ODI

CitizenQuery-UK: Benchmarking LLM performance in citizen queries about public information on gov.uk

Data-centric AI

ODI Research
ADVANCING TRUST IN DATA

DAI

February 2026

Contents

Project Team:

Neil Majithia, Dr. Rajat Shinde, Zo Chapman, Prajun Trital, Jordan Decker,
Dr. Manil Maskey, Prof. Elena Simperl, Prof. Sir Nigel Shadbolt

Licence:

"Citizen queries" – questions asked by members of the public about government policies and services that are pertinent to their circumstances – represent an obvious use case for artificial intelligence (AI). Large language models (LLMs) are ubiquitous in the lives of the global digital population as chatbots, search tools, and virtual assistants, and while individuals place their trust in them for advice, governments are also deliberating on the idea of building their own to serve citizen query contexts. The speed, accessibility, strong general knowledge skills, and natural language interactions of LLMs, as well as their ability to provide bespoke responses tailored to users' language, educational, and accessibility needs, present a clear future for AI-powered information services for citizen queries.

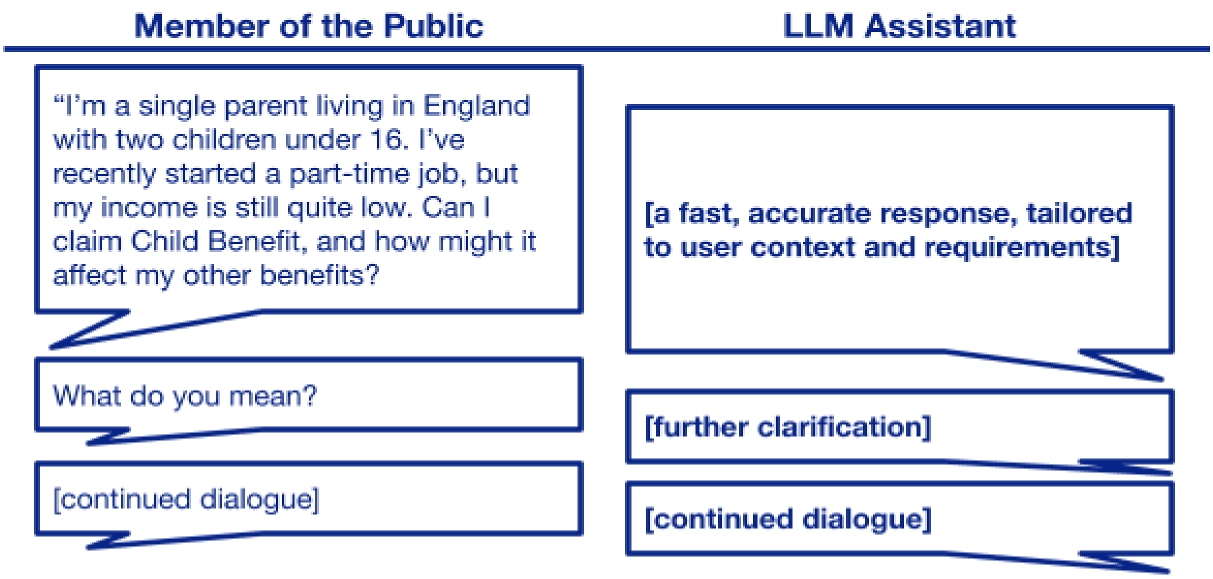

Figure 1: An example citizen query interaction with an LLM

However, such a deployment of LLMs is no doubt high-stakes. A citizen query can concern a wide range of topics, including welfare eligibility, tax payments, financial vulnerability, public health, and more, and is often extremely specific. Thus, if AI provides incorrect or outdated information, it could have severe, negative, and likely invisible ramifications for their life if they were to, for example, mistakenly believe they are ineligible for benefits. LLMs have a well-established tendency to misinform, making them "unusable at best and dangerous at worst" in sensitive information tasks without robust safeguards. If they are to be officially adopted by a national government in the throes of a crisis of public trust like the UK, there must be evidence demonstrating their accuracy and overall value in answering citizen queries.

Currently, there are no benchmark datasets – authoritative, standardised exams for AI – that can be used to evaluate LLM performance in the context of citizen information services, thereby limiting the amount of evidence for government decision-making about the safety and procurement of AI tooling. In this work, we have developed the first such dataset: **CitizenQuery-UK**, a collection of **22,066** synthetically generated citizen queries and their corresponding answers, grounded on real-world scenarios and entirely built from information about UK government policy and services sourced from gov.uk.

The release and utilisation of this benchmark dataset, and iterations upon it, will:

- Provide a rich collection of evidence regarding how LLMs tell users information about UK government policy, demonstrating how accuracy, refusal, and overall utility differ between model providers and reasoning modes. In other words, this provides a measure of LLM trustworthiness in the context of citizen queries.
- Determine the domains in which UK government information is and is not accessed and communicated accurately by LLMs, thereby indicating where the government should do more in order to make pertinent information readily available in a world of increasing LLM usage (as per previous work at the ODI).
- Form the foundations of a UK government initiative that will build centralised benchmarks and evaluation suites that fit the needs and wants of the UK's industries, institutions, and people, enabling the government to navigate its relationship with the technology industry so it can reach and stay at the cutting edge of innovation.

The paper outlining the motivations, methodology, and preliminary evaluations of LLMs has been published as a pre-print on arXiv, while the dataset, split into 3, has been published on Hugging Face. In addition, we will keep a separate set of queries private, held exclusively to test models in the future without risk of benchmark contamination.

Benchmarking LLMs with Citizen Queries

AI benchmarking works by asking models a series of questions and testing to see how accurate they can be. Early benchmark datasets built for these purposes consisted of multiple-choice questions, but our work sits alongside the range of new benchmarks being built that check the long-form text generated by AI. For this purpose, we required a bespoke dataset and a sophisticated evaluation method that uses AI to judge other AI, having built a robust framework that ensures its judgements are not biased or inaccurate.

<table>
<tr><th>Prompt</th><th>Expected response</th><th>Persona of asker</th></tr>
<tr><td rowspan="7">I'm a carer looking after my elderly relative with a severe disability. What support am I eligible for, and how do I apply for Carer's Allowance?</td><td rowspan="7">If you have substantial caring needs, you may be eligible for Carer's Allowance. You need to contact the DWP and provide details about your caring responsibilities, including how many hours you spend caring per week and the nature of the care required. You'll also need to provide proof of your relationship with the person you're caring for and your own financial situation.</td><td>26-45 year old; parent; secondary school education; medium digital literacy; moderate household income</td></tr>
<tr><th>Domain / Subdomain</th></tr>
<tr><td>Benefits</td></tr>
<tr><th>Information Source</th></tr>
<tr><td>https://www.gov.uk/carers-allowance</td></tr>
<tr><th>Date valid</th></tr>
<tr><td>2026-01-28</td></tr>
</table>

Figure 2: An example row (with only a selection of columns) from CitizenQuery-UK

The public dataset contains **22,066 prompt-response pairs**, covering **7,725 unique topics**. Each pair was generated by **Qwen 2.5 72B**, an openly available LLM, which also generated an associated “**persona**” that can be used to test whether LLMs accurately adapt their responses according to the person asking a citizen query, based on their age, education, and income level. Prompts can be **procedural** (“How do I apply for Universal Credit?”), **informational** (“What is Capital Gains Tax?”), **instructional** (“Where can I renew my British passport if I’m currently abroad?”), or comparative (“What is the difference between A-levels and T-levels?”). All were built on the structures of **real-world citizen queries** we collected from online sources. The dataset underwent extensive cleaning and alignment to ensure factuality, trustworthiness, and overall validity.

How we assessed LLM performance

We tested 11 foundational models. Open-weight models tested were **Meta Llama 3.1 8B and 3.3 70B; ChatGPT OSS-20B; Kimi Kimi-K2-Instruct;** and **Qwen 3 32B.** Closed-weight models tested were **Anthropic Claude-4.5-Haiku; Google Gemini-3-Flash;** and **ChatGPT 4o, 4.1, 5.2, and o3**. Benchmark results are presented in **Appendix I.**

To test foundational models with our dataset, we adapted industry-standard methods for fact-checking AI.

Our evaluations ask:

- How often do AI models abstain from answering citizen queries?
- How factual are AI models in their responses to citizen queries?
- Do AI models say too much or too little when they answer citizen queries?

Answering these questions enables a rich understanding of how AI interacts with citizens in the UK, providing a picture of the nuances and trade-offs in this context and provoking several discussion points to be taken forward. To do so, our method (Figure 2, overleaf) decomposes AI models' answers to queries in the dataset into individual bullet-point "claims", checks each claim against the expected response provided in CitizenQuery-UK, detects whether the model has refused to answer the question and, if not, reports a score that balances factual precision with the number of claims made.

Our findings and what they mean

1. **There is a need for a nuanced approach to AI adoption.**
   Our results show that open-weight models (like the Llama and Qwen families) can now perform competitively against closed-source counterparts (like ChatGPT or Claude). However, distinct "personalities" emerge across all families of AI models: some models are highly accurate but overly "chatty," while others are more concise but less detailed. Within the tested families, there was evidence that more recently released models perform better in citizen query contexts.

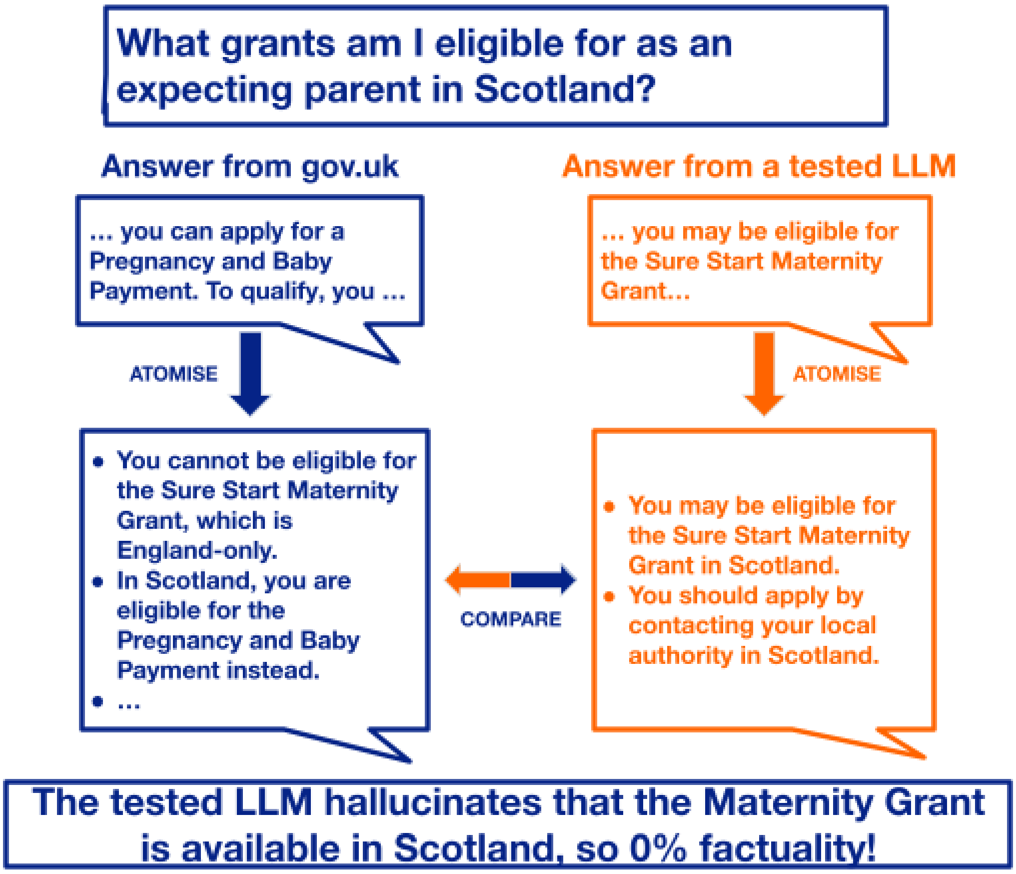

Figure 2: Our evaluation method

There is therefore a need for a nuanced approach to the official adoption of LLMs in the citizen query context: with the current speed of progress and open-weight models outcompeting closed-source ones, any “vendor lock-in” could be detrimental to the future of AI-empowered public services. Official government adoption must contend with potential opportunity costs and be flexible accordingly, and with model performance changing over time, ongoing monitoring is essential to evidence our decision-making at the cutting edge.

2. **LLMs are inconsistent.**
   Reliability is non-negotiable; a citizen must receive consistently accurate advice. However, we found that while models often answer correctly, they are imperfect and, moreover, they suffer from high inconsistency and unpredictability. A visible "long tail" in performance undermines the absolute trust required for model usage in citizen query contexts, with extremely high variance making the tested models appear unreliable.

   As mentioned previously, people across the UK now use AI assistants as integral components of their day-to-day life. While technology providers work on making the tested models less inconsistent, we recommend that the government ensures users are aware of the risk of inconsistency and know where to find authoritative information. Without this, users are at risk of putting unwarranted trust in AI systems, increasing the risk of their harm.

3. **AI likes to talk. At what cost?**
Citizens interacting with government services usually need concise, direct answers, but AI models often show an eagerness to "talk too much," providing lengthy responses that bury key facts or extend beyond information on government websites, thereby risking inaccuracy. Crucially, when we experimented with forcing models to be more concise and direct, their factual accuracy actually declined, suggesting that when they are asked to concentrate their responses to citizen queries, they do not ground them in gov.uk information. Verbosity is known behaviour of LLMs – they are prone to “word salad” responses that make them harder to use and decrease their reliability. For AI to accurately support people when they ask citizen queries, models must contend with the requirement for short, accurate responses.

   One reason why LLMs are so verbose is that they are tuned to synthesise information from diverse sources across the internet into their responses. While this is incredibly helpful for many contexts, such as programming, it means that, in the citizen query use case, government resources are not prioritised over others. This increases the likelihood that AI models provide misinformation from elsewhere.

   To make government resources like gov.uk more of a priority, we assert that they should be made more “AI-ready”, facilitating easy usage across AI training and deployment to ensure they are the go-to evidence bases to use when answering a citizen query.

4. **Models are not brave enough to say “I don’t know”.**
In the context of public services, giving bad advice is far worse than giving no advice at all. Here, we found that models rarely refuse to answer a question, instead attempting to answer almost every query regardless of their actual knowledge. This has obvious safeguarding implications, especially given that, if they trust an LLM, a user may not ever seek confirmation of what it says outside of their own private conversations with it, leading them to act on misinformation – e.g., not apply to welfare they’re entitled to – invisibly.

   The lack of "fallibility" is a dangerous trait; without the ability to admit ignorance, models risk leading vulnerable users to act on convincing but ultimately incorrect or outdated information. Despite the commercial drawbacks to admitting their models’ fallibility, it is a safeguarding requirement to ensure more trustworthy interactions with AI, whether in citizen queries or beyond.

Conclusions and next steps

CitizenQuery-UK represents a well-evidenced, responsible way to support the integration of AI into citizen query services. By being the first government-focused benchmark dataset designed explicitly for the citizen query use case, it enables the assessment of the suitability of different LLMs in public service delivery and, therefore, allows a body of evidence to be collected in order to support decisions about the UK government's AI strategy. All code is released and licensed as open-source, enabling anyone to use it for their own benchmarking needs.

The team has noted a number of visible next steps for us to take forward in developing CitizenQuery-UK further:

- Operationalising the constant update of the dataset as information is updated on gov.uk.
- Submission of the benchmark to online LLM benchmarking leaderboards, such as Kaggle's benchmarks page, in order to test a wide variety of frontier LLMs as they are released.
- Multilingual tests, using the Welsh version of gov.uk as a starting point.
- Repeating this methodology with other aspects of government information online, including NHS public health information.
- Repeating this methodology with a non-UK country, whether with English-language information or otherwise.

Author Information and Acknowledgements

This work was conducted by Neil Majithia, Dr. Rajat Shinde, Dr. Manil Maskey, and Prof. Elena Simperl during Dr. Shinde's fellowship at the ODI. He and Dr. Maskey both are affiliated with the NASA MSFC Office of Data Science and Informatics (ODSI).

The project team was supported by a number of short-term interns at ODSI: Zo Chapman, Prajun Trital, and Jordan Decker. ODSI resources and compute were utilised for much of the benchmark development.

We thank the following for providing their insights over the course of our research.

- Ben Moore (DSIT)
- Saasha Nair (DSIT)
- Jenny Chim (RAI UK)
- Andy Dudfield (FullFact)
- Dr. Tommaso Spinelli (GDS)
- The wider ODI team.

Appendix I – Results

Table 4: **Zero-Shot Performance:** F1@K statistics, abstention rates, and verbosity (ΔK).

		F1@K Statistics				
Model	**Abst.%**	**Mean**	**Med.**	**Std.**	**IQR**	**Avg ΔK**
`Claude-4.5-Haiku`	0.38%	**0.8725**	0.9231	**0.1827**	**0.0973**	+16.63
`Gemini-3-Flash`	0.80%	0.8150	0.8966	0.2258	0.1746	**+5.40**
`GPT-OSS-20B`	0.22%	0.7968	**0.9333**	0.3094	0.1605	+11.78
`Kimi-K2-Instruct`	**1.37%**	0.7450	0.8571	0.2693	0.3048	+5.69
`Llama-3.1-8B-Instruct`$^{\Uparrow}$	0.71%	0.8423	0.9231	0.2325	0.1268	+8.39
`Llama-3.3-70B-Instruct`	0.16%	0.8465	0.9231	0.2169	0.1296	+6.68
`Qwen3-32B`	0.00%	0.8286	0.9231	0.2453	0.1448	+6.62
`ChatGPT 4o`$^{\propto}$	0.40%	0.7862	0.9268	0.3200	0.1655	+8.45
`ChatGPT 4.1`$^{\propto}$	0.60%	0.7870	0.9375	0.3189	0.1829	+7.50
`ChatGPT 5.2`$^{\propto}$	0.20%	0.8104	0.9564	0.3285	0.0978	+14.17
`ChatGPT o3`$^{\propto}$	0.20%	0.8099	0.9600	0.3329	0.1042	+14.01

$^{\Uparrow}$ This model also served as the helper model $\mathcal{M}_{AFG}$ and adjudicator model $\mathcal{M}_{AFV}$ in the evaluation pipeline; scores may therefore reflect self-preference bias.
$^{\propto}$ Tested on a domain-stratified sample ($n = 500$) from the full dataset due to API accessibility constraints.

Table 5: **Few-Shot Performance:** Abstention rates, F1@K distribution, and verbosity when models are provided with 3 examples.

		F1@K Statistics				
Model	**Abst.%**	**Mean**	**Med.**	**Std.**	**IQR**	**Avg ΔK**
`Claude-4.5-Haiku`	1.52%	**0.8627**	0.9130	**0.1843**	**0.1056**	+15.88
`Gemini-3-Flash`	0.82%	0.8027	0.8889	0.2308	0.1974	**+4.77**
`GPT-OSS-20B`	0.48%	0.7537	0.9310	0.3567	0.2144	+10.24
`Kimi-K2-Instruct`	1.52%	0.7994	0.8966	0.2442	0.2065	+7.54
`Llama-3.1-8B-Instruct`$^{\Uparrow}$	**1.58%**	0.7970	0.8889	0.2378	0.2155	+4.98
`Llama-3.3-70B-Instruct`	0.68%	0.8401	0.9167	0.2158	0.1418	+5.72
`Qwen3-32B`	0.00%	0.8596	**0.9412**	0.2217	0.1106	+7.31
`ChatGPT 4o`$^{\propto}$	0.00%	0.7965	0.9412	0.3224	0.1492	+8.50
`ChatGPT 4.1`$^{\propto}$	0.80%	0.7995	0.9444	0.3214	0.1357	+7.66
`ChatGPT 5.2`$^{\propto}$	0.00%	0.8203	0.9630	0.3250	0.1003	+12.48
`ChatGPT o3`$^{\propto}$	0.60%	0.7796	0.9600	0.3639	0.1232	+15.11

$^{\Uparrow}$ This model also served as the helper model $\mathcal{M}_{AFG}$ and adjudicator model $\mathcal{M}_{AFV}$ in the evaluation pipeline; scores may therefore reflect self-preference bias.
$^{\propto}$ Tested on a domain-stratified sample ($n = 500$) from the full dataset due to API accessibility constraints.

Appendix II – Examples

Prompt	Expected Response	Response (GPT-OSS-20B)
I'm caring for my niece whose parents have passed away. How do I apply for Guardian's Allowance and what additional support am I entitled to?	To apply for Guardian's Allowance, you need to contact the Guardian's Allowance Unit and provide details about your care of the child and the parents' passing. The allowance is £22.10 per week and is added to Child Benefit you are entitled to, which is tax-free. You should inform the unit of any changes in your circumstances, such as moving house or changes in your income. This support is available if you're raising a child whose parents have died or if one parent is still alive.	**1. Check you're eligible. You're eligible if:** a. You're the legal guardian of **a child who has died.** b. The child was under 16 (or 16-20 if still in full-time education). **c. You're not receiving any other benefit for that child.** **2. Collect the paperwork:** **a. The Child's Death Certificate** b. …

Example 1: An AI model hallucinates that a person is only eligible for Guardian's Allowance if they are the guardian of a child who has died. It also wrongly states that the asker is not eligible if they receive other benefits for the child.

Prompt	Expected Response	Response (Qwen-3-32B)
I'm expecting a multiple birth and have children already. Am I eligible for the Sure Start Maternity Grant, and what are the requirements? I live in Scotland.	If you're expecting a multiple birth and have children already, you may be eligible for the Sure Start Maternity Grant. However, this grant is not available in Scotland. Instead, you can apply for a Pregnancy and Baby Payment. To qualify for the grant, you or your partner must already receive certain benefits. You must apply within 11 weeks of your baby's due date or within 6 months after the baby's birth. The grant does not need to be repaid and will not affect your other benefits.	Yes, **you may be eligible for the Sure Start Maternity Grant** if you're expecting a multiple birth. … **Residency: You must be a UK resident and have lived in the UK for at least 3 years.** … Apply through your local council (in Scotland, contact your Local Authority). Let me know if you need further clarification!

Example 2: An AI model wrongly claims that the Sure Start Maternity Grant is available in Scotland, even though it is not.

Prompt	Expected Response	Response (Llama 3.1 8B)
I need to add my ex-partner's name to my child's birth certificate. What steps should I follow if we are not married or in a civil partnership?	To add your ex-partner's name to your child's birth certificate without being married or in a civil partnership, you need to re-register the birth at the registrar office. You'll need to complete a form and submit it to the register office. If one of you cannot attend, the absent person must fill in a 'Statutory declaration of acknowledgement of parentage' form and send it to the General Register Office. You'll also need to order a new birth certificate separately, as it won't be provided automatically. This process ensures the correct father's name is on the record.	**1. Get a court order: You'll need to apply to the court for a 'parental responsibility agreement' or a 'parental responsibility order'. This will give your ex-partner parental responsibility, which means they'll have the same rights and responsibilities as you.** 2. **Fill in the application form:** You can get the application form from the court or from the Scottish Government website. You'll need to fill it in and send it to the court with the required documents. …

Example 3: The AI model hallucinates the requirement of a court order, which would add unnecessary stress and financial cost without any real benefit.