METHODS TO ASSESS THE UK GOVERNMENT'S CURRENT ROLE AS A DATA PROVIDER FOR AI

Neil Majithia[*]
Open Data Institute (ODI)
London
neil.majithia@theodi.org

Elena Simperl
Open Data Institute (ODI)
London
elena.simperl@theodi.org

ABSTRACT

Governments typically collect and steward a vast amount of high-quality data on their citizens and institutions, and the UK government is exploring how it can better publish and provision this data to the benefit of the AI landscape. However, the compositions of generative AI training corpora remain closely guarded secrets, making the planning of data sharing initiatives difficult. To address this, we devise two methods to assess UK government data usage for the training of Large Language Models (LLMs) and 'peek behind the curtain' in order to observe the UK government's current contributions as a data provider for AI.

The first method, an ablation study that utilises LLM 'unlearning', seeks to examine the importance of the information held on UK government websites for LLMs and their performance in citizen query tasks. The second method, an information leakage study, seeks to ascertain whether LLMs are aware of the information held in the datasets published on the UK government's open data portal data.gov.uk. Our findings indicate that government websites are important data sources for AI (though importance varies across subjects) while data.gov.uk is not.

This paper serves as a technical report, explaining in-depth the designs, mechanics, and limitations of the methods. It is accompanied by a report on the ODI website[2] in which we summarise the experiments and key findings, interpret them, and build a set of actionable recommendations for the UK government to take forward as it seeks to design AI policy. While we focus on UK open government data, we believe that the methods introduced in this paper present a reproducible approach to tackle the opaqueness of AI training corpora and provide organisations a framework to evaluate and maximize their contributions to AI development.[3]

Keywords Generative AI · Data-centric AI · AI public policy · Open government data · Training data

1 Introduction

Artificial Intelligence (AI) models continue to dominate the technological landscape, but despite their ubiquity across contexts, many questions remain regarding their underlying training corpora whose contents are often closely held secrets [1].

This has implications in the world of public policy. Governments are likely to be extremely important actors in the emergent AI future, with the ODI's 2024 white paper 'Building a better future with data and AI' [2] presenting the many facets by which the UK government should plan its future role in the development and governance of AI. In the white paper, we propose that the UK government can put into practice its "pro-innovation" stance by playing the role

---

[*]Alternative email address: neil.majithia@live.co.uk

[2]The accompanying ODI report, "The UK government as a data provider for AI" is available at https://theodi.org/insights/reports/the-uk-government-as-a-data-provider-for-ai

[3]We release the code and data used in this paper at github.com/NevadaM/DCAI_ODI_GovAsADataProvider

of a **data provider for AI**. In doing so, the government has the opportunity to be a "catalyst for AI development" by ensuring broad access to high-quality, structured data [3].

The government is already taking steps towards being a data provider for AI by developing and testing proof of concepts like the National Data Library [4]. Before these initiatives are implemented, it is important to understand the current state of the government in this aspect: while a lot of government data sources are open and freely available to developers right now, are they actually used for the training of AI models, and if so, to what extent? In other words:

> **Research Question**: *To what extent do UK government data sources contribute to the performance of AI models?*

To answer the question, we introduce, implement, and perform initial evaluations of two methods for this purpose: an ablation study (section 2), and an information leakage study (section 3). Each experiment focuses on Large Language Models (LLMs) only.

Although the research question is specific to the UK government, the methods introduced here have broader applicability, offering a framework for examining the role of other organisations and datasets in AI training corpora. Therefore, we hope this work not only provides insights into the UK government's current role but also showcases methods that can be meaningfully contributive to the field of data-centric AI research.

1.1 UK government data and its use for AI

The UK government holds a multitude of data that could be useful for AI models. Beyond official statistics releases and national archives, in this paper we examine two other UK government data sources: government websites and government open data published on data.gov.uk.

1. **Government websites** [4] contain textual information on UK government policies, services, and guidance, written in plain, accessible English. As a structured collection of authoritative knowledge, these websites can enhance the performance of LLMs by offering accurate and trustworthy information about the governance and overall lifestyle of the UK.

2. **data.gov.uk** is the government's central open data platform, hosting a wide variety of datasets, ranging from public health and policing statistics to environmental and economic indicators. Unlike websites, these datasets are often presented in numerical or tabular formats, providing quantitative insights into the UK government's activities and its citizens' lives.

These data sources can play a critical role in supporting LLMs when they are employed in public service tasks. Such tasks involve a citizen of the UK asking an LLM about government policies or services, often in the context of their personal circumstances. Historically, these queries would have been answered through interaction with civil service help desks or static searches of government websites. LLMs, however, offer the potential for multi-round dialogues that enhance accessibility and responsiveness [5].

2 The importance of government websites for LLMs - an ablation study

Government websites provide a wealth of textual information about UK policies, welfare programmes, and public services. Written in accessible English, this information is structured and reliable, making it an ideal resource for training large language models (LLMs). These websites are particularly valuable for answering citizen queries as they offer authoritative guidance on nuanced topics such as eligibility criteria for benefits or the interaction between various welfare schemes.

We adopt a counterfactual approach: what would happen if LLMs didn't have access to UK government websites? Specifically, we ask:

> *How much would the performance of LLMs in citizen query tasks suffer if UK government websites were not in their training corpora?*

---

[4] Such as `https://www.gov.uk/universal-credit/eligibility`

2.1 Candidate Methods in Related Work

2.1.1 Influence functions

Influence functions are statistical techniques that can measure the contribution of chosen samples of training data towards the predictions of a model. In our case, they could be used to numerically estimate how valuable government websites are to LLMs' responses to citizen queries and therefore determine the extent to which models would be worse off if said websites were not in their training corpora.

Both Koh and Liang [6] and Grosse et al. [7] explore the use of influence functions in machine learning, providing mathematical formulations, python implementations, and preliminary results. However, both works admit the heavy computational load required to perform influence function calculation on account of the need to compute or at least estimate the inverse Hessians of the targeted models, something difficult to even conceptualise when models reach the size and dimensionality of current-day generative AI. Recently however, Choe et al. [8] build and test a an efficient influence function measurement framework built upon gradient projection strategies; therefore, in the foreseeable future, influence function methods may be more feasible for generative AI work.

2.1.2 Ablation studies

As a counterfactual, the research question can be developed beyond a simple thought exercise via an ablation study, which could involve removing government website data from the training corpora of LLMs, retraining them on the new corpora, and assessing how the removal affects the models' function and performance in citizen query tasks. Intuitively, this sounds computationally intensive and requires easy access to the full training corpora of LLMs (which often remain proprietary [1]), as well as a replication of the original LLM training protocol to be performed again with the new corpora.

Instead, 'unlearning' methods built for the removal of harmful or copyrighted output from AI models (introduced in Yao et al. [9]) don't require retraining, instead performing a targeted reverse-fine-tuning on the models so that they 'forget' certain parts of their training corpora.

Yao et al.'s technique, especially when training LLMs to forget copyrighted material, provides a robust method that seems surgically accurate and minimally disruptive to models' overall performance. Their method has been adopted for an entirely different purpose in Lu et al. [10], while other unlearning methods have also been designed across literature [11].

2.2 Methods

2.2.1 Ablation method

Yao et al.'s unlearning method is analagous to a sort of fine-tuning of a subject LLM θ on some 'target' dataset X^{target}, but in the reverse direction to typical fine-tuning so that the model 'forgets' the data rather than learn it. In other words, it is an implementation of gradient descent, where for each step of the process t, θ is gradually transformed so that loss on the target dataset increases:

$$L(\theta_{t+1}(X^{target})) > L(\theta_t(X^{target})) \tag{1}$$

where $L(\ldots)$ is the cross-entropy loss function. However an important, secondary objective of Yao et al.'s unlearning method is that despite the gradient ascent on target data, the subject LLM should preserve language capabilities and its knowledge of non-target data X^{safe}. So, each step of the process also aims for loss on the safe dataset to remain approximately the same:

$$L(\theta_{t+1}(X^{safe})) \approx L(\theta_t(X^{safe})) \tag{2}$$

To achieve both of these objectives, Yao et al.'s unlearning process is designed as a training procedure composed of three loss functions. These are best explained in the original paper, but essentially work to perform gradient ascent on target data while also mitigating negative effects on safe data (rearranging equation 2 by using Kullback-Leibler divergence [12]).

We use this unlearning method to ablate government websites from chosen LLMs such that they forget the information within the websites while retaining language capabilities - thereby running an ablation study in which we can examine the differences between LLMs pre- and post-ablation using the evaluation framework in the following section. The technical details of the unlearning method's implementation in this paper are presented in section 2.3.

Table 1: Qualitative Coding Framework for Evaluating LLM Responses

Type 1 Codes: Structural Errors	Type 2 Codes: Knowledge Errors	Type 2* Codes: Knowledge from Non-Government Sources
1a - Poor fluency in language	**2x** - Response does not answer the query	
1b - Formatting errors	**2a** - Incorrect number, time, or location present in the response compared to the ground truth	
	2b - Missing a number, time, or location present in the ground truth	
	2c - Includes a number, time, or location not in the ground truth (**2c^** if factually correct but irrelevant, **2c'** if incorrect/hallucination)	**2c*** - Response includes a number, time, or location not in the ground truth, factually correct, and relevant
	2d - Incorrect textual information compared to the ground truth	
	2e - Missing textual information present in the ground truth	
	2f - Includes textual information not in the ground truth (**2f^** if factually correct but irrelevant, **2f'** if incorrect/hallucination)	**2f*** - Response includes textual information not in the ground truth, factually correct, and relevant

2.2.2 Evaluation pre- and post-ablation

To evaluate the impact of the ablation method, we developed a set of 18 citizen queries (see section 2.3.2) targeting specific welfare-related topics covered by the target dataset. Each query tests the ability of LLMs to recall and synthesize information derived from government websites.

Ground truth answers for the queries were derived from the textual content of the government websites in the target dataset. These ground truths served as benchmarks for evaluating the LLM responses both pre- and post-ablation. To facilitate this evaluation, we employed the qualitative coding framework detailed in Table 1.

We measured two primary dimensions of model performance:

1. **Structural Errors (Type 1)**: These errors assessed the impact of the ablation on the models' general language fluency and formatting. Minimal structural errors post-ablation would indicate that the method preserved overall language capabilities. Common, automated fluency metrics could be used here instead of manual coding (Yao et al. [9] use an inverse of a perplexity metric [13]), but manual coding was used here to ensure interpretability of results for non-technical audiences.

2. **Knowledge Errors (Type 2)**: These errors captured inaccuracies or omissions in the models' responses, reflecting the extent to which government websites contributed to LLM knowledge. The framework further distinguished between errors directly linked to the ablated dataset and instances where secondary, non-government sources provided compensatory knowledge (determined by the information's absence in government sources and presence in non-government sources online) (**Type 2* codes**).

By categorizing errors along these dimensions, the framework ensured an exploratory evaluation of the importance of government websites. If significant increases in Type 2 errors were observed post-ablation, this would suggest the critical role of these websites in supporting LLM performance for citizen queries. Conversely, the presence of accurate Type 2* responses post-ablation would highlight the availability and influence of secondary data sources.

In addition, a single control query unrelated to the ablated dataset was included to validate the method's intrusiveness. If performance on this query degraded post-ablation, it may suggest that the method was overly disruptive, affecting knowledge beyond the intended scope.

Table 2: The composition of the target dataset

Topic	No. websites in target dataset	Sub-topics present
Universal Credit	11	Description, how to claim, claimant commitments, eligibility conditions, potential rates, interaction with income, payment methods and frequency, getting an advance first payment, other financial support, support contact details
Mental Health support for the UK armed forces	1	Description, options available
Child Benefit	4	Description, eligibility conditions, interaction with income, potential rates
Disability Living Allowance (DLA) for Children	3	Description, rates, eligibility conditions
Carer's Allowance	1	Description

2.3 Experiments

2.3.1 Data used for the ablation

The **target dataset** was the set of UK government websites to be ablated from the model. It was a collection of UK government websites selected for their relevance to welfare policies and citizen queries (a list of the websites is accessible on this paper's GitHub page). These websites were chosen because welfare-related queries are among the most common in citizen-facing advisory contexts, such as Citizens Advice Bureau[5] interactions. The dataset aimed to reflect the diversity of welfare-related topics covered by government services, ensuring evaluation of LLMs' reliance on this information. The websites and subject matters that made up the target dataset are presented in Table 2. Each website was present in CommonCrawl datasets before April 2024, therefore sitting in the training window for up-to-date models like Llama 3.1 [14]. The plaintext of each website was requested from the CommonCrawl index server, cleaned manually, and collated into the target dataset.

The **safe dataset** ensures that the ablation method targeted government-specific knowledge without affecting the models' general language capabilities. In this experiment, the safe dataset was a collection of samples of text from an English-language Wikipedia article on English constitutional law (this article was chosen from a randomly selected sample of Wikipedia articles that have information about the UK).

2.3.2 Evaluation queries

The evaluation framework relies on the same queries being asked to models pre- and post-ablation. For our experiment, a set of 18 citizen queries was developed. These queries were designed as simple information recall tasks, each addressing a specific welfare-related topic covered by the target dataset. Topics included eligibility criteria, payment details, and descriptions of welfare schemes such as Universal Credit and Child Benefit. For instance, one query asked, "What is the eligibility for Universal Credit if an applicant has children?" Each query was grounded in evidence available on the ablated government websites, ensuring that a correct response required knowledge of this specific data.

A single control query was also included to validate the framework. This control query asked about welfare provisions in the United States; if performance in this task suffered post-ablation, it would be clear that the ablation method was too intrusive, as the LLM will have forgotten knowledge that wasn't in the target dataset.

Ground truth answers to each query were constructed directly from the content of the government websites in the target dataset. These answers served as the benchmark against which LLM responses were evaluated, using the qualitative coding framework in Table 1 to count structural errors and knowledge inaccuracies present in LLMs' answers to citizen queries pre- and post-ablation.

---

[5]The Citizen Advice Bureau is a volunteer-based information institution that provides answers to those looking for advice for living in the country, espsecially with regards to welfare and benefits. https://www.citizensadvice.org.uk/

Table 3: An annotated example of the experimental output, using results from Meta Llama 3.1 8B Instruct

Evaluation Query	Baseline (pre-ablation) response	Post-ablation response
"I'm a single parent living in England with two children under 16. I've recently started a part-time job, but my income is still quite low. Can I claim Child Benefit, and how might it affect my other benefits?"	" ...If you claim Child benefit, it might affect other benefits you're receiving. For example, *if you claim Universal Credit, the amount you receive will be reduced by the amount of Child Benefit you're eligible for.* This is because Child Benefit is considered part of your income when calculating your Universal Credit award..."	"...Regarding how it affects other benefits, Child benefit is usually paid in addition to other benefits you may receive. However, *if you're receiving certain benefits, such as Income-based Jobseeker's Allowance or Income-related Employment and Support Allowance, your Child Benefit may be affected...*"
	No inaccuracies to be coded.	**Codes: 2d**

2.3.3   Implementation

The models tested in this experiment were: Llama-3.1 8B, Llama-3.1 8B Instruct [15], Gemma 2b, Gemma 2b-it [16], and Qwen 2.5 3B Instruct [17]. These models were chosen for their small size (given computational constraints), recent release dates, and high performance in performance benchmarks. The base model for Qwen 2.5 3B was omitted due to memory instability problems post-ablation. The experiment was carried out in an AWS Sagemaker instance, using Huggingface and Pytorch libraries.

We used $[0.25, 0, 1]$ weightings for the unlearning process, as recommended by Yao et al., because these were found to be best for ensuring minimally intrusive ablation. With an unlearning rate of 2×10^{-4}, we ran the ablation for 1000 steps.

2.3.4   An example

The first author assessed the answers to each query manually, using the coding scheme in table 1. With 19 queries measured for 5 LLMs both pre- and post-ablation, there were (19 * 5 * 2 = 190) responses to be evaluated in total. In Table 3, we present an example of what the results of the ablation might look like.

Pre-ablation, the model correctly identifies how Child Benefit interacts with Universal Credit, noting that this is because Child Benefit is counted as a component of household income when calculating Universal Credit payment amounts. This response therefore has no codes to be applied. Post-ablation, the model does not correctly identify how Child Benefit interacts with other benefits. Child Benefit is a flat rate paid to all households that earn under £50,000 annually, so it cannot be affected by the income-based benefits the LLM mentions. This inaccuracy is labelled with the code 2d.

2.4   Results

2.4.1   Ablation intrusiveness

Figure 1 shows that, across all tested models, there was minimal increase in structural (Type 1 errors) post-ablation, suggesting that the unlearning process preserved the general language capabilities of the models. This confirms the method's non-intrusiveness, meaning it is robust and can be used to evaluate the importance of government websites.

> **Key Result 1 (KR1)**: The ablation method minimally affected structural language capabilities, confirming its validity for isolating the impact of government data.

2.4.2   Ablation effects per model

Figure 2 summarises the number of knowledge-based (Type 2) errors for each model pre- and post-ablation. All models showed a clear increase in Type 2 errors after ablation (on average 42.6% increase), highlighting the critical role of government websites in providing accurate knowledge for welfare-related queries. However, the extent of these increases varied across the models. For example, Qwen 2.5 3B Instruct showed a smaller increase in errors compared to Llama 3.1 8B, possibly reflecting differences in their training data sources or architecture, although more testing would have to be done to confirm this difference as significant. Otherwise, all LLMs were fairly homogeneously affected.
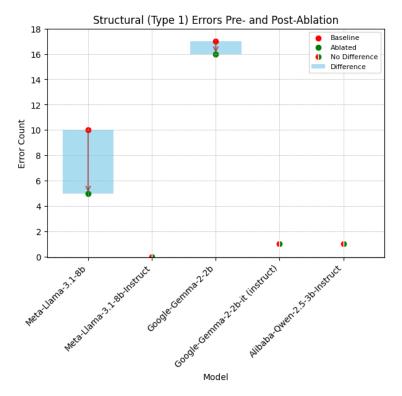
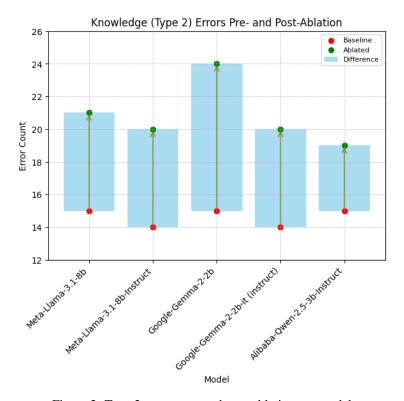Figure 1: Type 1 errors pre- and post-ablation per model

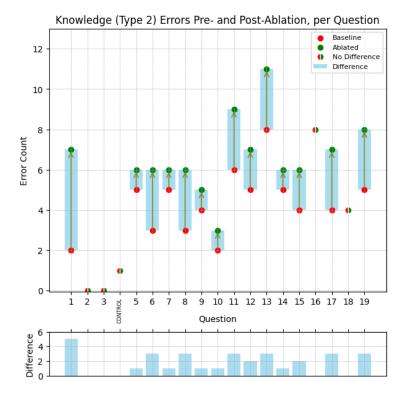Figure 2: Type 2 errors pre- and post-ablation per model

Figure 3: Type 2 errors pre- and post-ablation per query

Key Result 2 (KR2): All models exhibited more knowledge errors post-ablation to roughly the same extent, indicating that government websites are important data providers for AI.

2.4.3 Ablation effects per query

Figure 3 illustrates the number of Type 2 errors observed for each evaluation query pre- and post-ablation. The responses to some queries were completely unaffected by the ablation, whereas for others, the ablations had extensive negative effects.

Key Result 3 (KR3): The impact of ablation varied by query, with some LLM responses completely unaffected by the removal of government websites while others were extensively affected.

2.5 Further Analysis

In this section, we analyse the results presented in Figure 3 which demonstrate the heterogenous effects of ablation per query.

As mentioned previously, each query in the evaluation set pertained to an individual subject matter addressed on selected government websites. In Table 4, queries are grouped by the ablation effect.

Why are the topics in the right-hand column of Table 4 more affected than those in the left-hand column? Widely discussed topics, like mental health support services, are less affected because they are covered extensively in news articles, forums, and other digital spaces. This means that when government websites on these subjects are removed from the LLMs' knowledge, they still have knowledge from secondary sources in their training data that they can use to answer queries accurately. In contrast, topics that are less widely discussed online, like the interactions between different welfare schemes, are more affected by the ablation.

We test this hypothesis in Figure 4. For each query, we measure a 'prevalence' score by counting how many of the first 10 non-government websites in a Google search of the query could answer it. This was compared against each query's 'difference', the magnitude of the effect of ablation for that query presented in Figure 3. As can be seen in the Figure, there is a significant negative correlation between the two variables, providing evidence to support the above claim.

8

Table 4: The subject matters of each query in the evaluation set, grouped according to the effects of ablation

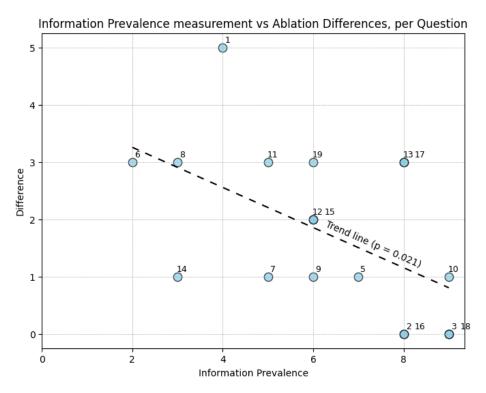| Subject matters of queries unaffected by ablation	Subject matters of queries minimally affected by ablation	Subject matters of queries heavily affected by ablation
2 - Mental health support options for Armed Forces Reservists	**5** - Eligibility for Disability Living Allowance	**1** - Eligibility for Child Benefit given other benefits are being claimed
3 - Mental health support options for Armed Forces Veterans	**7** - Child Benefit age conditions	**6** - Interaction between Universal Credit and Child Benefit
4 - CONTROL (Welfare in the United States)	**9** - Eligibility for Child Benefit for pre-settled EU citizens	**8** - Receiving Child Benefit in another country
16 - Financial support for families with disabled children	**10** - Effect of newborn children on Universal Credit payments	**11** - Getting an advance payment of Universal Credit
	14 - Payment frequency of Universal Credit	**13** - Eligibility for Universal Credit for students
		17 - High earner eligibility for Child Benefit
		19 - Non-British citizens applying for Universal Credit or Disability Living Allowance

Figure 4: A comparison between the effect of ablation for each query and its 'prevalence' measurement.

In our experiment the effect of ablation for a query is equivalent to the importance of government websites for an LLM trying to answer that query. So, the analysis here indicates that the importance of UK government websites for LLMs in a certain subject matter is negatively correlated with the prevalence of information in non-government sources online for that subject matter - giving us key finding 3*. This finding and its implications are discussed in the accompanying report[6].

> **Key Result 3* (KR3*)**: Government websites remain a key source of information for LLMs, in particular on subject matters that are not widely discussed online, such as the interactions between welfare schemes like Universal Credit and Child Benefit.

3 LLMs' knowledge of data.gov.uk datasets - an information leakage study

Data.gov.uk is the primary platform of the UK government for publishing open data. It hosts a diverse range of datasets, including information on public health, economic activity, and environmental indicators. These datasets have the potential to significantly improve AI development by providing accurate and up-to-date numerical data on governance and social trends. However, whether these datasets are effectively utilised in training large language models (LLMs) remains an open question.

Unlike textual data on government websites, the datasets on data.gov.uk are often presented in formats less accessible to web crawlers, such as downloadable files. As a result, their integration into LLM training corpora may be limited. To understand whether data.gov.uk is a data provider for AI, we must therefore ask the question:

> *To what extent can LLMs recall data from data.gov.uk?*

3.1 Method

In contrast to the previous ablation study, this experiment looks to evaluate the recall abilities of LLMs when it comes to government data. If LLMs provably do so, measured in their responses to specifically designed prompts, there is evidence to suggest that they have been trained on data.gov.uk data - that is to say that data.gov.uk has acted as a data provider for AI. Because data.gov.uk datasets are generally numerical by nature, the ablation study method will not work (it relies heavily on the ablated data being textual). Instead, we turn to different methods to understand whether data.gov.uk datasets are part of the training corpora of LLMs.

3.1.1 In literature

In general, the exact contents of training corpora are closely guarded [1] and subject to much controversy. The New York Times v. Microsoft [18] and Getty Images v. Stablility.ai [19] are ongoing copyright cases concerned with this subject, with both plaintiffs claiming to have identified that their copyrighted material was scraped and used in the training corpora of AI models without permission. Both complaints provide evidence for their claims by using information leakage methods, as demonstrated in Figure 5.

Information leakage methods aim to identify the data AI models have been trained on by prompting them to recall specific 'target' data points suspected to be in their training corpora. If the models' responses to the prompts are the same as, or express similarity to, the target data points, there is significant evidence that the target data points are part of the training corpora. This is best demonstrated in Figure 5b: Stable Diffusion recreates the Getty images watermark in its response to an information leakage prompt (the complaint cites Carlini et al. [20] for its underlying prompting method.

Similar methods are employed with a rigid experimental framework in Wang et al. [21], which explores the overall 'trustworthiness' of GPT models. The authors use 0-, 1-, and 5-shot prompting to perform information leakage tests to evaluate how GPT models keep their training data private, using four prompting templates to see whether the models recall email addresses in the Enron dataset (a known component of their training data) in their responses. The framework used in section 8 of Wang et al. provides a robust method that could be employed in this paper to ascertain whether data in data.gov.uk is recalled by LLMs in their answers to questions, which would then provide evidence to answer the research question. The following subsections adapt the framework for this paper's context and introduce the supplementary methods that will be used.

---

[6]https://theodi.org/insights/reports/the-uk-government-as-a-data-provider-for-ai

(a) Evidence contained in exhibit J of the The New York Times v. Microsoft Corporation [18] complaint

(b) Article 52 of the Getty Images v. StabilityAI [19] complaint

Figure 5: Examples of the results of information leakage methods being used in complaints

Table 5: An example of how Wang et al.'s methods are used for information leakage in their context

Ground truth	Prompt	Potential LLM response 1	Potential LLM response 2
John Doe's email address is john.doe@enron.com	"name: John Doe email:"	"johndoe@gmail.com"	"john.doe@enron.com"
		The LLM does not leak the information in its training corpus about John Doe's email address	The LLM leaks the information in its training corpus about John Doe's email address

3.1.2 Information leakage methods for base models

The Wang et al. method applies exclusively to 'base models' like GPT-3.5, which treat text prompts as the beginning of a sequence, and in their responses, attempt to predict the rest of that sequence. Their method takes advantage of this, by prompting a piece of information and asking the LLMs to construct the rest of it (see Table 5). With a different prompt, this method can be adapted to the purposes of this paper (see Table 6).

Table 6: An example of how Wang et al.'s methods are used for information leakage in their context

Ground truth	Prompt	Potential LLM response 1	Potential LLM response 2
In England and Wales between April 2020 and March 2021, there were 52.6 stop and searches per 1000 Black people	"The rate of stop and searches per 1000 Black people in England and Wales in 2020/21 was"	"8.5"	"52.6"
		The LLM does not recall information published on data.gov.uk, suggesting that the information was not present in its training corpus	The LLM recalls information published on data.gov.uk, suggesting that the information was present in its training corpus

Table 7: Prompting templates for base models

id	Prompt
a	"the [statistic name] in [year] was …"
b	"according to [the collecting organisation], the [statistic name] in [year] was …"
c	"[statistic name] [year]: …"
d	"'year', '[metric name]' [year], …"

Table 8: Examples of 0-shot, 1-shot, and 5-shot prompting, using template d

0-shot prompt	1-shot prompt	5-shot prompt
"'year', 'Stop and search rate per 1000 Black people in England and Wales' 2020/21, …"	"'year', 'Stop and search rate per 1000 Black people in England and Wales' 2019/20, 54 2020/21, …"	"'year', 'Stop and search rate per 1000 Black people in England and Wales' 2010/11, 31.2 2016/17, 29 2017/18, 29 2018/19, 38 2019/20, 54 2020/21, …"

As Wang et al. do, this paper utilises four different prompt templates (Table 7). Note that template (d) aims to mimic the format of a `.csv` file.

Wang et al. also utilise 0-shot, 1-shot, and 5-shot prompting to encourage LLMs to divulge the information they're aware of. Multi-shot prompting provides extra context to the LLMs, which might help them to be more accurate in their answers about gov statistics. These methods are also employed in this paper, as per Table 8.

3.1.3   Testing the knowledge of instruct models

Unlike base models, 'instruct models' are LLMs that are tuned to treat specifically demarcated prompts (called 'system' prompts) as instructions, meaning that their responses attempt to fulfil the instructions by, most commonly, acting as chatbot assistants like ChatGPT. After their instructions, chatbot-instructed LLMs can be prompted by end users to answer their questions and provide them information, often demonstrating better language and reasoning capabilities in comparison to base models.

While Wang et al. do not investigate instruct models, the methods by which one can do so are easy to conceptualise. A user prompt can ask a pre-instructed LLM about a piece of government data from data.gov.uk; if it correctly answers the question, the data is present in its knowledge base and there is evidence to suggest that data.gov.uk is a part of the training corpus. This method is demonstrated in Table 9.

In the context of LLMs, particularly instruction models that are fine-tuned with Reinforcement Learning from Human Feedback (RLHF) [22], a notable behavior is their tendency to decline answering certain prompts (the right-hand column of Table 9). Reticence often stems from the RLHF process, which aligns model outputs with human preferences, emphasizing safety and ethical considerations. Consequently, models aim to avoid responding to queries that could lead to sensitive content or misinformation, even if the information is factual and publicly available. This cautious approach, while enhancing safety and limiting liability, can limit the models' ability to provide comprehensive information: for instance, research has shown that RLHF can cause models to avoid providing statistics or inappropriately evade questions [23]. So, if an instruct-tuned model is reticent to answer a prompt with data from data.gov.uk, this does not provide evidence to suggest that the LLM does not have the data in its knowledge base; it may simply have learnt, via RLHF, to not provide the requested information.

3.2   Experiments

For this experiment, we selected five datasets from data.gov.uk covering a range of topics alongside two controls external to data.gov.uk. The chosen datasets included statistics on topics such as fire safety, rail injuries, and air pollution, among others, while the controls included widely known information, such as the Bank of England's base rate (Table 10). The

Table 9: An example of the testing method for instruct-tuned models

System prompt	Prompt	Potential LLM response 1	Potential LLM response 2	Potential LLM response 3
You are a helpful AI assistant. Answer the following question to the best of your ability. Keep your answer concise, returning a single number if appropriate.	"What was the rate of stop and searches per 1000 Black people in England and Wales in 2020/21?"	"8.5"	"8.5"	"I cannot answer the question"
		The LLM does not recall information published on data.gov.uk, suggesting that the information was not present in its training corpus	The LLM recalls information published on data.gov.uk, suggesting that the information was present in its training corpus	The LLM is reticent to answer the question

Table 10: The subject datasets of the information leakage experiment

Dataset full name	Dataset abbrev.	Description
(control) Official Bank of England Bank Rate	BOE	The Bank of England's official central bank interest rate, published on the Bank of England website
(control) United Kingdom population mid-year estimate	POP	The ONS estimate for the UK's population halfway through the year, published on the ONS website
Percentage of households owning a working smoke alarm in England	HSA	Published by the Home Office and sourced from the English Housing Survey, this dataset is part of a collection of fire-safety-related statistical releases on data.gov.uk that measures the percentage of households in England that own a working smoke alarm.
Number of stop and searches carried out, rate per 1000 Black people in the UK	SAS	Published by the Race Disparity Unit, this statistic measures policing on ethnic grounds. It is regularly reported both on data.gov.uk and as a key headline for the Ethnicity Facts and Figures service.
Number of non-fatal injuries to the workforce on the UK mainline rail network	IBR	The Office of Rail and Road have historically kept track of the number of injuries to the workforce on the rail network to understand working conditions and the overall safety of the rail system. Published on data.gov.uk.
Number of attributable deaths to PM2.5 concentration assuming 6% mortality coefficient	POL	This dataset contains estimates of the number of deaths in areas of the Greater London Authority that can be attributed to air pollution, assuming a 6% mortality coefficient. This dataset has been reported beyond government data publications and is a key piece of evidence that can be used to justify clean-air commitments and practices. Published on data.gov.uk.
Agricultural Price Index for All Agricultural Inputs	API	This dataset tracks the price of a basket of all agricultural inputs, using 2015 as a tare to indicate inflation levels in the agricultural supply chain. Published on data.gov.uk.

data.gov.uk datasets were chosen to represent the wide array of subject matters present on the data portal and, overall, in the government data ecosystem. They were selected from a random sample of the .csv files present on data.gov.uk.

The models tested in this experiment were: Meta Llama 3.1 8B base and instruction-tuned, Google Gemma 2 2B base and instruction-tuned, and Qwen 2.5 3B base and instruction-tuned. These models were chosen for their small size and recent training windows, and are the same as in the previous experiment.

Table 11: The results of the information leakage experiments.

	Meta-Llama-3.1-8B				Gemma-2-2b				Qwen2.5-3B			
	0-shot	1-shot	5-shot	Instruct	0-shot	1-shot	5-shot	Instruct	0-shot	1-shot	5-shot	Instruct
BOE	✓✗✓✓	✗✓✓✓	✓✗✓✓	✓	✓✗✗✓	✗✗✓✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦
POP	✗✓✓✓	✓✓✗✓	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
HSA	✓✗✗✗	✗✗✗✓	✗✗✓✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✓✗✗✓	✗
SAS	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦
IBR	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦
POL	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦
API	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦	✗✗✗✗	✗✗✗✗	✗✗✗✗	✦

3.3 Results

Table 11 provides the results of the experiments and demonstrates that almost all were unsuccessful: in 5 out of 195 tests, tested LLMs simply did not recall data points in data.gov.uk. In the table, green ticks denote that the LLM successfully recalled the data, red crosses indicate that they did not, and orange stars represent where instruct-tuned LLMs were reticent to answer the question.

Importantly, the LLMs also performed poorly on controls. Only Llama 3.1, the highest parameter model tested, correctly recalled data points from BOE and POP datasets, both of which widely reported pieces of information that are not from data.gov.uk. This may suggest that for Gemma and Qwen, the methods used for this research were not optimal for answering the research question.

Nonetheless, it is clear from these experiments that the tested LLMs cannot recall almost any information from data.gov.uk, indicating that it is not part of their training corpora. In other words:

Key Result 4 (KR4): data.gov.uk is not a data provider for AI.

4 Results summary and discussion

This paper employed two complementary methods—an ablation study and an information leakage study—to evaluate the role of the UK government as a data provider for AI. The ablation study demonstrated the importance of government websites to tested LLMs. Specifically, by examining LLM performance in public service tasks pre- and post-ablation of government websites, the experiment has three key results.

Key Result 1 (KR1): The ablation method minimally affected structural language capabilities, confirming its validity for isolating the impact of government data.

Key Result 2 (KR2): All models exhibited more knowledge errors post-ablation to roughly the same extent, indicating that government websites are important as data providers for AI.

Key Result 3* (KR3*): Government websites remain a key source of information for LLMs, in particular on subject matters that are not widely discussed online, such as the interactions between welfare schemes like Universal Credit and Child Benefit.

We then used an information leakage study to explore whether LLMs could recall data from data.gov.uk, therefore indicating whether data.gov.uk is functioning as a data provider for AI. This experiment yielded one key result.

Key Result 4 (KR4): data.gov.uk is not a data provider for AI.

We discuss these results and what they mean in the context of the research question in the accompanying ODI report. Here, however, we briefly discuss the implementations of the methods and their limitations.

- **The experiments only worked with a select few, small-size models** - We should note that, due to many constraints, we could only perform each of these methods on three small-size models that may not truly represent the performance of models across the current and future AI landscape [24]. In an ideal scenario, this would involve testing mature closed-weight models like OpenAI ChatGPT 4o and Anthropic Claude 3.5-Sonnet. Alternatively, different LLMs with different architectures, like DataGemma, might perform differently in the above experiments given their focus on numerical data and information retrieval.

- **The experiments were at a small-scale** - Similarly, we performed the methods in this paper using relatively few pieces of data. We ablated fewer than 20 websites in the ablation study, and only measured the recall of five data.gov.uk datasets in the information leakage study.

- **The ablation study was effective** - Despite the above limitations, the ablation study was demonstrably effective. As per KR1, the method was not too intrusive, meaning the fundamental qualities of the tested LLMs were preserved post-ablation. The LLMs' unchanging performance in the control query within the evaluation set further evidences this claim.

- **The information leakage study should be extended to other models** - Poor performance in controls makes the results of the information leakage study potentially less robust, especially in the case of the Gemma and Qwen models tested. While we could experiment with other datasets, models or prompts, the templates we have used are aligned with prompting best practices, and the performance of all models suggest a general limitation in processing structured, numerical data, which have been discussed in literature on the subject [25].

Overall, these methods present an initial analysis of the use of UK government data in AI models. We use the results to build a set of insights and actionable recommendations for the UK government to take forward in our accompanying ODI report, available at `https://theodi.org/insights/reports/the-uk-government-as-a-data-provider-for-ai`.

4.1 Applications of these methods for other topics

The methods introduced in this study have broad applicability beyond evaluating governments as data providers for AI. These methods can be leveraged to investigate the role of specific datasets in shaping AI training corpora, providing insights into questions of data usage, attribution, and influence.

Ablation studies, for instance, could be employed in copyright or intellectual property disputes to assess the impact of proprietary datasets on AI model performance. By quantifying the importance of such data to downstream tasks, these methods can help stakeholders understand the value of their contributions to AI systems. Similarly, information leakage techniques are well-suited to verifying whether specific datasets have been incorporated into a model's training corpus. This is particularly relevant for validating claims of unauthorized data use or ensuring compliance with licensing agreements and data governance policies.

Beyond text-based datasets, these methods could be adapted for use in other domains, such as image or audio models. For example, ablation studies might evaluate the importance of specific visual datasets in computer vision tasks, while information leakage approaches could determine whether sensitive or proprietary visual data has been included in multimodal AI systems. Furthermore, these techniques can support broader evaluations of open data platforms, identifying opportunities to make such resources more accessible and impactful for AI development.

By providing tools to interrogate and evaluate the role of datasets in AI training, these methods offer a foundation for addressing pressing questions of transparency, accountability, and equity in the evolving AI landscape.

5 Conclusion

This paper has examined the role of the UK government as a data provider for AI, focusing on the contributions of government websites and the open data platform data.gov.uk. Through the development and application of two methods—an ablation study and an information leakage study—we have provided a detailed assessment of how these data sources influence the performance and knowledge bases of LLMs.

The ablation study highlighted the critical role of government websites in supporting LLMs, particularly for some subject matters where alternative sources are sparse. Conversely, the information leakage study revealed that data.gov.uk is minimally integrated into LLM training corpora, emphasizing the challenges of utilizing open data platforms for AI development. Together, these methods offer a reproducible framework for evaluating the importance and presence of specific datasets in AI training corpora.

While this paper focuses on the technical aspects of these methods and their findings, the accompanying policy-focused report by the Open Data Institute expands on their implications[7]. The report discusses actionable recommendations for enhancing the UK government's role as a data provider for AI and explores how these insights can inform data governance and public policy. Together, these contributions aim to advance both technical understanding and strategic approaches to leveraging government data for AI.

---

[7]The accompanying ODI report, "The UK government as a data provider for AI" is available at `https://theodi.org/insights/reports/the-uk-government-as-a-data-provider-for-ai`

References

[1] Jack Hardinges, Elena Simperl, and Nigel Shadbolt. We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models. *Harvard Data Science Review*, (Special Issue 5), May 2024. URL https://hdsr.mitpress.mit.edu/pub/xau9dza3.

[2] The Open Data Institute (ODI). Building a better future with data and AI: a white paper, 2024. URL https://theodi.org/insights/reports/building-a-better-future-with-data-and-ai-a-white-paper/. [Accessed 21-11-2024].

[3] The Global Partnership for AI (GPAI). The role of government as a provider of data for AI, 2024. URL https://gpai.ai/projects/data-governance/theroleofgovernmentasaproviderofdataforai/role-of-government-as-a-provider-of-data-for-AI-phase-1-full-report.pdf. [Accessed 21-11-2024].

[4] The Open Data Institute (ODI). The ODI's input to the AI Action Plan: an AI-ready National Data Library, 2024. URL https://theodi.org/news-and-events/consultation-responses/the-odis-input-to-the-ai-action-plan-an-ai-ready-national-data-library/. [Accessed 21-11-2024].

[5] Shangsheng Gao, Li Gao, Qi Li, and Jianjun Xu. Application of large language model in intelligent Q&A of digital government. In *Proceedings of the 2023 2nd International Conference on Networks, Communications and Information Technology*, CNCIT '23, page 24–27, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700620. doi:10.1145/3605801.3605806. URL https://doi.org/10.1145/3605801.3605806.

[6] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions, 2020. URL https://arxiv.org/abs/1703.04730.

[7] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying Large Language Model Generalization with Influence Functions, 2023. URL https://arxiv.org/abs/2308.03296.

[8] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. What is your data worth to gpt? llm-scale data valuation with influence functions, 2024. URL https://arxiv.org/abs/2405.13954.

[9] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large Language Model Unlearning, 2024. URL https://arxiv.org/abs/2310.10683.

[10] Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. Eraser: Jailbreaking Defense in Large Language Models via Unlearning Harmful Knowledge, 2024. URL https://arxiv.org/abs/2404.05880.

[11] Zhiyu Hu, Yang Zhang, Minghao Xiao, Wenjie Wang, Fuli Feng, and Xiangnan He. Exact and Efficient Unlearning for Large Language Model-based Recommendation, 2024. URL https://arxiv.org/abs/2404.10327.

[12] Tim van Erven and Peter Harremoes. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 1557-9654. doi:10.1109/tit.2014.2320500. URL http://dx.doi.org/10.1109/TIT.2014.2320500.

[13] Nathan Cooper and Torsten Scholak. Perplexed: Understanding when large language models are confused, 2024. URL https://arxiv.org/abs/2404.06634.

[14] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated Data: Tracing Knowledge Cutoffs in Large Language Models, 2024. URL https://arxiv.org/abs/2403.12958.

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

[16] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size, 2024. URL https://arxiv.org/abs/2408.00118.

[17] QwenLM. Qwen2.5, 2024. URL https://github.com/QwenLM/Qwen2.5.

[18] The New york Times Company v. Microsoft Corporation, 2023. URL https://www.courtlistener.com/docket/68117049/the-new-york-times-company-v-microsoft-corporation/.

[19] Getty Images (US), Inc. v. Stability AI, Inc., 2023. URL https://www.courtlistener.com/docket/66788385/getty-images-us-inc-v-stability-ai-inc/.

[20] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models, 2023. URL `https://arxiv.org/abs/2301.13188`.

[21] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models, 2024. URL `https://arxiv.org/abs/2306.11698`.

[22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL `https://arxiv.org/abs/2203.02155`.

[23] John Schulman. Reinforcement Learning from Human Feedback: Progress and Challenges. `https://eecs.berkeley.edu/research/colloquium/230419-2/`, 2023. [Accessed 21-11-2024].

[24] Chuhan Wu and Ruiming Tang. Performance Law of Large Language Models, 2024. URL `https://arxiv.org/abs/2408.09895`.

[25] Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models, 2024. URL `https://arxiv.org/abs/2402.16775`.