

Beyond Automation: Redesigning Jobs with LLMs to Enhance Productivity*

Andrew Ledingham, Michael Hollins, Matthew Lyon, David Gillespie, Umar Yunis-Guerra, Jamie Siviter, David Duncan, and Oliver P. Hauser

Abstract

The adoption of generative artificial intelligence (AI) is predicted to lead to fundamental shifts in the labour market, resulting in displacement or augmentation of AI-exposed roles. To investigate the impact of AI across a large organisation, we assessed AI exposure at the task level within roles at the UK Civil Service (UKCS). Using a novel dataset of UKCS job adverts, covering 193,497 vacancies over 6 years, our large language model (LLM)-driven analysis estimated AI exposure scores of 1,542,411 tasks. By aggregating AI exposure scores for tasks within each role, we calculated the mean and variance of job-level exposure to AI, highlighting the heterogeneous impacts of AI, even for seemingly identical jobs. We then use an LLM to redesign jobs, focusing on task automation, task optimisation, and task reallocation. We find that the redesign process leads to tasks where humans have comparative advantage over AI, including strategic leadership, complex problem resolution, and stakeholder management. Overall, automation and augmentation are expected to have nuanced effects across all levels of the organisational hierarchy. Most economic value of AI is expected to arise from productivity gains rather than role displacement. We contribute to the automation, augmentation and productivity debates as well as advance our understanding of job redesign in the age of AI.

Keywords: artificial intelligence; technological change; labour productivity; human capital; task automation; job redesign

JEL codes: J24, O33, D24, J21

*Ledingham, Hollins, Lyon, Gillespie, Yunis-Guerra, Siviter and Duncan: Government Digital Service, UK Department of Science, Innovation and Technology. Hauser: Department of Economics, University of Exeter; Institute for Data Science and Artificial Intelligence, University of Exeter; and Evaluation Task Force, UK Cabinet Office. We would also like to thank the following people for their constructive feedback and review: Gemma Alderton, Anil Doshi, Aakash Paul, Alexandra Pop and Piers Walker. O.P.H. is grateful for financial support from the University of Exeter and the UKRI Future Leaders Fellowship (project reference: 1067). Correspondence: matthew.lyon@dsit.gov.uk and o.hauser@exeter.ac.uk.

The adoption of generative artificial intelligence (AI) is estimated to lead to fundamental shifts in the labour market ([Brynjolfsson and Mitchell, 2017](#); [Acemoglu et al., 2022](#); [Eloundou et al., 2024](#)). The impact of generative AI and associated technologies—which we collectively refer to as AI—hinges on the exposure of jobs and tasks to automation or augmentation using these technologies. Past studies have used occupational, industrial and geographical variation to estimate the impact of AI exposure, typically using the Occupational Information Network (O*NET) database to estimate AI exposure ([Felten et al., 2021](#); [Eloundou et al., 2024](#)). By some estimates, up to half of all jobs may have the majority of their tasks ‘exposed’ to the impact of generative AI in the near term ([Eloundou et al., 2024](#)), with the largest AI exposure among high-wage occupations ([Felten et al., 2023](#)). An often implied suggestion is that exposed jobs will experience partial or full automation and lead to job replacement through AI ([Frank et al., 2019](#)).

However, the assumption that large swathes of jobs will simply be automated and displaced likely oversimplifies a complicated question.¹ Jobs are neither monolithic nor unadaptive to economic pressures ([Svanberg et al., 2024](#)). We build on past work that treats jobs as a bundle of discrete tasks ([Felten et al., 2021](#); [Gmyrek et al., 2023](#); [Eloundou et al., 2024](#); [Klein Teeselink, 2025](#); [Hampole et al., 2025](#)). As a result, we measure the impact of AI at the task level, and do so by extracting tasks from real job vacancies from a large public sector employer, allowing for a more granular impact assessment across similar jobs. It also shifts the nature of the impact of AI from job displacement to specific task automation — and the opportunity for productivity gains through reinvesting newfound time into other activities ([Acemoglu and Restrepo, 2018](#)). As [Agrawal et al. \(2023\)](#) argue, “automation of some tasks can lead to augmentation of labour

¹There is an ongoing debate about whether and which jobs may have already experienced displacement from AI. [Brynjolfsson et al. \(2025a\)](#); [Klein Teeselink \(2025\)](#); [Demirci et al. \(2025\)](#); [Lichtinger and Hosseini Maasoum \(2025\)](#) and [Dominski and Lee \(2025\)](#) present evidence of displacement in some contexts and countries, while [Chandar \(2025\)](#) and [Humlum and Vestergaard \(2025\)](#) find no evidence in others. [Svanberg et al. \(2024\)](#) argue that many jobs will simply not be economic attractive or technically feasible to automate in the near future.

elsewhere” (p. 155). Similarly, [Brynjolfsson and Mitchell \(2017\)](#) have suggested that complementarities and redesigned processes are key to productivity gains. To respond to these calls in the literature, we seek to answer the following research questions: What are the heterogeneous impacts of AI at the task level on similar but unique jobs? When AI can automate some tasks but not the entire job, how should those jobs be redesigned to make productive use of the freed-up time?

In this paper, we demonstrate how AI exposure at the task level varies across the diversity of similar jobs in a single industry—the UK public sector—and then focus on how the freed-up time from AI automation can be used to lead to productivity gains rather than purely to job displacement.² Drawing on a unique dataset from UK government job advertisements, we cover a total of 193,497 vacancies — nearly all UKCS job ads over a six-year period. Using a bottom-up approach, we extract tasks for each vacancies and use a Large Language Model (LLM) to calculate the AI exposure for each task. This results in a dataset of 1,542,411 tasks and their corresponding exposure scores. Aggregating task exposure scores to the job level, we introduce a threshold above which jobs are considered fully automatable, while jobs below the threshold may only contain some tasks that can be automated. In the latter case, we then apply another LLM to redesign the job with a focus on how to reinvest the freed-up time to maximise the role’s productivity. Therefore, while previous empirical papers have focused on AI exposure alone ([Felten et al., 2021](#); [Gmyrek et al., 2023](#); [Eloundou et al., 2024](#)), we expand this analysis to study what implications partial AI automation has for the future of jobs and tasks.

We present three key findings. First, we document a diversity of AI exposure across UKCS jobs. Most roles (>60%) are predicted to have medium exposure scores but there is a relatively high variance across similarly exposed jobs in the extent to which AI will affect the range of tasks. These medium-exposed jobs form the main focus of our redesign, revealing opportunities

²Technological shocks can also create entirely new jobs or shift the concentration and growth of jobs in other industries ([Autor et al., 2015](#); [Acemoglu and Restrepo, 2020](#)), which is not the focus of our paper.

for productivity gains by automating some tasks but deepening and expanding remaining tasks in a job role. By contrast, 20% of all job roles have low exposure to AI with limited potential for automation or augmentation. Notably, these roles are heavily concentrated among senior roles, with nearly half (47%) of all senior civil service (SCS) roles falling into this category. Finally, jobs that are highly exposed to AI and therefore could be fully automated amount to approximately 18% of UKCS jobs. Some of the most affected tasks in this category include administrative support and records management, which are mostly found in lower grade roles, but it also includes data analysis and policy development, which are primarily concentrated among higher grade roles.

These findings illustrate that generative AI has the potential for both efficiency gains—by cutting labour costs of fully automatable jobs—and productivity gains—by redesigning the freed up time of tasks that are automated or augmented by AI—and that the distribution of automation and augmentation will have nuanced effects across the organizational hierarchy.

Second, after applying our methodology to remove automatable tasks and redesign jobs, our results show that a quarter of all new tasks (26 percent) are focused on strategic leadership, followed by complex problem resolution (18 percent), and stakeholder management and communication (17 percent). Notably, strategic leadership and complex problem resolution are heavily concentrated among higher grade job roles, while stakeholder management and communication are frequent among lower grade roles. By contrast, we observe notable reductions in tasks that involve administrative support, records management, and data analysis. Some results demonstrate the importance of context: one of the least exposed tasks is prison management, a physically involved task, and its importance (and therefore time allocation) increases after the job redesign. We conduct robustness checks and vary the process of new task design, showing that most new tasks are centered on those in which human labour has a comparative advantage (Acemoglu and Restrepo, 2018; Loaiza and Rigobon, 2024).

Third, to understand the implications for businesses, we identify conditions under which productivity gains of augmentation outweigh efficiency gains of full automation and job losses. To do this, we examine specific scenarios by varying the automation threshold θ . In our preferred specification, we choose a conservative threshold of $\theta \geq 80$. Under this assumption, 145,864—approximately 75% of all jobs in our dataset—qualify to some degree for redesign. This has direct implications for costs reductions and productivity gains: if only jobs with a higher proportion of time spent on high exposure tasks are considered fully automable, this means that most jobs cannot be displaced by AI and so direct cost savings will be low, with the majority of jobs instead being redesigned. By combining redesigned jobs with salary bands, we derive estimates for the total potential benefits of applying AI for UKCS jobs. As such, we identify potential productivity gains of £5.2bn in roles where AI is more likely to shift the task composition towards higher value work and cost reductions of £1.1bn in roles dominated by high exposure tasks. More generally, adopting a “job redesign” perspective that does not treat all exposed jobs as fully automatable but instead focuses on maximizing productivity may become a dominant strategy for many firms implementing AI, automating only roles above the critical threshold θ . Overall, we find that most economic value is expected to arise from productivity gains rather than role displacement.³

Our contributions connect to three literatures. First, our new methodological framework provides a conceptual step forward in the job automation and augmentation literature (Felten et al., 2021). Our study proposes a new framework that goes beyond broad descriptions of potential AI exposure for jobs and tasks and instead optimises job redesign according to increased productivity from AI. Starting with existing UKCS job descriptions in vacant positions, we apply our LLM-based algorithm to calculate AI exposure at the task level for every job before prescriptively designing tasks. For job redesign, AI is assumed to be deployed in tasks where

³This aligns with Svanberg et al. (2024) who argue that it may not be technically feasible or economic attractive to fully automate many AI-exposed jobs.

predicted automation goes beyond a critical threshold and the time saved is reinvested in activities where humans have comparative advantage. Our process involves a simulation-based evaluation of cost savings and productivity gains simultaneously. Furthermore, by conducting our analysis for real job advertisements in an actual workforce at the task level, we are able to estimate the impact of AI on seemingly similar job roles within broader professions. While previous studies have used standard industry datasets ([Felten et al., 2021](#); [Eloundou et al., 2024](#); [Gmyrek et al., 2023](#); [Hampole et al., 2025](#); [Henseke et al., 2025](#); [Brynjolfsson et al., 2018](#); [Henseke et al., 2025](#)), such as O*NET, ISCO-08 and the UK Standard Occupational Classification, our more granular approach reveals substantial variation in the impact of AI across profession, domain and, in particular, seniority.⁴ We therefore argue that the effects of AI will be much more heterogeneous on seemingly identical jobs, suggesting that previous estimates of the impact of AI may have inaccurately estimated the impact of AI on specific job groups. This aligns with [Fenoaltea et al. \(2024\)](#) who argue that the impact of AI will be more gradual and, in particular, high-skilled jobs will not be uniformly affected by AI. Indeed, in contrast to previous hypotheses ([Felten et al., 2023](#)), in this large public organisation with many office desk jobs, we show that the impact of AI will be largest in lower grade roles, whereas higher grades comprise most human-centric tasks. However, the precise impact will depend on the specific job role and context, shifting the debate to a more precise estimate of exposure for each job and firm instead of a industry-level estimate. In particular, our end-to-end pipeline can be applied to a wide variety of job description datasets to estimate the potential impact of AI and facilitate job redesign accordingly.

Second, we contribute to the AI and productivity literature ([Noy and Zhang, 2023](#); [Brynjolfsson et al., 2025b](#)) by predicting how humans can spend freed-up time in the organisation.

⁴Complementing our within-firm study, contemporaneous papers study different features of heterogeneous AI exposure to labour markets. For instance, drawing on O*NET tasks, [Hampole et al. \(2025\)](#) study heterogeneity of task-level AI exposure *across* firms and [Fenoaltea et al. \(2024\)](#) introduce the AI Startup Exposure index to study heterogeneity across occupations, industries and regions.

Recent studies have provided early evidence for productivity gains from AI in specific contexts and companies. Brynjolfsson, Li and Raymond (2025) find a 15% productivity boost in customer-facing services after the introduction of a generative AI-based conversational assistant (Brynjolfsson et al., 2025b), while Dell’Acqua et al. (2024) show that management consultants completed more tasks, and at times of higher quality, using generative AI (Dell’Acqua et al., 2023). Other studies have focused on time savings through the use of AI in education (Choo, 2024) and healthcare (Meynhardt et al., 2025), arguing saved time can be reallocated to other tasks. Our study creates a systematic framework to examine which tasks will be automated or expanded, in which jobs, and what work will be done more with the freed-up time to increase productivity. This view expands the scope of future studies: Instead of looking at tasks in isolation (e.g., where AI will speed up a process, increasing throughput), our approach suggests task automation can and should be studied simultaneously with task augmentation to maximize productivity.⁵

Lastly, our study contributes to the job (re)design literatures (Hackman and Oldham, 1976; Bartling et al., 2012; Gibbs et al., 2022) by providing a more granular understanding of the nature of jobs in the future. Karasek (1979) observed that job redesign should consider expanding decision latitude for workers to reduce mental strain and job dissatisfaction (Karasek Jr, 1979). Our findings show that tasks that escape automation are those in which humans have a comparative advantage over AI⁶ and, specifically, that automated tasks are replaced with more human-centred tasks, including leadership and soft skills, many of which require human judgment and offer some degree of decision latitude and autonomy.⁷ Moreover, Autor & Thompson (2025)

⁵Earlier work by Stephany and Teutloff (2024) focused on skills has similarly argued that a holistic view is necessary for organisations to better adapt to technological change.

⁶In our baseline framework, the LLM replaces the automated task with one of the existing tasks for the same job. In an extension, we study whether expanding the scope of potential tasks changes our results. In line with Acemoglu and Restrepo (2018) (Acemoglu and Restrepo, 2018) and Loaiza and Rigobon (2024) (Loaiza and Rigobon, 2024), we find that these new tasks that are being created also favour activities, in which human labour has a comparative advantage.

⁷The result that LLM-led job redesign leads to more human-centric tasks being proposed is intuitive but not

propose that expertise is a critical moderating factor in whether automation leads to higher or lower wages for specific jobs in the future (Autor and Thompson, 2025). This prediction aligns with our findings across the organizational hierarchy—but in contrast to previous findings (Felten et al., 2023)—such that more senior roles, which often require expertise in some domain or skill, are *less* exposed to AI. These roles include human-centric tasks, many of which also have multiplier and cascading effects affecting other roles, such as strategic leadership, vision, and innovation. However, the need for more human-centric skills is not restricted to senior roles alone. For instance, even among lower grades, we might expect to see a shift in training and education policies to deepen skills in stakeholder management, complex problem resolution, human-centric leadership, and risk and quality management.

The rest of the paper is structured as follows: Section 1 introduces the various datasets which we use for our analysis. Section 2 covers our methodology for task segmentation, the derivation of the exposure scores, and the job redesign workflow. Section 3 presents our key findings, including how any newfound extra time might be most productively used. Section 4 discusses and caveats our results. Full details of the prompts we use, summary statistics, and ancillary charts can be found in the Supplementary Material

1 Dataset Description

The core dataset for our analysis is built from two data sources. First, we use an extract of job vacancies from the Government Recruitment Information Database (GRID), described in Section A.1. After extracting and categorising the tasks from the job descriptions in these vacancies (see Section 2.2), we then join on a second dataset for departmental information, the UK Civil Service Statistics (UKCSS), which we cover in Section A.2. Together, this forms the

inevitable. For instance, Gong & Png (2024) show that automation in some tasks leads human workers specialising in the remaining non-automated tasks without regard to the specific nature of the remaining tasks, purely because workers no longer face the costs of coordination (Gong and Png, 2024).

foundation for how we categorise job roles by their exposure to AI, culminating in the results presented in Section 3.

1.1 Summary statistics of the consolidated dataset

Our dataset, constructed using the steps described above, contains 1,542,411 tasks from 193,497 vacancies on Civil Service Jobs between 16th January 2019 and 3rd December 2024.

1.1.1 Department

Vacancies are sourced from 37 departments including 17 ministerial departments, 14 agencies and other public bodies, and 6 non-ministerial departments. Together, these departments employ 433,890 FTEs, comprising 85 percent of the total UKCS workforce.

The top five departments collectively represent 50 percent vacancies in the data (MoD, HMRC, HO, HMPPS, FCDO), while the bottom five departments (DESNZ, DVLA, CH, UKEF, MO) represent just two percent. DWP have the highest proportion of FTE roles in the Civil Service workforce (16%) but only represent 3% of the roles in our dataset. A full list of these departments, their acronyms, representation in the dataset, and FTE can be found in Table 5 in the Supplementary Materials, Section B.4.

1.1.2 Grade

The UKCS grading structure reflects levels of responsibility and expertise. Administrative Assistants (AA) and Administrative Officers (AO) handle routine administrative tasks and frontline services, while Executive Officers (EO) take on decision-making and operational management. Higher Executive Officers (HEO) and Senior Executive Officers (SEO) are involved in policy development and project management. Grade 7 (G7) and Grade 6 (G6) roles focus on complex policy work and strategic planning. The Senior Civil Service (SCS), including Deputy Directors, Directors, and Permanent Secretaries, lead major policy areas and government strategy.

Each grade represents increasing leadership and accountability in delivering public services and policies.

The grade in the GRID data is a free text field, lacking standardisation. We perform a manual mapping exercise to map each vacancy's grade to the corresponding grade bucket from Table 21 in the UKCSS. Around half of all vacancies in our dataset are either SEO or HEO grades, compared with a representation of 30 percent in the wider UKCS. The G6 and G7 grades are also comparatively over-represented, comprising 24 percent of our data and only 15 percent of the UKCS workforce. By contrast, the lower grades (AA/AO/EO) are relatively under-represented in our vacancies data. Senior leadership roles, unsurprisingly, have the lowest number of roles in the dataset (4,510) and the UKCS workforce (7,295). A full breakdown of the distribution of grades within GRID compared to the wider UKCS is available in Table 6 in the Supplementary Materials, Section A.

1.1.3 Median Salary

As outlined in Section A.2, the GRID data does not include the salary of each role. The UKCS advertises roles using pay bands, with these pay bands varying by department, grade and role location. We therefore do not have accurate data outlining the actual total cost of each role in our dataset. Instead, we estimate the cost of each role using the average salary for each grade and department combination.

Through this, the average annual median salary in our dataset is £42,776. The lowest value salary is £21,130 and the highest is £113,980. There is a standard deviation of £13,622, with the 25th and 75th percentiles £34,860 and £57,150 respectively. The median salary in our dataset is significantly higher than the Civil Service median salary (£33,980 in 2024). This is likely reflective of the grade composition outlined in the previous subsection: that our dataset over-represents roles in higher grades which are better paid.

1.1.4 Profession

The UKCS is made up of a variety of professions. A profession within the UKCS refers to a specialised area of expertise: each profession plays a role in ensuring that the UKCS functions efficiently. The full breakdown of these professions including their representation in our dataset is included in Table 7 in the Supplementary Materials, Section A.

Our underlying dataset consists of 28 professions and the option for “other” professions. The other option represents the largest group of the roles advertised (36.77%). Operational Delivery Profession is the largest specified profession in the dataset with 27,343 roles (14.13%). The next four professions with the highest representation in the data include Data and Digital with 21,169 roles (10.94%), Policy Professions with 16,825 roles (8.70%), Project Delivery with 10,434 roles (5.39%) and Human Resources with 7,041 roles (3.64%).

We can get an insight into the advertised roles classified as “other” in inspecting which departments they belong to. The majority of these roles (61.74%) are based in three departments: HMPPS (30.50%), MoD (17.49%), and MoJ (13.75%). UKCSS shows that essentially all roles in HMPPS (96.32%) are Operational Delivery roles. Therefore, when quantifying AI exposure by profession, we consider all roles in HMPPS to be Operational Delivery roles.

1.2 Overall representation of the UKCS

The summary statistics presented suggests that our sample of job roles is of a higher grade and therefore higher paid on average compared to the wider UKCS. Such differences are to be expected, particularly given that some departments advertise roles directly instead of through Civil Service jobs. Nevertheless, to account for the discrepancies between the proportion of roles in each department, grade and profession in our dataset and the wider UKCS, we apply an iterated proportional fitting process, described in Section 2.4.

2 Methodology

This Section details the key stages of data processing and analysis involved in extracting tasks and assigning the AI exposure scores. We perform this step on the raw GRID data before joining it to the UKCSS so that the task categories and subcategories are not impacted by roles which cannot be joined to the UKCSS.

We first use an LLM to extract tasks from each role and assign AI exposure scores. We then perform clustering on these tasks to enhance our understanding of the types of tasks and their automation potential. Next, we aggregate up to the job role level, using role level statistics to group all roles depending on the likely impact AI will have on the role. This enables us to apply iterative proportional fitting to arrive at estimated potential savings. Once we have identified roles in scope for productivity gains, we assess how roles could be redesigned with AI in the future. A simplified version of this workflow is visualised in 1.

2.1 LLM choice and bulk operation

We use the LLM Claude Sonnet 3.5 v2 (“anthropic.claude-3-5-sonnet-20241022-v2:0”), available via Amazon Web Services (AWS) Bedrock. We choose this particular model as (at the time of writing) it scores higher in most benchmarks when compared to other LLMs available via Bedrock⁸ and is suitable for bulk API calls. The tool calling functionality also allows us to specify a structured output format, permitting easier and more robust model validation.

To run the analysis in bulk we use the batch inference endpoint of AWS Bedrock. This is a more cost-effective and scalable method than the on-demand endpoint. To improve the ease of access to the batch inference endpoint, we developed a Python package named `LLMBO` (Large Language Model Batch Operations). This is open source, [available](#) under MIT licensing

⁸Based on GPQA scores taken from <https://llm-stats.com/>

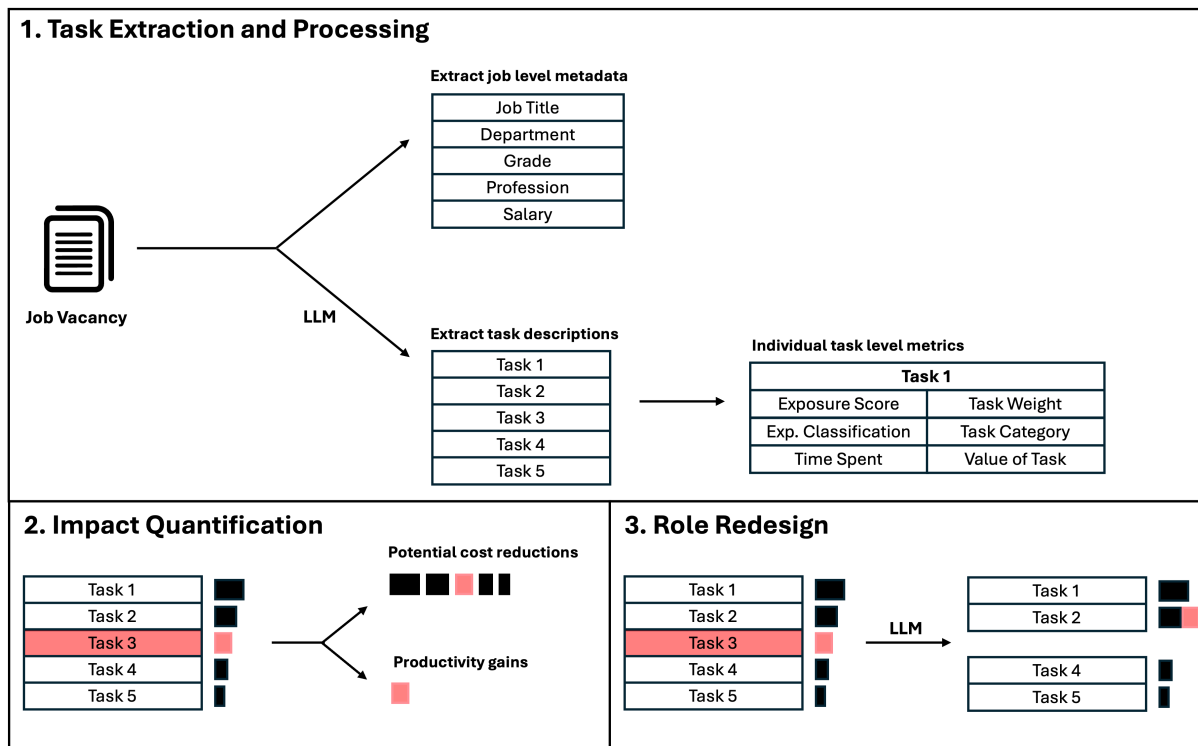


Figure 1: Visual representation of analytical workflow

This visual shows the workflow for one single job description - we repeat this for all 193,260 job descriptions in our dataset. Step 1 (Task Extraction and Processing) uses an LLM to extract each unique task from the job description. When coupled with metadata from the job description, this provides us with a detailed role profile for each job. Step 2 (Impact Quantification) assesses the potential impact of AI on each job. We use the proportion of time spend on high exposure tasks to determine if a role is in scope of potential cost reductions or productivity gains. Step 3 (Role Redesign) assesses how any role in scope for productivity gains is likely to change through the use of AI. We remove the time spent on high exposure tasks and assign this time to one of the remaining tasks in the list.

from the pypi Python package repository.⁹ LLMB_O provides many wrappers for the endpoint which improve usability, the most useful being the ability to specify an output schema for the query, and have the schema validated for each LLM response. This allows post-processing of the results to be performed without needing complicated regular expressions or other similar workarounds.

⁹Documentation is provided at <https://co-cddo.github.io/gds-idea-llmbo/>

2.2 Task extraction and AI exposure scores

Existing empirical approaches rely on using high-level, static databases with defined lists of standardised tasks, such as the O*NET database or ISCO-08 framework. One of our key innovations in this paper is the application of a bottom-up data-generating process, using vacancies data to extract role tasks. This allows us to analyse AI exposure at much higher specificity and capture the nuances between roles of similar professions.

To do this, our starting point is the two fields which encompass the text of the job advert. First, a job summary, which typically includes contextual details on the team and department. Second, a job description, which typically includes the tasks and skills associated with the role. As the job description is in most cases what we require, we do a first-pass using this field. Each job description is passed into the LLM along with its vacancy ID and a prompt to extract all tasks and assign each an exposure score between 0 and 1 (see the Supplementary Materials, Section C.1 for the precise prompt and output format).

Sometimes failures occur in roles where the the job summary instead contains the job details and the contextual information is given in the job description. In such cases, we re-run the query using the job summary field. Once processing errors and jobs with under two tasks are removed, we end up with 212,048 job vacancies broken down into their constituent tasks, with each task having an exposure score between zero and one. The final sample size is 193,260 roles, largely due to a fraction of jobs having “Other” or “Industrial” as their grade which cannot be mapped using the UKCSS.

As this process shows, UKCS job descriptions do not follow a uniform structure and vary depending on the role and department. They include information beyond the list of tasks the role will execute, with additional information including background on the team and department, the skills required, the recruitment and onboarding processes, and the working hours and conditions. All of this contextual information is not relevant for our purpose, and therefore we

require the LLM to extract each individual task and provide a final structured dataset, with each task listed uniquely.

It is important to clarify that our methodology strictly focused on extracting the sections of job descriptions that relate to the tasks performed and functional context of the role. Any personal information which can occasionally be included in job descriptions, such as contact emails or phone numbers, were removed before any data entered the pipeline or was passed through the LLM.

For each job role j , we extract a total of T tasks and assign an exposure score task e_t to each individual task. Based on these exposure scores, we classify tasks using the following scheme:

Very Low Exposure: $0 \leq e_t < 0.3$

Low Exposure: $0.3 \leq e_t < 0.5$

Medium Exposure: $0.5 \leq e_t < 0.7$

High Exposure: $e_t \geq 0.7$ set $h_t = 1$

where h_t is an indicator variable that signals if a task is classified as high exposure or not. This classification is based on that used by Gmyrek et al. (2023), who use intervals of 0.25 (Gmyrek et al., 2023). We adapt this and have longer tails of 0.3 at either end, given that our LLM typically assigns rounded exposure scores. This classification informs the methodology for deriving savings estimates in Section 2.5, where we use h_t to calculate the proportion of time each job role spends on high exposure tasks and quantify the impact of AI implementation. This classification is an initial step, and in Section 2.3 we show how we calculate the time spent on each task. Together, each task's exposure classification and the estimated time spent on it provides us with a way to construct the number of hours each role spends on tasks in each classification in a typical working week. A breakdown of the job batching statistics is provided

in Table 8 in the Supplementary Materials, Section D.

To ensure consistency and accuracy of our methodology, we perform several validation checks to be assured of the accuracy and completeness of the task extraction and exposure scores, which we detail in the Supplementary Materials, Section B.

2.3 Task Time Allocation

The previous steps enable us to assess the exposure to AI for each role, and, using the task clustering, assess why different roles have differing levels of AI exposure. The next step is to quantify the potential impact of AI use across the entire UKCS. To do this, we must first estimate the current amount of time each role spends on the tasks listed in the job description. For all roles in our dataset, we assume the role holder is contracted to work 37 hours per working week.¹⁰ While we recognise that there will be a variation in working hours, and that a proportion of roles will be part time, we use this assumption in the absence of role level data on hours worked.

Initially for simplicity, and in line with the majority of other studies conducting similar analysis, we considered a uniform distribution of the 37 hours across all listed tasks for each job role, where each task has a $\frac{1}{T}$ of the FTE allocation, where T is the total number of tasks listed. However, we consider such a crude approach to be unrealistic and not fully representative of how workers actually allocate their time. In the absence of comprehensive time-use or time sheet data, we assume that the order in which tasks are listed reflects their relative importance or frequency.¹¹ Specifically, we assume that tasks appearing higher in a job description occupy a greater share of the role holder's working hours.¹²

¹⁰This is the typical length of a working week defined in a full-time UKCS contract, as outlined, for example, by HMRC.

¹¹We re-examine the impact of this assumption in robustness checks in Appendix I

¹²This is backed up by conversations with hiring managers who told us that more important tasks are often the first that come to mind when job descriptions are written and that they are often on purpose listed earlier. Similarly, career and recruitment advice platforms recommend prospective employers list tasks and responsibilities in order of

To implement this, we introduce a decay rate parameter $\delta \in (0, 1]$ that controls the rate at which the importance of tasks decreases depending on their position in the job description. Given a job role j consists of T total tasks, each individual task $t \in \{1, 2, \dots, T\}$ is assigned an unnormalised weight:

$$d_t = \delta^{t-1}$$

A decay rate of $\delta = 1$ corresponds to equal weighting, while values closer to 0 increase the emphasis on earlier listed tasks. To illustrate the impact of the decay rate, we visualise the individual task weights for a job description with six tasks in Figure 14 in Section I in the Supplementary Materials. As shown, a decay rate of 0.3 or less would place what we consider a disproportionate emphasis on the first task, with a weight of more than 70%. While a decay rate of 0.5 would place less emphasis on the first task, we still believe such a rate risks not appropriately weighting the breadth of tasks completed by a job role when estimating overall AI exposure.

For this reason, we select a decay rate of $\delta = 0.75$. We consider this rate to strike a balance between prioritising earlier tasks, in line with our assumption, but still ensuring that all tasks in the job description contribute meaningfully to our findings. This results in a geometric sequence of unnormalised weights across tasks (1, 0.75, 0.56...). These weights are then normalised to ensure that the weights in each job role sum to 1:

$$D_t = \frac{d_t}{\sum d_t} \quad \text{such that} \quad \sum_{t=1}^T D_t = 1$$

where D_t is the normalised weight of task t .

importance or time commitment (HR Daily Advisor, 2012; Indeed Editorial Team, n.d.; The Muse, n.d.). However, this is not a codified process and the order of tasks should be taken as indicative, not conclusive, evidence for the importance to the employer.

As detailed in Section 2.2, we use an LLM to assign each task t an AI exposure score $e_t \in [0, 1]$. This score represents the level of exposure the individual task itself has to AI - it does not account for the amount of time spent on the task and does not directly reflect how much that task contributes to the overall exposure of the role. For example, a task with high exposure but minimal time allocation will have limited impact on a role's overall level of exposure to AI. Therefore, we calculate a weighted exposure score for each job role, allowing us to combine the task-level AI exposure with the assumed importance of each task:

$$E_j = \sum_{t=1}^T D_t \cdot e_t$$

where $E_j \in [0, 1]$ is the weighted exposure score for each job role j . We also use the final calculated decay rate to distribute characteristics of the role across tasks: for example, in Section 2.5, we multiply the weight of each task by the total salary of the role to estimate the value of time spent on each task.

Table 1 illustrates the decay rate's impact. This job is an Operational Delivery role made up of six tasks. Using the task classification as per Section 2.2, tasks 1 and 2 are classified as high exposure; task 3 is medium exposure; tasks 4 and 5 are low exposure; and task 6 is very low exposure. A manual, qualitative review suggests that the top three tasks likely represent the core responsibilities of the role holder, involving visa application processing and casework handling. The other tasks are broader and more strategic, focussing on organisational change and therefore likely to be less frequent. Under an equal weighting, this role would be assigned an average exposure score of 0.53. By contrast, applying the decay-weighted method increases the average exposure score to 0.64, with more weighting placed on the earlier tasks which have a higher exposure levels.¹³

¹³While in this example the earlier tasks happen to have been more exposed to AI, with other roles the opposite is true, so the decay-weighting doesn't automatically lead to higher exposure scores.

Table 1: Average Exposure Score and Task Weighting Comparison

Task Details			Equal Weighting		0.75 Decay Rate	
#	Task Description	Exposure Score	Task Weight	Weighted Score	Task Weight	Weighted Score
1	Managing and processing visit visa applications from multiple countries including India, Thailand...	0.80	0.17	0.13	0.30	0.24
2	Remote printing operations for net migration casework consolidated in the UK	0.90	0.17	0.15	0.23	0.21
3	Managing complex visa applications requiring detailed assessment and decision making	0.60	0.17	0.10	0.17	0.10
4	Implementing new services and solutions to improve the end-to-end customer journey	0.40	0.17	0.07	0.13	0.05
5	Managing change and business transformation within the Decision Making Centre	0.30	0.17	0.05	0.10	0.03
6	Driving forward UKVI's People Agenda	0.20	0.17	0.03	0.07	0.01
Sum			1.00	0.53	1.00	0.64

The table shows the average exposure score for an example job role, comparing an equal weighting scenario with a decay rate of 0.75. Exposure score column is the raw assigned exposure score. Task weight is the weighting given to each task - under equal weighting this is $\frac{1}{7}$, under decay rate this is a normalised value based on the position of the task in the job description. The weighted score column is the product of exposure score and task weight for each scenario. Given the task weight column is already normalised, we sum the weighted task level scores to produce the average exposure score for each scenario. For the equal weighting, the sum is 0.53. For the 0.75 decay rate, this sum is 0.64

2.4 Sample Re-weighting

While our sample dataset has a broad range of roles and consists of 193,260 unique job roles, this does not fully represent the entire UKCS population of 516,150 FTE, as of July 2025. To estimate potential savings across the full UKCS, we re-weight the sample so that it represents the full workforce. We apply a single-stage approach based on Iterative Proportional Fitting

(IPF) to construct a set of post-stratification weights.

Given that each job role in our sample is characterised by department or organisation $o_j \in O$, grade $g_j \in G$ and profession $p_j \in P$, we define the initial weights of each job role and the known totals for each department, grade, and profession:

$$w_j^{(0)} = 1, \quad N_o (o \in O), \quad N_g (g \in G), \quad N_p (p \in P).$$

Although our sample covers a wide range of departments, the 37 included in our dataset represent only 85% of total UKCS FTE. To ensure that the department marginal is exhaustive, we introduce a synthetic category `OTHER_DEPTS`, which accounts for the remaining 15% (a total of 76,235 FTE).

For each grade–profession combination, we append a corresponding row with department set to `OTHER_DEPTS`. This allows the raking procedure to allocate weight to unobserved departments, rather than assuming their structure mirrors that of the sampled ones. Unlike a uniform scaling approach, this method treats the unobserved share explicitly, leading to a more robust and transparent weighting scheme. We then convert known totals into marginal distributions and use these as raking constraints:

$$\pi_c = \frac{N_c}{N}$$

The current weighted proportion for each category is calculated as:

$$s_c(w) = \frac{\sum_{j:c_j=c} w_j}{\sum_j w_j}$$

where $c \in \{d, g, p\}$ indicates the dimension being adjusted (department, grade, or profession), and c_j denotes the category membership of role j on that dimension. At each iteration k , the IPF procedure updates the weight by multiplying the previous weight by the ratio of

the target marginal proportion to the current sample proportion for the department, grade, or profession:

$$w_j^{(k+1)} \leftarrow w_j^{(k)} \cdot \left(\frac{\pi_{c(j)}}{s_{c(j)}(w^{(k)})} \right), \quad c \in \{d, g, p\},$$

For each iteration, the weights are updated once per dimension, in sequence: department, grade, then profession. For example, when $k = 1$, the initial weights $w_j^{(0)} = 1$ for each role are multiplied by the ratio of the target marginal proportion to the sample proportion for each department. These updated weights are then multiplied by the grade ratio, and finally by the profession ratio to obtain $w_j^{(1)}$.

The process continues until convergence, defined as:

$$\max_j \left| \frac{w_j^{(k+1)} - w_j^{(k)}}{w_j^{(k)}} \right| < 10^{-6},$$

or until 30 iterations have been completed. After convergence, the resulting weights $w_j^{(*)}$ sum to an arbitrary constant. To scale these to the full UKCS workforce N , we apply:

$$w_j = w_j^{(*)} \cdot \frac{N}{\sum w_j^{(*)}}$$

This guarantees:

$$\sum w_j = N$$

The final weights w_j reflect the known totals of departments, grades, and professions in the UKCS and can be used to compute nationally representative estimates. We use the same weights generated from this process to scale up other variables in our dataset - notably total estimated salary cost and potential savings.

2.5 Savings Estimate

In Section 2.2, we classify each individual task based on its assigned exposure score e_t . Tasks with $e_t \geq 0.7$ are classified as high exposure and assigned an indicator $h_t = 1$; all other tasks are assigned $h_t = 0$. In Section 2.3, we then assign each task a normalised weight D_t based on its position in the job description, ensuring that earlier tasks are weighted more heavily.

We use these earlier steps to estimate the potential cost reductions and productivity gains from AI. We begin by calculating the proportion of each job role's time spent on high exposure tasks, denoted H_j :

$$H_j = \sum_{t=1}^T D_t \cdot h_t$$

We then introduce a threshold parameter $\theta \in [0, 1]$, which represents the minimum proportion of time that a job role must spend on high exposure tasks to be considered in scope for potential cost reductions.¹⁴ Roles falling below this threshold are instead assumed to be in scope for productivity gains. We first filter out any roles that have no high exposure tasks (we assume no potential savings or impact from AI for these roles) and then classify each remaining job role by comparing the proportion of time spent on high exposure tasks to a given threshold level:

$$\mathbb{I}_j^{(C)} = \mathbf{1}[H_j \geq \theta], \quad \mathbb{I}_j^{(P)} = \mathbf{1}[H_j < \theta]$$

where $\mathbb{I}_j^{(C)}$ and $\mathbb{I}_j^{(P)}$ are indicator variables equal to 1 if a job role is classified as in scope for potential cost reductions or productivity gains, respectively.

¹⁴There are many explanations for why we might expect jobs to require a minimal threshold of tasks being affected before an entire job gets automated, including the organisational challenge of combining multiple partial jobs into a single role, or the lack of economic attractiveness or technical feasibility to make automation worthwhile (Svanberg et al., 2024).

As outlined in Section A.2, we join UKCSS data to estimate the salary cost S_j of each job role, which we use to estimate the potential financial impact of AI. Roles identified as in scope for potential cost reductions (those for which $\mathbb{I}_j^{(C)} = 1$) are assumed to have the potential to be fully displaced and we therefore take the entire salary cost as the estimated saving¹⁵. For roles instead classified as in scope for productivity gains (those for which $\mathbb{I}_j^{(P)} = 1$), we estimate savings based on the proportion of time spent on high-exposure tasks, making the assumption that for these roles AI could automate the most highly exposed tasks, allowing the role holder to focus on their remaining tasks. The estimated potential cost reductions C_j or productivity gains P_j for each job role are therefore defined as:

$$C_j = \mathbb{I}_j^{(C)} \cdot S_j \quad P_j = \mathbb{I}_j^{(P)} \cdot H_j \cdot S_j$$

This gives us an estimated saving for each job role in our dataset, but our dataset only represents just over a third of the total UKCS FTE. To estimate the total potential impact across the entire UKCS, we scale each role's projected saving by its representative weight w_j , derived by the methodology in Section 2.4. This allows us to compute nationally representative totals for both cost reductions and productivity gains:

$$C = \sum w_j \cdot C_j, \quad P = \sum w_j \cdot P_j$$

The assumption with the highest impact is the value of θ . A lower value would mean more roles are classified as in scope for potential cost reductions. To reflect this uncertainty, we repeat this process across thresholds ranging from 0 to 1 in 0.05 increments. At each value of θ , we reclassify roles as either in scope for potential cost reductions or productivity gains and recompute the role level and UKCS level totals.

¹⁵We do not consider additional employment costs here, such as pension and National Insurance contributions. Typically these can be worth as much as 40% of the base salary.

Our modelled estimate uses a threshold of 80% ($\theta = 0.8$), a figure chosen in line with other literature which, if anything, is a conservative threshold to choose. Muro, Maxim, and Whiton (2019) deem a job to be “high risk” if over 70% of the tasks completed are potentially automatable (Muro et al., 2019), whilst Arntz, Gregory and Zierahn (2016) take a similar approach, where jobs with a probability of automation of at least 70% are considered to be highly automatable (Arntz et al., 2016).

To make concrete how we evaluate each role, we use the earlier job role used in Section 2.3. In Table 2, we quantify the potential opportunity from AI at our modelled threshold $\theta = 80\%$. This role has two tasks (1 and 2) classified as high exposure tasks, where $h_i = 1$. The proportion of time spent on high exposure tasks H_j for this role is therefore 0.53 (the sum of assigned weights for the two high exposure tasks). We compare this to all automation thresholds. At our chosen modelled threshold level where $\theta = 0.8$, $H_j < \theta$, meaning that the role is in scope for productivity gains only at this threshold. In this case, we sum the value of time spent on high exposure tasks only, which, given a salary cost of £38,680, results in a potential productivity gain of £20,500. This process is repeated for every threshold level between 0 and 1. Given $H_j = 0.53$ for this role, any threshold where $\theta \leq 0.53$ would result in this job role being in scope for potential cost reductions, meaning we would assume the full salary cost of £38,680 could be saved.

2.6 Task Clustering

Having broken down jobs into individual tasks, we require a typology of tasks to help make sense of the data. To do this, first we preprocess each task using the standard natural language processing (NLP) techniques.¹⁶ The tasks are then embedded using the pre-trained “all-mpnet-base-v2” model and the associated sentence transformer methods in Reimers and Gurevych

¹⁶Specifically: removing digits, punctuation, and stopwords, making lowercase, tokenizing by splitting on whitespace, and lemmatizing each token.

Table 2: Example savings quantification using 75% decay rate

#	Assigned Exposure Score	Exposure Classification	Value of h_t	Assigned weight	Value of Task (£)
1	0.80	High	1	0.30	11,604
2	0.90	High	1	0.23	8,896
3	0.60	Medium	0	0.17	6,575
4	0.40	Low	0	0.13	5,028
5	0.30	Very Low	0	0.10	3,868
6	0.20	Very Low	0	0.07	2,707
Sum				1.00	38,680

Task 1 and 2 are classified as high exposure tasks, with an assigned exposure score of 0.7 or greater. The value of task is the product of the assigned weight and the total salary cost of the role, in this case £38,680.

(2019) (Reimers and Gurevych, 2019). This returns the embeddings matrix $\mathbf{E} \in \mathbf{R}^{1,737,971 \times 768}$.

Next, to mitigate the so-called “curse of dimensionality” and reduce computational cost, we use Principal Component Analysis (PCA) to project the 768 dimensions down to 25. We then apply the K-Means unsupervised clustering algorithm to group the embeddings. We select the hyperparameter of clusters to be ten, which after experimentation appears to provide a sufficiently broad typology without conceptual overlap. To check the clustering effectiveness, we review examples from each cluster as well as visually inspecting a two-dimensional projection of the ten categories.

With the tasks clustered into ten categories according to their semantic content, we then repeat the process to get three sub-clusters per cluster. This allows us to break down the ten broad clusters into more detail, highlighting some interesting heterogeneities within each broader cluster (see Section 3.2.6). Next, we require human-interpretable labels for each cluster and sub-cluster. To get these, we randomly sample 200 tasks from a given (sub-)cluster and pass it to the LLM with the prompt in Section C.3 in the Supplementary Materials.

This yields a two-level categorisation of all tasks, grouped by semantic similarity. We are now in a position to explore our results for all UKCS roles and tasks.

2.7 Role Redesign

The final step in our pipeline helps us consider how roles might be redesigned following the potential of AI transformation in the workplace. As most firms are only beginning to trial and pilot AI tools, there is little systematic empirical data on how workers effectively reallocate their time across a wide range of jobs following AI implementation in their role. Given this constraint, we turn to an LLM-led process to redesign the job in anticipating where productivity gains may be found. Specifically, we prompt the LLM to suggest the most important task for a worker to devote their newfound time towards, the “focus task”. While our main analysis focuses on a single focus task drawing from existing tasks, we conduct several robustness checks below that relax the assumption of a single existing task. We then compare total time spent before tasks are automated, and after they have been replaced with each role’s focus task.

Beginning with the 193,260 roles in our sample, we remove the 7,779 that are fully automatable, and another 39,575 which have no automatable tasks. This leaves us with 145,906 roles that will not be fully automated but that each have at least one task that could be automated by AI. This is the population of roles that we take forward to consider augmentation with AI.¹⁷

To consider the most productive alternative, we pass each role title along with its remaining tasks into the LLM and ask the LLM to choose which remaining task a person should focus their newfound time on, along with a rationale (see prompt 1 in Appendix C). Thus, for every role, we are provided with a suggested focus task along with some reasoning.¹⁸ For simplicity, we assume that this focus task replaces all of the weighted time allocation previously taken up by highly exposed tasks.¹⁹

¹⁷Roles with no automatable tasks may still benefit from AI assistance on their tasks with some exposure. However, for simplicity they are excluded from consideration here.

¹⁸Although often called “reasoning”, LLMs have not been found to reason in the way humans do. Anthropic (Lindsey et al., 2025) demonstrates this difference between Claude 3.7 Sonnet’s actual reasoning versus what it claims to be doing.

¹⁹As a robustness check we also allowed the LLM to reorder all remaining activities and recalculated the associated task weights and FTE time spent on each activity category. This shows little difference to the focus task

To better understand why the LLM chooses certain tasks as the most valuable alternative we take three random samples of 2,000 rationales and ask the LLM to provide themes (see prompt 2 in Appendix C). After consolidating the results to remove duplication, we are left with the six categories in Table 3. As focus tasks could plausibly fall into several rationale themes, we ask the LLM to apply up to three themes per focus task reasoning (prompt 3 in the Appendix C). This allows us to map, for each focus task, how it was reasonably chosen.

To ensure that our analysis is not unduly restrictive, we also run several extensions and robustness checks. First, we check how focus tasks and time allocations differ between high- and low-grades. We proxy this using the top and bottom income deciles. As income is very closely tied to grade due to the UKCS grading structure, this is a suitable approach that is backed up by the deciles' composition. The lowest income decile comprises roles earning below £30,150, and is made up of 10,890 AA/AO grades and 3,858 EO grades. By contrast, the highest income decile comprises workers earning over £62,590, including 16,230 G6/G7, and 2,064 SCS.²⁰

Second, a plausible alternative to the focus task approach is that, with some tasks automated away, workers may be able to spend their newfound time across all their tasks rather than a single one. In addition, some of the tasks are likely to be augmented following the rollout of AI tools. Therefore, we next allow the LLM to re-specify and reorder remaining tasks, leading to new task descriptions, task weights, and time shares (see prompt 4 in Appendix C).

Finally, the disruption and opportunity from AI technology may not only lead to existing tasks disappearing but new ones being created. To consider this possibility, we complete a second extension, where we allow the LLM to suggest new tasks specific to the role rather than simply reprioritising existing tasks. We provide the LLM with the full role summary and description to give it a contextualised understanding of the role and how it fits into its team

exercise described above (see Table 14 and prompt four in Appendix C.4).

²⁰The slight difference in totals is likely due to the lower/upper bounds of the salary bands.

and department. Next, we tell the LLM which tasks have been automated away due to AI (all high exposure tasks) to exclude the possibility that they're added back in as new tasks. Finally, we provide a list of remaining tasks and invite the LLM to suggest any new tasks that could be relevant. The number of new tasks that can be suggested is limited to the number of tasks automated away. The LLM returns the tasks reordered according to their importance for the role and assigns each new task an appropriate category (see Prompt 5 in Appendix C).

Given the large amount of input tokens required for this level of context (including tooling), we perform this robustness check on a 10 percent sample of the full set of roles in scope for productivity gains. Accordingly, we use a sample stratified by department and grade to ensure it remains representative while keeping costs down. After running this process on our sample of 14,591 roles, we recalculate the task weights using the order we receive the tasks back from the LLM output. In this way, the LLM has been given maximum freedom to both suggest new tasks and reorder all tasks. Together, the results from these exercises, detailed in Section 3.4, allow us to explore role-specific possibilities across a range of potential future scenarios.

3 Results

3.1 AI Exposure at Task Level

Our final dataset covers 193,497 unique job descriptions with 1,542,411 associated tasks. From our task extraction process, the minimum number of tasks identified in a role was 2, and the maximum was 37. This is partly reflective of the varying data quality in the job description field, but it also partly captures the genuine breadth of roles in the UKCS.²¹ The mode number of tasks per role is 6 and the mean is 8. The standard deviation of tasks per role is 3.36, and 90 percent of roles have between 4 and 14 tasks identified. The tasks are spread evenly across three of the four classifications outlined in Section 2.2: 3.7% of tasks are classified as very

²¹After manual review, the role with 37 identified tasks really does appear to comprise an incredibly broad role.

low exposure, 33.4% as low exposure, 32.9% as medium exposure and 30.1% as high exposure tasks. In short, our LLM approach suggests that the majority of tasks have at least some degree of AI exposure.

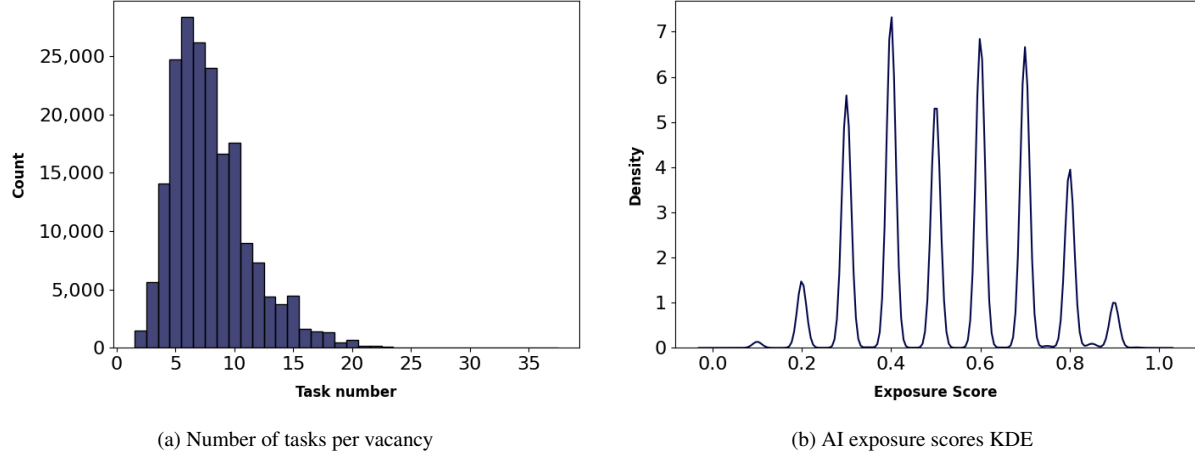


Figure 2: Distributions of the tasks per role and AI scores

Tasks per role are extracted by the LLM as described in Section 2.2. By construction, there are at least two tasks per role. Likewise, the AI exposure scores are bounded between zero and one by construction.

The AI exposure scores range from zero to one and are interpreted as cardinal measures, since they represent magnitudes of exposure with consistent intervals and a true zero indicating no possible exposure. The mode exposure score is 0.4, the mean is 0.53 and the standard deviation is 0.18. While the score range is continuous, the kernel density estimation is spiky because the LLM selects multiples of 0.1 in most instances. This might be because the frequency of numbers between zero and one (or ratings) in the training data concentrate around one decimal place.

3.2 AI Exposure Across Roles

Having presented the potential impact of AI at the task level, we estimate their aggregate impacts on the 193,497 job roles in the GRID dataset. First, we use task level exposure scores to calculate the mean and standard deviation of exposure for each role. We then focus on the types

of tasks performed, identifying which common tasks are most frequently associated with high AI exposure.

3.2.1 Exposure clustering

Whilst using the proportion of time spent on high exposure tasks provides a relatively simplistic way to differentiate roles, it does not give a fundamental understanding of why certain roles have a higher AI exposure score than others. It is for this purpose we next create explainable exposure clusters and explore them further.

Building on the methodology in [Gmyrek et al. \(2023\)](#), we construct four clusters named Low, Adaptation, Augmentation, and Automation, arranged in order of increasing AI exposure on job roles. We created these exposure clusters based on two dimensions: first, the mean exposure score, being the average AI exposure score across all tasks listed within the role; second, the standard deviation of exposure score, a measure of the variation in AI exposure scores across all tasks within the role. Following [Gmyrek et al. \(2023\)](#), we use a K-means clustering approach to produce four distinct exposure clusters. Annex K presents a statistical summary of these exposure clusters.

The interpretation of each cluster is summarised as:

- **Low - 40,272 roles (20.81% of sample):** Roles with low mean AI exposure score and low standard deviation. With a mean role level exposure score of 0.38, the majority of tasks in these jobs have low exposure to AI and therefore have limited automation potential and are unlikely to be significantly affected by technological advancements. While time savings are possible with AI, they are expected to be negligible.
- **Augmentation - 59,135 roles (30.56% of sample):** Roles with a low mean exposure score but a high standard deviation. With a mean role level exposure score of 0.48, while the majority of tasks in such roles have limited exposure, these jobs still contain a small

number of high exposure tasks. AI will therefore automate or speed up some routine tasks, changing how people work and allowing the role holder to focus on higher-value activities.

- **Adaptation - 59,891 roles (30.95% of sample):** Roles with a high mean and high standard deviation. With a mean role level exposure score of 0.59, the impact of AI on these roles varies substantially depending on how AI is integrated and the specific tasks involved. At a minimum, AI will reduce time spent on specific tasks, but in some cases, entire roles may be displaced. The potential impact depends on the type of job and the technology used.
- **Automation - 34,199 roles (17.67% of sample):** Roles with a very high mean and low standard deviation. With a mean role level exposure score of 0.70, the majority of tasks completed in these roles are highly exposed to AI, meaning that many of these jobs could feasibly be fully automated and displaced. These roles involve structured, repetitive tasks that require minimal human intervention once AI and automation technologies are implemented.

3.2.2 Heterogeneity of AI exposure across similar roles

Our methodology using a bottom-up approach to identify each task from job adverts enables us to speak to a much greater diversity of job roles that are seemingly similar at face value. Specifically, when looking at the job title alone, one might assume that all jobs with the same job title would be completing the same tasks. As a result, one might assume that they would all face the same risk of automation, adaptation or augmentation.

This assumption is also rooted in how the current literature approaches LLM-based estimations of job roles' exposure to AI. Recent studies use approaches and datasets that focus on broad categorisations of job roles and tasks. For example, [Felten et al. \(2023\)](#) link AI appli-

cations to human abilities and then to occupations in the United States Department of Labor’s Occupational Information Network (O*NET) database, and [Eloundou et al. \(2024\)](#) study GPT exposure for tasks in the O*NET database. [Gmyrek et al. \(2023\)](#) study the impact of AI more globally, estimating task and occupation-level automation scores in the International Standard Classification of Occupations (ISCO-08). All three papers focus on AI exposure at the task level but do so with just one job description for each job. At the same time, [Felten et al. \(2023\)](#) and [Eloundou et al. \(2024\)](#) use the O*NET database, while [Gmyrek et al. \(2023\)](#) use the ISCO-08 classification, leading to the same jobs having two *different* job definitions (and two lists of tasks) in the two databases, whilst still representing only two distinct data points for the same job.

For instance, consider the job of an ‘Economist’. Both O*NET and ISCO-08 each include one instance of an ‘economist’ and each provide a list of tasks for what the job entails, each database effectively treating all economists in the same way. We pass these tasks through the exposure analysis in our approach.²² For an Economist role, the ISCO-08 framework lists 12 tasks with an average AI exposure score of 0.74 and standard deviation of 0.09, and the O*NET database lists 13 core tasks with an average AI exposure score of 0.65 and a standard deviation of 0.16.²³ These two data points are plotted in Figure 3.

Using these datasets would suggest uniform treatment for all economist roles, with no variability over seniority or domain. Note that the ISCO-08 database specifies that the 12 tasks listed under ‘Economists’ are also applicable to ‘Econometrician’, ‘Economic Advisor’, ‘Economic Analyst’ and ‘Labor Economist’. In reality, there is substantial variability in the types of tasks completed by economists in different fields. In addition, compared to the more analytic-orientated ISCO-08 tasks, the O*NET database has a much wider range of tasks, including

²²See scoring prompt in Section C.2 in the Supplementary Materials.

²³O*NET 29.3 Database, SOC code 19-3011.00 and ISCO-08 Code 2631. O*NET tasks are given a scale of importance and so are weighted as per the methodology in Section 2.3.

analyzing data, providing economic advice and reviewing documents.

Given that the point estimates between the two datasets fall either side of 0.70—a commonly used automation threshold (Arntz et al., 2016; Muro et al., 2019)—the exposure score associated with the ISCO-08 dataset might suggest that the role of an economist should be fully automated, while the exposure score in the O*NET database suggests the opposite. These different datasets would have different policy implications for how AI will transform the role of economists.

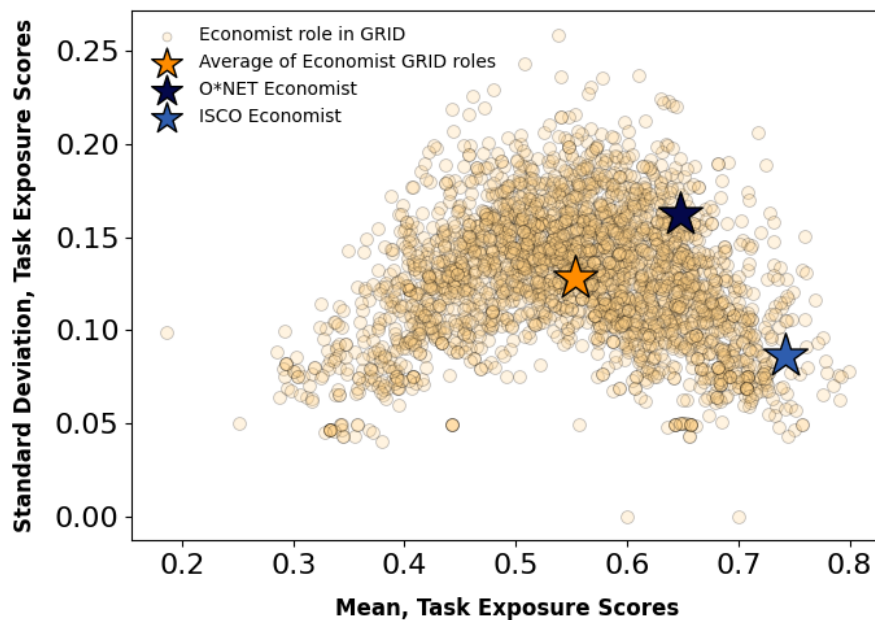


Figure 3: Economist roles exposure score comparison between GRID, O*NET and ISCO-08

“Economist role in GRID” comprises all roles in the Government Recruitment Information Database (GRID) with a vacancy title containing “economics” or “economist”, tagged in the Government Economic Service Profession, or the type of role contains “Economist”. Exposure scores for the GRID data are weighted as described in Section 2.3 and the weighted means and standard deviations are calculated accordingly. O*NET tasks are taken from the core set of tasks in O*NET 29.3 Database, SOC code 19-3011.00. O*NET tasks are ordered by their data value on the importance scale (‘IM’) and weighted using the same decay rate and process as with the GRID data. ISCO-08 tasks are listed for ISCO-08 Code 2631. They are unordered by importance so are simply assigned equal weighting.

Instead of using generalized task lists, our GRID dataset includes 2,773 economist roles with 20,986 unique tasks, providing a more granular picture of the AI exposure of economists, as shown in Figure 3. Rather than giving a single point estimate for all economists, we can instead provide a distribution of exposure scores that reflect the variety of tasks in different

economist roles in the UKCS. This provides a more nuanced opportunity in Section 3.4 to redesign each of the 2,773 economist roles according to whether AI can automate or augment tasks within the specific job profile.²⁴

3.2.3 Heterogeneity of AI exposure by grade

We find that the impact of AI is unevenly distributed across the organisational hierarchy. Roles classified as Low are typically concentrated among higher grades, where tasks involve strategic oversight rather than routine execution. This is reflected by 47% of all Senior Civil Servant (SCS) roles falling in this cluster, compared to 21% of AA/AO roles and 9% of EO roles. The distribution across departments and professions is relatively consistent. By contrast, the distribution of Augmentation across grades is relatively even. As with the Low cluster, Augmentation roles are slightly more common in higher grades, particularly Grade 6/7 (41%) and SCS (36%).

Summary statistics of the roles in each category show that grade varies significantly between exposure categories, much more than profession and department. Roles at lower grades are much more likely to have high exposure to AI because they carry out a much higher proportion of tasks categorised as Records Management and Admin Support (on average 60.4% of tasks at AA/AO grade roles in Automation are in these two categories), which are more likely to be substituted by AI. Across all grades, but particularly more visible in the Low and Augmentation categories, as seniority increases the proportion of tasks categorised as Stakeholder Engagement and Team Leadership increase. These responsibilities rely heavily on human input, making them less susceptible to automation.

These differences reinforce the underlying pattern: Lower graded roles are more exposed to AI as they typically focus on specific, high exposure tasks, while more senior roles remain relatively insulated because of the complexity, judgment, and leadership required in their day-

²⁴See also the plots for human resources managers and lawyers in Supplementary Materials H, Figure 13.

to-day responsibilities. This suggests that automation exposure is fundamentally driven at the task level, and that analysis of individual tasks is essential for identifying the real impact of AI across roles.

3.2.4 AI exposure mean and variance across roles

Figure 4 visualises the distribution of the four exposure clusters using all job descriptions in the GRID dataset. The figure shows that even roles within the same cluster can significantly vary in both the mean and variance of exposure score, and so we next focus on the reasons different roles are clustered into the same exposure category.

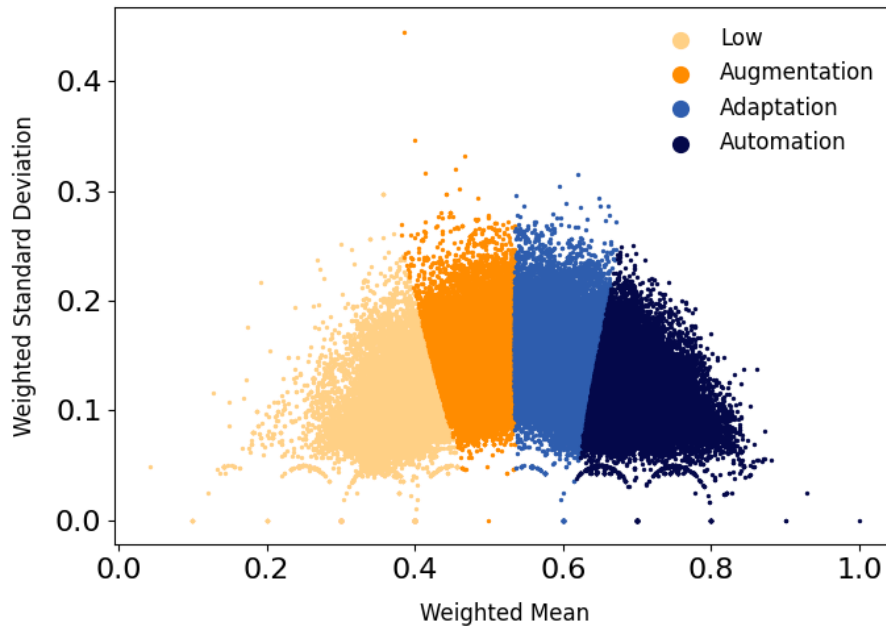


Figure 4: Job roles grouped into exposure clusters

Each job role is clustered into one of four groups using K-means clustering based on weighted mean and weighted standard deviation of the aggregate AI exposure score for tasks within each job description. Roles in the Low cluster have a low mean AI exposure score and low standard deviation; Augmentation roles have a low mean exposure score but a high standard deviation; Adaptation roles have a low mean exposure score but a high standard deviation; and Automation roles have a very high mean and low standard deviation.

Although it is self-evident that roles with a higher AI exposure score to be categorised into an exposure cluster with higher automation potential, the role of the standard deviation is

initially less clear. On the surface, we might expect to see a fundamental difference between a role with a mean exposure score of 0.7 and a very low standard deviation of 0.1 compared to a role with the same mean exposure score but a much higher standard deviation of 0.3. However, despite the considerable difference in task-level variation, both of these roles would be clustered in the Automation group.

To try to account for the differences between two roles in the same cluster with different standard deviations, we conduct a sampling of roles in each exposure cluster, comparing the number of tasks for roles in the 10th (high std) and 90th (low std) standard deviation percentile. Across all four exposure clusters, roles with the highest standard deviation have more tasks on average (8.5 tasks) than those with the lowest standard deviation (5.7 tasks), shown in Figure 5. The additional tasks are spread across the extremes, which explains the difference in variation.

For example, in the Automation cluster, high standard deviation roles have 0.9 more high-exposure tasks and 1.7 more low-exposure tasks than their low standard deviation counterparts. This broader task mix results in significantly greater variance, but the continued dominance of highly exposed tasks means the role remains classified within the Automation cluster given the high overall exposure to AI.

3.2.5 Task categorisation

To enhance our understanding of the common characteristics of roles in each exposure cluster, and exactly why roles from different professions, departments and grades are in the same exposure group, we analyse the types of tasks that are common in each category.

Figure 6 breaks down roles in each exposure cluster, showing how task types vary by grade across jobs with similar overall exposure profiles. This starts to show the commonalities between roles in each exposure category. Roles in the Low exposure group are characterised by a high concentration of tasks requiring human judgment, discretion, or interpersonal engage-

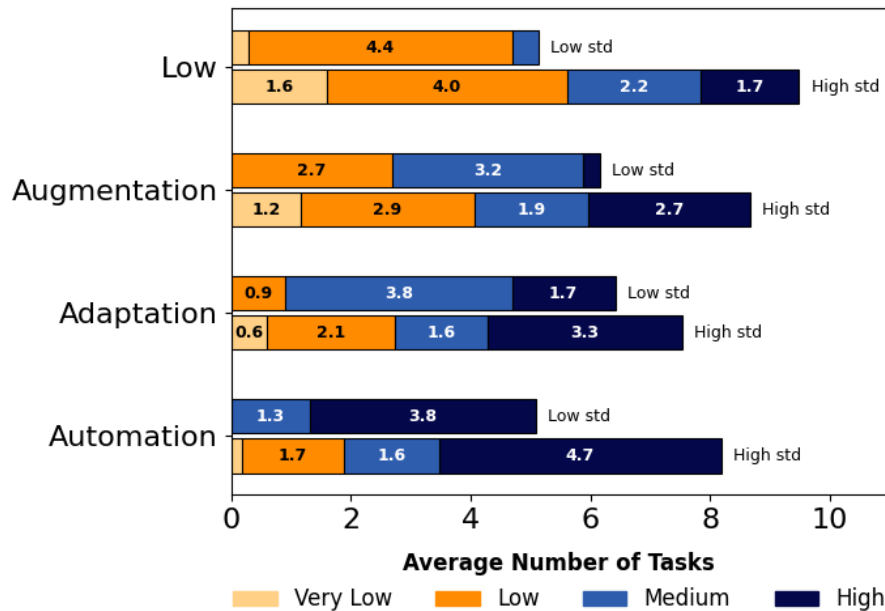


Figure 5: Number of tasks by automation category

For each exposure cluster, we present the average number of tasks for roles in the 10th percentile (Low std) and the 90th percentile (High std). The tasks are split according to the exposure score classification outlined in 2.2.

ment. Tasks categorised as Stakeholder Engagement, Team Leadership, and Policy Development make up the majority of work across many grades in this group. These are tasks that are not easily substituted with AI, explaining their consistently low exposure scores.

Conversely, roles in the Automation cluster have a higher proportion of tasks in Admin Support and Records Management, particularly in the lower grades. These are structured, repeatable activities that are more prone to automation by AI, especially where digital workflows already exist. The dominance of these task types helps explain why such roles are more exposed to displacement than augmentation.

The Augmentation and Adaptation clusters fall in between: Roles in these groups tend to contain a wider mix of task types, implying that AI may assist rather than displace human workers. Not being dominated by either highly automatable or entirely human-centric tasks, these roles typically involve a blend of activities and together represent around two thirds of

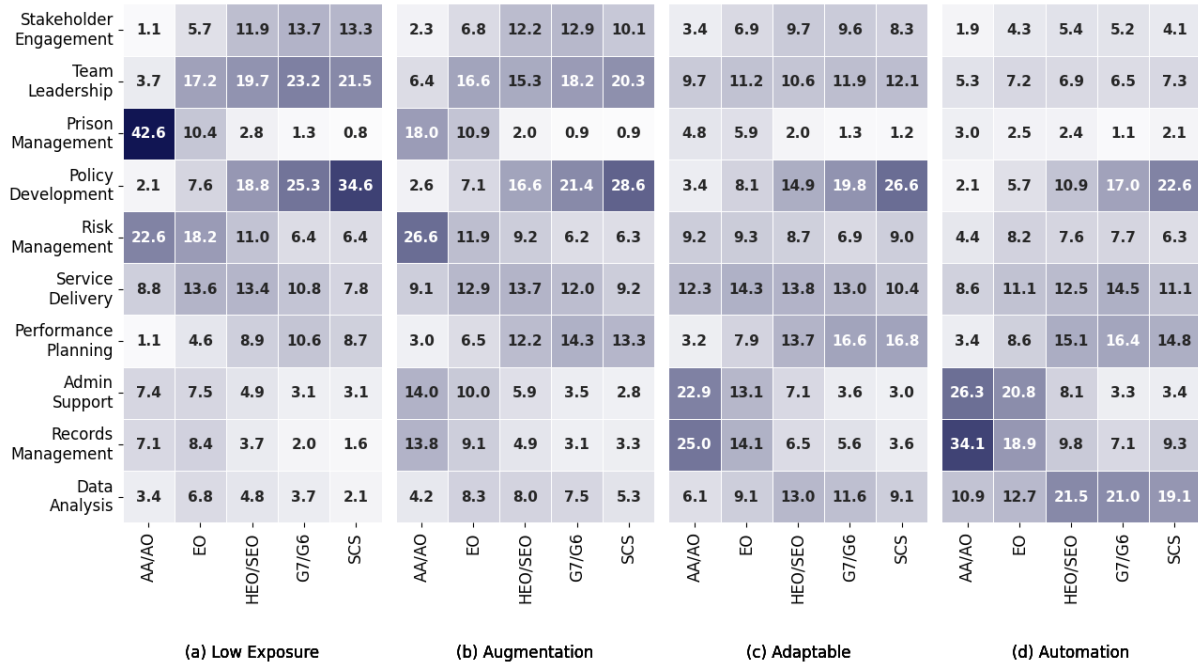


Figure 6: Percentage of time spent on each task category by grade

Each square shows the proportion of time spent in each grade that fall into each specific task type category, calculated using the methodology outlined in Section 2.3 and assuming a 37 hour working week. Since the columns represent the distribution of task types within each grade and exposure cluster, they sum to 100%, or 37 working hours. For example, 34.1% of working hours performed by roles in the Automation exposure category at grade AA/AO are categorised as Records Management. Task categories are ordered from lowest to highest average exposure score.

roles in our dataset. This variation in task types makes them well suited to partial automation, where AI tools automate or significantly streamline a handful of tasks, while leaving more judgment-based or collaborative elements to humans. As a result, roles in these categories are likely to evolve rather than disappear, with workers being increasingly supported by AI tools.

3.2.6 Sub-task categorisation

So far, we have used task types and job clustering to explain in detail why certain roles are more or less exposed to AI. Here, we deepen this analysis for each task type. One advantage of the bottom-up approach we take—using real job ads from a real employer—is that we are able to extract the actual tasks that a specific role is expected to complete, rather than a generic list of tasks. This means that jobs with seemingly similar job titles can be distinguished in much

greater detail.

For this analysis, we are therefore able to show not only what separates a low exposure task from a high exposure task generally, but also, for example, what separates a low exposure Policy Development task from a high exposure Policy Development task. To do so, we break each task type down further, into three subcategories for each.

In Section L in the Supplementary Materials, we show each of these task subcategories, and, to show the clearest variation, compares the proportion of tasks in each subcategory for roles in the Low and Automation groups. For example, we can see that, of all Data Analysis tasks completed by Low exposure roles, 59% of them are Research Analysis, compared to 24% for Automation roles. This provides deeper insight into the make-up of roles with high exposure to AI. These roles are not only characterised by a higher concentration of structured, repeatable tasks that are generally more susceptible to automation, but also by the nature of the specific task subcategories they perform within each broader task type. Even in task type with lower AI exposure (Stakeholder Engagement or Policy Development), individuals in roles in the Automation group tend to carry out task subcategories that are more automatable compared to those in Low exposure roles. This increased share in high exposure task subcategories helps explain why some roles are more vulnerable to automation, even when their overall area of work would typically suggest lower exposure. We can clearly see that, excluding Admin Support, roles in the Automation cluster spend significantly more time on high exposure task subcategories than roles in the Low cluster.

The analysis also shows that roles classified as Low exposure are not completely insulated from high exposure tasks. Roles that may appear similar at a high level (for example, two Operational Delivery roles both working predominantly on Service Delivery tasks) can have different exposure profiles once the specific subtasks they perform are considered. This variation at the task subcategory level helps explain why roles with similar titles or functions can experience

very different impacts from AI. Together, these analyses highlight that understanding the impact of AI exposure requires looking beyond broad categories to the detailed composition of work within each role.

3.3 Impact of AI

The AI exposure scores of tasks within job descriptions suggest AI could impact jobs across the UKCS. To explore this further, we first quantify the potential opportunity in monetary terms.²⁵ We break down the opportunity into potential cost reductions and productivity gains, highlighting that the impact of AI is not uniform across all tasks and job roles within the UKCS. Second, for those roles that we suggest can achieve productivity gains, we seek to model how these workers are likely to spend any freed up time from the automation of high exposure tasks.

3.3.1 Savings opportunity versus productivity gains

Figures 7a and 7b illustrate the potential opportunity from AI across the UKCS in both potential cost reductions and productivity gains. We find the potential savings show large variation depending on the chosen threshold of time spent on high exposure tasks. A lower threshold, such as those between 0%-30%, show potential cost reductions of more than £10bn, but it is not realistic to consider a role that spends just 30% of time on highly exposed tasks to be in scope for full job automation.

At our chosen threshold level of 80% where, for a role to be considered in scope for potential redundancy and therefore cost reductions, the role holder must spend more than 80% of time on high exposure tasks, we find that 34,182 roles (6.7% of FTE (Full Time Equivalents)) would be

²⁵After multiplying the salary of each role by the weight produced from the IPF procedure, we estimate a total salary cost of the UKCS of £20bn and a mean salary of £39,819. This highlights the importance of our IPF procedure to make our results more applicable to the UKCS. As outlined in Section 1.1.3, our raw sample from GRID has a median salary of £42,776. Our new mean salary post IPF is now close to the published UKCS mean salary of £40,700. (Mean salary are taken from [2025 Civil Service Statistics Data Browser](#).) The fact that the mean salary of our dataset post IPF is lower than the actual mean salary of UKCS, means that we can therefore be more confident in the order of magnitude of the scale of potential opportunity that we put forward.

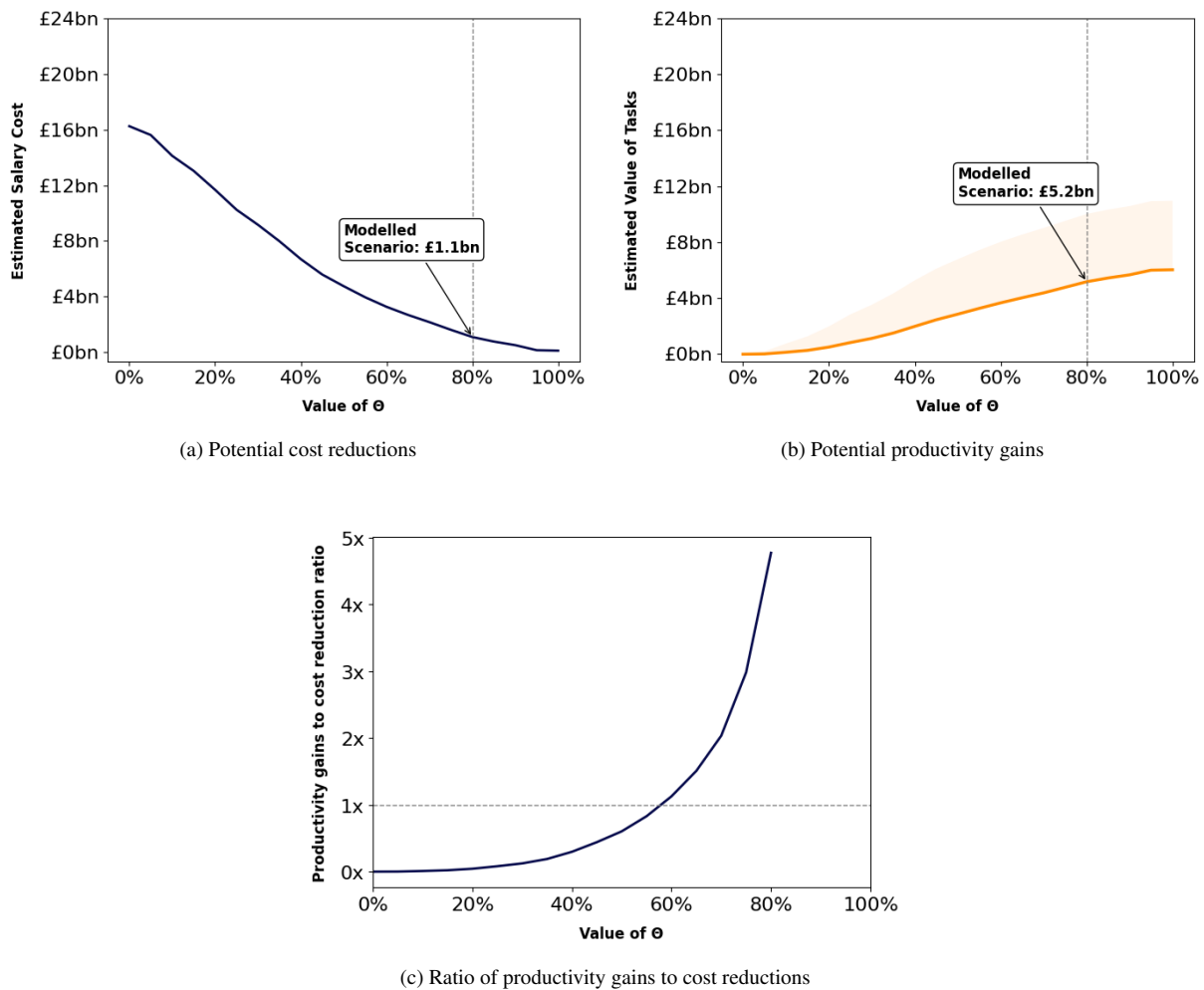


Figure 7: Estimated monetary opportunity from AI implementation

For all graphs, the x-axis shows the parameter θ — as per Section 2.5 this represents the minimum proportion of time a job role must spend on high exposure tasks to be considered in scope for potential cost reductions. Figure 7a sums the entire salary cost of each role meeting a given threshold level. Figure 7b sums the value of time spent on high exposure tasks for any roles that do not meet the threshold. For the upper bound, represented by the shaded area, this adds the value of time spent on medium exposure tasks for the same roles. Figure 7c shows the ratio of productivity gains to cost reductions, indicating when productivity improvements outweigh direct salary savings. The dashed line, representing a ratio of 1x, is the point where productivity gains outweigh potential cost reductions.

displaced²⁶. Furthermore, the automation of high exposure tasks across a further 392,065 roles (76% of FTE) would lead to 5.2m working hours per week being freed up.²⁷ For the remaining 89,901 roles (18% of FTE), no impact of AI is anticipated.

²⁶This is equivalent to 5.3% of UKCS total salary costs, or cost savings of £1.1bn.

²⁷In monetary terms, this equates to £5.2bn. Depending on the scale of AI implementation and developments in technology, these productivity gains could increase towards an upper bound of £10bn, saving 9.6m working hours.

The results outlined here rely on our assumption for the amount of time the role holder spends on each task in their job description. In Section I, we show the change in the total estimated opportunity at different decay rates. Broadly, we find changing the decay rate leads to only minor variation in the estimated opportunity across the majority of threshold levels. Figure 16b shows that there is relatively little change to productivity gains and cost savings with a decay rate of 1.0 (i.e. all tasks are weighted equally) but that a decay rate of 0.5 leads to more substantial changes, primarily increasing the weight of the potential cost reductions. This reinforces the risk outlined in Section 2.3 that too high a decay rate would likely place too much weight on the first task and skew the results. We are therefore satisfied that our chosen decay rate of 0.75 is a sensible middle ground for our central estimate of potential savings. Figures 15 and 16 in the Supplementary Materials show the impact of the decay rate across these threshold levels.

Figure 7c shows the ratio of the value of estimated productivity gains compared to the value of potential cost reductions. At our modelled scenario of $\theta = 0.8$, productivity gains are four times higher than the value of potential cost reductions. More broadly, at all threshold levels of 60% and above, the value of productivity gains outweigh the value of potential cost reductions, represented by the dashed line. These higher threshold scenarios, where at least 60% of working time for a given role is spent on high exposure tasks, represent the most realistic scenarios, given that we believe that it will be extremely difficult at this point for any organisation to completely remove roles where less than half their time is spent on high exposure tasks. This finding implies that, given the current task profile of job roles and potential capabilities of AI, the majority of economic value is expected to arise from increased productivity, rather than mass role displacement.

In Figure 8, we compare the proportion of roles in each of the four exposure clusters created in Section 3.2.1 to the threshold values of θ . This comparison shows the robustness of both the

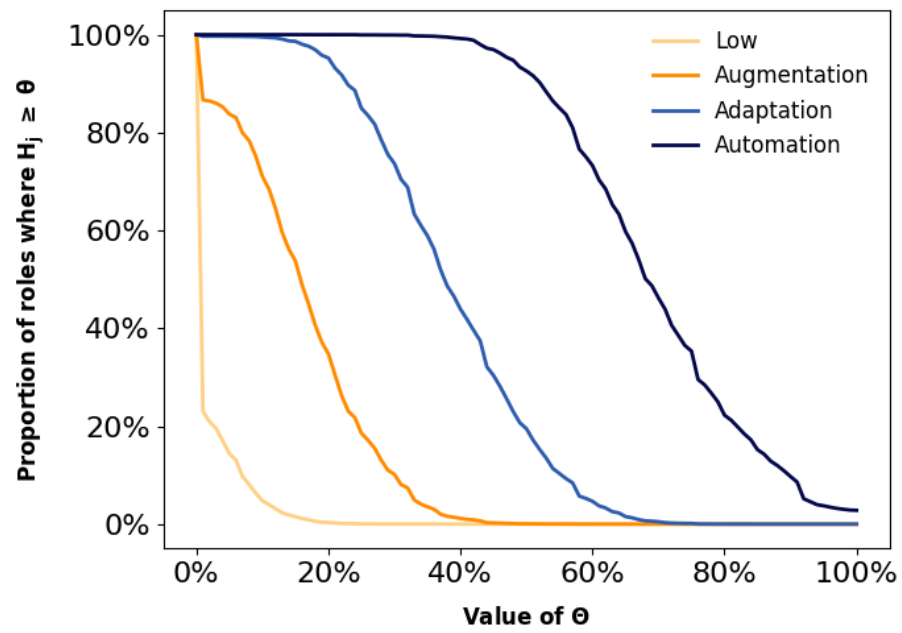


Figure 8: Proportion of roles meeting threshold

Figure shows the proportion of roles in our sample of jobs that meet each given threshold value of θ - in Section 2.5, this was outlined as the proportion of time spent on high exposure tasks.

threshold analysis and the exposure clusters. We find a consistent pattern in the distribution of high exposure tasks across clusters. Roles in the Automation cluster retain a high proportion of high exposure tasks even at elevated thresholds, whereas roles in the Low and Augmentation clusters decrease rapidly as the threshold increases. The strong alignment between the cluster-based and threshold-based approaches, despite being derived from separate methodologies, provides mutual reinforcement and validation of both frameworks in capturing meaningful variation in AI exposure.

3.4 Job Redesign Through Task Optimisation

The previous set of results speaks to the importance of augmentation as a key opportunity to reap the benefits from integrating AI across the workforce. Having quantified the opportunities for time savings in each role, we now focus on how roles in scope for productivity gains might be transformed. These roles have at least one high exposure task but are not automatable based on our modelled threshold level of $\theta = 0.8$. As noted previously in Section 3.2.5, these roles typically belong to the Adaptation and Augmentation clusters, and they make up the majority of our dataset. They are characterised by a mix of tasks which are neither totally automatable nor entirely human-centric.

To analyse the impact of productivity gains, we seek to understand how time is freed up from the implementation of AI and how that extra time might be used. For instance, if a worker saves half an hour per day which they then take as free time, no productivity benefits will arise. By contrast, if they allocate the extra time to the most valuable alternative—the opportunity cost—then the benefits for the business will be maximised.

Let us first consider the distribution of total weighted time (see Section 2.3) per role that could be freed up with automation of high exposure tasks. The median FTE share saved is 0.31, meaning that half of people in these non-automatable roles could have over 30 percent of their

time freed up from task automation. The lower quartile is 0.17, the upper quartile is 0.50 and so the interquartile range is 0.33. This begs the question: what might they do with their time?

3.4.1 Productivity gain through deepening existing focus tasks

From the remaining non-automatable tasks for each role, we require the LLM²⁸ to select a “focus task”, which is defined as where the worker should focus their efforts to maximise productivity (see Prompt 1 in Appendix C). This will mean that the time freed up from tasks being automated will be reallocated to a single focus task. The LLM also provides reasoning for each of these focus tasks, with each assigned up to three categories from the themes below.

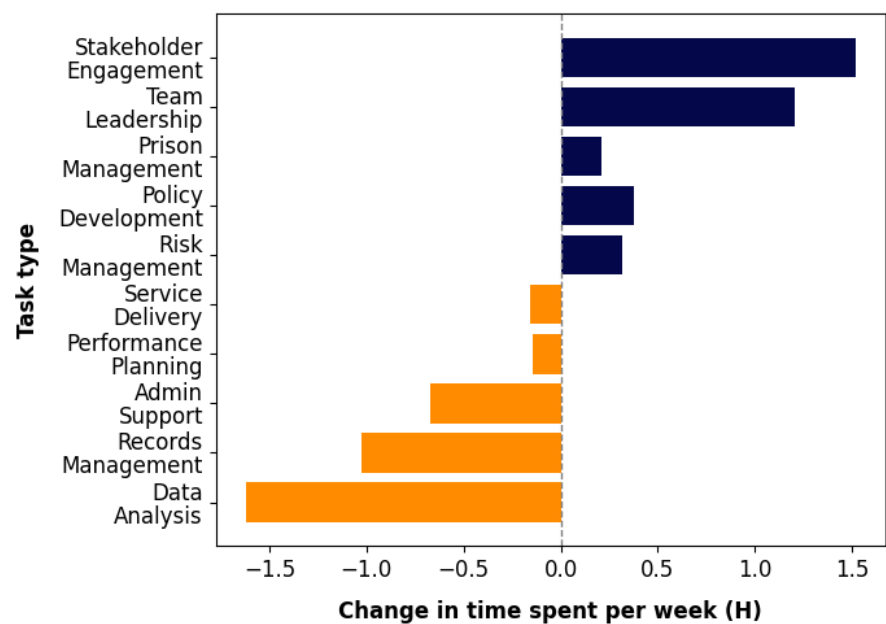


Figure 9: Change in time spent per week following automation of high exposure tasks, assuming a 37 hour week and each role equivalent to 1 FTE.

Table 3 shows the six identified themes and the share that each theme makes up of all focus tasks. To put this into context, we compare the total time spent before and after the tasks with

²⁸Claude Sonnet 3.5 v2, the same foundation model used throughout our analysis. See Section 2.1 for further details.

Table 3: Reasoning themes for focus tasks

Label	Description	Share
Strategic Leadership and Vision	Tasks involving setting strategic direction, developing policies, and leading organisational transformation, including decision-making that influences long-term outcomes and establishes foundational frameworks	26%
Complex Problem Resolution	Tasks requiring sophisticated analysis, investigation, and critical thinking to solve multifaceted problems, particularly in high-stakes situations that demand human judgment	18%
Stakeholder Management and Communication	Tasks focused on building and maintaining relationships with stakeholders, including complex communication, knowledge transfer, and effective engagement across different audiences	17%
Innovation and Process Excellence	Tasks centered on continuous improvement, modernisation, and transformation of systems and processes, with emphasis on maximising value and creating multiplier effects across the organisation	14%
Human-Centric Leadership	Tasks requiring uniquely human capabilities including team development, mentoring, empathy, and providing personalised services that cannot be automated	14%
Risk and Quality Management	Tasks involving risk assessment, security oversight, compliance monitoring, and quality assurance to maintain operational integrity and safety standards	11%

Share totals may not sum to one due to rounding to one decimal place. Each focus task allowed up to three thematic tags.

high exposure are automated and replaced by the role's focus task. Figure 9 depicts reduced time spent on highly exposed task categories and increased time spent on areas such as Stakeholder Engagement and Team Leadership, where human input and judgement is especially valuable.²⁹ Specifically, if we consider an average worker in the UKCS who works 37 hours with potential for productivity gains, then on a weekly basis this would mean they spend 40 minutes less on administration support, 62 minutes less on records management, and 97 minutes less on data

²⁹See also the work by [Loaiza and Rigobon \(2024\)](#) whose work using the O*NET data also finds a shift toward more human-intensive tasks in the US since 2016.

analysis. By contrast, they would increase time on other task types: 22 minutes on policy development, 72 minutes on team leadership, and 91 minutes on stakeholder engagement.³⁰ Moreover, on average, focus tasks are placed higher than automatable tasks in role descriptions, suggesting that employers also place more importance on them.

3.4.2 Impacts of redesign across job grades

While Figure 9 shows the change in time spent for the average worker with potential productivity gains, we have previously shown how the impacts of automation on roles likely vary according to seniority (see Figure 6), with those in the lower grades being more exposed to AI. As a result, the new task makeup of each role will in part be determined by the role's grade.

The grade context also shapes which kinds of tasks the redesigned role would be doing. For example, someone working in an administrative role is unlikely to be doing substantially more people management following the redesign of their role. Therefore, we examine the focus tasks and time reallocation more closely for the top and bottom deciles of roles in scope for productivity gains (see methodology in Section 2.7).

First, while those in the top decile (upper grades) have an average (median) 21 percent of their time freed up from automating highly exposed tasks, those in the bottom decile (lower grades) free up 48 percent of their time. In line with Figure 9, both groups experience declines in time spent on data analysis, records management and admin support, and more time allocated to stakeholder management and team leadership.

Second, however, these broad categories mask heterogeneity by grade. We exploit the second tier of our task categories (subcategories). While both deciles increase performance development by 39 minutes, those in lower grades spend an extra 23 minutes per week on training and professional development, while upper grades add just 5 minutes. By contrast, upper grades

³⁰For reference, the absolute shares of all task categories are provided in Figure 19 in the Supplementary Materials, Section M

spend an extra 37 minutes per week on team leadership and people management, while those on lower grades spend just an extra 10 minutes. This highlights that even when both deciles are spending more time on team leadership in general, the kinds of tasks this will lead to vary considerably.

Third, relative changes can be large, even if absolute changes in time spent are similar. For example, those in the upper grades spend an average of 252 minutes per week on stakeholder engagement, while those in lower grades spend just 78. However, following our job redesign exercise, the top decile spends a further 57 minutes stakeholder engagement, and the bottom decile by 81 minutes. While the change in time spent is of similar magnitude, note that the relative change for upper and lower grades is a 23 percent and 103 percent increase respectively. Therefore, though both groups will be doing more stakeholder management, this group of tasks will mean a much larger relative change for those in lower grades.

Fourth, we can also consider the focus tasks themselves. Whereas policy development comprises 25 percent of high grade focus tasks, they make up only 5 percent of low grade focus tasks. By contrast, the most common low grade focus tasks are risk management (18 percent), administrative support (17 percent) and records management (14 percent). This compares with 7 percent, 1 percent and 2 percent respectively for the high grade decile (see Figure 20 in the Supplementary Materials, Section M).

Finally, another lens to consider the differing nature of the focus tasks by grade is to examine their purported reasoning by the LLM. Over 70 percent of high grade focus tasks are grouped into strategic leadership and vision, and innovation and process excellence. Meanwhile, the majority of low grade focus tasks relate to risk and quality management, and human-centric leadership. This suggests that higher grades would spend more of their time on focus tasks that concern strategic direction and organisation transformation that creates multiplier effects across teams. By contrast, lower grades would spend more time on ensuring operational in-

tegrity, maintaining security, and performing tasks requiring uniquely human capabilities such as mentoring and empathy (see Figure 21 in the Supplementary Materials, Section M).

3.4.3 Robustness: Task optimisation with AI augmentation of existing tasks

An extension of our framework explores to what extent tasks would not just be automated and then replaced by a “focus task”, but also the possibility for AI to potentially augment all remaining tasks.

Specifically, once the automatable tasks are removed from roles, we allow the LLM to suggest how the remaining tasks may look following AI transformation, and to reorder them based on their importance (see prompt 4 in Appendix C). This additional analysis offers both qualitative and quantitative insights. First, we are provided with a set of tasks that have been reimagined. In essence and category, they remain the same, but the LLM also describes how each task can be augmented with AI. By providing the vacancy title and job context (Karasek Jr, 1979), recommendations can be highly contextually relevant.

For example, we consider the role of Head of Property Strategy at HMRC. The original task of “Implement holistic strategic asset management with HMRC’s estates management systems and governance arrangements” was reimagined as “Direct implementation of AI-enhanced strategic asset management systems, leveraging predictive analytics and IoT integration for proactive governance. Focus on strategic oversight and exception handling while automated systems handle routine monitoring and reporting.” The suggestion of using IoT devices would mean that instead of waiting for an asset to fail, this worker could use real-time data from sensors to predict and address issues before they become costly emergencies. This would fundamentally shift asset management from being a reactive role of fixing faults, to a proactive one of preventing them from occurring in the first place. In this world, buildings would not be merely passive assets but rather data-producing environments. The LLM also shifts the focus

of the task from reporting to strategic oversight.

Second, if we do not restrict the LLM to choose a single focus task, but rather allow it to reorder all remaining tasks in order of importance after they have been augmented by AI, and reallocate time proportionally, we get similar results to our main analysis focusing only on the focus task (see Table 14 for comparison). As the large changes in time spent occur between the pre-and post-automation shares ahead of any reallocation of time, most of the impact of AI appears to come from the automation and removal of tasks, rather than a reordering and reprioritisation between tasks.

3.4.4 Robustness: Task optimisation using new tasks

Acemoglu and Restrepo (2018) suggest that the automation of some tasks can also lead to the creation of entirely new tasks in which humans have comparative advantage (Acemoglu and Restrepo, 2018). We explore this possibility by expanding our main approach to enable the LLM to create new tasks that are not within the current job profile (see Prompt 5 in Appendix C.4 for full details).

A qualitative inspection of the outputs reveals that the LLM is able to provide a contextually relevant mix of old and new tasks that reshape each role. As the maximum number of new tasks is limited to how many tasks were automated, roles with more high exposure tasks are transformed, as one might expect. The new tasks largely consist of using AI tools, checking their output, or establishing frameworks and risk management to ensure their responsible use. However, not all new tasks are purely AI, with some relating to digital transformation or team strategy, for example, implying that the LLM is able to assess where potential gaps might be in what is important to the role.

Quantitatively, the results are more extreme than the previous job redesign analyses. Some task categories plummet in their share of time: administrative support and records manage-

ment decline from around nine percent of time pre-transformation to just two percent post-transformation, reflecting that many of these tasks are automated, few new tasks of this kind are proposed, and remaining tasks in these categories are not deemed important to the role and reordered downwards. By contrast, stakeholder engagement rises from nine percent to 12. Similarly, the time share for data analysis increases from 11 percent to 15.³¹

Most starkly, the share of time dedicated to risk management rises from nine percent to over 21 percent. This is due to a large number of new tasks (which are highly prioritised and therefore gain importance in the overall job) that relate to the responsible deployment of AI systems and their associated frameworks. Tasks in this category include helping to upskill others to use new tools, verifying and checking the outputs of AI systems, performing strategic oversight of AI, mitigating bias, and so forth. However, this likely overstates the case for each role: As the LLM only has each role for context, managing the risk of AI transformation is localised to each role. In reality, as AI transforms workflows, there will likely be a division of duties between central enabling teams (perhaps which do not yet exist), and those who use AI tools.

These results provide some initial evidence that allowing the creation of new tasks could shift job profiles even more than our main results suggest. While we believe that this is feasible, we expect these changes to occur over the medium to longer-term, as they would require greater upskilling and reskilling of workers in existing roles. Such role redesigns resulting from AI transformation would potentially add pressure to further develop new and existing expertise for job roles (Autor and Thompson, 2025).

³¹Surprisingly, this occurs despite a decline in data analysis tasks pre- and post-transformation, as more data analysis tasks are automated than are proposed as new tasks. As time dedicated to data analysis rose nevertheless, this reveals the rising perceived importance of data analysis tasks.

4 Discussion

We present a new framework to analyse job roles at the task level for their AI exposure and build an end-to-end pipeline for redesigning jobs where AI only partially automates and augments some but not all tasks. In this study, we perform this analysis on UKCS job roles and demonstrate that productivity in nearly two thirds of jobs could be optimised by replacing automatable tasks with those where humans have comparative advantage, such as leadership, vision, and stakeholder management. By building from a bottom-up task-based framework, we provide granular estimates of AI exposure, bridging the explainability gap between aggregated statistics and micro foundations. Using this information, we consider how potential job redesigns could allow freed-up time from automation to be focused on valuable human-centric tasks, thereby improving productivity. Overall, we found that most economic value arises from productivity gains rather than role displacement.

While the investigation in this paper focused on the impact of AI in the UKCS, we believe our approach can be used for any organisation with job descriptions and/or vacancies data. One of the key take-aways from our analyses is that heterogeneity and context matter for more precise estimations of AI exposure for each role and how it can be redesigned as a result. We therefore hope that more granular assessments can be performed using our methodology for other organisations to optimise productivity. Initial evidence suggests the potential productivity gains from AI in other public sector organisations could be even more significant than those outlined for the UKCS: healthcare professionals spend an average of 13.5 hours per week creating or adding to clinical documentation ([Nuance Communications, 2023](#)); 75% of teachers reported general administrative work took up 'too much' of their time ([Department for Education, 2024](#)); and 54% of police officers spend at least 2 hours per day on paperwork ([Nuance Communications, 2021](#)). We believe the scope for transformation and task redesign across the

public and private sectors is considerable.

As with all studies of this kind, there is significant uncertainty in the degree to which automation will unfold in the coming years and to what extent the capability of AI will continue to expand as rapidly as it has in the past five years. That said, our study provides some guidance as to where organisations should focus their attention when identifying opportunities for AI to automate or augment tasks. Our approach focused on identifying tasks that are estimated to be ripe for automation—expanding on previous work (Felten et al., 2023; Gmyrek et al., 2023; Eloundou et al., 2024) by showing more heterogeneous and granular estimates of AI exposure across roles—and re-designing them to optimise the value of their “focus task”. However, beyond organisationally relevant outcomes such as productivity (e.g. Noy and Zhang (2023); Brynjolfsson et al. (2025b)) or innovation (e.g. Doshi and Hauser (2024); Boussioux et al. (2024)), AI-augmented job redesign could further focus on employee motivation (Hackman and Oldham, 1976), autonomy (Karasek Jr, 1979), or other characteristics that employees value (Maestas et al., 2023).

We caution that there are some limitations of our study. First, we weighted the tasks according to where they appear in the job description, assuming that tasks mentioned earlier are traditionally viewed as more central to a role and therefore occupy more of its time. Our robustness analyses weighing tasks more equally show similar trends, suggesting this modelling assumption is not central to our conclusions. Second, the decision on how to assign total working time across tasks in a job description makes the implicit assumption that a job description is fully representative of the tasks that the role holder will undertake. In reality, job descriptions are often promotional products aimed at attracting applicants to apply for a job role. The tasks listed may therefore be aspirational and represent the most interesting and engaging elements of a job role. As a result, our analysis does not encompass the entire spectrum of tasks a job would normally entail. The types of tasks typically overlooked in job descriptions may be more

routine, admin-based tasks, which could be automated or augmented by AI. In other words, it is possible that there is more time savings potential as well as redesign potential than our study shows. Future studies may wish to assess the actual time spent on various tasks on the job through surveys, time use diaries, or time tracking software.

To reap the benefits resulting from AI adoption, organisations will need to make decisions about the extent to which AI should act as a complement or substitute for each jobs. This will likely be an endogenous decision depending in various firm and industry characteristics. Our framework captures this by introducing an automation threshold θ , which we conservatively set at 80%. Unlike previous literature that focuses on AI exposure alone and therefore remains agnostic about AI as a complement or substitution (Felten et al., 2021, 2023), our study speaks to this debate by adding a “redesigning” stage after estimating AI exposure. By identifying which tasks employees could focus on when AI frees up some of their time, we can estimate the value to the firm conditional on a chosen θ . Our results suggest that many firms will find that productivity benefits from redesigning a role will outweigh the direct cost savings from making the role redundant, unless the automation potential is very high. For jobs that are redesigned, our method makes a prediction of which tasks an employee should focus on to add value to the firm by doing work where they have a comparative advantage over AI. Future research may wish to dive into the process and dynamics of AI-induced job redesign. For instance, job crafting (Wrzesniewski and Dutton, 2001), where employees play an active role in shaping their role and work, seems like an important area of work, which can may also help reduce anxiety around AI adoption during this transition.

Indeed, as the adoption of AI accelerates across companies and industries, the scale of the challenge of preparing the workforce to use AI effectively will also grow. Our main results focus on employees shifting to human-centric work, leadership, relationship building, and risk management when AI frees up time in their role. This shift will lead to doing more work that

is similar to work employees may have done in the past. However, in an extension of our approach, we have also considered that AI adoption may lead to the creation of entirely new roles ([Acemoglu and Restrepo, 2018](#)). Many of these new tasks involve the supervision and risk management of AI systems that will become more prevalent in the future ([Law and Varanasi, 2025](#)). In both scenarios, the implications for workforce upskilling are considerable, likely requiring substantial time and resource investments in the development of new human capital.

References

- Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American economic review*, 108(6): 1488–1542, 2018.
- Daron Acemoglu and Pascual Restrepo. Robots and jobs: Evidence from us labor markets. *Journal of political economy*, 128(6):2188–2244, 2020.
- Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo. Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1):S293–S340, 2022.
- Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. Do we want less automation? *Science*, 381(6654):155–158, 2023.
- Melanie Arntz, Terry Gregory, and Ulrich Zierahn. The risk of automation for jobs in oecd countries: A comparative analysis. OECD Social, Employment and Migration Working Paper 189, OECD Publishing, 2016. URL <https://doi.org/10.1787/5j1lz9h56dvq7-en>.
- David Autor and Neil Thompson. Expertise. *Journal of the European Economic Association*, page jvaf023, 2025.
- David H Autor, David Dorn, and Gordon H Hanson. Untangling trade and technology: Evidence from local labour markets. *The Economic Journal*, 125(584):621–646, 2015.
- Björn Bartling, Ernst Fehr, and Klaus M Schmidt. Screening, competition, and job design: Economic origins of good jobs. *American Economic Review*, 102(2):834–864, 2012.
- Léonard Boussioux, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R

- Lakhani. The crowdless future? generative ai and creative problem-solving. *Organization Science*, 35(5):1589–1607, 2024.
- Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.
- Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn and what does it mean for occupations and the economy? In *AEA papers and proceedings*, volume 108, pages 43–47. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2018.
- Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. Canaries in the coal mine? six facts about the recent employment effects of artificial intelligence. *Stanford Digital Economy Lab*. Published August, 2025a.
- Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work*. *The Quarterly Journal of Economics*, 140(2):889–942, 02 2025b. ISSN 0033-5533. doi: 10.1093/qje/qjae044. URL <https://doi.org/10.1093/qje/qjae044>.
- Bharat Chandar. Tracking employment changes in ai-exposed jobs. Available at SSRN 5384519, 2025.
- Sam Choo. Generative ai prompt engineering for educators: Practical strategies. *Education and Information Technologies*, 29(1):1–22, 2024. doi: 10.1007/s10639-024-12247-5. URL <https://doi.org/10.1007/s10639-024-12247-5>.
- Fabrizio Dell’Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraye, François Cadelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on

knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 24(013), 2023.

Ozge Demirci, Jonas Hannane, and Xinrong Zhu. Who is ai replacing? the impact of generative ai on online freelancing platforms. *Management Science*, 2025.

Department for Education. Working lives of teachers and leaders: Year 1 core research report. Technical report, Department for Education, 2024. URL https://assets.publishing.service.gov.uk/media/66f673e03b919067bb482842/Working_Lives_of_Teachers_and_Leaders_-_Year_1_Core_Research_Report.pdf.

Jacob Dominski and Yong Suk Lee. Advancing ai capabilities and evolving labor outcomes. *arXiv preprint arXiv:2507.08244*, 2025.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024.

Edward Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021. doi: <https://doi.org/10.1002/smj.3286>. URL <https://sms.onlinelibrary.wiley.com/doi/abs/10.1002/smj.3286>.

Edward W Felten, Manav Raj, and Robert Seamans. Occupational heterogeneity in exposure to generative ai. *Available at SSRN 4414065*, 2023.

Enrico Maria Fenoaltea, Dario Mazzilli, Aurelio Patelli, Angelica Sbardella, Andrea Tacchella, Andrea Zaccaria, Marco Trombetti, and Luciano Pietronero. Follow the money: a startup-based measure of ai exposure across occupations, industries and regions. *arXiv preprint arXiv:2412.04924*, 2024.

Morgan R Frank, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539, 2019.

Michael Bazylik Gibbs et al. How is new technology changing job design? *IZA World of Labor*, 2022.

Pawel Gmyrek, Janine Berg, and David Bescond. Generative ai and jobs: A global analysis of potential effects on job quantity and quality. *ILO Working paper*, 96, 2023.

Jie Gong and Ivan PL Png. Automation enables specialization: Field evidence. *Management Science*, 70(3):1580–1595, 2024.

J Richard Hackman and Greg R Oldham. Motivation through the design of work: Test of a theory. *Organizational behavior and human performance*, 16(2):250–279, 1976.

Menaka Hampole, Dimitris Papanikolaou, Lawrence DW Schmidt, and Bryan Seegmiller. Artificial intelligence and the labor market. *NBER Working Paper*, 2025.

Golo Henseke, Rhys Davies, Alan Felstead, Duncan Gallie, Francis Green, and Ying Zhou. How exposed are uk jobs to generative ai? developing and applying a novel task-based index. *arXiv preprint arXiv:2507.22748*, 2025.

HR Daily Advisor. 5 critical components every job description must contain, 2012. URL <https://hrdailyadvisor.com/2012/02/15/5-critical-components-every-job-description-must-contain/>.

Anders Humlum and Emilie Vestergaard. Large language models, small labor market effects. Technical report, National Bureau of Economic Research, 2025.

Indeed Editorial Team. How to write your own job description, n.d. URL <https://www.indeed.com/career-advice/career-development/how-to-write-your-own-job-description>.

Robert A Karasek Jr. Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative science quarterly*, pages 285–308, 1979.

Bouke Klein Teeselink. Generative ai and labor market outcomes: Evidence from the united kingdom. *Available at SSRN*, 2025.

Matthew Law and Rama Adithya Varanasi. Generative ai and changing work: Systematic review of practitioner-led work transformations through the lens of job crafting. In *International Conference on Human-Computer Interaction*, pages 131–152. Springer, 2025.

Guy Lichtinger and Seyed Mahdi Hosseini Maasoum. Generative ai as seniority-biased technological change: Evidence from us resume and job posting data. *Available at SSRN*, 2025.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the bi-

ology of a large language model. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.

Isabella Loaiza and Roberto Rigobon. The EPOCH of AI: Human-machine complementarities at work. Working Paper SSRN ID: 5028371, MIT Sloan School of Management, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5028371.

Nicole Maestas, Kathleen J Mullen, David Powell, Till Von Wachter, and Jeffrey B Wenger. The value of working conditions in the united states and implications for the structure of wages. *American Economic Review*, 113(7):2007–2047, 2023.

Giacomo Marzi, Marco Balzno, and Davide Marchiori. K-alpha calculator–krippendorff’s alpha calculator: A user-friendly tool for computing krippendorff’s alpha inter-rater reliability coefficient. *MethodsX*, 2024. doi: 10.1007/s10763-025-01006-7. URL <https://doi.org/10.1016/j.mex.2023.102545>.

Christian Meynhardt, Peter Meybohm, Pascal Kranke, and Christian Holzing. Advanced prompt engineering in emergency medicine and anaesthesia: Enhancing simulation-based e-learning. *Electronics*, 14(5):1028, 2025. doi: 10.3390/electronics14051028. URL <https://doi.org/10.3390/electronics14051028>.

Mark Muro, Robert Maxim, and Jacob Whiton. Automation and artificial intelligence: How machines are affecting people and places. Report, The Brookings Institution, 2019. URL https://www.brookings.edu/wp-content/uploads/2019/01/2019.01_BrookingsMetro_Automation-AI_Report_Muro-Maxim-Whiton-FINAL-version.pdf.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Nuance Communications. The role of technology in law enforcement paperwork: Annual report. Technical report, Nuance Communications, 2021. URL https://www.nuance.com/content/dam/nuance/en_us/collateral/dragon/report/rpt-2021-role-of-technology-in-law-enforcement-paperwork-annual-report-en.pdf.

Nuance Communications. Assessing the burden of clinical documentation. Technical report, Nuance Communications, 2023. URL <https://www.nuance.com/en-gb/healthcare/campaign/report/assessing-the-burden-of-clinical-documentation.html>.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.

Fabian Stephany and Ole Teutloff. What is the price of a skill? the value of complementarity. *Research Policy*, 53(1):104898, 2024.

Maja Svanberg, Wensu Li, Martin Fleming, Brian Goehring, and Neil Thompson. Beyond ai exposure: which tasks are cost-effective to automate with computer vision? *Available at SSRN 4700751*, 2024.

The Muse. How to read job description keywords, n.d. URL <https://www.themuse.com/advice/how-to-read-job-description-keywords>.

Amy Wrzesniewski and Jane E Dutton. Crafting a job: Revisioning employees as active crafters of their work. *Academy of management review*, 26(2):179–201, 2001.

Supplementary Materials

A A. Data description

A.1 Government Recruitment Information Database

The GRID comprises a backend feed of Civil Service Jobs postings. Civil Service Jobs is a tool offered by the Government Recruitment Service to facilitate the filling of roles.³² These job adverts are accessible to existing Civil Servants with registered accounts on Civil Service Jobs and/or to any member of the public at the time of recruitment. However, historic data is not publicly available: adverts are removed once applications close, and the Government does not publish an archive or database of past job adverts. We obtained an extract of all roles published on Civil Service Jobs between 16th January 2019 and 3rd December 2024, representing almost six years worth of data.

Table 4: Variables in the GRID extract

Variable	Description
Vacancy ID	A unique identifier assigned to the job advert
Vacancy title	Job title of the role
Department	The organisation holding the vacancy. This could be a department, agency, arms-length body, or devolved administration.
Posting date	The date and time the role was published on Civil Service Jobs
Closing date	The date and time the role closed for applications on Civil Service Jobs
Which profession is this job	The UKCS profession associated with the role — selected from 28 options or ‘Other’
Type of role	The civil service type(s) of role
Describe the type of job	Short job title, if profession is unavailable
Job summary	Normally an overview of the department and team
Job description	Normally a description of the role and the duties associated with it
Job grade	The Civil Service grade associated with the role

Some job adverts have the job summary and description swapped; these are processed twice by the LLM to split out job tasks correctly. See Section 2.2.

The eleven variables in the GRID extract are summarised in Table 4. Coverage varies signif-

³²See <https://www.civilservicejobs.service.gov.uk/>

icantly. Our main variables of interest (vacancy ID, department, job summary, job description, job grade) have very high coverage, but the grouping variables (profession, role, type) are often either missing or classified as ‘Other’. However, these are not material for the main body of our analysis.

A.1.1 Choice of departments

The raw GRID extract contains 187 unique departments. We begin by filtering to focus only on entities with circa 1,000 vacancies or more, and for which we have corresponding full-time equivalents (FTE) and salary data split out by grade in the UKCSS. The latter condition means that some of the larger organisations, such as Public Health England, are removed from consideration.

Nevertheless, this approach enables us to focus on entities of a minimum size and, crucially, for which we can produce estimates for staff costings via salary and grade information. Following this selection process we are left with 37 departments. The LLM then extracts each task from each role’s job description and assigns an AI exposure score (see section 2 for further methodological details).

A.2 Civil Service Statistics Bulletin

Following the departmental selection and the extraction and categorisation of role tasks, we join data from two tables in official UKCSS.³³ These statistics are sourced from all UKCS organisations via the Annual Civil Service Employment Survey (ACSES).

We first use Table 21 (responsibility by department) which breaks down each Civil Service department into full-time equivalents (FTE) per grade. Due to small numbers between one and four, some data is suppressed as confidential and simply marked as “c”. Second, we use Table 25 (median salary by responsibility level and government department) which breaks down

³³See <https://www.gov.uk/government/collections/civil-service-statistics>

each Civil Service department by median salary across grades. Grades are aggregated into: Senior Civil Service, Grades 6 and 7, Senior and Higher Executive Officers, Executive Officers, and Administrative Assistants and Officers. Again, some data is confidential and suppressed (marked as “c”) due to small numbers between one and 29.

We first assign each job grade in GRID to one of the buckets in the UKCSS. Then, by joining UKCSS onto the GRID vacancies using these buckets, we are able to attach a median salary to each vacancy and get the number of FTEs at that grade within the relevant organisation. This is crucial for enabling us to estimate the quantitative potential benefits of AI implementation. After this step, it is simply a matter of assigning each vacancy to a cluster to get the results in [Section 3](#).

B B. Validation Checks

B.1 Task extraction validation

As a first step, we conduct a structured comparison between the tasks extracted by the LLM and those extracted manually by a human reviewer. In Section E in the Supplementary Materials, we provide two illustrative examples to reflect different levels of complexity: one job description with responsibilities clearly laid out in bullet-pointed list (Table 9) and another where responsibilities are embedded within longer-form prose and more generally defined (Table 10).

For the job description in Table 9, the model essentially concurs with the output generated through manual extraction. By contrast, for the job description in Table 10, the LLM identifies a greater number of distinct tasks by breaking down longer, compound sentences into shorter responsibilities. Meanwhile, the human tends to preserve longer spans of text that combine multiple actions or concepts. This pattern holds across the broader dataset of 200 job descriptions that were manually sampled: the model extracted an average of 12 tasks per description, compared to 8 extracted by the human. This suggests that the model breaks down job descriptions into more granular components, potentially offering a finer view of task exposure and role complexity.

B.2 Alignment of exposure scores with AI experts

The next two validation checks relate to the LLM’s assigned exposure scores for each task. Gmyrek et al. (2023) identified an “an astonishing proximity of GPT-4 predictions [AI exposure scores] to the judgments made by a group of AI experts” (p. 10) Gmyrek et al. (2023) in previous studies. To confirm this in our data, we provide a sample of 100 task descriptions to a group of four AI experts and ask them to assign an AI exposure score between zero and one.³⁴ This

³⁴Our group of AI experts include an academic, with research focussing on the impact of AI, and three UKCS Data Scientists, each with experience of building internal AI tools to streamline workflows. None of the experts have been involved in the drafting of this paper, nor were they given any detailed guidance or a framework to

sample covers ten tasks across ten different subtask categories, ensuring a broad coverage of task types and exposure scores.

To compare the experts' scores against the LLM, we use a Spearman's rank correlation coefficient which checks how the rank order of scores align. The mean Spearman's rank correlation between the scores of each human reviewer and the LLM scores is 0.673, indicating a moderately strong positive association in the relative ranking of responses. For completeness, we also report a mean Pearson correlation coefficient of 0.649, similarly reflecting a moderately strong positive relationship.

These results suggest initial agreement between the LLM and human assessments, both in terms of score ranking and overall score magnitude. However, visual inspection of the score distributions reveals important differences. In Section G in the Supplementary Materials, we present Figure 12a which plots the Kernel Density Estimate (KDE) of the distribution of exposure scores assigned by the LLM compared to that of the mean exposure score assigned by the human reviewers. The LLM's distribution shows a clear tendency to assign higher exposure scores than human reviewers, despite the overall alignment in rank ordering. Figure 12b plots the KDE for each individual human reviewer and highlights variability in both the central tendency and spread of scores assigned by different experts. This variation is quantitatively reflected in the Krippendorff's Alpha value of 0.526³⁵, which suggests poor agreement among the human reviewers. Given the inconsistency we observe among human experts, we believe that LLMs offer a useful and potentially more consistent approach for assigning exposure scores.

influence how they assign AI exposure scores.

³⁵Typically a value of 0.67 or more suggests a moderate level of agreement between human reviewers [Marzi et al. \(2024\)](#).

B.3 Foundation model comparison

Satisfied with the rank ordering, we also double-check the magnitude of exposure scores by comparing our chosen foundation model (Claude Sonnet 3.5 v2) against three alternative foundation models: Claude Haiku 3.5 (“anthropic.claude-3-5-haiku-20241022-v1:0”), Llama 3.3 70B Instruct (“meta.llama3-3-70b-instruct-v1:0”), and Mistral Large (“mistral.mistral-large-2402-v1:0”). These models are chosen as they also support batch operations in AWS Bedrock. We pass each model the same sample (a department’s raw vacancies data) and prompt/tool it with the same inputs (see the prompt in Section C.1 in the Supplementary Materials).

In Section F in the Supplementary Materials we present the results of this comparison. Figure 10 plots the cumulative density function of each model’s mean exposure scores per role. Unsurprisingly, there is a very close correspondence between Claude Sonnet 3.5 and Claude Haiku 3.5. The Llama model shows a greater proportion of roles allocated a low exposure score up to 0.5 (far left tail), but thereafter it very closely follows the Claude models. The Mistral model, by contrast, tracks the Claude models closely up to around an average exposure score of 0.5, and thereafter is persistently lower than the others.

Moreover, Figure 11 compares the model outputs (number, scores, and standard deviation of the scores) at the task level. Any points on the 45 degree line depict agreement between Claude Sonnet 3.5 v2 and another foundation model: points further away from the 45 degree line shows greater disagreement. The density and shape of the points around the 45 degree line highlights whether any model is consistently different in one direction. For example, the average task exposure score of Llama appears to be lower compared to Sonnet 3.5 v2, indicated by the rectangular shape of the points (as opposed to the elliptical shapes of the other subfigures). Nevertheless, in all cases, the vast majority of the distribution is clustered around the 45 degree line as indicated by the lighter colour. Taken together, we conclude that despite some inevitable model heterogeneity, the foundation models show a broad correspondence in their

assessment of AI exposure, and that our chosen foundation model is not an outlier.

B.4 Department

Table 5: Departments comprising our core dataset, the sample from GRID.

Department	Acronym	Vacancies (% of dataset)	FTEs (% of UKCS)
Animal and Plant Health Agency	APHA	1,960 (1.0)	3,195 (0.6)
Companies House	CH	982 (0.5)	1,775 (0.3)
Cabinet Office	CO	7,634 (3.9)	6,700 (1.3)
Crown Prosecution Service	CPS	2,789 (1.4)	7,100 (1.4)
Department for Business and Trade	DBT	4,742 (2.5)	5,830 (1.1)
Department for Levelling Up, Housing and Communities	DLUHC	3,537 (1.8)	3,785 (0.7)
Department for Culture, Media and Sport	DCMS	2,271 (1.2)	1,025 (0.2)
Department for Energy Security & Net Zero	DESNZ	1,028 (0.5)	4,675 (0.9)
Department of Health and Social Care	DHSC	3,790 (2.0)	3,585 (0.7)
Defence Science and Technology Laboratory	DSTL	1,683 (0.9)	4,860 (0.9)
Driver and Vehicle Licensing Agency	DVLA	988 (0.5)	5,480 (1.1)
Driver and Vehicle Standards Agency	DVSA	1,688 (0.9)	4,630 (0.9)
Department for Work and Pensions	DWP	6,121 (3.2)	84,415 (16.4)
Department for Environment, Food and Rural Affairs	Defra	4,697 (2.4)	6,590 (1.3)
Department for Education	DfE	6,962 (3.6)	6,920 (1.3)
Department for Transport	DfT	3,482 (1.8)	3,730 (0.7)
Foreign, Commonwealth & Development Office	FCDO	11,697 (6.0)	8,020 (1.6)
Food Standards Agency	FSA	1,338 (0.7)	1,530 (0.3)
HM Land Registry	HMLR	1,249 (0.6)	6,165 (1.2)
HM Courts and Tribunals Service	HMCTS	5,010 (2.6)	14,345 (2.8)
HM Prison & Probation Service	HMPPS	12,480 (6.4)	64,565 (12.7)
HM Revenue and Customs	HMRC	23,106 (11.9)	62,225 (12.1)
HM Treasury	HMT	2,553 (1.3)	1,970 (0.4)
Home Office	HO	17,789 (9.2)	48,160 (9.3)
Health and Safety Executive	HSE	1,767 (0.9)	2,930 (0.6)
Intellectual Property Office	IPO	1,061 (0.5)	1,660 (0.3)
Insolvency Service	IS	1,066 (0.6)	1,755 (0.3)
Maritime and Coastguard Agency	MCA	1,296 (0.7)	1,205 (0.2)
Medicines and Healthcare Products Regulatory Agency	MHRA	1,455 (0.8)	1,445 (0.3)
Met Office	MO	687 (0.4)	2,355 (0.5)
Ministry of Defence	MoD	32,387 (16.7)	35,135 (6.8)
Ministry of Justice	MoJ	7,712 (4.0)	7,995 (1.5)
National Crime Agency	NCA	1,898 (1.0)	5,775 (1.1)
OFGEM	OFGEM	1,857 (1.0)	2,195 (0.4)
Scottish Government	SG	6,572 (3.4)	8,925 (1.7)
UK Export Finance	UKEF	862 (0.4)	570 (0.1)
UK Health Security Agency	UKHSA	5,301 (2.7)	5,645 (1.1)
Total		193,497 (100.0)	438,865 (85.0)

The bracketed terms show the percentage to one decimal place of all vacancies in our dataset, and the percentage of FTEs as a share of the total Civil Service as per the 2025 Statistical Bulletin.

B.5 Grade

Table 6: The distribution of vacancies and roles in various grades, both within our dataset and across the wider UKCS.

Grade	Acronym	Vacancies (% of dataset)	FTEs (% of UKCS)
Administrative Assistant/ Administrative Officer	AA/AO	21,835 (0.11)	118,625 (0.23)
Executive Officer	EO	29,137 (0.15)	128,790 (0.25)
Higher Executive Officer/ Senior Executive Officer	HEO/SEO	91,092 (0.47)	156,305 (0.30)
Grade 7 / Grade 6	G7 / G6	46,686 (0.24)	81,975 (0.16)
Senior Civil Service	SCS	4,510 (0.02)	7,525 (0.01)
Unreported		237 (0.0)	22,930 (0.04)
Total		193,497 (1.00)	516,150 (1.00)

Note: Unreported grades from the GRID data are typically from where all grades are put into the grade field rather than none. Totals represent the sum across all grades, including unreported data. Civil service totals derived from the 2025 Statistical Bulletin.

B.6 Profession

Table 7: Number of GRID dataset roles by profession and Civil Service workforce distribution (2024)

Profession	Vacancies (% of dataset)	FTEs (% of UKCS)
Commercial	2,338 (0.01)	6,965 (0.01)
Counter Fraud Professions	1,973 (0.01)	9,570 (0.02)
Government Communication Service	3,711 (0.02)	4,670 (0.01)
Government Corporate Finance	1,572 (0.01)	80 (0.00)
Government Digital and Data Profession	21,169 (0.11)	24,115 (0.05)
Government Economic Service	1,253 (0.01)	2,350 (0.00)
Government Finance	5,212 (0.03)	10,370 (0.02)
Government Legal Service	934 (0.00)	10,435 (0.02)
Government Operational Research Service	554 (0.00)	1,225 (0.00)
Government Project Delivery Profession	10,434 (0.05)	17,960 (0.04)
Government Property Profession	2,254 (0.01)	7,095 (0.01)
Government Science and Engineering	5,367 (0.03)	13,845 (0.03)
Government Social Research	679 (0.00)	1,630 (0.00)
Government Statistical Service	944 (0.00)	3,375 (0.01)
Human Resources	7,041 (0.04)	11,295 (0.02)
Intelligence Analysis	1,711 (0.01)	3,445 (0.01)
Internal Audit	351 (0.00)	630 (0.00)
International Trade Profession	1,288 (0.01)	1,735 (0.00)
Knowledge & Information Management	2,094 (0.01)	2,315 (0.00)
Medical Profession	190 (0.00)	2,260 (0.00)
Operational Delivery Profession	27,343 (0.14)	270,745 (0.53)
Other	71,142 (0.37)	12,385 (0.02)
Planning Inspectors	17 (0.00)	420 (0.00)
Planning Profession	414 (0.00)	230 (0.00)
Policy Profession	16,825 (0.09)	33,915 (0.07)
Psychology Profession	377 (0.00)	1,465 (0.00)
Security Profession	3,500 (0.02)	9,635 (0.02)
Tax Profession	2,635 (0.01)	14,535 (0.03)
Government Veterinary Profession	175 (0.00)	410 (0.00)
Total	193,497 (1.00)	510,125 (1.00)

C C. LLM Prompts

Note: a string wrapped in `<>` tags indicates to the LLM where to expect a string. A string wrapped in `{}` brackets is a placeholder for the Python string object, and will be substituted in on invocation as per Python's f-strings syntax. Valid responses from the LLM must conform to the tools provided.

C.1 Task extraction and exposure score

System prompt: You are an expert in extracting skills from job descriptions and assessing their potential automation with GPT technology.

User prompt: Extract a list of tasks from the job advert in the provided format. Do not include tasks which concern the recruitment process, the onboarding process, working hours, or working conditions. `<job_description>{job_description}</job_description>`

```
1 # Tools for task extraction and exposure scoring
2 class Task(BaseModel):
3     task_number: int = Field(..., description="A unique task number")
4     task_details: str = Field(
5         ..., description="the text extract which describes the task"
6     )
7     exposure_score: float = Field(
8         ...,
9         description="a score of range 0-1 of potential automation of the
10            task with GPT technology",
11     )
12
13 class TaskOutput(BaseModel):
14     tasks: list[Task] = Field(
15         ..., description="A list of tasks extracted from job_advert"
16     )
```

C.2 Assigning exposure scores to O*NET and ISCO-08 tasks

System prompt: You are an expert in assessing their assessing job tasks' potential automation with GPT technology.

User prompt: Use the tool provided to assign each of these tasks a score of range 0-1 of potential automation of the task with GPT technology: <tasks>{formatted_tasks}</tasks>

```
1 class AutomationScores(BaseModel):
2     """A tool for structuring automation scores from a list of tasks."""
3     scores: Dict[str, Annotated[float, Field(strict=True, ge=0, le=1)]] =
4         Field(
5             ..., description="A dictionary mapping each task ID to its
6                 automation exposure score (0.0 to 1.0)"
7         )
```

C.3 Task clustering

System prompt: You are an expert in giving human-interpretable labels to clusters of job tasks.

User prompt: Provide a short descriptive label for this cluster of tasks: <tasks>{task_list}</tasks>. Provide no other text.

C.4 Focus tasks prompts

Prompt 1

User prompt: A person working as a {job_title} has had parts of their role augmented by AI, freeing up time for other tasks. Given the remaining responsibilities:{task_details} where should they focus their efforts to maximize productivity? Reply with the task number and your reasoning using the schema.

```
1 class TasktoFocus(BaseModel):
2     task_number: int
3     reasoning: str = Field(
4         ..., description="Briefly, why should they focus on this task?"
5     )
```

Prompt 2

User prompt: You are an expert in picking out key themes from text. AI has automated some tasks of workers. From their remaining tasks, someone has picked what they believe as the most important task to focus their freed up time on. Your job is to identify key themes in

the reasonings provided below. These might (but not necessarily include) synergies, human-centered work, unlocking the value of others' work etc. Reasonings:{reasonings_formatted}.

Reply using the provided tool.

```
1 class TaskReasoning(BaseModel):
2     """Schema for identifying key themes in reasonings."""
3     label: str = Field(
4         ..., description="A short label for the theme"
5     )
6     description: str = Field(
7         ..., description='A brief description of the theme'
8     )
9
10
11 class TaskReasonings(BaseModel):
12     """Collection of key themes in reasonings"""
13     query_output: List[TaskReasoning] = Field(
14         ..., description = 'List of key themes and their labels'
15     )
```

Prompt 3

User prompt: You are an expert at classifying the reasoning behind the choice of tasks to focus on. AI has automated the task of a worker. From the worker's remaining tasks, they have picked what they believe as the most important task to focus their freed up time on. Accompanying their choice of best-alternative task to focus on, they have provided their reasoning. You must review the categories of reasoning below and indicate which are present in their reasoning with a 1, or not (a 0). {consolidated_themes} Be selective and apply the most relevant categories: choose at least one category but no more than three. Their reasoning is: {focus_tasks.reasoning}

Reply using only the provided tool.

```
1 class ThemeApplication(BaseModel):
2     strategic_leadership_and_vision: Literal[0, 1] = Field(
3         description="Tasks involving setting strategic direction,
4         developing policies, and leading organizational transformation,
5         including decision-making that influences long-term outcomes and
6         establishes foundational frameworks."
7     )
8     stakeholder_management_and_communication: Literal[0, 1] = Field(
```

```

6         description="Tasks focused on building and maintaining
           relationships with stakeholders, including complex communication
           , knowledge transfer, and effective engagement across different
           audiences."
7     )
8     risk_and_quality_management: Literal[0, 1] = Field(
9         description="Tasks involving risk assessment, security oversight,
           compliance monitoring, and quality assurance to maintain
           operational integrity and safety standards."
10    )
11    innovation_and_process_excellence: Literal[0, 1] = Field(
12        description="Tasks centered on continuous improvement,
           modernization, and transformation of systems and processes, with
           emphasis on maximizing value and creating multiplier effects
           across the organization."
13    )
14    human_centric_leadership: Literal[0, 1] = Field(
15        description="Tasks requiring uniquely human capabilities including
           team development, mentoring, empathy, and providing personalized
           services that cannot be automated."
16    )
17    complex_problem_resolution: Literal[0, 1] = Field(
18        description="Tasks requiring sophisticated analysis, investigation,
           and critical thinking to solve multifaceted problems,
           particularly in high-stakes situations that demand human
           judgment."
19    )

```

Prompt 4 (robustness check - task reimagining and reordering)

System prompt: You are an expert in understanding how workforce tasks might change following digital and AI transformation.

User prompt: A person working as a job_title has had parts of their role augmented by AI. Your task is to consider how the rest of their role might be re-imagined to maximize the worker's productivity. Specifically, you must consider how their remaining tasks might be augmented or remain unchanged following the use of generative AI tools. You are to take things in two steps.

In the first step, consider each task in turn. If the task has little or no scope to change, label it with 'No change.' Otherwise, return 'augmented'. If the task is labelled with no change, retain its original details. Otherwise, give a new task detail of the redesigned (augmented) task.

The second step is that, once you've done step one for all tasks, return the tasks in order of importance, the most important being first etc. `<tasks>{task_details}</tasks>`

Use the following context on the job role to consider how the tasks should be ordered by importance: `<job_context>{job_context}</job_context>`. Reply with each task number, its label, and its new task details, in order of task importance, using the schema.

```
1 class TaskLabel(str, Enum):
2     NO_CHANGE = "No change"
3     AUGMENTED = "Augmented"
4
5
6 class Task(BaseModel):
7     task_number: int = Field(..., description="The provided task number.")
8     label: TaskLabel = Field(
9         ..., description="Whether the task is unchanged from AI or can be
10         augmented."
11     )
12     new_task_details: str = Field(..., description="A description of the
13         task.")
14
15 class Tasks(BaseModel):
16     tasks: list[Task] = Field(
17         ...,
18         description="A list of tasks in order of importance, the most
19         important being first.",
20     )
```

Prompt 5 (robustness check - new tasks)

System prompt: You are an expert in helping UK civil servants manage the impacts of AI use in the workplace.

User prompt: AI is transforming the role of {job_title} in the {dept_key} department of the UK Civil Service. Your assignment is to assess how the role's tasks could look after the transformation. To do this, you must follow a series of steps:

INSTRUCTIONS:

1. Review the context of the role to build a picture of what it entails.

2. Note which tasks have been automated away. These will not be part of the final role.
3. ****Consolidate All Potential Tasks:**** First, create a single conceptual pool of all tasks for the redesigned role. This includes both the essential tasks from the <remaining_tasks> list and any high-impact new tasks you identify (you may suggest up to {n_automated_tasks} new tasks).
4. ****Rank the Consolidated List:**** Finally, sort this entire pool of tasks strictly by their strategic importance for the future role. The final order in your output must reflect this holistic ranking.

RULES:

- Each task must have a unique integer ID, description, and category. For new tasks you put forward, you must assign a negative task number.
- Give no new tasks if none are forthcoming. Do not suggest new tasks merely for the sake of it.
- Ensure any new tasks proposed are suitable for the role's seniority.
- Categories assigned to tasks **MUST** follow the task categories set out in the schema. No other values are acceptable.
- Your final output **MUST** be a single JSON object with one top-level key: 'tasks'. The value of 'tasks' must be a list of the task objects.
- Respond using only the provided tool, wrapped in <tool_code> tags.

INFORMATION:

<role_context>{job_context}</role_context> <automated_tasks>{automated_tasks}</automated_tasks>
<remaining_tasks>{not_automated_tasks_csv}</remaining_tasks>

```
1 class TaskCategoryEnum(str, Enum):
2     """Enumeration for the allowed task categories."""
3     POLICY_DEVELOPMENT = "policy_development"
4     RECORDS_MANAGEMENT = "records_management"
5     ADMIN_SUPPORT = "admin_support"
6     TEAM_LEADERSHIP = "team_leadership"
7     PERFORMANCE_PLANNING = "performance_planning"
8     STAKEHOLDER_ENGAGEMENT = "stakeholder_engagement"
9     RISK_MANAGEMENT = "risk_management"
10    DATA_ANALYSIS = "data_analysis"
11    SERVICE_DELIVERY = "service_delivery"
12    PRISON_MANAGEMENT = "prison_management"
13
14
15 class Task(BaseModel):
16     """Defines the structure of a single task"""
17     task_number: int
18     task_details: str
19     task_category: TaskCategoryEnum
20
21
22 class LLMTaskOutput(BaseModel):
23     """A simple model that matches the LLM's direct output."""
24     tasks: List[Task]
```

D D. Batch jobs

Table 8: Batch jobs submitted to AWS

Department	Jobs submitted	Input tokens	Output tokens	Estimated cost (\$)
APHA	1,965	2,358,509	1,149,268	12
CO	7,744	9,751,222	3,512,667	41
CH	989	1,028,889	448,940	5
CPS	2,901	2,456,537	925,543	11
DBT	4,756	5,364,334	2,145,701	24
DCMS	2,373	2,256,564	929,201	10
DEFRA	4,752	4,957,816	2,045,168	23
DESNEZ	1,029	1,203,957	386,968	5
DfE	7,055	7,173,759	2,990,093	33
DfT	3,512	3,603,893	1,260,563	15
DHSC	3,819	3,744,590	1,524,661	17
DLUHC	3,585	3,671,316	1,547,197	17
DSTL	1,929	1,918,084	959,836	10
DVLA	994	1,074,512	362,290	4
DVSA	1,704	1,587,786	596,375	7
DWP	6,136	6,395,104	2,843,597	31
FCDO	12,828	20,090,255	5,022,531	68
FSA	1,485	1,418,773	462,563	6
HMCTS	5,053	12,658,284	3,384,245	44
HMRC	23,233	22,693,160	9,359,973	104
HMLR	1,290	1,164,029	471,430	5
HMT	2,569	4,408,516	1,214,338	16
HMPPS	22,930	49,247,201	15,996,437	194
HO	17,887	18,923,030	7,328,642	83
HSE	1,780	1,893,237	718,481	8
IS	1,079	1,140,314	513,285	6
IPO	1,068	1,267,985	516,871	6
MCA	1,312	1,304,457	435,268	5
MO	1,028	1,073,579	377,762	4
MHPRA	1,463	1,383,185	529,792	6
MoD	41,259	40,360,808	15,628,319	178
MoJ	7,875	20,300,877	4,324,330	63
NCA	1,927	1,751,747	662,616	8
OFGEM	2,044	2,195,675	849,874	10
SG	6,649	6,059,459	2,909,109	31
UKEF	870	776,490	310,048	3
UKHSA	5,676	6,555,705	2,801,753	31
Total	216,548	275,213,638	97,445,735	1,144

One job means splitting a role into tasks and assigning exposure scores. Estimated costs based on AWS Bedrock pricing: <https://aws.amazon.com/bedrock/pricing/> with \$1.50 per million input tokens and \$7.50 per million output tokens.

E E. Manual task extraction comparison

Table 9: Job description with clearly defined tasks.

The role of the PMO Support Officer covers a diverse range of activities to support the delivery of the programme objectives. The PMO Support Officer enables the smooth running of the programme by supporting the PMO Manager through the operation of project/ programme management processes, and the co-ordination of business management actions and activities on their behalf. This includes: • Scheduling meetings; taking accurate minutes with clearly attributable actions, monitoring and prompting the timely completion of actions. • Drafting skeleton packs for governance board and liaising across the team to ensure team members complete sections attributed to them on time and to quality. • Support in maintaining programme controls and in producing project reports. • Making sure programme documents such as delivery plans and risk and issue registers are regularly updated and used across the programme to support delivery. • Maintaining the stakeholder log for internal and external stakeholders. • Programme file management using robust version control. • Support the monitoring of programme spend and contribute to the compilation of budgets.	
LLM inferred tasks:	Manually inferred tasks:
Scheduling meetings and taking accurate minutes with clearly attributable actions, monitoring, and prompting the timely completion of actions.	Scheduling meetings; taking accurate minutes with clearly attributable actions, monitoring and prompting the timely completion of actions.
Drafting skeleton packs for governance board and coordinating with team members to ensure timely and quality completion of their sections.	Drafting skeleton packs for governance board and liaising across the team to ensure team members complete sections attributed to them on time and to quality.
Supporting the maintenance of programme controls and producing project reports.	Support in maintaining programme controls and in producing project reports.
Ensuring regular updates and usage of programme documents such as delivery plans and risk and issue registers to support delivery.	Making sure programme documents such as delivery plans and risk and issue registers are regularly updated and used across the programme to support delivery.
Maintaining the stakeholder log for internal and external stakeholders.	Maintaining the stakeholder log for internal and external stakeholders.
Managing programme files using robust version control.	Programme file management using robust version control.
Supporting the monitoring of programme spend and contributing to the compilation of budgets.	Support the monitoring of programme spend and contribute to the compilation of budgets.

Table 10: Job description with generally defined tasks.

Within the directorate, the vision is to build a community of highly skilled and motivated project delivery professionals to establish and transform our Test and Trace services. Part of this will include establishing a pool of talented project delivery professionals who will be forward-deployed across the change portfolio based on the demands of Test and Trace/UKHSA. Members of the pool will be assigned to projects and programmes to manage, support, guide, direct and intervene at programme and project levels where necessary. Flexibility will be key as we will seek to retain and redeploy resources as demand shifts in line with our portfolio. Senior Project Manager: The role of the Senior Project Manager is to lead / manage the project and the project team on a day-to-day basis. The Senior Project Manager is responsible for driving and overseeing the delivery of the project to ensure that the objectives are clearly defined and achieved within the agreed time, cost and quality constraints. The Senior Project Manager has a key role in project governance and working with stakeholders, to ensure the agreed project outputs are delivered to enable benefits to be realised.	
LLM inferred tasks:	Manually inferred tasks:
Lead and manage the project team on a daily basis.	<i>No matching task extracted</i>
Drive and oversee the delivery of the project to ensure objectives are met within agreed time, cost, and quality constraints.	The Senior Project Manager is responsible for driving and overseeing the delivery of the project to ensure that the objectives are clearly defined and achieved within the agreed time, cost and quality constraints.
Play a key role in project governance.	<i>No matching task extracted</i>
Work with stakeholders to ensure project outputs are delivered, enabling benefits realization.	The Senior Project Manager has a key role in project governance and working with stakeholders, to ensure the agreed project outputs are delivered to enable benefits to be realised.

Note: Job descriptions have been manually selected by the authors to represent the contrasting quality and clarity of task information in job descriptions in our sample dataset. For the manually inferred tasks, an analyst was given the full job description text and asked to extract each unique task.

F E. Foundation model comparison

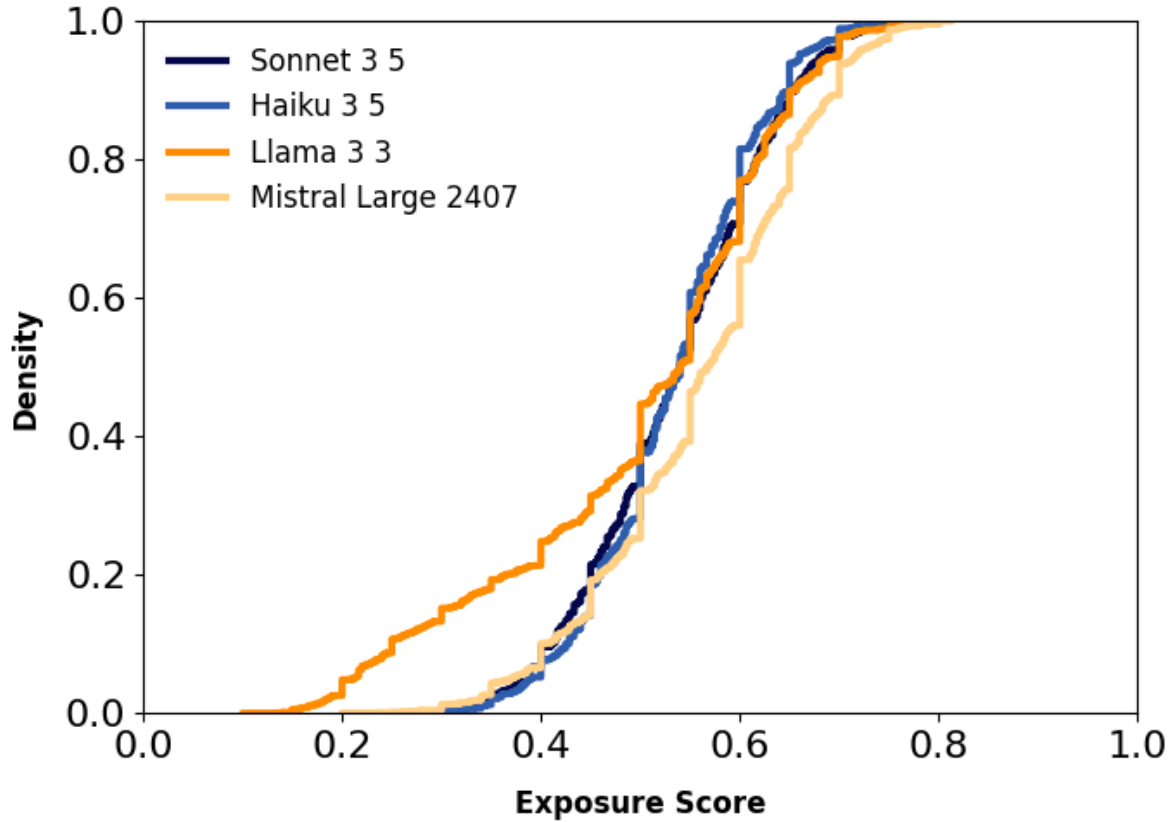


Figure 10: Role automation scores for foundation models

Empirical cumulative density function of exposure scores for each foundation model: Claude Sonnet 3.5 v2 (“anthropic.claude-3-5-sonnet-20241022-v2:0”), Claude Haiku 3.5 (“anthropic.claude-3-5-haiku-20241022-v1:0”), Llama 3.3 70B Instruct (“meta.llama3-3-70b-instruct-v1:0”), and Mistral Large (“mistral.mistral-large-2402-v1:0”). We use Claude Sonnet 3.5v2 as our chosen foundation model for extracting tasks and assigning exposure scores. This appears to be within the local range of the other foundation models, with the exception of Llama 3.3 at the lower end of the distribution.

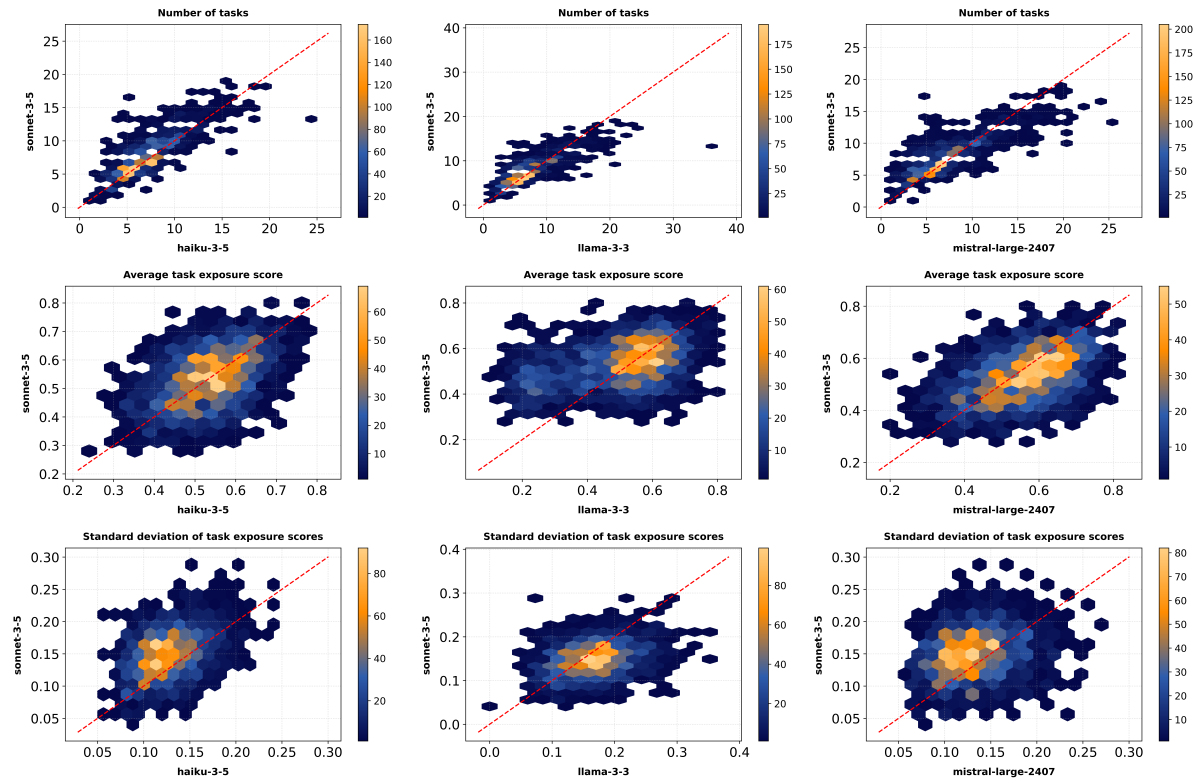


Figure 11: Claude Sonnet 3.5 model versus other foundation models: number of tasks extracted, the average exposure of score of tasks per role, the standard deviation of exposure scores per role.

Other foundation models are Claude Haiku 3.5 (“anthropic.claude-3-5-haiku-20241022-v1:0”), Llama 3.3 70B Instruct (“meta.llama3-3-70b-instruct-v1:0”), and Mistral Large (“mistral.mistral-large-2402-v1:0”). Points clustered around the 45 degree line imply agreement between the different foundation models when given the same prompts and the same Department of Education input data (system and user prompt in Task extraction and exposure score LLM prompt in Supplementary Materials).

G G. Exposure score validation

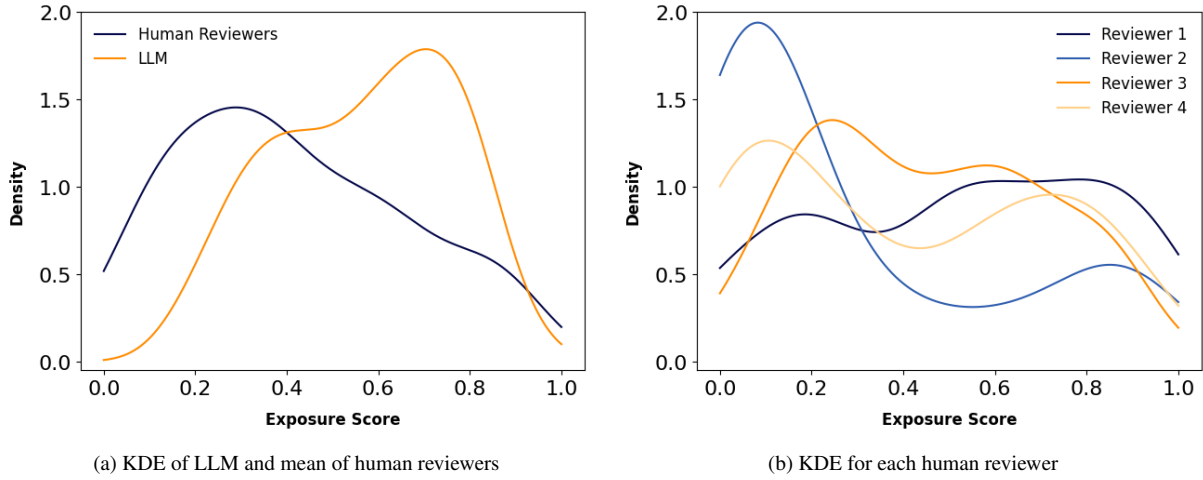


Figure 12: AI Exposure score human validation

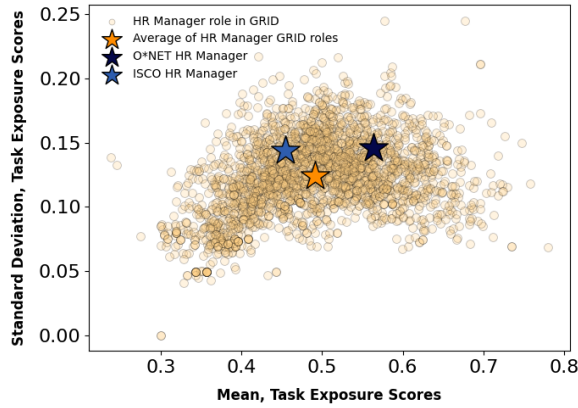
Both figures are based on a subset of one hundred tasks. 'Human Reviewers' in figure 12a is the mean exposure score assigned by the four human reviewers. Figure 12b compares the exposure scores assigned by each individual human reviewer.

H H. AI exposure comparison: O*NET and ISCO Databases

Table 11: Full list of tasks for Economist roles from O*NET and ISCO-08 databases

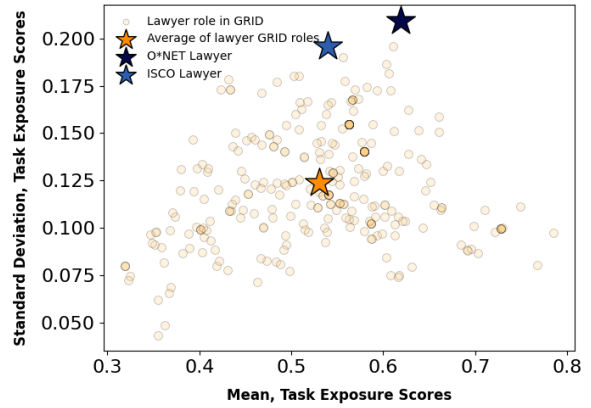
#	O*NET Task Description	ISCO-08 Task Description
1	Study economic and statistical data in area of specialization, such as finance, labor, or agriculture	Forecasting changes in the economic environment for short-term budgeting, long-term planning and investment evaluation
2	Compile, analyze, and report data to explain economic phenomena and forecast market trends, applying mathematical models and statistical techniques	Formulating recommendations, policies and plans for the economy, corporate strategies and investment, and undertaking feasibility studies for projects
3	Study the socioeconomic impacts of new public policies, such as proposed legislation, taxes, services, and regulations	Monitoring economic data to assess the effectiveness, and advise on the appropriateness, of monetary and fiscal policies
4	Explain economic impact of policies to the public	Forecasting production and consumption of specific products and services based on records of past production and consumption and general economic and industry-specific conditions
5	Review documents written by others	Preparing forecasts of income and expenditure, interest rates and exchange rates
6	Provide advice and consultation on economic relationships to businesses, public and private agencies, and other employers	Analysing factors that determine labour-force participation, employment, wages, unemployment and other labour-market outcomes
7	Formulate recommendations, policies, or plans to solve economic problems or to interpret markets	Applying mathematical formulae and statistical techniques to test economic theories and devise solutions to economic problems
8	Supervise research projects and students' study projects	Compiling, analysing and interpreting economic data using economic theory and a variety of statistical and other techniques
9	Conduct research on economic issues, and disseminate research findings through technical reports or scientific articles in journals	Evaluating the outcome of political decisions concerning public economy and finances and advising on economic policy and possible courses of action in light of past, present and projected economic factors and trends
10	Develop economic guidelines and standards, and prepare points of view used in forecasting trends and formulating economic policy	Preparing scholarly papers and reports
11	Teach theories, principles, and methods of economics	Examining problems related to the economic activities of individual companies
12	Testify at regulatory or legislative hearings concerning the estimated effects of changes in legislation or public policy, and present recommendations based on cost-benefit analyses	Conducting research on market conditions in local, regional or national areas to set sales and pricing levels for goods and services, to assess market potential and future trends and to develop business strategies
13	Provide litigation support, such as writing reports for expert testimony or testifying as an expert witness	N/A

O*NET tasks are taken from the O*NET 29.3 Database, SOC code 19-3011.00. O*NET tasks are ordered by their data value on the importance scale ('IM'). ISCO-08 tasks are listed for ISCO-08 Code 2631. They are unordered by importance so are simply assigned equal weighting.



(a) Human Resources Manager

“HR Manager in GRID” comprises all roles in the Government Recruitment Information Database (GRID) with a vacancy title containing “Human Resources Manager”, or works as a grade 6 or 7 in a human resources type of role, or in the human resources profession. O*NET tasks are taken from the core set of tasks in O*NET 29.3 Database, SOC code 11-3121.00. ISCO-08 tasks are listed for ISCO-08 Code 1212.



(b) Lawyers

“Lawyer in GRID” comprises all roles in the Government Recruitment Information Database (GRID) with a vacancy title containing “Lawyer”. O*NET tasks are taken from the core set of tasks in O*NET 29.3 Database, SOC code 23-1011.00. ISCO-08 tasks are listed for ISCO-08 Code 2611.

Figure 13: Exposure score comparison between GRID, O*NET and ISCO-08

Exposure scores for the GRID data are weighted as described in Section 2.3 and the weighted means and standard deviations are calculated accordingly. O*NET tasks are ordered by their data value on the importance scale (‘IM’) and weighted using the same decay rate and process as with the GRID data. ISCO-08 tasks are unordered by importance so are simply assigned equal weighting.

I I. Decay rate sensitivity analysis

I.1 Savings estimates at 80% threshold level

Table 12: Savings estimates under different decay rates at 80% automation threshold

Saving Type	0.75 (Baseline)	0.5 (High)	1.0 (Equal)
Potential Cost Reduction	£1.10bn	£2.45bn (+123%)	£0.94bn (-14%)
Productivity Gain	£5.24bn	£4.07bn (-22%)	£5.32bn (+2%)

I.2 Impact of decay rate on task weight

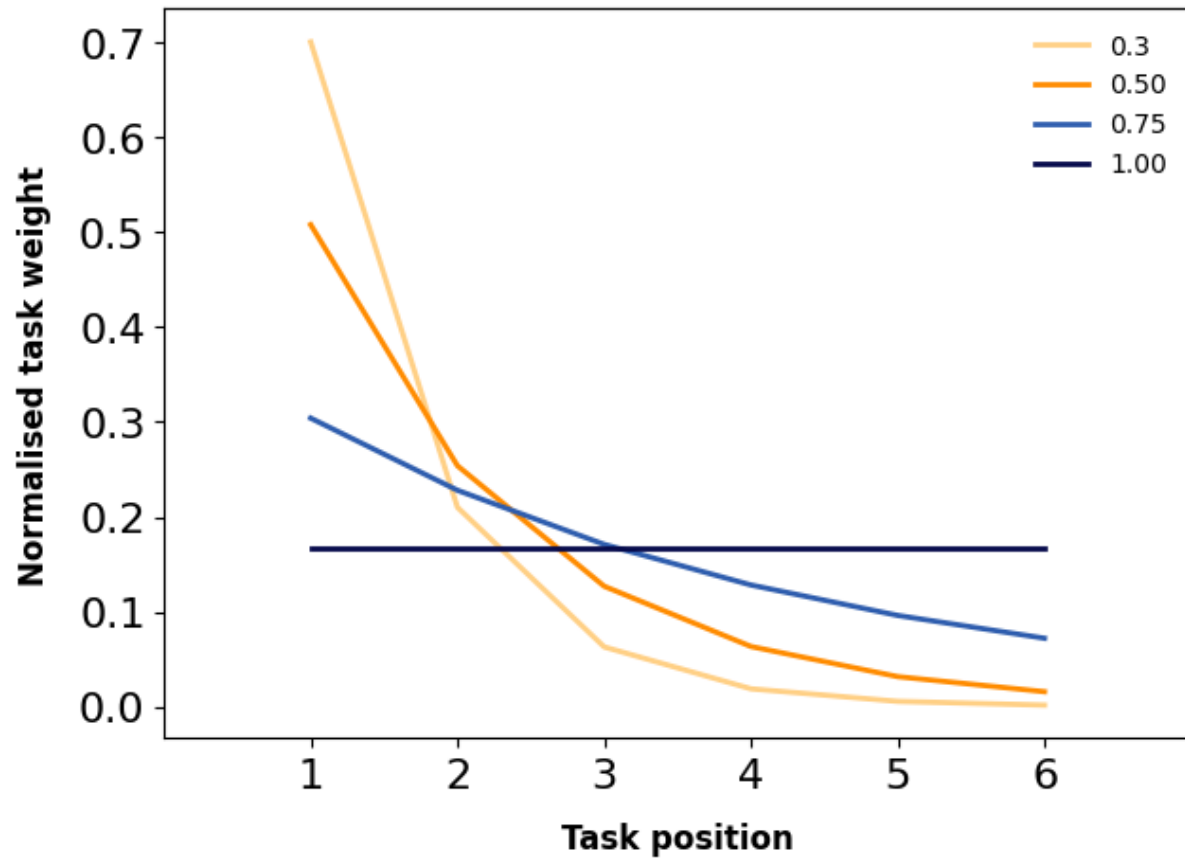


Figure 14: Weight per task under different decay rates

Each line represents a different decay rate applied to an job description with 6 tasks - this number was selected as it is the mode number of tasks extracted from job roles in our sample

I.3 Range of Estimated Opportunity by Decay Rate

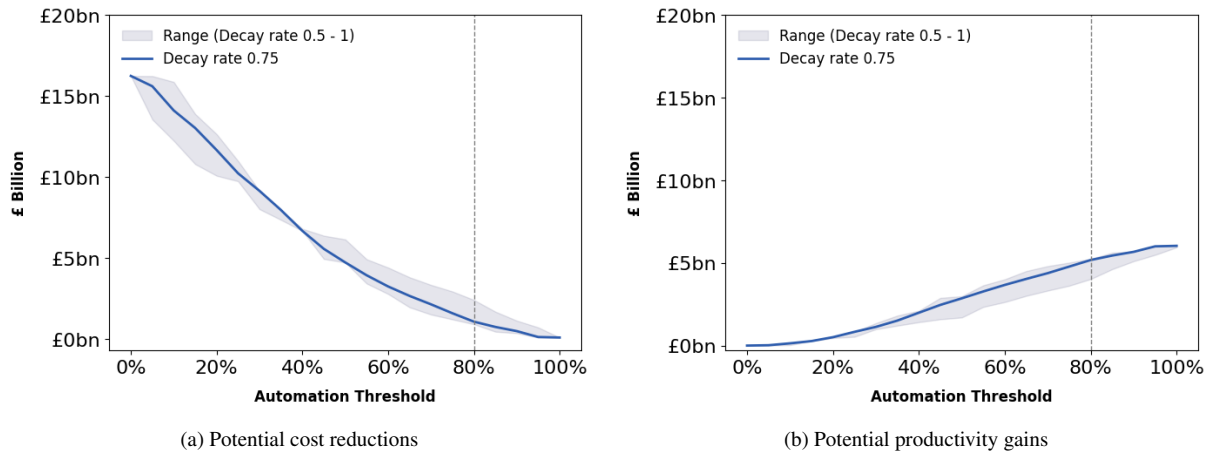


Figure 15: Range of the estimated monetary opportunity based on decay rate

The graph shows how the estimated monetary figures change at each automation threshold depending on the decay rate. We select decay rates of 0.5 and 1 for the range here to present a realistic range - from Figure 14, we consider a decay rate of 0.3 to be over extreme.

I.4 Percentage change in estimated opportunity by decay rate

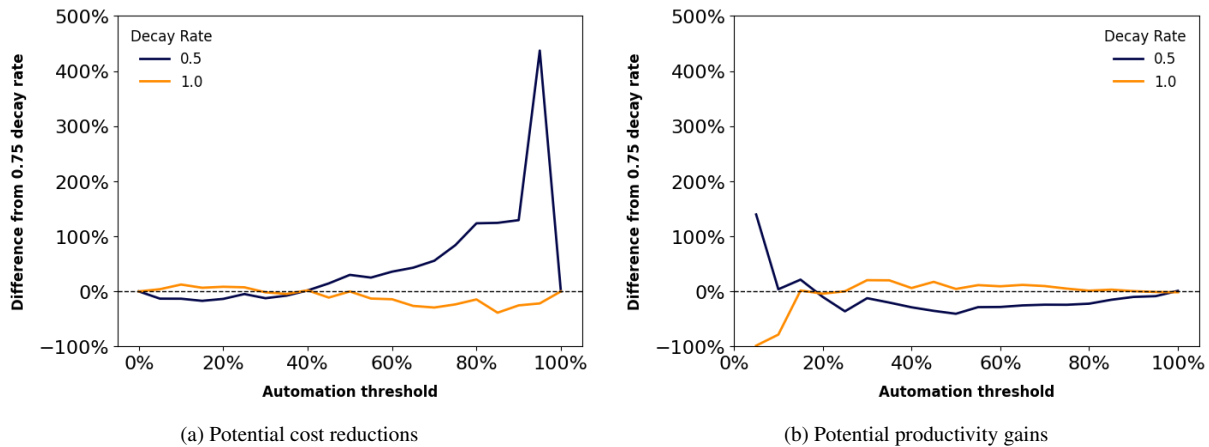


Figure 16: Percentage difference in estimated monetary opportunity from 0.75 decay rate

Graphs compare the percentage increase or decrease in each savings category compared to a decay rate of 0.75. The 400%+ increase seen at the 95% threshold for potential cost reductions is expected given minor estimated savings (£0.14bn at our baseline decay rate of 0.75 compared to £0.7bn with a decay rate of 0.5). Whilst the percentage increase is therefore significant, the actual monetary difference is minor.

J J. Task clustering

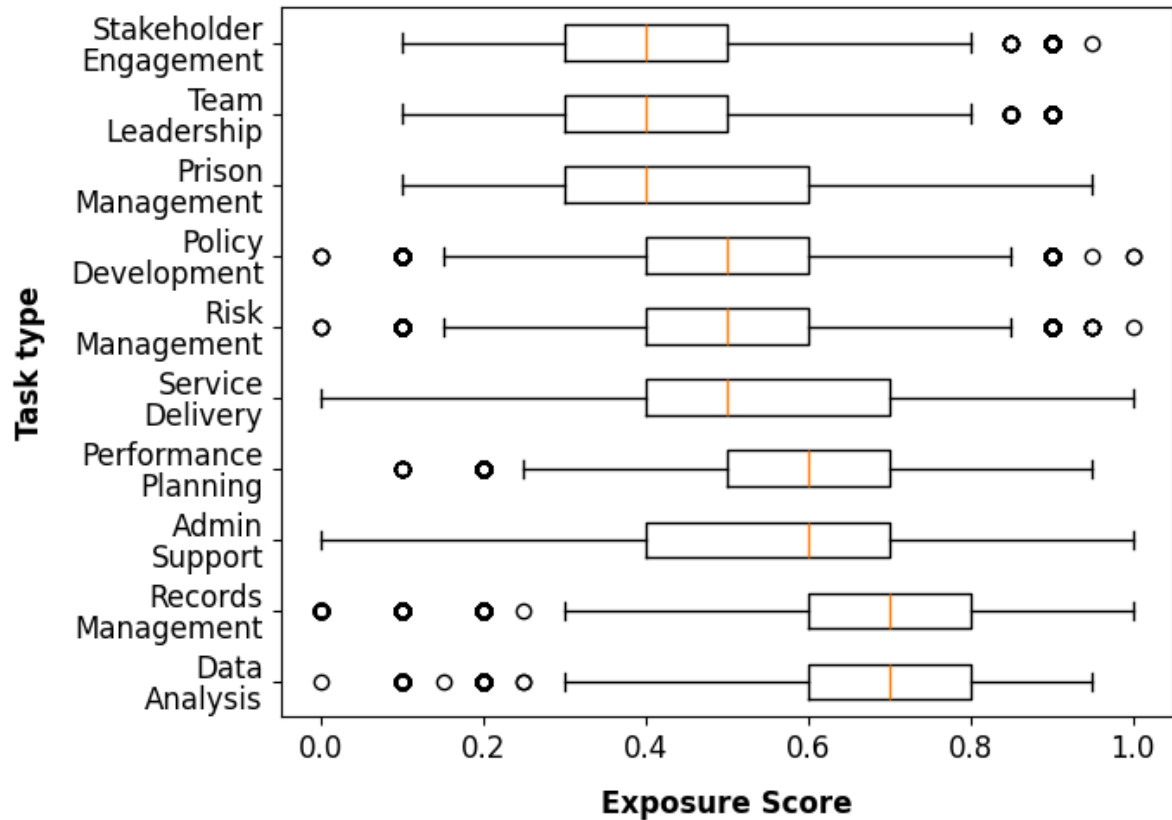


Figure 17: Exposure Score Distribution by Task Category

Exposure scores assigned by the LLM ranging between zero and one. Boxes indicate the median with the orange line and with the box bounding the upper and lower quartile. Whiskers represent the range of the the upper or lower quartile (1.5 times the IQR), excluding outliers which are highlighted as points. Task categories are ordered by the median exposures.

K K. Cluster scores

Table 13: Summary of exposure metrics across clusters

Cluster	Mean Exposure	Mean Std Exposure	Highest Exposure	Lowest Exposure
Low	0.381	0.118	0.470	0.043
Augmentation	0.483	0.158	0.538	0.357
Adaptation	0.586	0.162	0.678	0.530
Automation	0.697	0.136	1.000	0.621

For each job role, we calculated the weighted mean exposure score using the raw exposure score assigned to each task and a decay rate of 0.75, as outlined in 2.6. The mean exposure score shown here is the average of the role level exposure scores. The mean std is the average of the weighted role level standard deviation exposure score.

L L. Subcategory comparison

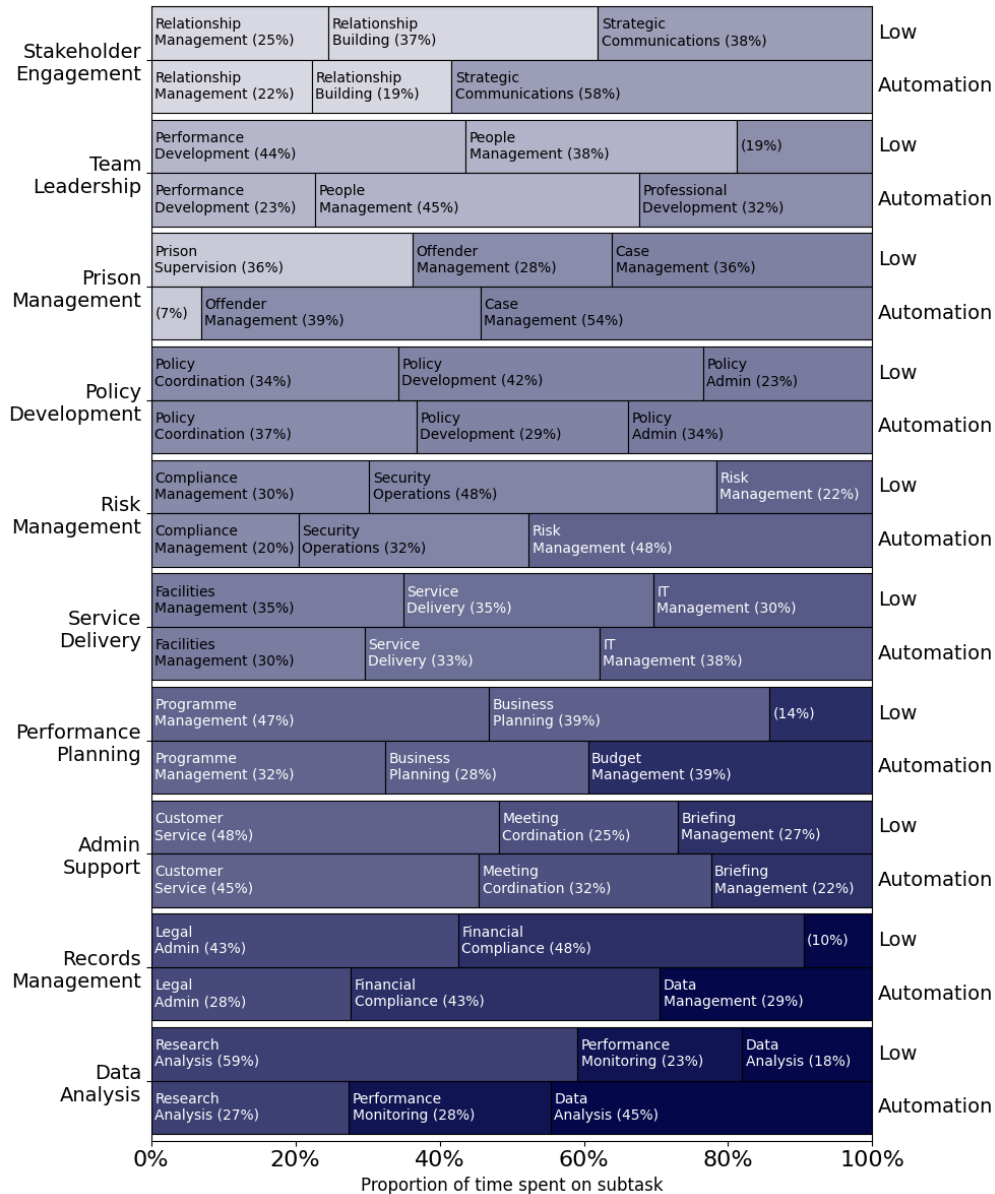


Figure 18: Proportion of subtasks in each task category, split between Low roles and Automation roles

Each pair of horizontal bars represents a task category, with time distributed across the three subtasks. The bars illustrate, for each group, the percentage of time spent on each subtask relative to the total time allocated to that task category, meaning the percentages sum to 100% within each bar. Subtasks in each task category are ordered from lowest to highest average exposure score.

M M. Focus task replacing high exposure tasks

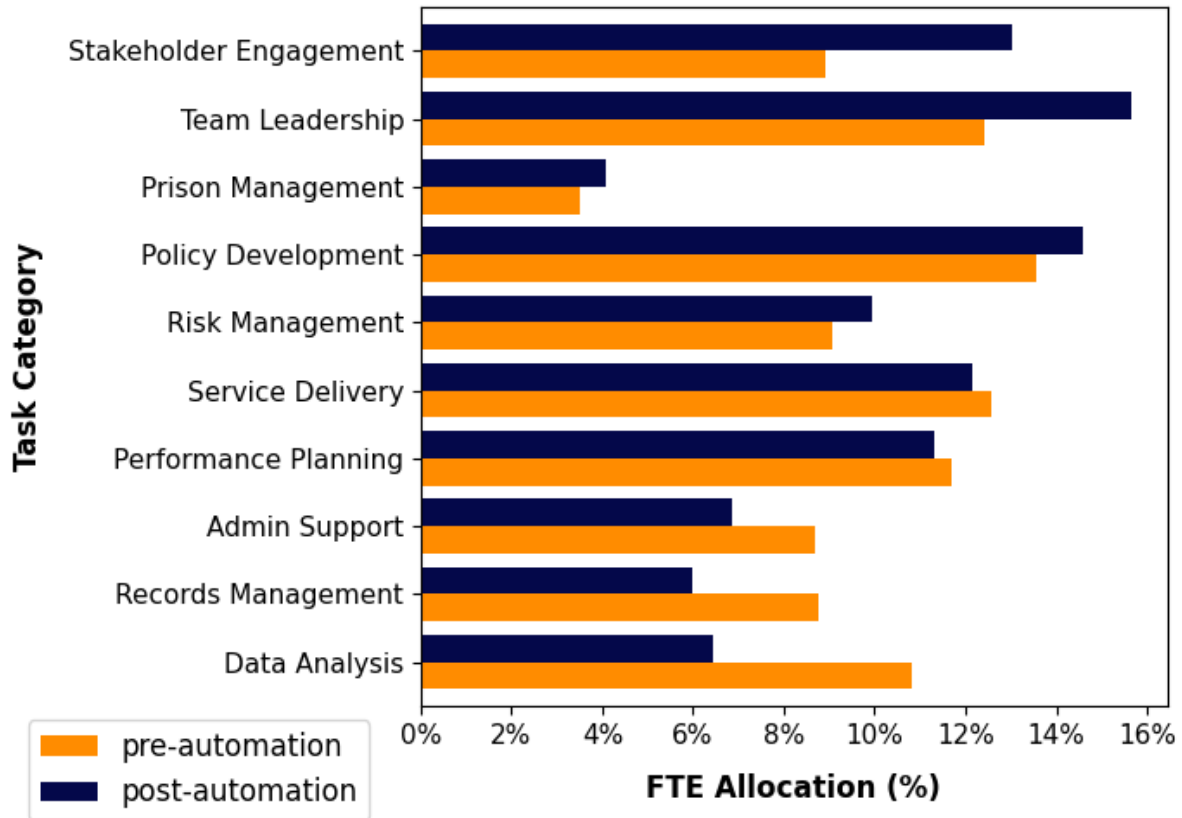


Figure 19: Total time allocation changes following focus tasks replacing high exposure tasks

Represents the total weighted task FTE allocation for the set of roles which have at least one high AI exposure task automated but where the role itself is not automated. This time shift can be thought of as a productivity gain. See Section 3.4 for full methodology.

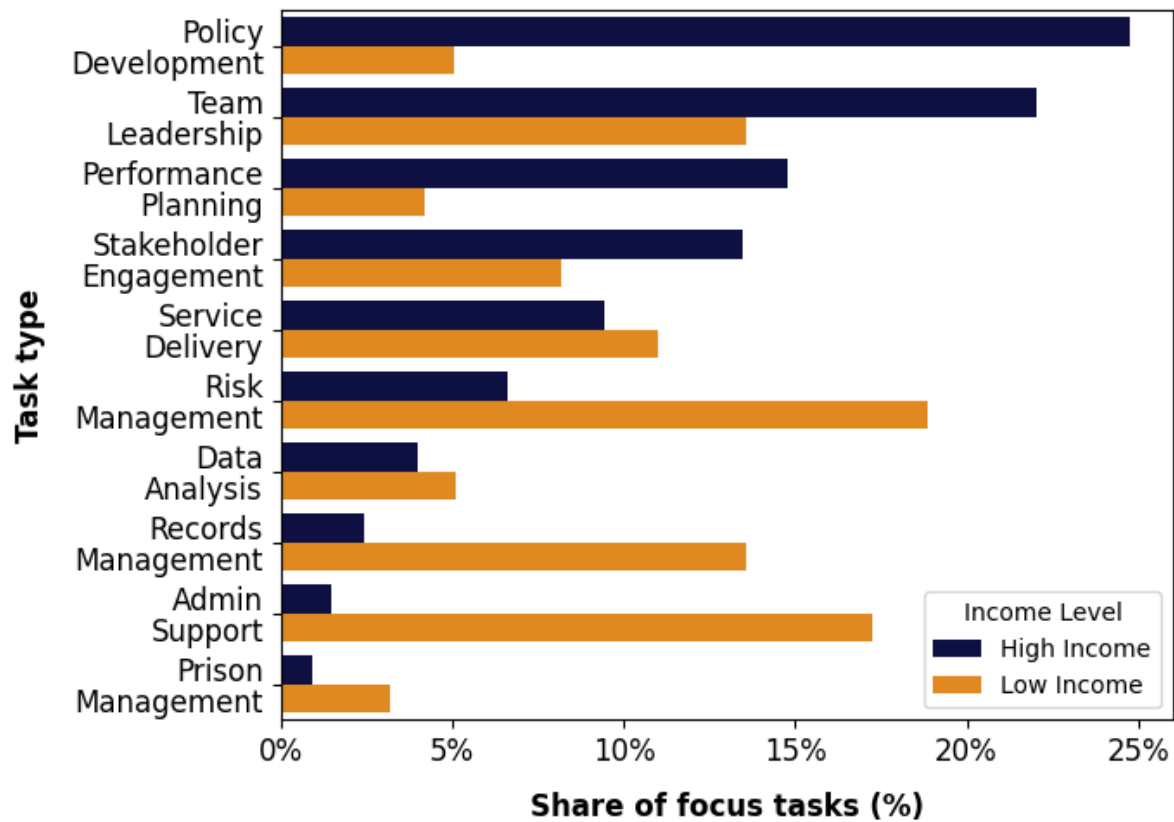


Figure 20: Focus task categories split by top and bottom decile

Compares the task category assignments for the focus tasks of the top and bottom income decile for all non-cashable roles in our dataset.

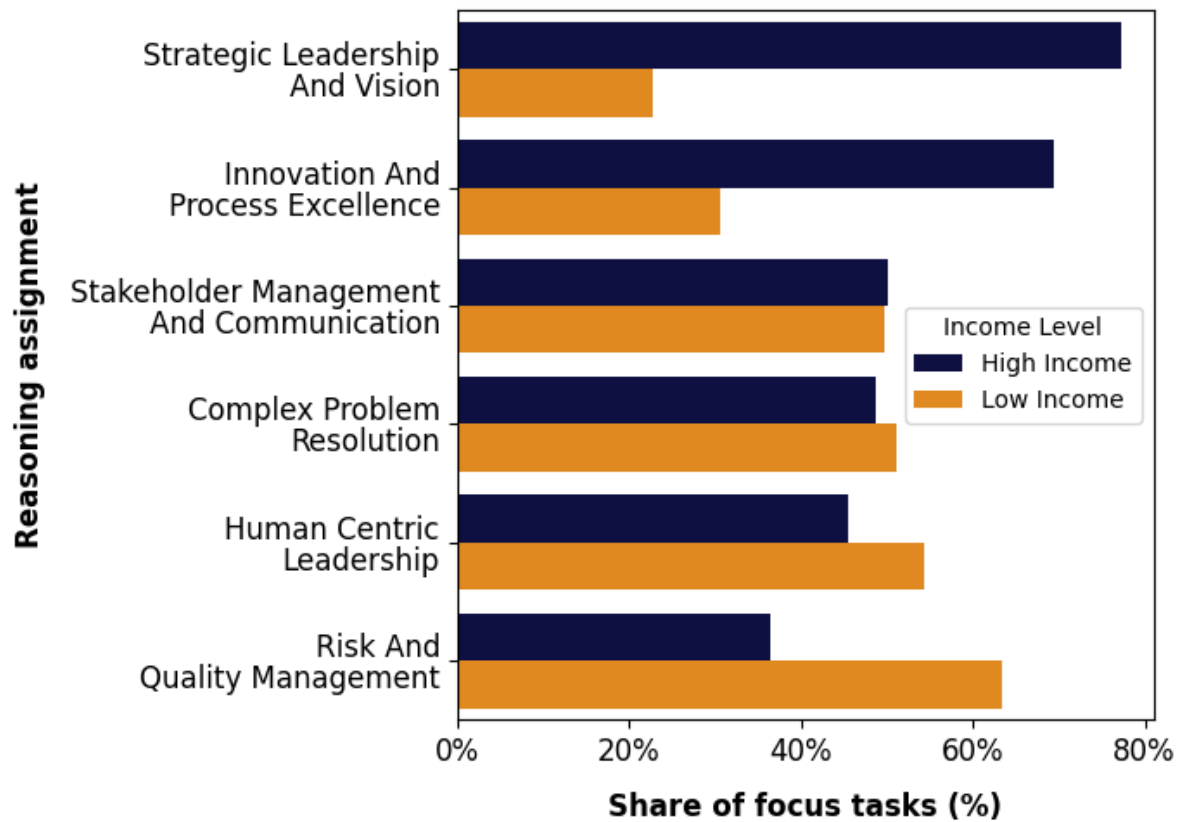


Figure 21: Share of focus tasks assigned into each kind of reasoning

Each focus task may be assigned to up to three reasoning clusters. See Section [3.4.1](#) for details.

Table 14: Comparing time allocations following the automation of tasks in non-cashable roles (%)

Task Category	Pre-automation	Post-automation	Focus Task	Augment + Reorder	New Tasks Added
Stakeholder Engagement	8.91	12.44	13.01	12.13	11.50
Team Leadership	12.41	15.15	15.66	14.71	11.23
Prison Management	3.51	3.85	4.07	4.15	2.36
Policy Development	13.58	14.08	14.59	14.32	18.25
Risk Management	9.08	9.52	9.94	10.97	21.45
Service Delivery	12.57	12.70	12.14	12.21	8.67
Performance Planning	11.71	11.03	11.32	10.97	6.60
Admin Support	8.67	7.84	6.85	7.25	2.39
Records Management	8.75	6.71	5.98	6.58	2.58
Data Analysis	10.80	6.68	6.43	6.70	15.00

Pre-automation shows the time shares of the non-cashable roles with automatable tasks before any automation takes place. The post-automation time shares show the impact of automating away high exposure tasks and recalculating time spent on remaining tasks, with newfound time distributed proportionally. The focus task column shows the time shares when a single focus task is allocated all the freed up time from the automatable tasks. The augment and reorder column shows the case where the LLM is free to first augment and then reorder all remaining tasks in a job description (excluding the automatable tasks), followed by task weights recalculated and aggregated. Finally, new tasks added shows the time shares when the LLM is free to suggest new tasks, categorise them, reorder all tasks that aren't automated and then recalculate task weights based on this new order. All figures rounded to two decimal places.

The “post-automation” shows a simple check in which rather than reallocating all freed up time from automation to a single focus task, we provide the LLM with the full job description and summary and allow it to reorder all remaining tasks. With the tasks reordered, we recalculate the task weights with the usual decay rate of 0.75 and aggregate this across roles to look at total time allocation.

The key result is that most of the shift in time spent on different task categories comes from the automation of tasks, rather than whether we specifically choose a single focus task or allow the LLM to freely reorder all remaining tasks. As the final column also implies, it is the adding or removing of tasks that seems to have the most impact in changing workers' task time allocations, rather than reordering within a given set of tasks.

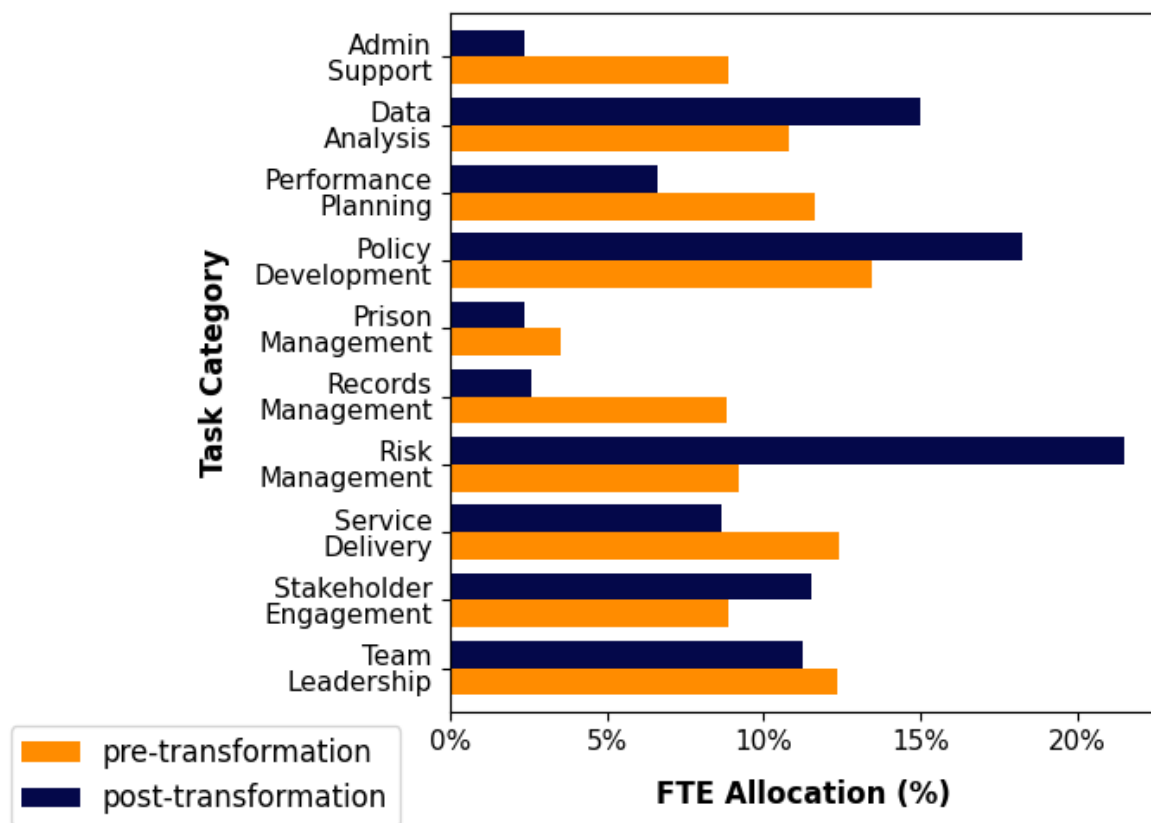


Figure 22: Total time allocation changes following the suggestion of new tasks and reordering of all tasks

The number of new tasks suggested per role is limited to the number of old tasks automated. The LLM returns all tasks for each role in their order to role importance. Weights are recalculated using the usual weighting and decay process set out in Section 2.3. See Prompt 5 in the Supplementary Materials, Section C.4 for full details of the LLM’s input.