using hoists or forklifts. **OPUS Routine score:** 0.8 **OPUS Rou-**

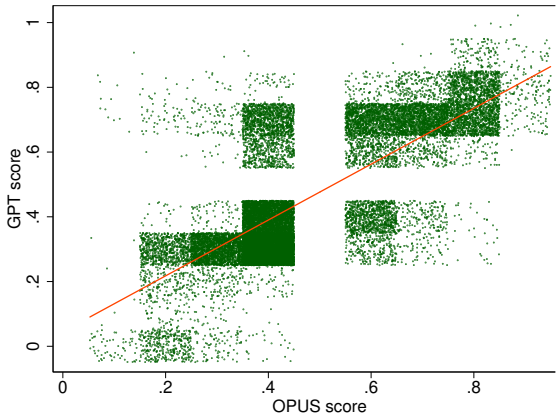
tine reasoning: Loading or unloading materials into containers or onto trucks using hoists or forklifts is a highly routine task that can be largely automated with current technology. The process involves repetitive movements and predictable steps, making it suitable for automation. Industrial robots and automated systems are already widely used in warehouses and loading docks to perform these tasks efficiently. However, I assign a score of 0.8 instead of 1 because there may still be some need for human oversight and intervention in certain situations. For example, if the materials being loaded are fragile or have irregular shapes, a human operator might need to make minor adjustments or use discretion to ensure proper handling. **GPT Routine score:** 0.7 **GPT Routine reasoning:** The task of loading or unloading materials into containers or onto trucks, using hoists or forklifts, is largely routine. It involves predictable, repetitive physical labor that can be automated to a significant extent. For instance, automated forklifts and conveyor belt systems exist that can perform these tasks with minimal human oversight. However, the score is not 1 because there may still be a need for human intervention in certain situations. For instance, if a container is not aligned correctly or if an item is too large or awkwardly shaped for the automated system to handle, a human may need to step in to resolve the issue. Therefore, while this task is primarily routine, there is still a minor non-routiness. **GPT Cognitive score:** 0.3 **GPT Cognitive reasoning:** The task of loading or unloading materials into containers or onto trucks, using hoists or forklifts, is primarily a manual task. It involves the use of hands and body, sensory-motor coordination, and possibly manipulation of tools or objects such as hoists or forklifts. However, there is a slight cognitive aspect to it as well, as the operator would need to make decisions about where to place materials, how to operate the machinery, and how to do so safely. Therefore, while this task is predominantly manual, it does incorporate some cognitive elements, hence the score of 0.3.

3.3 Scatter Scores

Since we have used both GPT-4 and OPUS-3, we can provide a scatter plot as shown below. Panel (a) of Figure 1 shows the original results,

Figure 1: Routine Scores

(a) Original



(b) Trimmed

which include many inconsistent values between the GPT-4 and OPUS-3 versions. We decided to delete these tasks, as shown in Panel (b) of the same figure. We provide an example of these inconsistencies to explain why the two LLMs have different opinions.

3.4 Inconsistent Example

Task description: Mark up, paste, and complete layouts and write typography instructions to prepare materials for typesetting or printing. **OPUS Routine score:** 0.4 **OPUS Routine reasoning:** The task of marking up, pasting, and completing layouts, as well as writing typography instructions for typesetting or printing, involves a combination of routine and non-routine elements. Routine aspects: Some parts of the layout process may be predictable and follow established templates or guidelines. Certain typography instructions might be standardized based on the type of material being prepared. Non-routine aspects: The specific content and design requirements for each layout may vary, requiring human judgment and creativity to ensure the final product meets the desired aesthetic and communicative goals. Writing typography instructions may involve considering the unique characteristics of the material and making decisions based on factors. **GPT Routine score:** 0.7 **GPT**

Routine reasoning: This task involves a mixture of both routine and non-routine characteristics. While aspects of the task, such as marking up and pasting layouts, can potentially be automated, the need for writing typography instructions suggests a level of human discretion and creativity. The judgment involved in deciding the appropriate typography for a given material is not easily replicable by current automation technologies, suggesting a non-routine characteristic. However, given that the bulk of the task appears to be routine in nature, with only a portion requiring human input, a score slightly leaning towards routine (0.7) is assigned. The score is not at the extreme end of the routine scale due to the presence of some non-routine elements.

4 Routineness versus Cognitivity

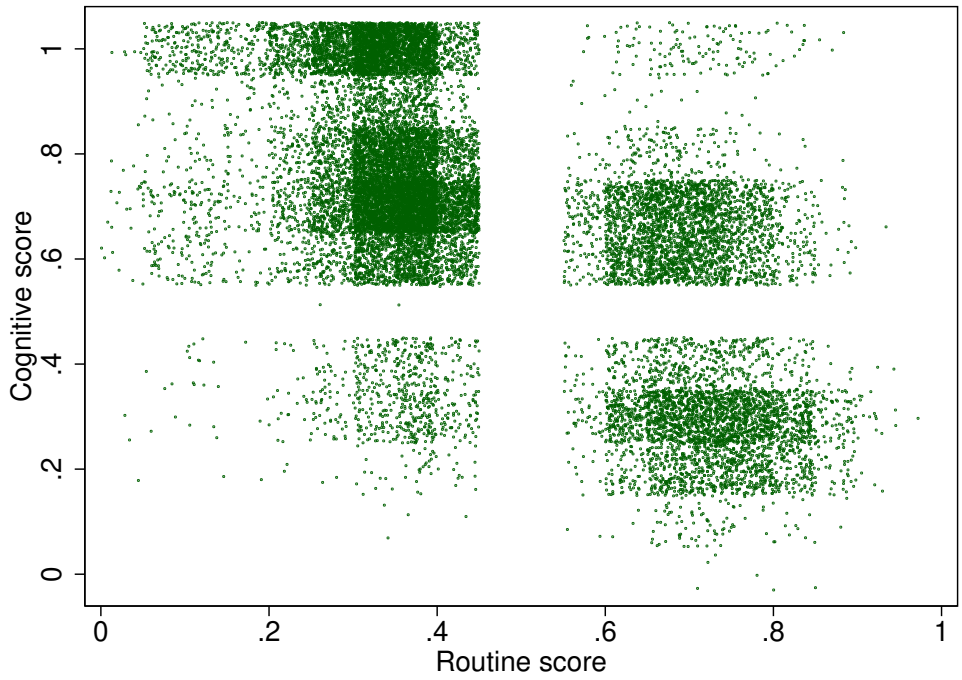
This section examines the levels of routineness and cognitiveness in occupations. As shown in Panel (a) of Figure 2, occupations at the six-digit SOC level can be categorized into four dimensions, such as routine-cognitive and non-routine-manual. It is important to note that each occupation comprises several tasks —typically ranging from ten to twelve, for instance. When calculating the routineness and cognitiveness values, tasks are weighted differently: those labeled as 'core' by O*NET are given a weight of 5, while those labeled as 'supplementary' are assigned a weight of 1.

Panel (b) of the same figure collapses the occupations into five SOC2010 categories following Autor and Dorn (2013). As one might expect, 'Management and Professional' occupations are highly cognitive and non-routine, while 'Production' and 'Construction' occupations are highly manual and routine. It is notable that 'Service' occupations are not classified as either non-routine or manual. The common narrative suggests that the hollowing out of middle-skilled occupations is leading to a shift towards non-routine service occupations. However, our RTI method indicates otherwise.

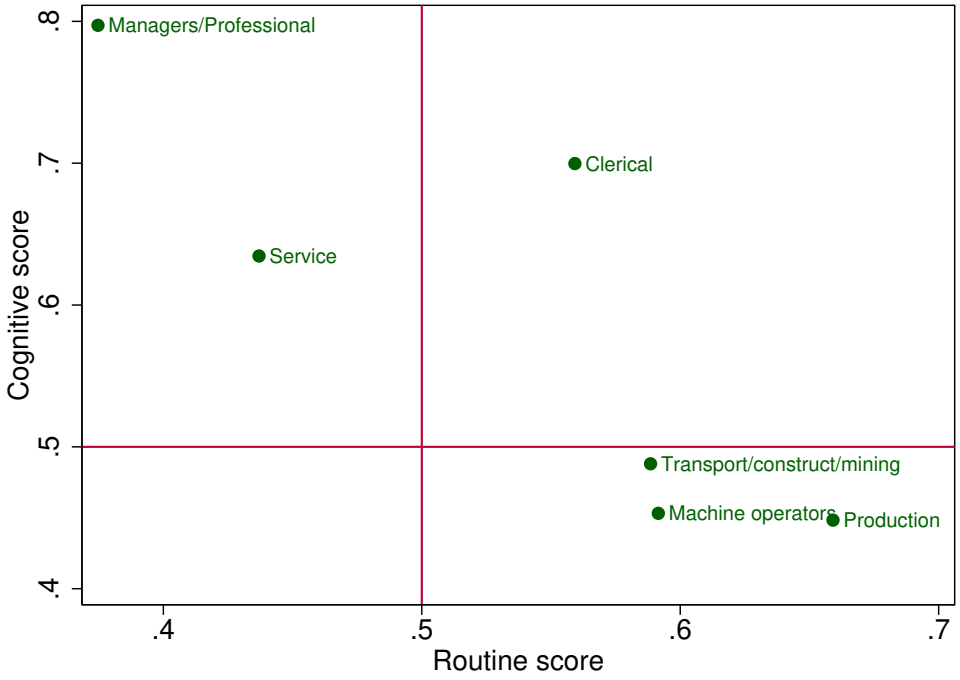
Remember that Autor and Dorn (2013)'s RTI measure is calculated by using three variables as shown in the equation below. This approach

Figure 2: Cognitive Score vs Routine Score

(a) SOC 6-digit Level



(b) Autor and Dorn (2013)'s Category using our RTI measure



to quantifying routine task intensity provides a comprehensive view of occupational characteristics. However, when comparing our results to theirs in Figure 3, a notable discrepancy emerges. Autor et al.'s service category aligns more closely with non-routine occupations, whereas our analysis shows no such correlation. This divergence underscores Walo (2023)'s contention that all RTI measurements in the literature have inherent limitations, despite their methodological sophistication.

$$RTI = \ln(\text{Routine}) - \ln(\text{Manual}) - \ln(\text{Abstract}) \tag{1}$$

Accordingly, we conduct a simple regression exercise using these three variables as follows.

$$AIRTI = \alpha_0 + \alpha_1 \ln(\text{Routine}) + \alpha_2 \ln(\text{Manual}) + \alpha_3 \ln(\text{Abstract}) + \varepsilon$$

We designate our measure of RTI as AIRTI (AI-decided RTI) to distinguish it from previous approaches. The regression results are presented in Table 1, with the corresponding fitted line illustrated in Figure 4. A notable aspect of this result is that the 'Manual' variable appears to be irrelevant to our AIRTI measure. It is important to emphasize that this discrepancy does not imply that either approach is definitively correct or incorrect.

Meanwhile, a notable finding from our analysis is that the signs of coefficients are all consistent with those used by Autor and Dorn (2013) in Equation (1). This consistency suggests that despite the differences in our methodologies, there is a fundamental alignment in how various factors contribute to the measure of routine task intensity. This alignment lends credibility to both approaches while highlighting the robustness of these relationships across different analytical frameworks.

5 Longitudinal Analysis

In this section, we briefly introduce the time series patterns based on our routine and cognitive measures. This analysis provides insights into how these key components of occupational characteristics have evolved over

Figure 3: Routine Score Comparison

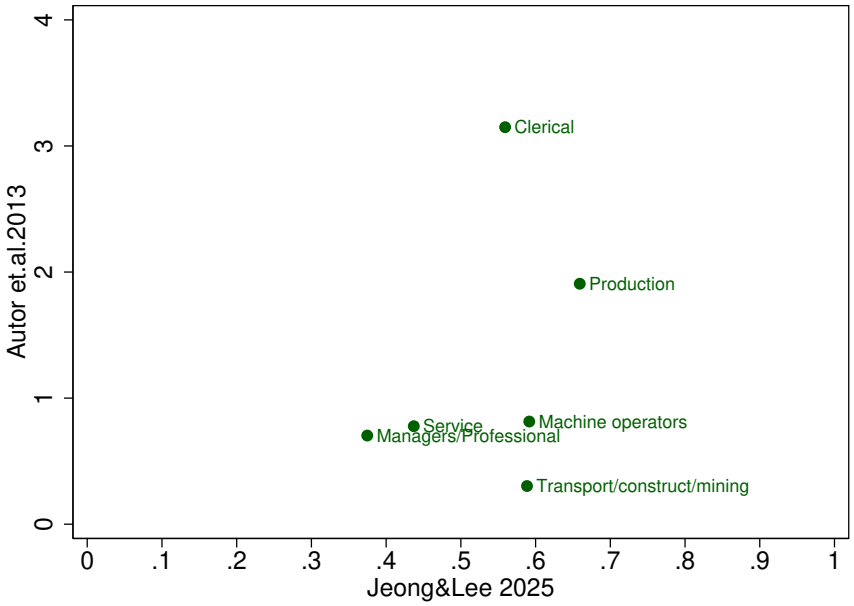


Figure 4: Regression Fitted Line

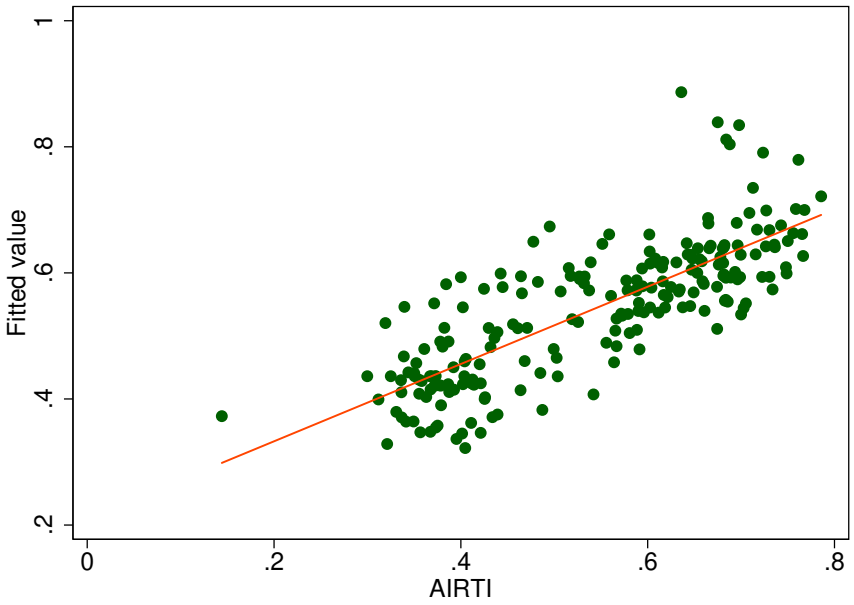


Table 1: Regressions

	Y=AIRTI
ln(routine)	0.079***
	(0.010)
ln(manual)	-0.004
	(0.004)
ln(abstract)	-0.104***
	(0.006)
N	213
R^2	0.613

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

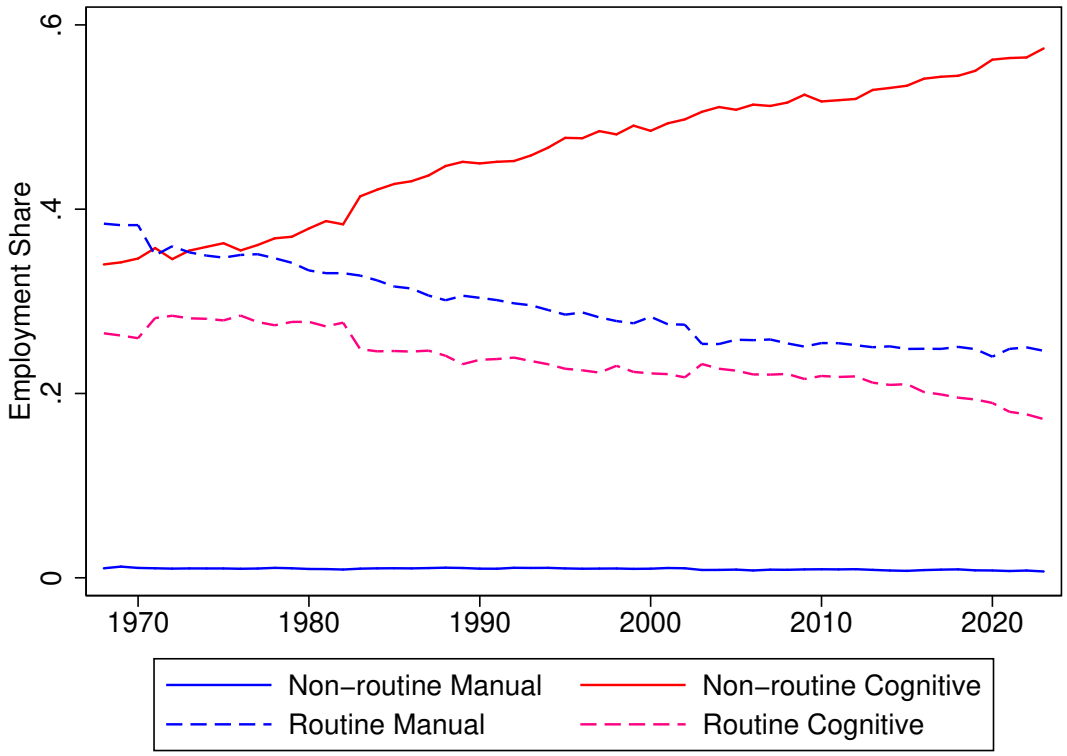
time. After this brief exposition, we will move on to the next section, where we analyze the findings of Goos et al. (2009) using our AIRTI (AI-decided Routine Task Intensity) measure. This upcoming analysis will provide a valuable comparison between our approach and established literature in the field.

The most critical aspect of our longitudinal analysis is the examination of employment shares across four categories: routine-cognitive, routine-manual, non-routine cognitive, and non-routine manual. Figure 5 illustrates these trends over time, with the blue line representing 'Manual' (i.e., non-cognitive) occupations and the dotted line indicating 'Routine' occupations. A well-known pattern emerges from this visualization. Regardless of the cognitive dimension, there is a clear and consistent decline in the share of routine occupations. The persistent nature of this decline in routine job shares implies a fundamental transformation in the nature of work, likely driven by technological advancements and changing economic structures. This decline in routine occupations, both cognitive and manual, may align with theories of RBTC and the increasing automation of routine tasks.

While we may attribute the decline in routine occupation shares dur-

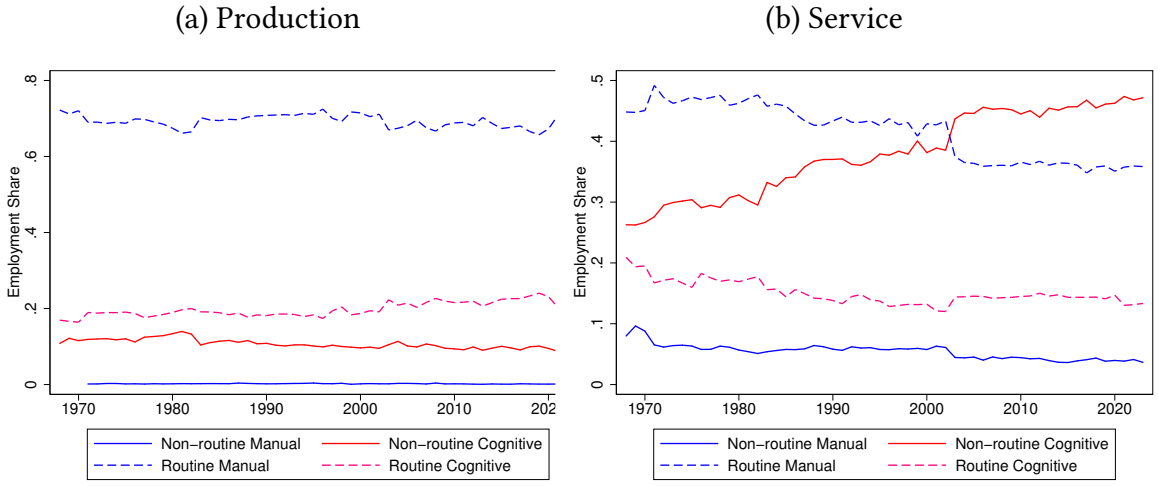
ing the 1980s to computerization, the persistence of this trend beyond 2000 presents a more complex puzzle. The continued decrease in routine job shares well into the 21st century challenges simple explanations based solely on the initial wave of computer adoption in the workplace. This persistent decline raises important questions about the driving forces behind labor market transformations in the digital age. Are we witnessing the effects of more advanced technologies, such as artificial intelligence and machine learning? Or are there other structural changes in the economy contributing to this ongoing shift? The lack of a clear explanation for this post-2000 trend highlights the need for further research into the evolving nature of work and the factors influencing occupational distributions in the modern economy.

Figure 5: Employment Share



To gain further insight into this phenomenon, we examine two additional figures (Figure 6 Panel (a) and (b)) that are identical to the previ-

Figure 6: Employment Share



ous one, but focus separately on 'Service' occupations and 'Production' occupations. These detailed visualizations reveal a crucial distinction: while other occupational categories, including 'Production', show relatively stable patterns, it is the 'Service' occupation category that exhibits the most notable change. This observation aligns with the emphasis placed by Autor and Dorn (2013) on the role of service occupations in shaping labor market trends. The unique trajectory of service occupations implies that this sector may be key to understanding the persistent decline in routine job shares beyond 2000.

The proportion of employment in non-routine cognitive tasks within the 'Service' occupation category has exhibited a steady increase from 1970 to the present. This trend aligns with the within-occupation changes highlighted by Fernández-Macías et al. (2023). A notable strength of our study is the application of the AIRTI scoring system at the six-digit level of Standard Occupational Classification (SOC) codes. This granular approach enables researchers to conduct detailed analyses of compositional changes within broader occupational categories.

6 Some Exercises Using Goos et al. (2009)'s Work

O*NET provides 165 variables, each scored from 0 to 7 based on survey responses. These variables cover a wide range of abilities, skills, and knowledge areas, such as Ability-Oral Expression, Ability-Reaction Time, Skills-Mathematics, Skills-Negotiation, Knowledge-Mathematics, and Knowledge-Building and Construction. Goos et al. (2009) manually selected 96 of these variables that they deemed relevant to routine intensity. However, we are skeptical that this selection may be arbitrary, and we are also uncertain about the appropriate weights and signs to assign to these variables.

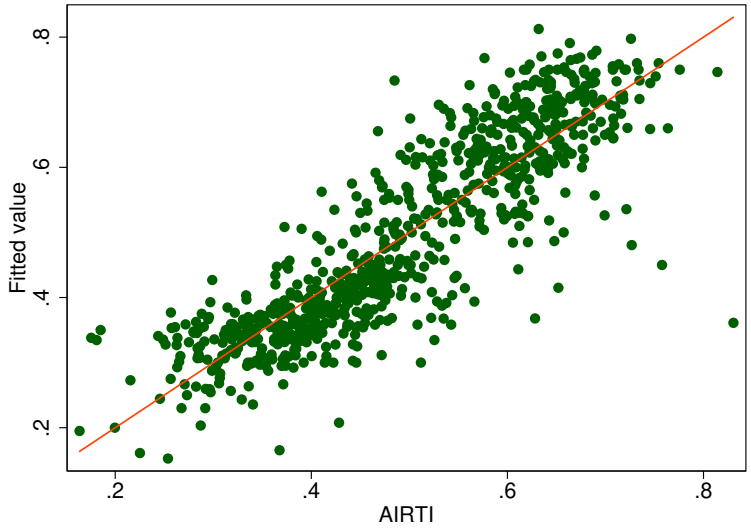
To address these concerns, we conducted a simple exercise assuming our AIRTI (AI-decided Routine Task Intensity) is the true value. Our approach involved the following steps: We performed an initial regression using AIRTI as the dependent variable and all 165 O*NET variables as explanatory variables. We then selected only those explanatory variables with p-values smaller than 0.05. Finally, we conducted a second regression using these remaining selected variables. The results of this analysis are presented in Table 2. The abbreviations used in this table are as follows: AB: Ability, KN: Knowledge, SK: Skill, and WA: Work Activities. Figure 7 presents the fitted regression line.

This straightforward analysis yields significant insights: Goos et al. (2009) omit the presentation of their variable selection process and calculation methods, even in appendices or replication files. Consequently, we cannot ascertain how our simple regression selections diverge from their choices. Nonetheless, these findings underscore the challenges inherent in constructing a RTI measure using the 165 O*NET variables in a manner that is both appropriate and methodologically sound.

Table 2: Selected Variables

	Dependent = AIRTI	
	Coefficient	Standard Errors
ABFluency of Ideas	-0.040***	(0.014)
ABMultilimb Coordination	-0.004	(0.005)
ABOral Expression	-0.056***	(0.008)
ABOriginality	-0.033**	(0.014)
ABPerceptual Speed	0.047***	(0.007)
ABProblem Sensitivity	-0.053***	(0.009)
ABSelective Attention	0.033***	(0.011)
ABTime Sharing	0.004	(0.009)
KNFine Arts	-0.029***	(0.005)
KNLaw and Government	-0.019***	(0.004)
KNPersonnel and Human Resources	-0.009*	(0.005)
KNPublic Safety and Security	0.005	(0.005)
KNSales and Marketing	0.015***	(0.003)
SKEquipment Selection	-0.017***	(0.005)
SKInstallation	0.028***	(0.004)
SKLearning Strategies	0.015**	(0.006)
SKSystems Analysis	-0.009	(0.007)
SKSystems Evaluation	0.002	(0.007)
WACoaching and Developing Others	-0.017**	(0.007)
WACommunicating with Supervisors	0.028***	(0.006)
WADocumenting/Recording Information	-0.017***	(0.004)
WAStaffing Organizational Units	0.007	(0.005)
WAThinking Creatively	-0.014**	(0.006)
WATraining and Teaching Others	0.018***	(0.006)
Constant	0.848***	(0.021)
Observations	798	
R^2	0.770	

Figure 7: Regression Line



7 Some Exercises Using Cortes et al. (2014)'s Work

Cortes et al. (2014) employed a classification system based on the Standard Occupational Classification (SOC) taxonomy to categorize occupations into four distinct groups. Their categorization is summarized in the following table:

Table 3: Cortes et al. (2014)'s Classification

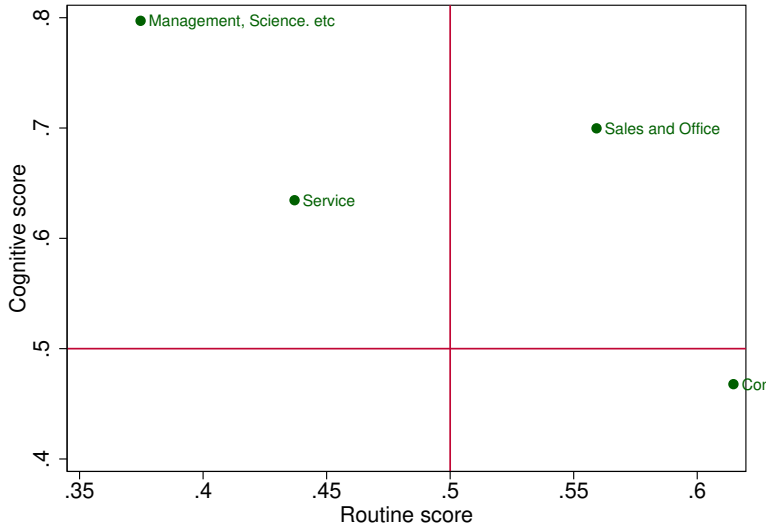
	Routine	Non-routine
Cognitive	Sales and Office	Management, Science. etc
Manual	Construction;Maintenance;Production	Service

Our findings, as presented in Panel (B) of Figure 8, offer a categorization that differs from previous research. The detailed breakdown of these categories is illustrated in Figure 8. A notable distinction between our classification and that of Cortes et al. (2014) is the categorization of 'Service' tasks. In our analysis, 'Service' is not classified under manual tasks, contrary to their approach.

Table 4: Our AIRTI

	Routine	Non-routine
Cognitive	Sales and Office	Management, Science. etc; Service
Manual	Construction;Maintenance;Production	

Figure 8: Our AIRTI categories



8 Conclusion

This paper introduces a novel approach to measuring Routine Task Intensity (RTI) using Large Language Models (LLMs) to analyze O*NET task descriptions. This method, which we term AI-decided Routine Task Intensity (AIRTI), offers several key contributions to the field. By leveraging LLMs to assess task routineness and cognitivity, we provide a new perspective on RTI measurement that complements existing approaches. Our method aims to overcome some of the shortcomings of previous RTI measures, such as the lack of capturing 'repetitiveness' in Autor et al. (2003) and Autor and Dorn (2013)'s approaches and the potential arbitrariness in variable selection when using O*NET data that appears in Goos et al. (2009).

We conducted a thorough comparison of our AIRTI measure with

established RTI metrics, revealing both consistencies and discrepancies that merit further investigation. Our analysis of employment share trends across routine and non-routine, cognitive and manual occupations reveals persistent declines in routine job shares extending beyond 2000, challenging simple explanations based solely on initial computerization. Notably, we identified the service occupation category as a key driver of changing labor market trends, aligning with and extending previous research by Autor and Dorn (2013).

Our exercise using Goos et al. (2009)'s approach highlighted the challenges in constructing RTI measures from O*NET variables, demonstrating the potential benefits of our LLM-based method. While our AIRTI measure offers new insights, it also has limitations, particularly regarding the reliability and replicability of LLM outputs. However, as LLM technology rapidly advances, these concerns may diminish over time. This research opens several avenues for future work. These include further validation and refinement of the AIRTI measure, exploration of the persistent decline in routine job shares post-2000, in-depth analysis of the unique trajectory of service occupations, and investigation of the discrepancies between AIRTI and traditional RTI measures.

In conclusion, our AI-driven approach to RTI measurement provides a valuable complement to existing methods, offering new perspectives on the changing nature of work in the face of technological advancement. As we continue to grapple with the complexities of labor market transformations, tools like AIRTI may prove instrumental in deepening our understanding and informing policy decisions.

References

Autor, D. H. and D. Dorn (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American economic review 103*(5), 1553–1597.

Autor, D. H., L. F. Katz, and M. S. Kearney (2006, April). The Polarization of the U.S. Labor Market. *American Economic Review 96*(2), 189–194.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics 118*(4), 1279–1333.

Cortes, G. M., N. Jaimovich, C. J. Nekarda, and H. E. Siu (2014). The micro and macro of disappearing routine jobs: A flows approach. Technical report, National Bureau of Economic Research.

Fernández-Macías, E., M. Bisello, E. Peruffo, and R. Rinaldi (2023). Routinization of work processes, de-routinization of job structures. *Socio-Economic Review 21*(3), 1773–1794.

Goos, M., A. Manning, and A. Salomons (2009, April). Job Polarization in Europe. *American Economic Review 99*(2), 58–63.

Goos, M., A. Manning, and A. Salomons (2014). Explaining job polarization: Routine-biased technological change and offshoring. *American economic review 104*(8), 2509–2526.

Haslberger, M. (2022). Rethinking the measurement of occupational task content. *The Economic and Labour Relations Review 33*(1), 178–199.

Marcolin, L., S. Miroudot, and M. Squicciarini (2016). The routine content of occupations: New cross-country measures based on PIAAC. *OECD Trade Policy Papers, No. 188, OECD Publishing, Paris.*

Walo, S. (2023, September). The link between routine tasks and job polarization: A task measurement problem? *LABOUR 37*(3), 437–467.

A Appendix: