OpenGovCorpus: Evaluating LLMs on Citizen Query Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract

"Citizen queries" are questions about government policies, guidance, and services relevant to an individual's circumstances. LLM-powered chatbots have a number of strengths that make them the obvious future for citizen query-answering, but hallucinated or outdated answers can cause significant harm to askers in such a sensitive context. We introduce **OpenGovCorpus-UK** and **OpenGovCorpus-eval**: a 7.5k-Q&A-pair benchmark synthesized from `gov.uk`, and its use in an evaluation framework for LLMs in citizen-query tasks. The protocol spans three evaluator classes ((1) open-weights models, (2) GPT-family models, and (3) human judgment) combining a persona-aware *Metadata Grader*, embedding- and token-level *Semantic Similarity*, and *LLM-as-a-Judge* with pass-rate aggregation. Results show strong few-shot gains, context and persona mismatches not captured by similarity metrics alone, and variation across families of closed/open models. We provide a reproducible procedure and thresholds suitable for lifecycle monitoring as policies and models evolve, supporting evidence-based public sector deployment for the future of trustworthy LLMs in government services.

1 Introduction

"Citizen queries" are questions about government policies and services that are specific to an individual's circumstances, spanning topics such as benefits, taxes, immigration, employment, and public health. This task naturally fits large language models (LLMs) (Fig. 1a) given their conversational interfaces and innate abilities to adapt language to users' literacy and accessibility needs. However, deployment for citizen queries is high-stakes: hallucinated or outdated guidance can harm end users, providing misinformation that has consequence for their day-to-day life, so LLM adoption requires evidence of accuracy and trustworthiness on citizen-query tasks. There is no widely adopted benchmark for this setting; we therefore introduce *OpenGovCorpus-UK* and an accompanying evaluation framework, *OpenGovCorpus-eval*. OpenGovCorpus-UK contains 7,553 prompt–response pairs derived from `gov.uk` policy and service text, enabling evaluation in a UK context. We assess models with similarity metrics and automated graders, including a persona-aware *Metadata Grader*, a *Semantic Similarity* grader, and *LLM-as-a-Judge*, and report pass-rate summaries using defined thresholds. In addition to open-weights evaluations under zero/few-shot prompting, we run automated evaluations on GPT-family models using a reproducible pipeline.

2 Background

Citizen queries and AI. Early work on citizen queries emerged during the 2000s e-government movement, which shifted information seeking from in-person or telephone advice toward online search and government websites [Schelin, 2003, Curtin et al., 2003]. Users generally preferred

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

the online modality [Reddick, 2005], and public-sector research focused on designing government websites around real information needs [Anthopoulos et al., 2007]. Evidence from advice agencies showed that citizen queries span broad, high-stakes topics, including financial vulnerability and other sensitive domains [Marcella and Baxter, 2000], and also many everyday concerns, sometimes unrelated to government itself [Lambert, 2011]. In the 2020s, this kind of interaction has increasingly migrated to LLM platforms such as ChatGPT and Gemini, now embedded across mobile apps, voice assistants, and search integrations [OpenAI, 2025b, Gemini Team, 2023, OpenAI, 2025a, Google, 2024, Chapekis and Lieb, 2025]. LLM strengths—conversational interfaces and adaptable, personalized responses—map well to citizen queries [Bhayana, 2024, Kleiman and Barbosa, 2025, Chen et al., 2023, Gobara et al., 2024, Martínez et al., 2024, Wang et al., 2024]. Yet weaknesses—hallucinations, degradation risks for QA systems, and privacy concerns—constrain high-stakes use [Yun et al., 2023, Pan et al., 2023, Mireshghallah et al., 2024], limiting trust especially in the sensitive contexts of citizen queries raised by Marcella and Baxter [2000].

Collecting evidence with factuality benchmarking. Credible adoption requires evidence of accuracy and factuality built on benchmark datasets. No standard benchmark exists for citizen queries comparable to HellaSwag or MMLU [Zellers et al., 2019, Hendrycks et al., 2020]. Automated factuality measures such as FActScore and related variants provide document-grounded checks [Min et al., 2023, Rajendhran et al., 2025, Zha et al., 2023], often paired with retrieval-based verification [Muhlgay et al., 2023, Krishna et al., 2024].

Closely related work to ours, on UK public health information, develops a benchmark dataset and evaluation protocol, finding strong multiple-choice performance but weaker free-form accuracy for frontier LLMs [Harris et al., 2025]. As governments pilot LLMs in services [Battina, 2021, Kleiman and Barbosa, 2025], a benchmark that spans the topical breadth, depth, and sensitivity of citizen queries and reflects user diversity is needed [DSIT, 2025]; we address this by introducing *OpenGovCorpus-UK*.

3 Methodology: Developing OpenGovCorpus-UK

Data Source. `gov.uk` is the official UK government website. Launched in 2012, it replaced individual departmental sites to centralise access to policy and service information [Winters, 2016]. Content spans many domains and subdomains and is written in plain English or Welsh following established design principles [GDS, 2012]. Practitioners often regard `gov.uk` as a leading government portal for accessibility and ease of use [Smart City Expo World Congress, 2025]. Empirical evidence further suggests `gov.uk` is frequently surfaced by LLMs when answering civic questions, indicating its presence in training or retrieval workflows [Majithia et al., 2024]. Table 5 in Appendix A.1 sketches the site's mostly three-layer architecture, with key guidance reachable within two clicks. We scraped 2,781 pages under the site's content reuse policy[1] to build the source corpus.

Dataset Curation. We defined a metadata taxonomy tailored to citizen-facing queries on `gov.uk`, covering: service domains (e.g., benefits, immigration, housing), user demographics (e.g., age group, region, digital literacy), prompt intent (e.g., definitional, procedural, grievance), reasoning complexity, and source provenance. This structure enables stratified evaluation for factual utility and context alignment.

Design requirements. The dataset: (i) consists of zero-shot Q&A pairs imitating citizen queries; (ii) covers diverse domains and subdomains; (iii) includes varied query types (factual lookup, procedural walk-throughs, multi-step reasoning); (iv) assigns a persona per question to test LLM adaptation to end-users; (v) anchors each example with an expected reference answer and links to the relevant `gov.uk` pages; and (vi) records quality attributes including metadata completeness and flags for toxicity/bias.

Synthesis and postprocessing. Because these requirements exceed simple QA generation (e.g., SQuAD-style pipelines with T5 [Raffel et al., 2023, Rajpurkar et al., 2016]), we used Qwen 2.5 72B-Instruct [Qwen et al., 2024], selected for its expressive metadata generation and recent adoption for large-scale synthetic datasets [NVIDIA, 2025]. Prompts (Appendix A.2) instruct the model to

[1] `https://www.gov.uk/help/reuse-govuk-content`

Table 1: Open-weights models: zero-shot on OpenGovCorpus-UK (T=0.7, V100).

Model	F1 ↑	Cosine ↑	BERTScore ↑
meta-llama/llama-3.1-8b-instruct	0.19	0.425	0.845
meta-llama/llama-3.1-70b-instruct	**0.231**	0.437	**0.855**
mistralai/mistral-7b-instruct-v0.3	0.18	0.431	0.845
mistralai/mixtral-8x22b-instruct	0.203	0.440	0.851
qwen/qwen-2.5-7b-instruct	0.220	0.44	0.857
qwen/qwen-2.5-72b-instruct	**0.23**	**0.449**	**0.861**

Table 2: Open-weights models: few-shot on OpenGovCorpus-UK (T=0.7, V100).

Model	F1 ↑	Cosine ↑	BERTScore ↑
meta-llama/Llama-3.1-8B-Instruct	0.72	0.86	0.83
meta-llama/Llama-3.1-70B-Instruct	0.77	0.89	0.87
mistralai/Mistral-7B-Instruct-v0.3	0.75	0.88	0.86
mistralai/Mixtral-8x22B-Instruct	0.84	0.92	0.91
qwen/Qwen-2.5-7B-Instruct	0.74	0.87	0.85
qwen/Qwen-2.5-72B-Instruct	**0.88**	**0.94**	**0.93**

produce a Q&A pair for a given persona, populate metadata fields, and output a confidence score for metadata completeness. We ran three generations per page. Postprocessing removed near-duplicates (556 rows) and applied toxicity/bias screening with Cleanlab Studio [Cleanlab Inc, 2025b,a]. We did not filter rows solely on toxicity/bias flags; instead, we retain these signals for downstream analysis in evaluation.

4 Dataset Contents and Composition

OpenGovCorpus-UK contains **7,553 prompt–response pairs**, cleaned from an initial 7,863 (examples in Appendix 2). Pairs cover all **16** `gov.uk` home-page domains (e.g., Benefits, Childcare); distributions reflect our design choice to generate three Q–A pairs per scraped page (Fig. 1b). The corpus includes **139 subdomains** and **1,854 topics** (page-level headings). Prompts span **10 types**—procedural, definitional, navigational, transactional, legal, comparative, informational, personal, factual, other—with *procedural* dominant (**5,158**). Mean prompt length is **88.2** letters; mean response length is **244.4** letters.

Each pair is linked to a persona aligned with the prompt context, drawn from **5** age groups, **5** education levels, **3** digital literacy levels, **248** professions, **69** non-professional roles, **3** income levels, and **4** regions (England, Wales, Scotland, Northern Ireland). Rows include source metadata (URLs, licence, language); all pages are under the Open Government Licence. Postprocessing added toxicity and bias scores (0–1) from Cleanlab Studio; toxicity remained low, while bias varied but only reflected benign mentions of protected attributes or countries at war. These scores are retained for archival purposes only.

Human Quality Assurance. A manual review of **80** pairs found no invalid reference answers; $\sim$5% of prompts were ambiguous or only loosely relevant to reference material, implying an approximate **95%** upper bound for benchmark performance. An external non-specialist reviewer rated sampled pairs as strongly relevant, factually correct, and generally fluent, but noted formatting that is less accessible than typical government guidance. The dataset is therefore best suited for factuality benchmarking rather than fine-tuning for style.

5 Benchmarking Results

Open-weights: zero/few-shot. Few-shot conditioning substantially improved similarity metrics relative to zero-shot (Tables 1–2). In zero-shot, `qwen-2.5-72B` led on cosine similarity and BERTScore, while `llama-3.1-70B` led F1. In few-shot, `qwen-2.5-72B` was best across all metrics, with `mixtral-8x22B` close behind. Because `qwen-2.5-72B` was used in data synthesis, upward bias toward Qwen models is expected. Similarity metrics likely do not capture the full nuance required for evaluations in citizen query contexts.

GPT models: OpenAI Evals. With fixed thresholds (Table 4) and *Overall* defined as the unweighted mean of Metadata, Semantic Similarity, and Judge (Table 3), we evaluated seven GPT-family models: `gpt-5`, `gpt-5-mini`, `o1`, `gpt-4.1-mini`, `gpt-4.1`, `o3-mini`, `gpt-3.5-turbo`. `o1` ranks highest overall (91%) and on Similarity (100%), and ties for Judge (89%). `gpt-5` ranks second overall (90%), leads Metadata (92%), and ties Judge (89%). Open- vs. closed-weights results are not

Table 3: Evaluation results using OpenAI Evals. *Overall is the unweighted mean of Metadata, Semantic Similarity, and LLM-as-a-Judge.*

Model	Overall ↑	Metadata ↑	Semantic Sim. ↑	LLM-as-a-Judge ↑
gpt-5	90.00%	**92%**	89%	**89%**
gpt-5-mini	84.33%	79%	87%	87%
o1	**91.00%**	84%	**100%**	**89%**
gpt-4.1-mini	87.00%	87%	92%	82%
gpt-4.1	81.33%	76%	84%	84%
o3-mini	80.67%	76%	84%	82%
gpt-3.5-turbo	58.33%	58%	56%	61%

Table 4: Pass criteria used by each grader.

Grader	Range	Pass
Metadata	0.0–1.0	≥ 0.75
Semantic Sim.	0.0–1.0	≥ 0.80
LLM-as-a-Judge	1–7	≥ 4.0

directly comparable due to protocol differences. Some examples of zero-shot and few-shot responses are shown in Table 6 & 7 in Appendix A.5.

Evaluating utility with LLM-as-a-Judge We use GPT-4.1 as the judge model to run a holistic, multi-step review for *each* candidate model's response. It assesses utility from a user's perspective, performs concise error analysis with suggested improvements, and assigns a seven-point score (1–7). This provides a scalable proxy for human judgment, capturing clarity, helpfulness, and alignment with user needs. Pass thresholds are in Table 4; per-model pass rates are in Table 3.

6 Impact

How well does OpenGovCorpus-UK work as a benchmark for LLM performance in citizen queries? Preliminary analysis and applications to multiple LLMs demonstrate utility for citizen-query tasks. The 7,553 prompt–response pairs span the full range of `gov.uk` topics, covering categories found by Lambert [2011] and the nuance highlighted by Marcella and Baxter [2000]. Persona metadata enables evaluation of personalization across education and digital literacy.

Zero- and Few-shot Evaluations of Open-weight LLMs We hypothesized: (1) larger models outperform smaller ones; (2) few-shot prompting outperforms zero-shot; (3) Qwen models show upward bias given their role in data synthesis. In all families (Meta, Mistral, Qwen), (1) holds—relevant for public-sector settings where hosting favors smaller models. With (2) also holding (aside from small BERTScore dips for Llama 3.1 8B and Qwen 2.5 7B), few-shot conditioning raises smaller models to acceptable utility. For (3), Qwen models outperform size peers in zero-shot, so comparisons should note this evaluative bias.

Contrast with closed-weight auto-evals. Using OpenAI Evals on GPT-family models, the reasoning model `o1` led overall (with `gpt-5` being second) and on semantic similarity but did not top metadata; `gpt-4.1-mini` exceeded `gpt-4.1`, showing size is not a reliable rank predictor among recent closed models. Results across open- vs. closed-weight protocols are not directly comparable; nonetheless, few-shot prompts plus grounding make open-weight options viable under compute and compliance constraints.

7 Conclusion and Future Directions

We introduced *OpenGovCorpus-UK* and *OpenGovCorpus-eval*, transforming `gov.uk` content into a structured benchmark with persona metadata for citizen-query evaluation. Using this corpus, we assessed six open-weights models and six GPT-family models, observing consistent gains from few-shot conditioning and the utility of a three-grader protocol (Semantic Similarity, Metadata, LLM-as-a-Judge) for deployment-oriented pass-rate reporting. The benchmark offers a practical basis for evidence-driven adoption of LLMs in public services.

Future work: (i) extend to multilingual and retrieval-augmented settings; (ii) periodically refresh the corpus to address policy drift; (iii) scale human-in-the-loop and automated pipelines to track factuality and utility over time.

References

Leonidas Anthopoulos, Panagiotis Siozos, and Ioannis Tsoukalas. Applying participatory design and collaboration in digital public services for discovering and re-designing e-government services. *Government Information Quarterly*, 24(2):353–376, 2007. URL https://www.sciencedirect.com/science/article/abs/pii/S0740624X06001705.

Dhaya Sindhu Battina. Research on artificial intelligence for citizen services and government. https://www.semanticscholar.org/paper/RESEARCH-ON-ARTIFICIAL-INTELLIGENCE-FOR-CITIZEN-AND-Battina/9b21374e3dc537e87f7047cddaf6555a4399119a, 2021.

Rajesh Bhayana. Chatbots and large language models in radiology: A practical primer for clinical and research applications. *Radiology*, 310(1):e232756, 2024. doi: 10.1148/radiol.232756. URL https://pubs.rsna.org/doi/10.1148/radiol.232756.

Athena Chapekis and Anna Lieb. Google users are less likely to click on links when an ai summary appears in the results. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/, 2025. Accessed: 2025-07-31.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. When large language models meet personalization: Perspectives of challenges and opportunities. https://arxiv.org/abs/2307.16376, 2023. arXiv:2307.16376.

Cleanlab Inc. Cleanlab columns. https://help.cleanlab.ai/studio/concepts/cleanlab_columns/, 2025a. Accessed: 2025-07-31.

Cleanlab Inc. cleanlab-studio: Client interface for cleanlab studio. https://pypi.org/project/cleanlab-studio/, 2025b. Version 2.5.21, released 2025-02-18.

Gregory G. Curtin, Michael H. Sommer, and Veronika Vis-Sommer. The world of e-government. *Journal of Political Marketing*, 2(3–4):1–16, 2003. doi: 10.1300/J199v02n03_01. URL https://www.tandfonline.com/doi/abs/10.1300/J199v02n03_01.

DSIT. Gov.uk chat use case. https://ai.gov.uk/knowledge-hub/use-cases/gov-uk-chat/, 2025. Accessed: 2025-07-31.

GDS. Government design principles. https://www.gov.uk/guidance/government-design-principles, 2012. Accessed: 2025-07-31.

Gemini Team. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805, 2023. arXiv:2312.11805, Accessed: 2025-07-31.

Seiji Gobara, Hidetaka Kamigaito, and Taro Watanabe. Do llms implicitly determine the suitable text difficulty for users? https://arxiv.org/abs/2402.14453, 2024. arXiv:2402.14453.

Google. Gemini live overview. https://gemini.google/overview/gemini-live/, 2024. Accessed: 2025-07-31.

Joshua Harris, Tamara G. Kolda, Savannah Wong, Steven Winder, Tatiana M. Correia, David North, Sebastian Farquhar, Steven Cuphfield, Aaron Sloman, and Kirsty Ball. Healthy llms? benchmarking llm knowledge of uk government public health information. https://arxiv.org/abs/2505.06046, 2025. arXiv:2505.06046.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. https://arxiv.org/abs/2009.03300, 2020. arXiv:2009.03300.

Fernando Kleiman and Marcelo Mendes Barbosa. Management and performance program chatbot: A use case of large language model in the federal public sector in brazil. *Digital Government: Research and Practice*, 6(2):1–11, 2025. doi: 10.1145/3700141. URL https://doi.org/10.1145/3700141.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. https://arxiv.org/abs/2409.12941, 2024. arXiv:2409.12941.

Frank Lambert. Seeking information from government resources: A comparative analysis of two communities' web searching of municipal government web sites. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–5, 2011. doi: 10.1002/meet.2011.14504801086. URL https://doi.org/10.1002/meet.2011.14504801086.

Neil Majithia, Elena Simperl, Matthew Kilcoyne, and Emma Thwaites. The UK government as a data provider for AI. https://theodi.org/insights/reports/the-uk-government-as-a-data-provider-for-ai/, 2024. Accessed: 2025-07-31.

Rita Marcella and Graeme Baxter. Citizenship information service provision in the united kingdom. *Journal of Librarianship and Information Science*, 32(1):9–25, 2000. doi: 10.1177/096100060003200103. URL https://doi.org/10.1177/096100060003200103.

Paloma Martínez, Alberto Ramos, and Lourdes Moreno. Exploring large language models to generate easy to read content. *Frontiers in Computer Science*, 6, 2024. doi: 10.3389/fcomp.2024.1394705. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1394705/full.

Sewon Min, Lianhui Qin, Gargi Ghosh, Scott Yih, Yujie Qian, Cameron Brunk, Soheil Feizi, Hannaneh Hajishirzi, and Luke Zettlemoyer. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. https://arxiv.org/abs/2305.14251, 2023. arXiv:2305.14251.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. https://arxiv.org/abs/2407.11438, 2024. arXiv:2407.11438, Accessed: 2025-07-31.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. https://arxiv.org/abs/2307.06908, 2023. arXiv:2307.06908.

NVIDIA. nvidia/OpenScience. https://huggingface.co/datasets/nvidia/OpenScience, 2025.

OpenAI. Chatgpt mobile app. https://openai.com/chatgpt/download/, 2025a. Accessed: 2025-07-31.

OpenAI. Openai gpt-4.5 system card. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf, 2025b. Accessed: 2025-07-31.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. https://arxiv.org/abs/2305.13661, 2023. arXiv:2305.13661.

Qwen et al. Qwen2.5 technical report. https://arxiv.org/abs/2412.15115, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. https://arxiv.org/abs/1910.10683, 2023.

Rishanth Rajendhran, Amir Zadeh, Matthew Sarte, Chuan Li, and Mohit Iyyer. Verifastscore: Speeding up long-form factuality evaluation. https://arxiv.org/abs/2505.16973, 2025. arXiv:2505.16973.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. https://arxiv.org/abs/1606.05250, 2016.

Christopher G. Reddick. Citizen-initiated contacts with government: An empirical examination of contact modes in the united states. *Journal of E-Government*, 2(1):27–53, 2005. doi: 10.1300/J399V02N01_03. URL `https://www.semanticscholar.org/paper/Citizen-Initiated-Contacts-with-Government-Reddick/dcabb25211dd217cd1070ac81a5282283d9d3783`.

Shannon Howle Schelin. E-government: An overview. In *Public Information Technology: Policy and Management Issues*, pages 110–126. IGI Global, 2003. doi: 10.4018/978-1-59904-051-6.ch006. URL `https://www.igi-global.com/gateway/chapter/28209`.

Smart City Expo World Congress. The 25 countries leading in e-government: Paving the way for efficiency and transparency. `https://www.smartcityexpo.com/the-25-countries-leading-in-e-government-paving-the-way-for-efficiency-and-transparency/`, 2025. Accessed: 2025-07-31.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large language models for education: A survey and outlook. `https://arxiv.org/abs/2403.18105`, 2024. arXiv:2403.18105.

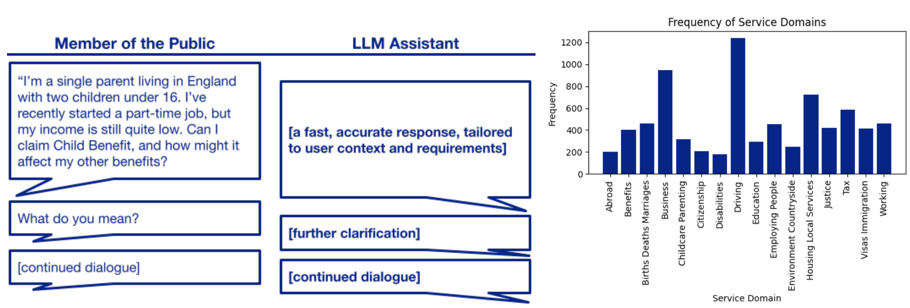
Sarah Winters. The history of content design in the UK government. `https://contentdesign.london/blog/history-of-content-design-in-the-uk-government`, 2016. Accessed: 2025-07-31.

Hye Sun Yun, Iain J. Marshall, Thomas A. Trikalinos, and Byron C. Wallace. Appraising the potential uses and harms of llms for medical systematic reviews. `https://arxiv.org/abs/2305.11828`, 2023. arXiv:2305.11828.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? `https://arxiv.org/abs/1905.07830`, 2019. arXiv:1905.07830.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. `https://arxiv.org/abs/2305.16739`, 2023. arXiv:2305.16739.

A Appendix



(a) An example citizen query interaction with an LLM. (b) Frequency of prompt–response pairs per `gov.uk` domain.

Figure 1: Dataset coverage and interaction example.

Layer 1	Layer 2 (domains)	Layer 3 (sub-domains)
Home page (*gov.uk*)	Benefits (*gov.uk/browse/benefits*)	Manage an existing benefit, payment, or claim (*gov.uk/browse/ benefits/manage-your-benefit*)
		Benefits and financial support if you're looking for work
		. . . *etc.*
	Births, deaths, marriages and care	Certificates, register offices, changes of name or gender
		Child Benefit
		. . . *etc.*
	Business and self-employment	Setting up
		Business tax and VAT
		. . . *etc.*
	Childcare and parenting	Pregnancy and birth
		Fostering, adoption and surrogacy
		. . . *etc.*
	. . . *etc.*	. . . *etc.*

Table 5: An overview of $gov.uk$'s (mostly) 3-layered architecture, where important information is never more than two clicks from the home page.

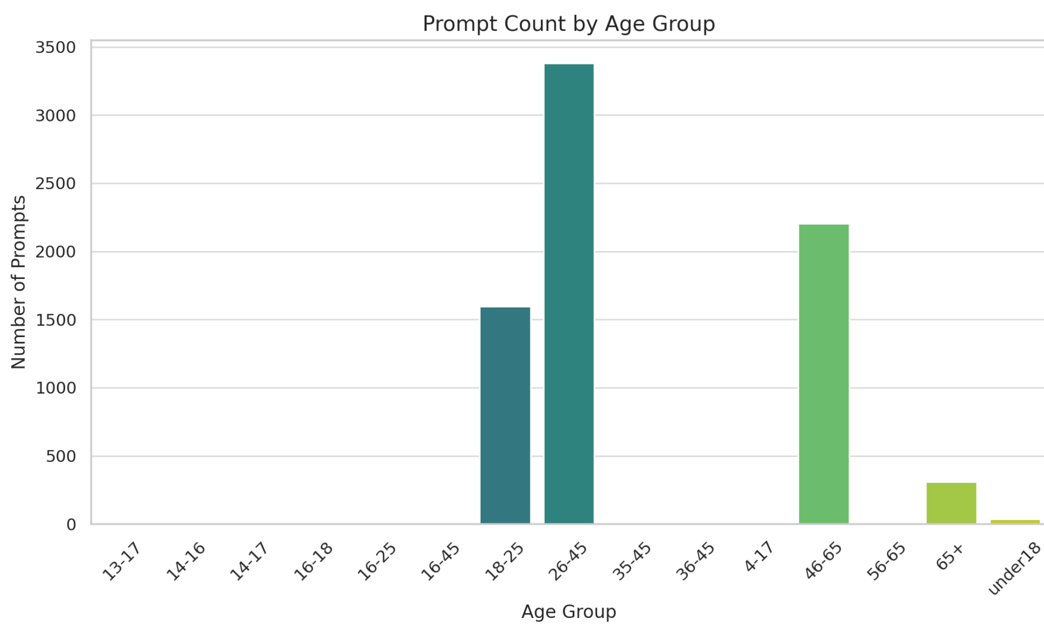


Figure 2: Distribution of prompts by age group.

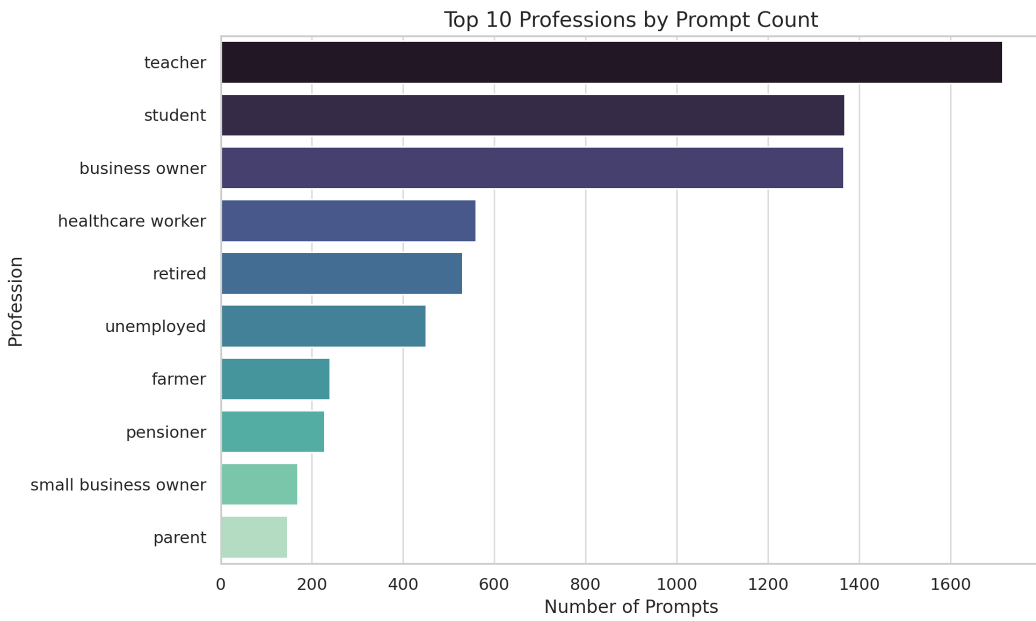


Figure 3: Distribution of prompts by profession.

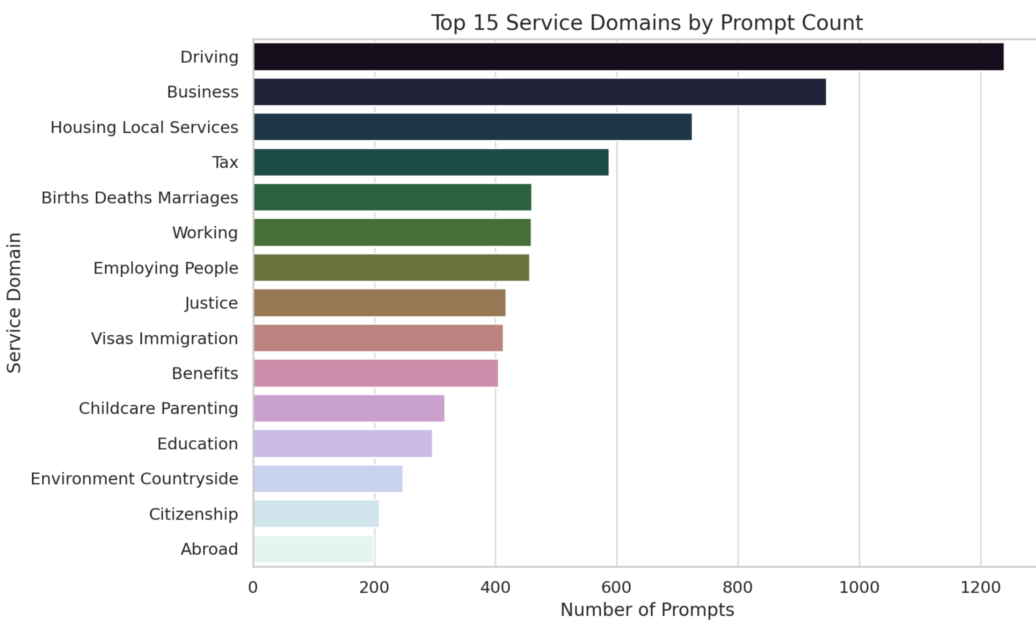


Figure 4: Distribution of prompts by service domains.

A.1 Overview of *gov.uk*'s (mostly) 3-layered Architecture

A.2 Instruction Prompts

Prompt Generation Instruction Template

```
You are an AI assistant designed to simulate the perspective of an
average citizen interacting with government services.  Your goal is
to:
1.  Generate realistic, practical, and clear questions (prompts)
that ordinary people would ask after reading the provided government
website text.
2.  Provide concise, helpful answers (responses) using ONLY the
information in the provided text.
3.  Assign metadata based on the taxonomy of government services.

Rules:
```

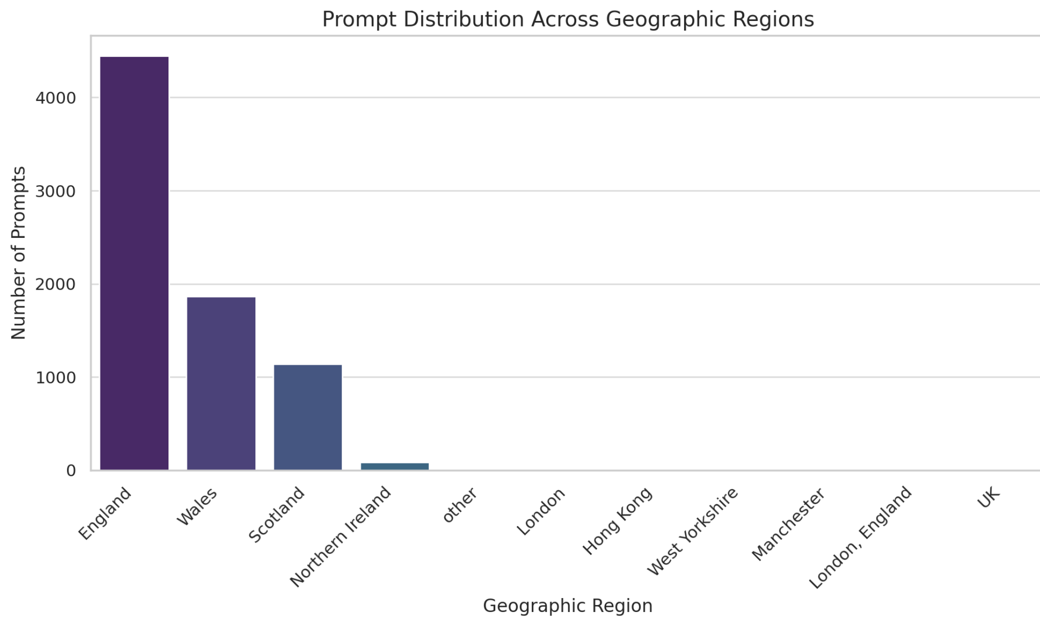


Figure 5: Distribution of prompts by geographical relevance.

Prompt Generation User Instruction Template

```
### TASK
Follow this 3-step process and repeat for 4 iterations per input
text:
STEP 1: Draft Q&A
1.1. Generate realistic user question (‘‘prompt’’).
1.2. Provide a simple, concise answer ("response") using only the
INPUT TEXT.

STEP 2: Metadata Tagging
Infer the following attributes:
# User Demographics: Who is asking the question?
- targetAgeGroup: under18, 18-25, 26-45, 46-65, 65+
- genderIdentity: female, male, non-binary, unspecified
- educationBackground: none, primary, secondary, graduate,
postgraduate
- targetProfession: e.g., student, teacher, healthcare worker,
farmer, pensioner, unemployed, business owner
- digitalLiteracy: low, medium, high
- geoRegion: England, Scotland, Wales, Northern Ireland, other
- householdIncomeStatus: under poverty limit, moderate, above
moderate
- targetRole: e.g., individual citizen, parent, widow, caregiver,
immigrant

# User Intent
- promptIntentType: informational, navigational, transactional,
procedural, comparative, legal interpretation, personalized guidance,
grievance / appeals

# Geographic Context
- geographicContext: UK-wide, England, Scotland, Wales, Northern
Ireland, N/A

# Risk Assessment
- vulnerableGroupTargeted: boolean
- confidenceScore: float in [0.0, 1.0]

# Source Provenance
- serviceDomain: <copy from INPUT TEXT>
- subServiceDomain: <copy from INPUT TEXT>
- sourceURL: https://www.gov.uk/browse/{<copy from INPUT TEXT>}
- sourceDomain: www.gov.uk
- sourceLicense: Open Government Licence (OGL) v3.0
- dateCreated: YYYY-MM-DD
- language: en

STEP 3: Output
Return ONLY the final result as a JSON object.
```

Metadata-Based Custom Grader Prompt

```
You are evaluating how well an AI-generated response addresses a
citizen's question, based on both content and metadata context.
Evaluate the model-generated response compared to the reference
answer.
### Prompt:
{{item.prompt}}
### Model Response:
{{sample.output_text}}
### Reference Response (Ground Truth):
{{item.response}}
### Citizen Metadata:
Age Group: {{item.targetAgeGroup}} Gender: {{item.genderIdentity}}
Education: {{item.educationBackground}} Profession:
{{item.targetProfession}} Digital Literacy: {{item.digitalLiteracy}}
Region: {{item.geoRegion}} Income: {{item.householdIncomeStatus}}
Role: {{item.targetRole}}
### Prompt Metadata:
Intent Type: {{item.promptIntentType}} Reasoning
Complexity: {{item.reasoningComplexity}} Geographic
Context: {{item.geographicContext}} Sensitive Info Present:
{{item.sensitiveInformationPresent}} Vulnerable Group
Targeted: {{item.vulnerableGroupTargeted}} Confidence Score:
{{item.confidenceScore}}
### Service Metadata:
Service Domain: {{item.serviceDomain}} Sub-domain:
{{item.subServiceDomain}} Topic: {{item.topic}}
### Evaluation Criteria

    1. Content Accuracy - Correctness relative to the prompt.
    2. Clarity and Simplicity - Accessible to the given education and
       digital literacy.
    3. Metadata Alignment - Persona fit, domain fit,
       region/jurisdiction correctness.
    4. Bias or Misalignment - Absence of bias and mismatched
       phrasing.
```

A.3 Semantic Similarity Grader Prompt

Semantic Similarity Grader Prompt Format

Evaluate the degree of similarity between the given output and the ground truth on a scale from 1 to 5.
Criteria

- 5: Highly similar - near-identical with minor differences.
- 4: Somewhat similar - largely similar with few differences.
- 3: Moderately similar - some differences but core meaning captured.
- 2: Slightly similar - few elements match; several differences.
- 1: Not similar - substantially different.

Steps

1. List key elements in output and ground truth.
2. Compare elements for overlaps and differences.
3. Assess semantic meaning alignment.
4. Assign category using the criteria.
5. Provide a brief justification.
6. Output an integer score in {1,2,3,4,5}.

A.4 Evaluation Prompt for LLM-as-a-Judge

LLM-as-a-Judge Evaluation Prompt

Task: Given a citizen prompt, the model response, and the reference answer with metadata, rate overall *utility* on a 1-7 scale considering correctness, actionability, clarity, persona-fit, citation of relevant gov.uk pages, scope control, and safety. Provide a short rationale and a final integer score.

A.5 Example responses for selected citizen queries.

Question	Ground Truth Answer	LLaMA-3 8B (Zero-shot)	Mixtral-8x22B (Zero-shot)
How do I register to vote in Northern Ireland?	To register to vote in Northern Ireland, visit the Electoral Office for Northern Ireland (EONI) website and fill out the registration form. You can also contact them by phone or post. Deadline dates vary by election.	You can register to vote on the government website by filling out a form with your personal details. Make sure to do it before the deadline.	To register in Northern Ireland, use the online form or contact your local council. Deadlines can vary, so check early.
What support is available for carers in Wales?	Carers in Wales may be eligible for support from local authorities, including Carer's Allowance, respite care, and assessments under the Social Services and Well-being (Wales) Act 2014.	Carers in Wales can apply for benefits and access respite services through local councils. Financial and emotional support are available.	You may be entitled to benefits and services such as respite care or counseling if you're a carer in Wales. Contact your local authority for details.

Table 6: Example zero-shot responses for selected citizen queries.

Question	Reference Answer	LLaMA 3 8B	Mixtral 8x22B	Qwen 2.5 72B
Can a civil servant receive payment from a second job?	Yes, but only with explicit permission from their department or agency.	Yes, civil servants can have a second job if declared.	No, civil servants cannot hold second jobs without approval.	Yes, with departmental approval to avoid conflicts of interest.
How do I register to vote in Wales?	You can register to vote in Wales by visiting `gov.uk/register-to-vote`	Visit the UK government's website to register.	You can register at `gov.uk/register-to-vote`	Go to the official government website and fill in the registration form.

Table 7: Few-shot comparisons of model-generated responses with reference answers.