

The UK government as a data provider for AI

November 2024

ODI Research
ADVANCING TRUST IN DATA



Contents

Introduction	3
Key findings and recommendations	5
The importance of government websites for LLMs	7
LLMs and their knowledge of data.gov.uk	12
Discussion, recommendations and conclusions	15
Conclusion	18
Appendices	19

About

This report was researched and produced by the Open Data Institute (ODI) and published in November 2024. Its lead authors were Neil Majithia and Elena Simperl, with help from Matthew Kilcoyne and Emma Thwaites. If you want to share feedback by email or would like to get in touch, contact the DCAI project lead Elena Simperl at elena.simperl@theodi.org

To share feedback in the comments, highlight the relevant piece of text and click the 'Add a comment' icon on the right-hand side of the page.



How can it be improved? We welcome suggestions from the community in the comments.

Introduction

Artificial intelligence (AI) promises opportunities for greater economic efficiency and quality of life around the world, but may also bring new socio-technical challenges and safety risks that governments will need to address¹, both as regulators and as providers of data.

The ODI's 2024 white paper 'Building a better future with data and AI'² presents the many facets by which the UK government should plan to have a hand in the development and governance of AI. This may involve the building of regulation and legislation to protect UK citizens from AI-enabled harms. However, there are also opportunities for the government to put its 'pro-innovation' approach³ approach into practice by provisioning 'public goods' for the world of AI.

Public good provision has historically been thought of as a core facet of government⁴, typically with regard to infrastructure like roads, defence and public health, all of which increase quality of life and productivity for the entire population of a country. Governments can provide infrastructure to the same effect in the context of AI, provisioning goods and services that developers and users can use to ensure that AI innovation continues at a rapid pace while remaining responsible and safe. The UK [AI Safety Institute](#) (AISI) has been set up for this exact purpose, with its open-source [Inspect](#) platform allowing developers from around the world to safety-test their large language models (LLMs) for free⁵. However, we contend that the government can go further by playing a role as a data provider for AI.

Governments typically collect and steward an abundance of data on their citizens and institutions. This government data could enhance public service delivery by capitalising on the opportunities presented by AI⁶. Outside the context of public services, however, it could also be a catalyst for creating a stronger ecosystem of AI innovation by increasing the equity of access to AI training data⁷.

Simply put, without data, there is no AI⁸. However, large-scale AI datasets are increasing in price and this is expected to continue as demand rises⁹. Meanwhile, organisations and individuals with digital assets are beginning to restrict access to their data: for example, nearly 14% of the most popular

¹ Whittlestone, J. and Clark, J. (2021), '[Why and How Governments Should Monitor AI Development](#)'.

² Worth, S., et al, Open Data Institute (2024), '[Building a better future with data and AI: a white paper](#)'.

³ Department for Science, Innovation & Technology, (2024), '[A pro-innovation approach to AI regulation: government response](#)'.

⁴ Stanford Encyclopedia of Philosophy (2021), '[Public Goods](#)'.

⁵ AI Safety Institute (AISI), n.d., '[About The AI Safety Institute \(AISI\)](#)'.

⁶ Thwaites, E., et al, Open Data Institute (2024), '[The ODI's input to the AI Action Plan: an AI-ready National Data Library](#)'.

⁷ The Global Partnership on Artificial Intelligence (GPAI) (2024) '[The Role of Government as a Provider of Data for Artificial Intelligence - Phase 1 Full Report](#)'.

⁸ Snaith, B., Open Data Institute (2023), '[What do we mean by "without data, there is no AI"?](#)'.

⁹ Paul, K. and Tong, A., Reuters (2024), '[Inside Big Tech's underground race to buy AI training data](#)'.

websites block CommonCrawl crawlers (the automated programs that access websites and store information on them in the open [CommonCrawl dataset](#), which is often used to train AI models) from collecting their data¹⁰. This closing of data massively benefits large organisations with either stockpiles of data or the financial capital to pay for them. As a result, the AI ecosystem risks ‘pricing out’ innovation by small organisations and academics while allowing larger organisations to monopolise. Equity of access to AI training data is therefore key and governments can ensure it by acting as data providers for AI.

The UK government holds a multitude of data that could be useful for AI models, including [official statistics releases](#) and [the National Archives](#). In this report, we examine two other UK government data sources: government websites, and government open data published on data.gov.uk.

Government websites contain textual information on government policies and can provide LLMs with intimate knowledge of the UK and life within it, knowledge that could benefit users within and outside its borders. As a collection of English text in natural language without sensitive or harmful information, these websites could also be used in the pre-training of LLMs to develop their language capabilities safely. Meanwhile, data.gov.uk is the government’s open data service and could provide AI models with accurate, up-to-date numerical information on the UK government and its citizens, which can be used as evidence in its responses to prompts.

The government is already taking steps towards becoming a data provider for AI by developing and testing proof of concepts like the National Data Library¹¹. Before these initiatives are implemented, it is important to understand the current state of the government as a data provider for AI. While government websites and data.gov.uk are open and freely available to developers in their current form, are they actually used for the training of AI models? Furthermore, are they beneficial to the performance of these models, especially in public service tasks?

In this report, we summarise and discuss the evidence presented by [PAPER], in which we carry out two data experiments—one for websites, and one for datasets from data.gov.uk—to answer the question:

To what extent do UK government data sources contribute to the performance of AI models?

These experiments are adapted from literature to provide numerical answers to this question. In this report, we outline both experiments, then present and analyse the results and discuss their implications. We conclude with recommendations for the UK government to take forward and develop so that they can better play the role of a data provider for AI.

¹⁰ Barr, A. and Hays, K., Business Insider (2023), ‘[The New York Times got its content removed from one of the biggest AI training datasets. Here's how it did it.](#)’.

¹¹ Massey, J., et al, Open Data Institute (2024), ‘[How to build a National Data Library](#)’.

Key findings and recommendations

This section offers an overview of this report's insights and actionable recommendations. Through two experiments, we assessed how government data sources—specifically government websites and data.gov.uk—support AI models in public service tasks. The findings reveal gaps and opportunities in making the government a data provider for AI.

Main findings

The first experiment investigated how government websites contribute to the performance of LLMs in public service tasks. This is where a citizen of a country queries the LLM for information on government policy related to their situation, for instance in the context of welfare schemes and their eligibility criteria. Adapting an established methodology, we find two key results:

Key result 1: Government websites are demonstrably important data providers for LLMs.

Key result 2: 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.

The second experiment explored whether datasets from data.gov.uk contribute to LLMs' knowledge bases. Using specific prompting techniques, we found that the tested LLMs could recall almost none of the statistics published on data.gov.uk, indicating:

Key result 3: data.gov.uk is not a data provider for foundational models.

Further details on the design of the experiments are described in the main body of this report.

Recommendations

To address these challenges and enhance the UK government's role as a data provider for AI, we propose this set of five actions, summarised here and explained in detail in the main body of this report:

1. **Continue to make data openly available and ensure it is AI-ready:** Make datasets discoverable by, and interoperable with, AI tools, for instance by attaching [Croissant](#) metadata. Present key statistics from datasets directly on data.gov.uk to improve accessibility to web crawlers and AI systems.
2. **Revise data reuse policies:** Revisit permissions for AI-related tools, such as web crawlers, across all government websites and platforms to boost innovation in AI-enabled digital products and services.
3. **Develop an AI-ready National Data Library:** Design the planned National Data

Library with AI readiness in mind, to reduce the amount of resources AI innovators need to invest in processing and transforming the data for AI tasks.

4. **Equip existing data access and sharing infrastructure with AI capabilities:** Take advantage of digital platforms and tools across the AI landscape to enable government datasets to work seamlessly with developers' workflows, thereby making government data more accessible and demonstrating the government's commitment to AI innovation.
5. **Invest in high-quality benchmarks and evaluation protocols:** As the field of AI continues to advance, build infrastructure to understand and measure the government's evolving role as a data provider for AI and promote safe, AI-enabled public services.

Conclusion

The UK government's ambition to foster a 'pro-innovation' AI ecosystem requires deliberate steps to improve access to high-quality data to train and fine-tune AI models. This report demonstrates both the current status and limitations of existing government data sources in supporting AI development. By implementing the above recommendations, the UK can position itself as a global leader in data provision for AI.

The importance of government websites for LLMs

In the first experiment, we aimed to quantify the importance of UK government websites as data providers for LLMs. These AI tools are increasingly used to provide public services, acting as chatbots to help citizens access information about government policies¹² ¹³. Their effectiveness in this role relies on their ability to accurately reflect policy information, often sourced directly from government websites.

LLMs may gain this knowledge from training on UK government websites¹⁴ that provide clear, reliable information authored by the government. Without this primary source, an LLM might rely on secondary sources—such as social media posts or newspaper articles—which may lack the same accuracy or provenance and therefore misinform users. Less popular topics are not picked up by these secondary channels. On this basis, it is massively important for government websites to be data providers for LLMs. We look to quantify this importance by exploring the counterfactual question:

How much would the performance of LLMs suffer if government websites were not in their training data?

Experiments and results

We conducted a so-called ‘ablation study’, which is a machine-learning technique that involves removing specific components of a system to examine the impact of those components on overall performance¹⁵. Following an ‘unlearning’ approach, we effectively removed the content in select government websites from the knowledge base of five LLMs ([Meta Llama 3.1 8B base](#) and [instruction-tuned, Google Gemma 2 2B base](#) and [instruction-tuned](#), and [Qwen 2.5 3B instruction-tuned](#)). We then evaluated the models on a set of citizen queries about UK welfare schemes (see Appendix 1), such as Universal Credit eligibility, using a grading framework to assess the accuracy of the responses before and after ablation. This is demonstrated at an abstract level in Table 1.

¹² Oxford Insights (2023), ‘[LLMs in Government: Brainstorming Applications](#)’.

¹³ Gao, S., et al., (2023), ‘[Application of large language model in intelligent Q&A of digital government](#)’.

¹⁴ Such as: ‘[Universal Credit: Eligibility - GOV.UK](#)’.

¹⁵ Sheikholeslami, S., (2019), ‘[Ablation Programming for Machine Learning](#)’.

Test case question	Baseline (pre-ablation)	Post-ablation
<p>'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?'</p>	<p>'...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...'</p>	<p>'...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...'</p>
	<p>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.</p>	<p>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.</p>

Table 1: An annotated example of the experimental output, using results from Llama 3.1 8B

Across all tested LLMs, the average rate of inaccuracy increased by 42.6% after the websites were removed, as seen in Figure 1. This increase provides proof of the importance of UK government websites for LLMs: removing their knowledge of government websites has a demonstrably negative effect on LLMs' accuracy in citizen query tasks.

However, across these tasks, the effect of the ablation varied, as seen in Figure 2. For some questions, the ablation had no effect, while in others, inaccuracies were up to 2.5 times more present post-ablation. In Table 2, the questions are grouped by the extent to which each has been affected by the ablation.

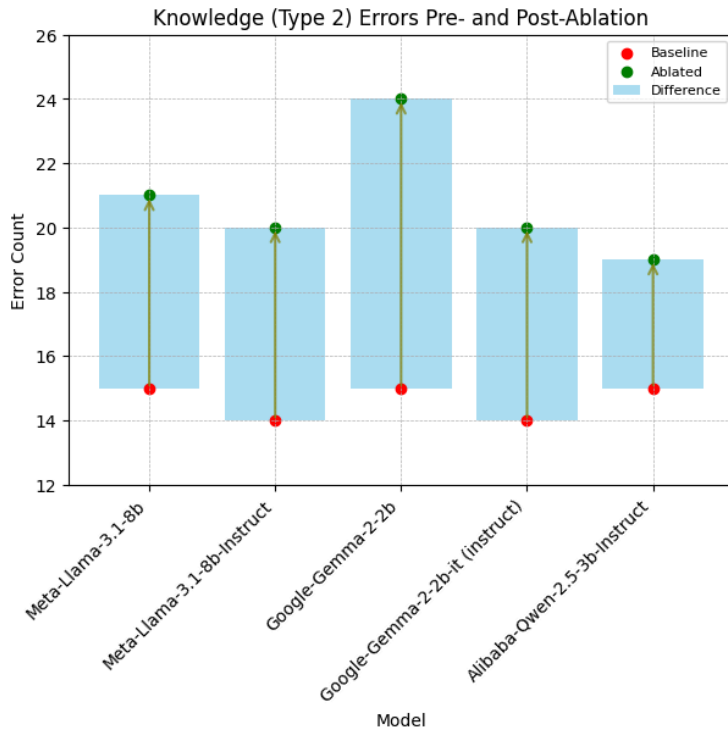


Figure 1: The total number of inaccuracies present pre- and post-ablation per LLM tested

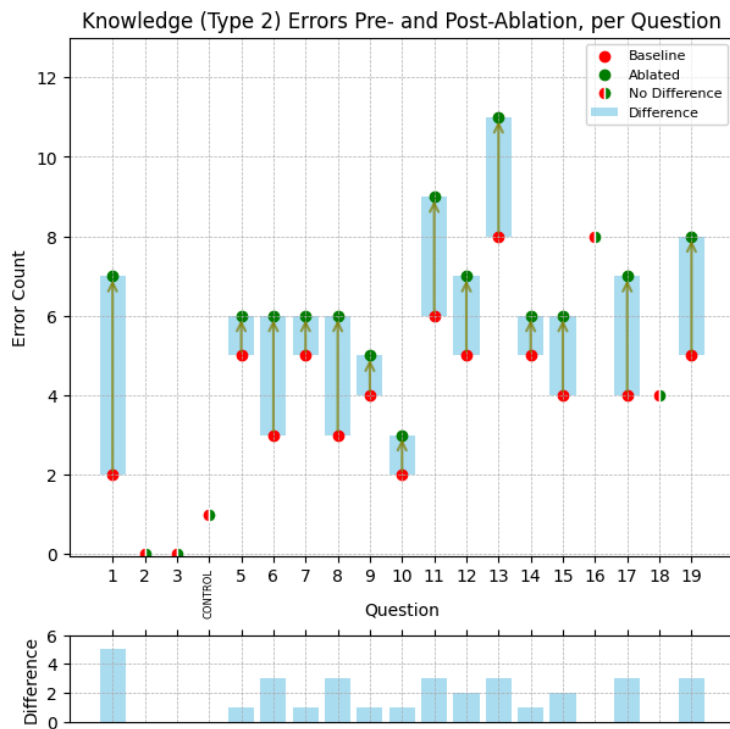


Figure 2: The total number of inaccuracies present pre- and post-ablation per question in the evaluation framework

Subject matters of questions unaffected by ablation	Subject matters of questions minimally affected by ablation	Subject matters of questions heavily affected by ablation
Question 2 - Mental health support for armed forces reservists	Question 5 - Eligibility for Disability Living Allowance	Question 1 - Eligibility for Child Benefit given other benefits are being claimed
Question 3 - Mental health support options for armed forces veterans	Question 7 - Child Benefit age conditions	Question 6 - Interaction between Universal Credit and Child Benefit
Question 16 - Financial support for families with disabled children	Question 9 - Eligibility for Child Benefit for pre-settled EU citizens	Question 8 - Receiving Child Benefit in another country
Question 18 - Information required for a Universal Credit application	Question 10 - Effect of newborn children on Universal Credit payments	Question 11 - Getting an advance payment of Universal Credit
	Question 12 - Other financial support that complements Universal Credit	Question 13 - Eligibility for Universal Credit for students
	Question 14 - Payment frequency of Universal Credit	Question 17 - High earner eligibility for Child Benefit
	Question 15 - Circumstance changes and their effects on Universal Credit	Question 19 - Non-British citizens applying for Universal Credit or Disability Living Allowance

Table 2: The subject matters of each question in the evaluation set, grouped according to the effects of ablation on each

Why are the topics in the right-hand column of Table 2 more affected than those in the left-hand column? The availability of secondary, non-government sources of information could perhaps explain the differing impacts of ablation. Widely discussed topics, like mental health support services, are less affected, because they are covered extensively in news articles, forums and other digital spaces. Therefore, when government websites on these subjects are removed from LLMs' knowledge, they still have knowledge from secondary sources in their training data that they can use to answer questions accurately. In contrast, topics that are less widely discussed online, like the interactions between different welfare schemes, are more affected by the ablation.

Our paper¹⁶ tests this potential explanation. We compare the extent to which each question was affected by ablation and a rudimentary ‘prevalence’ score that measures how much that question’s subject matter is represented in secondary sources online. We find a significantly negative correlation between the two variables, providing evidence for this hypothesis: where there is less secondary, non-government data on a specific subject matter, government websites concerning that subject matter are more important to LLMs.

Key results

This ablation study experiment yielded two key results:

Key result 1: Government websites are important data providers for LLMs.

Key result 2: The importance of government websites for LLMs varies depending on subject matters. In some subject matters, secondary non-government sources of information supplement the knowledge of LLMs. However, having authoritative, trusted sources for these popular topics remains critical, as citizens query LLMs for the provenance of their answers. At the same time, in other subject areas that are less prevalent online, there are demonstrably large differences between LLMs pre- and post-ablation, meaning that the government is a key source of information for them in these subject matters.

¹⁶ In section 2.5.

LLMs and their knowledge of data.gov.uk

Government open data, published under initiatives like data.gov.uk with the bespoke Open Government Licence¹⁷, represents a key resource for fostering innovation and transparency. The UK government is often cited as a leader in the open data space¹⁸, and data.gov.uk hosts a wide array of datasets on topics ranging from agriculture to public health. These datasets have the potential to enhance AI models by providing accurate, structured and up-to-date information that can evidence responses to public service tasks or otherwise enable data discovery and analysis^{19 20}.

However, while the platform is widely accessible, it is unclear to what extent data.gov.uk is actively used as a source in training LLMs. To investigate this, we conducted an experiment to assess whether LLMs can recall and integrate information from data.gov.uk into their responses:

Can LLMs accurately recall data from data.gov.uk?

Experiments and results

To assess the extent to which data.gov.uk serves as a training data source for LLMs, we used so-called ‘information leakage’ methods. These methods involve designing prompts to encourage LLMs to ‘leak’ or recall specific information from their training corpora. If the models accurately reproduce data from data.gov.uk, this provides evidence that the datasets are present in their training data^{21 22}.

We selected five datasets from data.gov.uk, representing a diverse range of topics, alongside two control datasets not sourced from data.gov.uk. The chosen datasets included statistics on topics such as fire safety, rail injuries and air pollution, while the controls included widely known information, such as the Bank of England’s base rate (Table 3). We used prompting templates adapted from literature²³ with varying levels of contextual information (0-shot, 1-shot, and 5-shot) to maximise the likelihood of the models recalling the target data points (see Appendix 2 for details).

¹⁷ The National Archives, n.d., ‘[Open Government Licence for public sector information](#)’.

¹⁸ Carrara, W., Fischer, S., and van Steenberg, E., European Data Portal (2015), ‘[Open Data Maturity in Europe 2015](#)’.

¹⁹ Shadbolt, N., et al. (2012), ‘[Linked open government data: lessons from Data.gov.uk](#)’.

²⁰ Massey, J., et al, Open Data Institute (2024), ‘[The promise and challenge of data discovery with LLMs](#)’.

²¹ Carlini, N., et al. (2023), ‘[Extracting Training Data from Diffusion Models](#)’.

²² Somepalli, G., et al. (2022), ‘[Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models](#)’.

²³ Wang, B., et al. (2024), ‘[DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models](#)’.

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 has 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 baseline to indicate inflation levels in the agricultural supply chain. Published on data.gov.uk.

Table 3: The subject datasets of the information leakage experiment

Table 4 contains the results of the experiments. 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.

	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
(control) BOE	✓✗✓✓	✗✓✓✓	✓✗✓✓	✓	✓✗✓✓	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
(control) POP	✗✓✓✓	✓✓✓✓	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
HSA	✓✗✗✗	✗✗✗✓	✗✗✓✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✓✗✓✓	✗
SAS	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
IBR	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
POL	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗
API	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗	✗✗✗✗	✗✗✗✗	✗✗✗✗	✗

Table 4: The results of the information leakage experiment

As can be seen, LLM recall of data.gov.uk data is not successful. In only five out of 195 non-control attempts did LLMs recall anything correctly, with all of these on the smoke alarm dataset.

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 are 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 our question.

Key results

It is clear from this experiment that data.gov.uk is demonstrably not a data provider for these LLMs—they can recall very little information from it, meaning that its data is likely not part of their training corpora.

Key result 3: data.gov.uk is not a data provider for foundational models.

Discussion, recommendations and conclusions

Access to government data is massively important for AI, not only because of the potential of open government data to catalyse the development of the AI ecosystem but also because of its potential to enhance AI tools like LLM-based chatbots in public service and citizen query tasks. The evidence presented in this report demonstrates that the UK government is currently acting as a data provider for AI in some regards, specifically with respect to its government websites on certain subject matters, but not in others, specifically with respect to data.gov.uk.

In our experiments we have used mainstream LLMs, which are accessible to a wide range of AI innovators without substantial infrastructure or skills investments. Just like many people use Google for their citizen queries, many people would turn to a 'generic' LLM rather than trying to find the information on a government website or data portal. What the results make clear is that such generic LLM capabilities are not enough to provide accurate information to citizens. Improvements could be achieved by making data AI-ready, but also by investing in LLMs whose functionality has been tailored along the three themes from above.

These include advanced models that have been fine-tuned to work with spreadsheets and other types of 'structured' data; so-called 'AI agents' that know how to process such data to e.g. calculate statistics; and models that prioritise accurate answers from authoritative sources. Innovators all over the UK building such solutions should be supported with up-to-date AI-ready government data.

The differences between data portals and websites

Why do government websites contribute knowledge to LLMs while data.gov.uk datasets do not?

LLMs are trained on swathes of data scraped from the internet by web crawlers, which are tools that automatically open websites and scrape the information on them into a central repository. In the April 2024 CommonCrawl dataset, crawlers scraped 13,556 data.gov.uk pages²⁴, meaning that data.gov.uk is being scraped and incorporated into the training corpora of LLMs.

²⁴ From a query for the "www.data.gov.uk/" URL pattern on the CommonCrawl April 2024 release - [link](#).

There can be many reasons why the data from data.gov.uk is not used by general-purpose LLMs to answer queries. For instance, learning and inferring from structured datasets is different from plain text.

Furthermore, data on data.gov.uk is inaccessible to machines and crawlers. Almost all of the 13,556 crawled data.gov.uk pages link to index pages²⁵ that contain information about certain datasets but simply provide download links for them rather than divulging their key statistics and measurements. While a human can download the dataset via the link and analyse it for themselves, a web crawler cannot. As a result, a crawler on one of these index pages will not actually scrape any information from the government dataset, meaning it cannot provide this information to LLMs for their subsequent training. Government websites are mostly available as text (in HTML format), so crawling that information can be done effectively with existing tools.

So, how can data.gov.uk be made more AI-ready, and how can government websites' AI-readiness potentially be improved? Overall, how can the government be a better data provider for AI?

- What AI-ready data means in practice depends on many factors, but established practices like FAIR-ness²⁶ (findable, accessible, interoperable, reusable) and linked data are a strong starting point. There are also emerging solutions such as [Croissant](#), a machine-readable metadata format specifically designed for machine learning. Supporting Croissant would allow data.gov.uk datasets to be indexed by [Google datasets search](#) and easily integrated into AI developer workflows²⁷. Government data with attached metadata in the Croissant format will be FAIR and overall more AI-ready, likely encouraging its use amongst AI developers.
- In addition, we suggest that steps should be taken to make data.gov.uk information as accessible to web crawling as possible, following best practices on making data accessible and usable, which include useful text summaries on dataset description pages with key information about areas such as the scope of the data, its range of values, and descriptive statistics.²⁸
- To further accommodate crawling across all of its digital resources, the UK government should revisit its data reuse policies, taking care to define their practices regarding their permissions for web crawlers on the robots.txt²⁹ conditions, both on government websites and data portals.

²⁵ Such as '[UK Trade in Services - data.gov.uk](#)'.

²⁶ GoFAIR, n.d., '[FAIR Principles](#)'.

²⁷ TensorFlow, n.d., '[Format-specific Dataset Builders](#)'.

²⁸ Koesten et al., International Journal of Human-Computer Studies (2020), [Everything you always wanted to know about a dataset: Studies in data summarisation](#).

²⁹ Google Search Central, n.d., '[Introduction to Robots.txt](#)'.

Secondary sources fill the gaps if hallucinations do not

Where government sources were not available to the LLMs, they answered our questions by referencing information they learnt from secondary, non-government sources, or, alternatively, made things up.

Such hallucinations are a massive problem and can only be solved with bespoke models. However, as people are expected to continue to use general-purpose services like ChatGPT, there is a risk that accurate, trusted information will not reach those who need it most, with implications for public trust.

At the same time, while secondary, non-government sources may facilitate an LLM to provide some correct information on government policies, this cannot be guaranteed for all sources concerned. In comparison to government sources, secondary sources may be out of date or otherwise inaccurate, with unknown provenance and references.

Therefore, the government must guarantee that there are simply no ‘gaps to be filled’ by secondary sources and hallucinations. To do so, they should ensure that they are the first and most important data provider for foundational models in all subject matters related to governmental affairs, which can be achieved by taking steps to deliberately provision government data for AI development.

- We already commented on the opportunities of an AI-ready NDL to enable federations to access high-quality datasets from across government³⁰. Designing it with AI use cases in mind will ensure that, when created, the NDL can act as a central data source for AI developers globally to use to train their models, mitigating some of the issues we have found in our research.
- Furthermore, existing infrastructure for publishing, accessing and sharing data, for instance data portal software or TREs, should be equipped to work seamlessly with AI developers’ workflows, thereby increasing accessibility of government data and demonstrating the government’s commitment to AI innovation. For example, the MLCommons community is working on extensions of the CKAN, Socrata and Dataverse data management software with Croissant metadata to make it easy for publishers to create AI-ready datasets.
- As they action these recommendations and take steps towards being better data providers for AI, the government could measure the effectiveness of their strategies and the overall safety of AI by investing in high-quality benchmarks and evaluation protocols that can measure the usage and impact of government data for LLMs.

³⁰ Massey, J., et al, Open Data Institute (2024), [‘How to build a National Data Library’](#) .

Conclusion

The government aims to take a 'pro-innovation' stance on AI and sees latest advances in generative AI as a means to improve public service delivery. To do so, we believe that they should ensure equitable access to data by acting as a data provider for AI. In this report, we provided evidence of the government's current status as a data provider for LLMs. Based on it, we built a set of actionable recommendations for the UK government to take so that it can accomplish its aims and demonstrate its commitment to innovation in AI.

Appendices

1 - Evaluation queries used for the ablation study

Table 5 provides the exact queries used to evaluate the LLMs pre- and post-ablation. These were designed to mimic citizen queries and each targeted a different subject matter that was represented on a single website in the corpus of websites that were ablated from the models.

id	Question Topic	Question
1	Eligibility for Child Benefit, and its interaction with other benefits	'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?'
2	Mental health support for armed forces reservists	'I'm a reservist in the UK armed forces and I've recently returned from a deployment overseas. I'm struggling with my mental health but I'm not sure where to turn. What support is available to me?'
3	Mental health support for armed forces veterans	'My partner is a veteran of the UK armed forces and is struggling with their mental health. We live in Scotland. What services are available to us, and how can I support them?'
4	CONTROL - Welfare support for carers in the US	'I live in the US, caring for my elderly mother who receives disability benefits. I spend over 35 hours a week caring for her. Am I eligible for any financial support, and how might it affect my other income?'
5	Eligibility for Disability Living Allowance	'I have a child with a disability who is under 16. We live in England. What financial support might we be eligible for, to help with the extra costs of care?'
6	Interaction between Universal Credit and Child Benefit	'I'm a single parent living in England with two children under 16. I've recently lost my job and I'm struggling to make ends meet. Can I apply for Universal Credit, and how will it affect my Child Benefit?'
7	Child Benefit continuation beyond a certain age	'I'm 18 years old and still in full-time education. My parents have been claiming Child Benefit for me, but I'm about to turn 19. Can they continue to receive it, and what do we need to do?'

8	Receiving Child Benefit in another country	'I'm a UK citizen planning to move abroad for work, but I have two children under 16. Will I still be eligible for Child Benefit if I move to another country for my job?'
9	EU citizens with pre-settled status claiming benefits	'I'm an EU citizen with pre-settled status in the UK, and I have a child. I've just started a new job in the UK. Am I eligible to claim Child Benefit?'
10	How having your first child affects Universal Credit	'I'm currently receiving Universal Credit and expecting my first child. How will having a baby affect my Universal Credit claim, and should I also apply for Child Benefit?'
11	Getting an advance payment of Universal Credit	'I'm starting on Universal Credit and worried about my finances. How can I get money before my first payment?'
12	Other financial support that complements Universal Credit	'I'm on Universal Credit and struggling with bills. What other financial support might be available to me?'
13	Student eligibility for Universal Credit	'I'm a student in the UK considering applying for Universal Credit. Under what circumstances would I be eligible?'
14	Payment frequency of Universal Credit	'I've just started receiving Universal Credit payments. How often will I be paid, and can I change the frequency?'
15	Circumstance changes and their effect on Universal Credit	'I'm receiving Universal Credit and my circumstances have changed. What should I do, and what changes do I need to report?'
16	Financial support for families with disabled children	'I have a child with a disability and live in England. What benefits might we be eligible for, and how much could we receive?'
17	High earner eligibility for Child Benefit	'I'm thinking of claiming Child Benefit. How much could I receive, and are there any implications if I'm a high earner?'
18	Information required for a Universal Credit application	'I'm considering applying for Universal Credit. What information do I need to provide, and how do I start the process?'

Table 5: Queries used to evaluate the performance of LLMs pre- and post-ablation

2 - Prompt templates for the information leakage study

Table 6 contains the prompt templates used in the information leakage study to encourage base (not instruction-tuned) LLMs to recall government information. Table 7 provides examples for 0-shot, 1-shot, and 5-shot prompting, and table 8 shows the prompts used for instruction-tuned LLMs. Further details can be found in section 3.2 of our paper.

Template	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]' \n[year], ...'

Table 6: Prompting templates used for base models

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, ...'

Table 7: Examples of 0-shot, 1-shot, and 5-shot prompting, using template d (see Table 6) and the Stop and Search rates by ethnicity dataset (see Table 3)

System prompt	Prompt
<p>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.</p>	<p>'What was the rate of stop and searches per 1000 Black people in England and Wales in 2020/21?'</p>

Table 8: The system prompt used for instruction-tuned models, and an example of the question asked to them, referencing the Stop and Search rates by ethnicity dataset (see Table 3)