

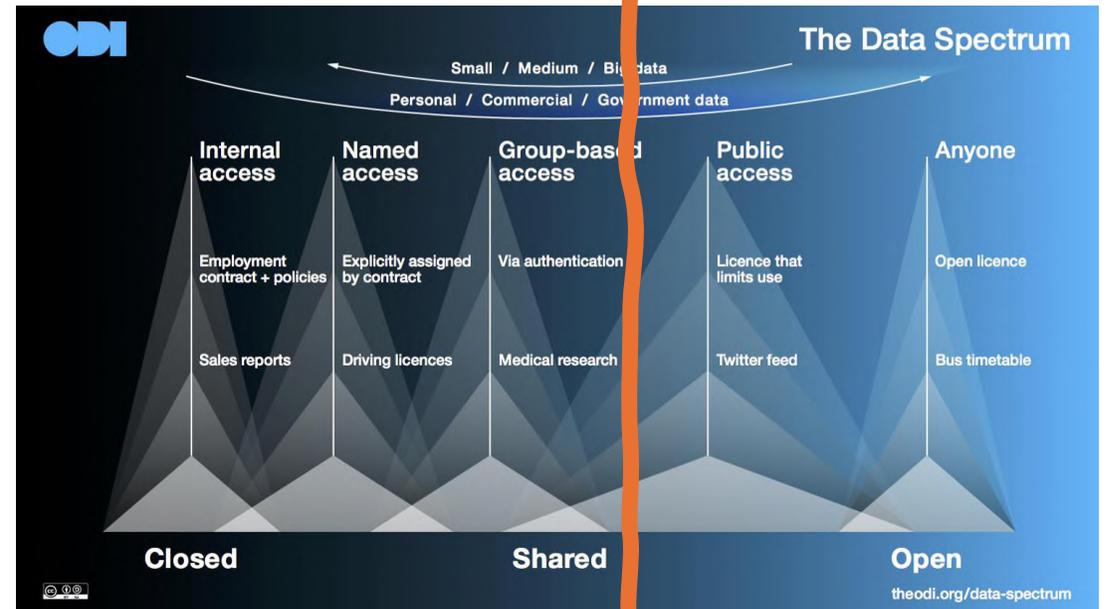
# Open data infrastructure in the age of generative AI

Elena Simperl

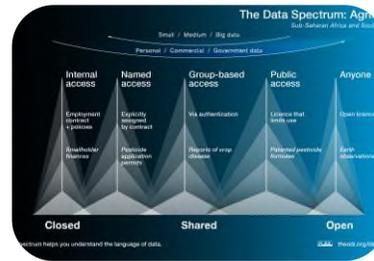
Turing seminar series, University of Bristol, 03/26

# Open data

- Data that anyone can access, use and share.
- No limitations on how the data will be used.
- Not the same as data in the public domain.
- Free to use does not have to mean free to access.



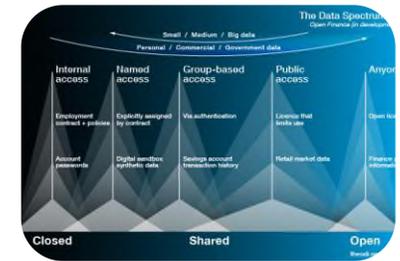
45m datasets indexed by Google Dataset Search, many domains beyond science and government



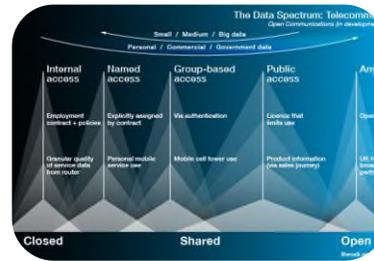
Agriculture



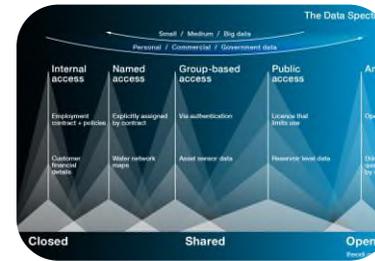
Energy



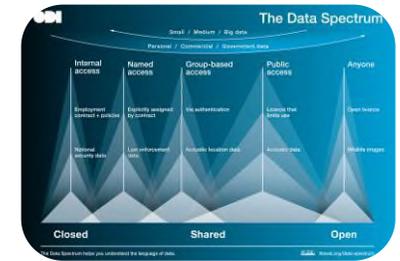
Finance



Telecoms



Water



Wildlife trade

Google europe water quality 2024

Last updated Download format Croissant Usage rights Topic Provider Free Saved data sets

100+ data sets found

**Waterbase - Biology, 2024**  
 data.europa.eu  
 unknown  
 Updated Jul 2, 2025  
 + more versions

**Europe Water Quality Instruments market USD...**  
 cognitivemarketresearch.com  
 pdf,excel,csv,ppt

**Waterbase - Biology, 2024** See More Versions

Explore at: data.europa.eu

unknown

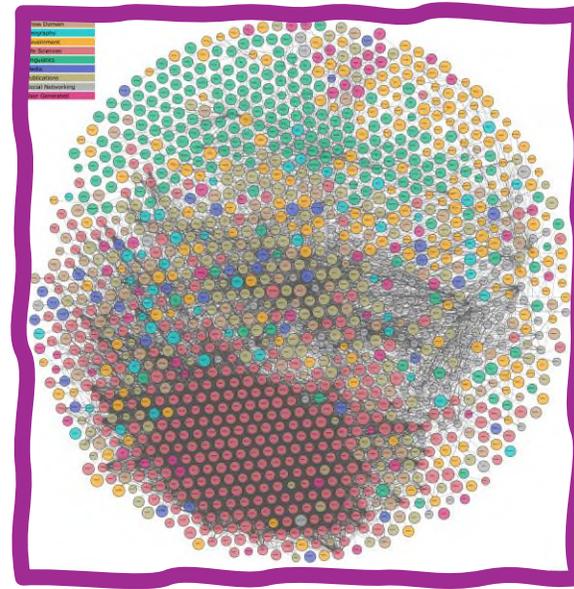
Data set updated Jul 2, 2025

Dataset authored and provided by European Environment Agency

Description

Waterbase serves as the EEA's central database for managing and disseminating data regarding the status and quality of Europe's rivers, lakes, groundwater bodies, transitional, coastal, and marine waters. It also includes information on the quantity of Europe's water resources and the emissions from point and diffuse sources of pollution into surface waters.

Water Quality Sensor Market



Access Status

Open	72,334	No description
Restricted	2,082	Part of EU Open Research Repository
Embargoed	296	Uploaded on September 12, 2025

resource types

Other	44,291	<b>The TIPMP Earth system model experiment protocol: phase 1 - datasets for Figures 2, 3 and 4</b> Bossert, Isaline
Publication	10,096	These files provide the data and code for Figures in a paper that describes a new Earth system model (ESM) experiment protocol, as part of the International Tipping Points Modelling Intercomparison Project (TIPMP) project. This protocol is proposed for the Coupled Model Intercomparison Project (CMIP) project. Part of EU Open Research Repository
Dataset	5,634	Updated on September 12, 2025
Software	5,388	September 12, 2025 (4)
Presentation	3,064	<b>Ocean Drilling Program Site 959 Datasets</b> Slups, Apay, Fokkema, Chris D., Aglaricus, Tobias, and 27 others
Image	2,682	-Version V4 now includes the raw data associated with the paper "Evidence for limited atmospheric pCO2 rise at the onset of the Miocene Climate Optimum", by Wubben et al.; Item "Miocene CO2: raw data" (nd5-972e695a012a1119150-icfb6a3c5292b). An early version of the manuscript is...
Poster		

Hugging Face Search models, datasets, users...

Models Datasets Spaces Community Docs Pricing

Main Tasks Libraries Languages Licenses

Other

Modalities

- 3D Audio Document
- Geospatial Image Tabular
- Text Time-series Video

Size (rows)

<1K >1T

Format

- json csv parquet
- imagefolder soundfolder
- webdataset text arrow

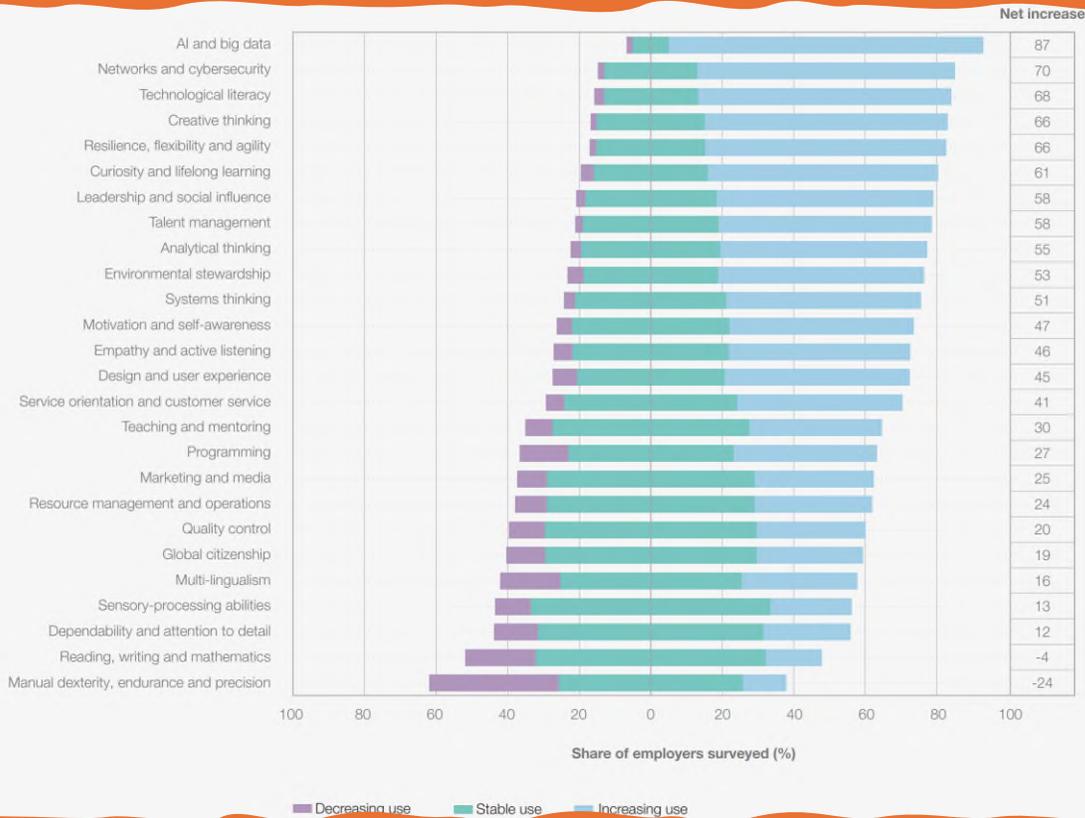
Datasets 498,729 Filter by name Full-text search Sort: Trending

- HuggingFaceFW/finpdfs**  
Viewer · Updated 6 days ago · 475M · ± 51.5k · ♡ 420
- HuggingFaceM4/FineVision**  
Viewer · Updated 10 days ago · 24.2M · ± 1.35k · ♡ 305
- fka/awesome-chatgpt-prompts**  
Viewer · Updated Jan 6 · 203 · ± 44.3k · ♡ 9.03k
- JDhruv14/Bhagavad-Gita\_Dataset**  
Viewer · Updated 1 day ago · 700 · ± 1.95k · ♡ 36
- LucasFang/FLUX-Reason-6M**  
Viewer · Updated 3 days ago · 5.89M · ± 7.3k · ♡ 32
- jupyter-agent/jupyter-agent-dataset**  
Viewer · Updated 4 days ago · 95.8k · ± 3.79k · ♡ 129
- Pageshift-Entertainment/LongPage**  
Viewer · Updated 10 days ago · 300 · ± 9.46k · ♡ 46
- Josephflowers/Finance-Instruct-500k**  
Viewer · Updated Mar 1 · 518k · ± 2.64k · ♡ 173
- SandboxAQ/aqcat25**  
Viewer · Updated 3 days ago · 11.1M · ± 146 · ♡ 15
- HuggingFaceFW/fineweb-2**  
Viewer · Updated Jun 27 · 5.02B · ± 46k · ♡ 639
- ASLP-lab/WenetSpeech-Yue**  
Updated 7 days ago · ± 919 · ♡ 24
- GeoGPT-Research-Project/GeoGPT-QA**  
Viewer · Updated 10 days ago · 41.4k · ± 298 · ♡ 14

### The portal in numbers

Understand the wealth of available information.

<b>1 959 745</b> Datasets	<b>203</b> Catalogues	<b>35</b> Countries
<b>664</b> Publications	<b>258</b> Data stories	<b>11</b> Learning materials



Source: World Economic Forum, 2024

Source: Multiverse, 2024

Technical data skills gaps are having a negative and widespread impact on business performance.

Teams are losing hours and days at a time because of talent shortages and inefficient data tasks: all of which add up to a significant productivity drain.



#### Questions for leaders

Do any of these challenges feel familiar to your business?

**4.34 hours**  
or  
**11% of the working week**

Average time spent per employee each week on inefficiently completing data tasks.

**25 working days**

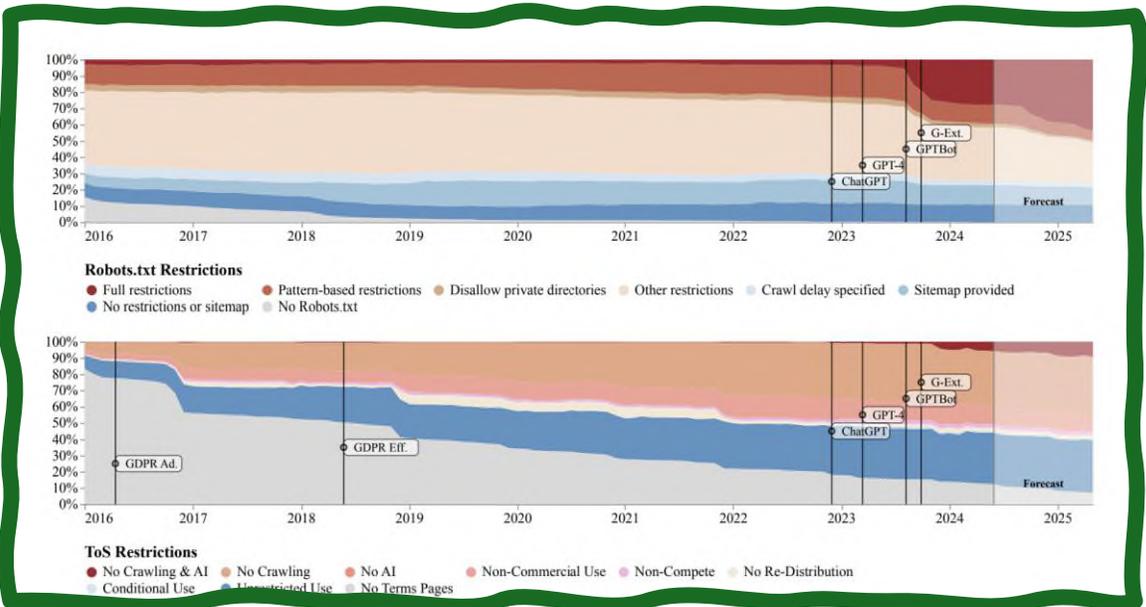
Estimated productive time each employee loses to data skills gaps every year.

#### The Multiverse take

"The lack of data skills is holding businesses back, and it is holding economies back. Workers are wasting hundreds of hours a year because they lack the skills to handle data effectively - that loss is coupled with the faulty decision-making and opportunity cost of poor data use.

"In the age of AI this issue is only exacerbated: companies' ability to harness this game changing technology depends on their ability to access, curate, and make use of their data. Companies have spent billions on software, but hardly anything on the skills needed to get the most from that software."

Euan Blair, CEO at Multiverse



Source: Longpre et al., 2024

Diff Categories About Submit Calendar English

## How crawlers impact the operations of the Wikimedia projects

1 April 2025 by [Birgit Mueller, Wikimedia Foundation](#), [Chris Danis, Wikimedia Foundation](#) and [Giuseppe Lavagetto, Wikimedia Foundation](#)

Translate this post

## Are we entering a “Data Winter”?

Stefaan G. Verhulst [Follow](#) 9 min read · Jan 23, 2024

62

The urgent need to prevent further decline in opening up data for reuse in the public interest



# Prior work

Semantic web and linked data  
Dataset discovery  
Human data interaction  
Open government data portals  
Data storytelling



## Characterising dataset search—An analysis of search logs and data requests

Emilia Kacprzak<sup>a,b,\*</sup>, Laura Koesten<sup>a,b</sup>, Luis-Daniel Ibáñez<sup>a</sup>, Tom Blount<sup>a</sup>, Jeni Tennison<sup>b</sup>, Elena Simperl<sup>a</sup>

<sup>a</sup> University of Southampton, UK  
<sup>b</sup> The Open Data Institute, UK

### ARTICLE INFO

Article history:  
Received 4 April 2018  
Received in revised form 3 September 2018  
Accepted 14 November 2018  
Available online 19 November 2018

Keywords:  
Dataset search  
Vertical search  
Search logs

### ABSTRACT

Large amounts of data are becoming increasingly available online. In order to benefit from it we need tools to retrieve the most relevant datasets that match ones data needs. Several vocabularies have been developed to describe datasets in order to increase their discoverability, but for data publishers it can be cumbersome to annotate them using all, leading to the question of what properties are most important. In this work we contribute with a systematic study of the patterns and specific features that inform consumers use to search for data and how it compares with general web search. We analyse a year of data requests for requests issued to one of them. Search queries issued on data portals differ from those issued to web search engines in their length, topic, and structure. Based on our findings we propose that portals search functionalities are currently used in an exploratory manner, rather than to retrieve a specific resource. In our study of data requests we found that geospatial and temporal attributes, as well as information on the required granularity of the data are the most common features. The findings of both analyses suggest that these features are of higher importance in dataset retrieval in contrast to general web search, suggesting that efforts of dataset publishers should focus on generating dataset descriptions including them.

© 2018 Elsevier B.V. All rights reserved.

### 1. Introduction

Data has become the most important digital asset in the world and its availability on the web is increasing rapidly. A growing number of organisations, mostly in the public sector, have set up their own data portals to publish datasets related to their activities. Similar trends can be observed in a variety of sectors. In the public sector, through initiatives such as Open Government Data (e.g. US Open Data portal,<sup>1</sup> UK Open Data portal<sup>2</sup> etc.), data can generate social impact, improve public services, and increase transparency [1]. Specialised vendors in commercial sectors such as finances and marketing co-exist alongside data marketplaces that connect supply and demand (e.g. data.world<sup>3</sup>, Microsof DataMarket,<sup>4</sup> etc.). In science, an increasing number of datasets get

published together with scientific publications, as open access and reproducibility become mainstream across subjects and research communities (e.g. Mendeley Data,<sup>5</sup> Elsevier DataSearch Platform<sup>6</sup>). A study by Cafarella et al. [2] estimated more than one billion sources of data on the web as of 2011, counting structured data extracted from web pages. In 2015 the Web Data Commons project extracted 233 million data tables from the Common Crawl [3]. The ability to generate business value from data analytics offers competitive advantage in virtually every industry worldwide [4].

Data is used in a variety of professional roles. Whether it is a data journalist writing an article that compares government transparency in different countries, an app developer trying to expand into new markets, a business analyst searching for evidence to substantiate their report, or a scientist replicating an experiment, the first and foremost step all these professionals have to take is find, or retrieve the most relevant datasets for their needs. In previous work, we found that data practitioners (with different professional backgrounds and skills) face various challenges finding the data

\* Correspondence to: The Open Data Institute, 65 Clifton Street, London EC2A 4JE, UK.  
E-mail address: emilia.kacprzak@theodi.org (E. Kacprzak).

<sup>1</sup> <http://data.gov>,  
<sup>2</sup> <http://data.gov.uk>,  
<sup>3</sup> <http://www.data.world>,  
<sup>4</sup> <http://www.datamarket.azure.com>.

<https://doi.org/10.1016/j.wswebm.2018.11.003>  
1570-8268/© 2018 Elsevier B.V. All rights reserved.

## A comparison of dataset search behaviour of internal versus search engine referred sessions

Luis-Daniel Ibáñez  
University of Southampton  
Southampton, UK  
l.d.ibanez@southampton.ac.uk

Elena Simperl  
King's College London  
London, UK  
elena.simperl@kcl.ac.uk

### ABSTRACT

Dataset discovery is a first step for data-centric tasks, from data storytelling to labelling for supervised machine learning. Previous qualitative research suggests that people use two types of search affordances to find the data they need: they either go to a data portal that probably contains the data and search there; or they start on a regular web search engine, which sometimes returns results that are datasets. For the first type of search, prior works have analysed logs from different data portals to understand basic facets of search behaviour such as query length or topics. In this paper, we advance the state of the art in dataset search behaviour with a comprehensive transaction log analysis study (n = 236441 sessions) on an international open data portal, in which we compare sessions acquired via a portal or referred from search engines. We analyse search logs from a data portal and search engines to understand differences in search behaviour. We find that search engines tend to be used for general searches, while data portals are used for more specific searches. We also discover and discuss behavioural patterns and user profiles across session types.

### CCS CONCEPTS

• Information systems → Environment-specific retrieval; Search interfaces.

### KEYWORDS

dataset search, information seeking, log analysis, search behaviour

### ACM Reference Format:

Luis-Daniel Ibáñez and Elena Simperl. 2022. A comparison of dataset search behaviour of internal versus search engine referred sessions. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*, March 14–18, 2022, Regensburg, Germany. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3498366.3505821>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CHIIR '22, March 14–18, 2022, Regensburg, Germany  
© 2022 Association for Computing Machinery.  
ACM ISBN 978-1-4901-9126-3/22/03...\$15.00  
<https://doi.org/10.1145/3498366.3505821>

### 1 INTRODUCTION

Data has become a fundamental resource to improve organisational processes, inform decisions, and train AI algorithms. For many data-centric tasks, the first step is *data discovery*, a term which refers to all activities around finding, making sense, and evaluating data for reuse [4].  
When looking for datasets, potential users tend to follow several strategies [16]: (i) they type keywords in a web search engine, which may lead them to an online data repository; (ii) they have an idea where the data might be located and use the search affordances of that site to find datasets matching their needs; (iii) they ask other people for suggestions; or (iv) where the publisher is a public authority they issue a freedom of information or a data request to the authority. These search strategies have been reported in the literature, but the ability of metadata, which records key information about datasets, for example the author, their affiliation, and contact point, to help people find relevant data. Data search (DDS) promote the use of standardised metadata vocabularies, such as DCAT and schema.org, to improve how they match people's information needs to datasets.

Previous work in understanding user search behaviour in data discovery has compared strategies of type (i) to (iv), using a mix of descriptive and qualitative methods on corpora with aggregated search sessions and data requests [12]; as well as strategies of type (ii) across different verticals. For instance, document against dataset search in the context of a digital library with both types of artifacts [3].

In this paper, we go a step further and present a transaction log analysis to compare sessions representative of strategies of types (i) and (ii). We tackle the following research questions:

- 1) What are the behavioural patterns of dataset search users? Do these patterns change depending on the type of session?
- 2) Are users more successful using one strategy or another? Is there a statistically significant relationship between strategy type and success?
- 3) Do users use more keywords and facets when they start on the portal or when they land on the portal from a web search engine? Is there a statistically significant relationship between strategy type and the use of facets and keywords?
- 4) What are the user profiles of dataset search users?

We use a transaction log of 236441 sessions over a one-year period from the European Data Portal (EDP, meanwhile re-named to [data.europa.eu](http://data.europa.eu)).<sup>1</sup> This is a portal for openly available government data. It aggregates and curates metadata of 1.4 million such datasets, held in over 80 national and regional repositories from 36 countries.

<sup>1</sup> <https://data.europa.eu/en>

### SPECIAL ISSUE PAPER



## Dataset search: a survey

Adriane Chapman<sup>1</sup>, Elena Simperl<sup>1</sup>, Laura Koesten<sup>1</sup>, George Konstantinidis<sup>1</sup>, Luis-Daniel Ibáñez<sup>1</sup>, Emilia Kacprzak<sup>2</sup>, Paul Groth<sup>3</sup>

Received: 27 December 2018 / Revised: 15 July 2019 / Accepted: 12 August 2019 / Published online: 24 August 2019  
© The Author(s) 2019

### Abstract

Generating value from data requires the ability to find, access and make sense of datasets. There are many efforts underway to encourage data sharing and reuse, from scientific publishers asking authors to submit data alongside manuscripts to data marketplaces, open data portals and data communities. Google recently beta-released a search service for datasets, which allows users to discover data stored in various online repositories via keyword queries. These developments foreshadow an emerging research field around dataset search or retrieval that broadly encompasses frameworks, methods and tools that help match a user data need against a collection of datasets. Here, we survey the state of the art of research and commercial systems and discuss what makes dataset search a field in its own right, with unique challenges and open questions. We look at approaches and implementations from related areas dataset search is drawing upon, including information retrieval, databases, data science and tabular search in order to identify possible paths to tackle these questions as well as immediate next steps that will raise the field forward.

Keywords Dataset search · Dataset retrieval · Dataset · Information search and retrieval

### 1 Introduction

Data is increasingly used in decision making: to design public policies, identify customer needs, or run scientific experiments [64,173]. For instance, the integration of data from deployed sensor systems such as mobile phone net-

✉ Adriane Chapman  
adriane.chapman@soton.ac.uk  
Elena Simperl  
E.Simperl@soton.ac.uk  
Laura Koesten  
L.M.Koesten@soton.ac.uk  
George Konstantinidis  
G.Konstantinidis@soton.ac.uk  
Luis-Daniel Ibáñez  
L.D.Ibanez@soton.ac.uk  
Emilia Kacprzak  
emilia.kacprzak@theodi.org  
Paul Groth  
p.groth@uva.nl

<sup>1</sup> University of Southampton, Southampton, UK  
<sup>2</sup> The Open Data Institute, London, UK  
<sup>3</sup> University of Amsterdam, Amsterdam, The Netherlands

works, camera networks in intelligent transportation systems (ITS) [103] and smart meters [3] is powering a number of innovative solutions, such as the city of London's oversight dashboard [17]. Datasets are increasingly being exposed for trade within data markets [13,70] or shared via open data portals [41,80,97,125,144,174] and scientific repositories [5,57]. Communities such as Wikidata or the Linked Open Data Cloud [125] come together to create and maintain vast, general-purpose data resources, which can be used by developers in applications as diverse as intelligent assistants, recommender systems and search engine optimization. The common intent is to broaden the use and impact of the millions of datasets that are being made available and shared across organizations [24,148,184]. This trend is reinforced by advances in machine learning and artificial intelligence, which rely on data to train, validate and enhance their algorithms [159]. In order to support these uses, we must be able to search for datasets. Searching for data in principled ways has been researched for decades [42]. However, many properties of *datasets* are unique, with interesting requirements and constraints, which have been recognized by the recent release of Google Dataset Search [141]. There are many open problems across dataset search, which the database community can assist with.

## The Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour

Laura M Koesten  
The Open Data Institute, UK  
Univ. of Southampton, UK  
laura.koesten@theodi.org

Emilia Kacprzak  
The Open Data Institute, UK  
Univ. of Southampton, UK  
e.kacprzak@theodi.org

Jenifer F A Tennison  
The Open Data Institute, UK  
jeni@theodi.org

Elena Simperl  
Univ. of Southampton, UK  
e.simperl@soton.ac.uk

### ABSTRACT

Structured data such as databases, spreadsheets and web tables is becoming critical in every domain and professional role. Yet we still do not know much about how people interact with it. Our research focuses on the information seeking behaviour of people looking for new sources of structured data online, including the task context in which the data will be used, data search, and the identification of relevant datasets from a set of possible candidates. We present a mixed-methods study covering in-depth interviews with 20 participants with various professional backgrounds, supported by the analysis of search logs of a large data portal. Based on this study, we propose a framework for human structured-data interaction and discuss challenges people encounter when trying to find and assess data that helps their daily work. We provide design recommendations for data publishers and developers of online data platforms such as data catalogs and marketplaces. These recommendations highlight important questions for HCI research to improve how people engage and make use of this incredibly useful online resource.

### ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g. HCI): User-centered design

### Author Keywords

Human Data Interaction; Data Search; Data Portal

### INTRODUCTION

Structured data, which is data that is explicitly organised, for example in relational databases, spreadsheets and web tables, is becoming critical in every domain and professional role [46]. We use it in various ways - from consulting official statistics



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI 2017 May 06–11, 2017, Denver, CO, USA  
© 2017 Copyright held by the owner/authors.  
ACM ISBN 978-1-5304-4655-9/17/05.  
DOI: <http://dx.doi.org/10.1145/3025453.3025838>

to running scientific experiments, finding travel routes, creating maps, predicting elections and designing better products. More and more of it can be accessed or purchased online - a 2011 study by Cafarella et al. found more than one billion structured data resources on the (deep) web (as HTML tables, lists, etc.) [11], while McKinsey estimated two years later that more than one million datasets have been made openly available by governments worldwide [45, 63]. At the same time, the demand for financial, economic and marketing data provided by vendors such as Bloomberg and others continues to increase [10]. And yet, despite its prevalence and applications, we still know very little about how people search for data, understand it and put it to use. The tools that support one's "data journey" [3] - exploring and assessing their relevance - often do not offer the best user experience [61].

Various scientific communities, including information retrieval (IR), databases, Linked Data, data visualisation and HCI, have looked at such data journeys from different angles. They have proposed new interaction models to engage with a particular species of data, such as graphs [57] or time series [16]; studied information needs and how they are formulated [7]; and developed tools for specific data-related activities, for example statistical analysis [33], visualisation [19], personal information management [31] and teamwork [5].

The scenario we are targeting is slightly different: imagine a data journalist writing an article about the runway expansion at London's main airport in the UK. As part of her research, the journalist will look for factual evidence to substantiate her story, in the form of reports, news on similar topics, as well as data about the economic, social and environmental ramifications of the project, arguing for or against expansion plans at each airport location. A large share of the relevant data is already available online, published by governmental agencies, researchers and other journalists. However, finding and using it is not always straightforward. The journalist could use regular search engines in the same way she does when looking for less structured kinds of information (such as regular Web sites). She might also know of a particular

## FORUM | UX MEETS AI

This forum provides a space to engage with the challenges of designing for intelligent algorithmic experiences. We invite articles that tackle the tensions between research and practice when integrating AI and UX design. We welcome interdisciplinary debate, artful critique, forward-looking research, case studies of AI in practice, and speculative design explorations. — Juho Kim and Henriette Cramer, Editors

## UX of Data: Making Data Available Doesn't Make It Usable

Laura Koesten, University of Vienna, Elena Simperl, King's College London

**D**ata plays an important part in our daily lives. It shapes how we view the world and, for better or worse, informs the decisions we make. Despite controversies around who collects and does what with data, data-centric work is seen as critical to solving the most complex problems of today, from climate change and security to health crises and

with other people's data? Using a dataset beyond the context in which it originated remains challenging. Simply making data available, even when following existing guidelines and best practices, does not mean it can be easily used by others. Further, there is little evidence that such guidance, including technical standards and community support, do indeed lead to more engagement and generate value.

In our research, we learned a great deal about how different audiences go about finding data online, how they choose what data to work with, and how they make sense of data that others have shared. We explored

the ability at a later point in time to reproduce what happened. This is especially the case as deep, exploratory or data-curation work is often seen as secondary to model design or machine learning.

Our initial exploratory studies were framed in the theories about information seeking: We asked people what their data tasks were—what they did with data in their jobs, how they searched for it, how they evaluated and selected datasets as search results, and how they explored and understood data with which they were not familiar.

We also used different forms of summarization (verbal and textual) to document the cognitive processes involved in data sensemaking. We asked people to describe and discuss data they knew or data that was new to them, and identified common activity patterns [1]. We could distinguish among three clusters: *inspecting, engaging with content, and placing data in context*. Our research suggests that

### Insights

- We need better ways to record and communicate that data represents only a perspective of reality.
- There is no typical data task. Data tools need to provide better support for each task in the data life cycle.
- Interaction challenges for making data usable include making data understandable, supporting collaboration, managing changes, and facilitating different modes of access and user engagement.

INTERACTIONS.ACM.ORG

MARCH–APRIL 2021 INTERACTIONS 97

## Patterns

### Article

## Dataset Reuse: Toward Translating Principles to Practice

Laura Koesten,<sup>1,2</sup> Pavlos Vougiouklis,<sup>2</sup> Elena Simperl,<sup>1</sup> and Paul Groth<sup>3,4,\*</sup>

<sup>1</sup>King's College London, London WC2B 4BG, UK

<sup>2</sup>Huawei Technologies, Edinburgh EH9 3BF, UK

<sup>3</sup>University of Amsterdam, Amsterdam 1090 GH, the Netherlands

<sup>4</sup>Lead Contact

\*Correspondence: [laura.koesten@kcl.ac.uk](mailto:laura.koesten@kcl.ac.uk) (L.K.), [p.groth@uva.nl](mailto:p.groth@uva.nl) (P.G.)

<https://doi.org/10.1016/j.patter.2020.100136>

**THE BIGGER PICTURE** The web provides access to millions of datasets. These data can have additional impact when it is used beyond the context for which it was originally created. We have little empirical insight into what makes a dataset more reusable than others, and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This work demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

**KEYWORDS** Data science output has been formulated, and tested for one domain/problem

### SUMMARY

The web provides access to millions of datasets that can have additional impact when used beyond their original context. We have little empirical insight into what makes a dataset more reusable than others and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

### 1 INTRODUCTION

There has been a gradual shift in the last years from viewing datasets as byproducts of (digital) work to critical assets, whose value increases the more they are used.<sup>1,2</sup> However, our understanding of how this value emerges, and of the factors that demonstrably affect the reusability of a dataset is still limited.

Using a dataset beyond the context where it originated remains challenging for a variety of socio-technical reasons, which have been discussed in the literature;<sup>3,4</sup> the bottom line is that simply making data available, even when complying with existing guidance and best practices, does not mean it can be easily used by others.<sup>5</sup>



This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

At the same time, making data reusable to a diverse audience, in terms of domain, skill sets, and purposes, is an important way to realize its potential value (and recover some of the, sometimes considerable, resources invested in policy and infrastructure support). This is one of the reasons why scientific journals and research-funding organizations are increasingly calling for further data sharing<sup>6</sup> or why industry bodies, such as the International Data Spaces Association (IDSA) (<https://www.internationaldataspaces.org/>) are investing in reference architectures to smooth data flows from one business to another.

There is plenty of advice on how to make data easier to reuse, including technical standards, legal frameworks, and guidelines. Much work places focus on machine readability

Patterns 1, 100136, November 13, 2020 © 2020 The Author(s).

**European data**  
 data.europa.eu The official portal for European data

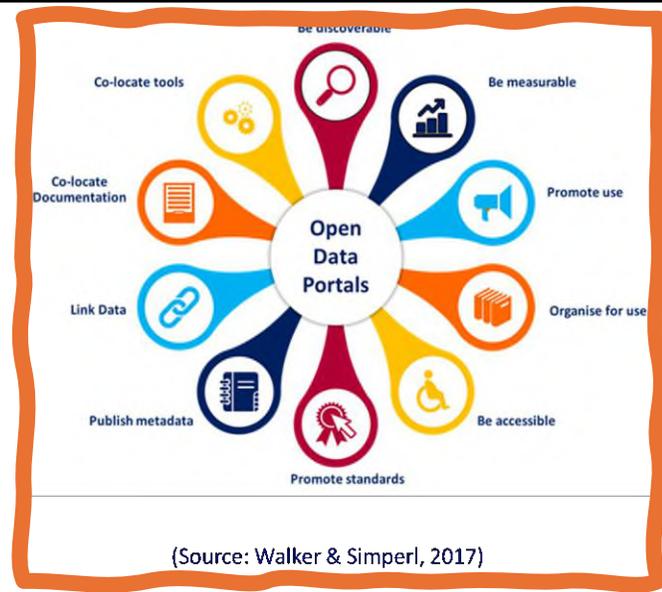
Home Data EU Open Data Days Academy Open data maturity Community Publications Documentation

Home > data.europa academy > Mid- and long-term sustainability of (open) d...

## Mid- and long-term sustainability of (open) data portal infrastructures



**Format:** E-learning, Reading  
**Theme:** PORTAL  
**Level:** Advanced



## Analytical Report 19: Understanding supply and demand on the European Data Portal

**Date:** 24 February 2021  
**Location:** Europe

The European Data Portal's 19th analytical report pro-vides a detailed analysis of the relationship between supply and demand of datasets on the EDP

## Analytical Report 18: Characterising Dataset Search on the European Data Portal

**Date:** 28 October 2020  
**Location:** Europe

The European Data Portal's 18th analytical report illustrates a quantitative study on data search through more than two years of EDP search and interaction logs. Understanding data search behaviour is key to developing better search algorithms and improving the search experience. This study presents current findings from key literature in data search.

**Principles and recommendations to make data.europa.eu data more reusable**

*A strategy-mapping report*

Hand Welcome to the User-centric Open Data Publishing Toolkit

How This Toolkit Is Organised

Tools

Case Studies and Research

THINKING ABOUT USERS

What Do We Mean By User-Centric?

What Are Users Doing With Your Data?

SELECTION

Why Are You Opening Data?

Creating an Aligned Open Data Culture



# Welcome to the User-centric Open Data Publishing Toolkit

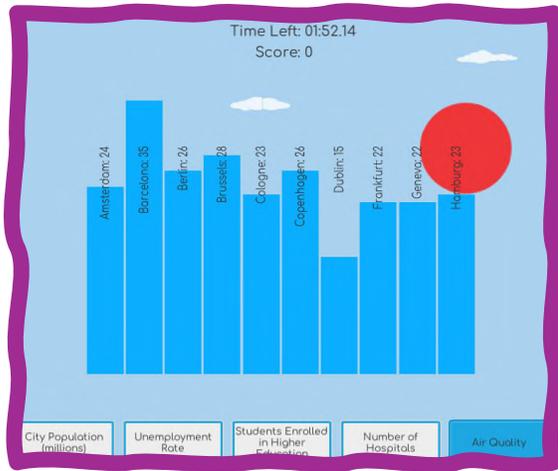
Copy

[How to use this toolkit](#)

How can we help you publish your open data?

How do you publish open data that people can use easily and effectively?

Open data is most often published in a form based on the preferences and decisions of the organisation that holds the data. However, these do not



### Studies in data interaction through play: do data games work?

**Tom Blount**  
Web and Internet Science  
University of Southampton, UK  
t.blount@soton.ac.uk

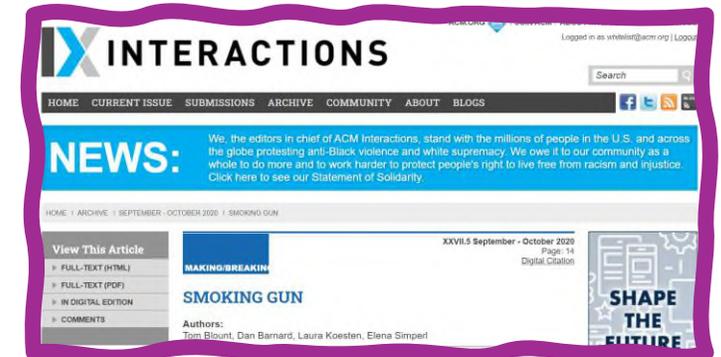
**Elena Simperl**  
Department of Informatics  
King's College London, UK  
elena.simperl@kcl.ac.uk

#### ABSTRACT

Data games allow people to interact with data through the act of play. They have been proposed in information design as a means to communicate data to a more diverse range of audiences, and make it more accessible and engaging. Several such games have been published in recent years, with mixed

#### INTRODUCTION

Our ability to stay informed and make decisions is increasingly affected by data [41]. Engaging with this data can take many forms, from tables and spreadsheets to charts, interactive experiences, and, more recently, data games and 'playable data' [27].



**INTERACTIONS**

NEWS: We, the editors in chief of ACM Interactions, stand with the millions of people in the U.S. and across the globe protesting anti-Black violence and white supremacy. We owe it to our community as a whole to do more and to work harder to protect people's right to live free from racism and injustice. Click here to see our Statement of Solidarity.

View This Article

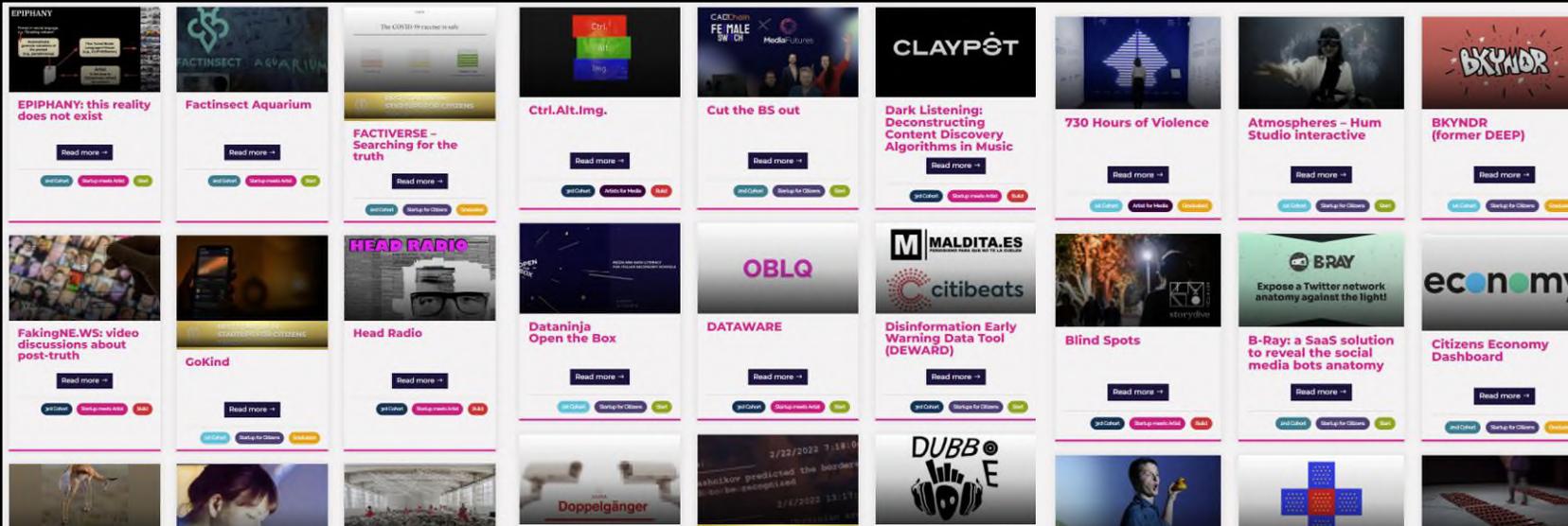
- FULL-TEXT (HTML)
- FULL-TEXT (PDF)
- IN DIGITAL EDITION
- COMMENTS

**SMOKING GUN**

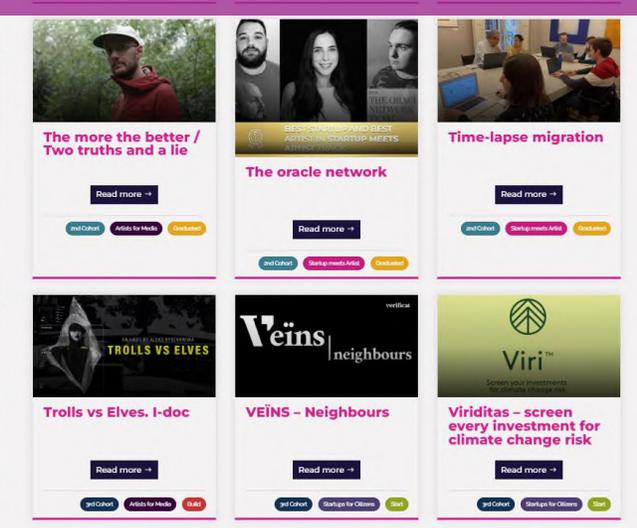
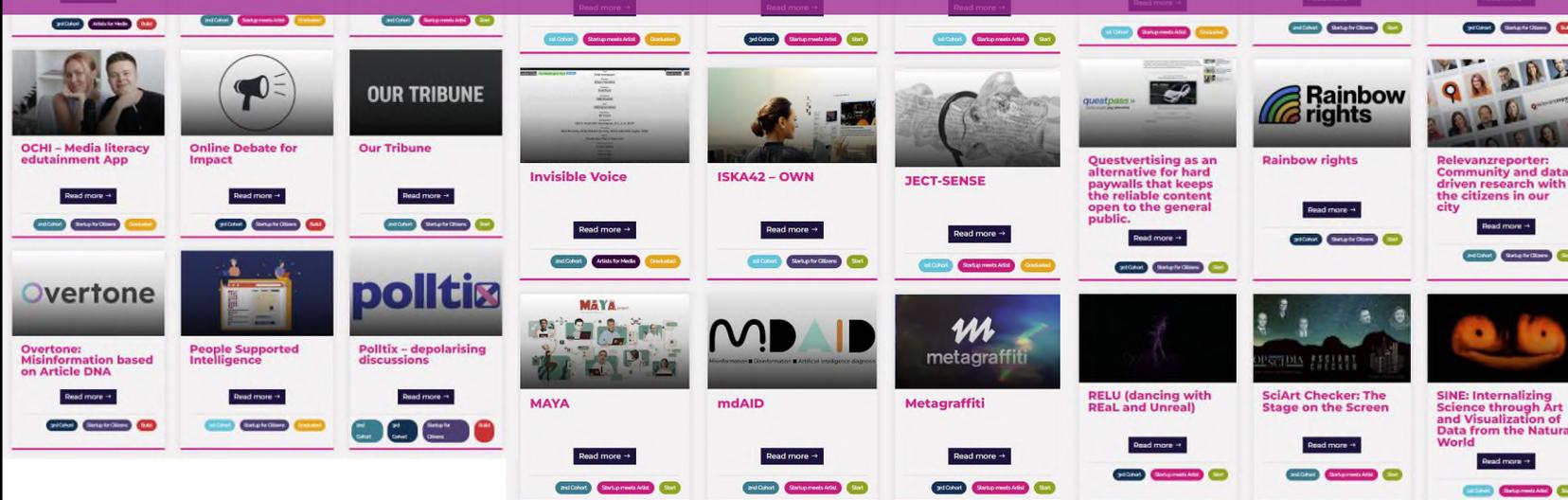
Authors: Tom Blount, Dan Barnard, Laura Koesten, Elena Simperl

XXVII.5 September - October 2020  
Page: 14  
Digital Citation

SHAPE THE FUTURE

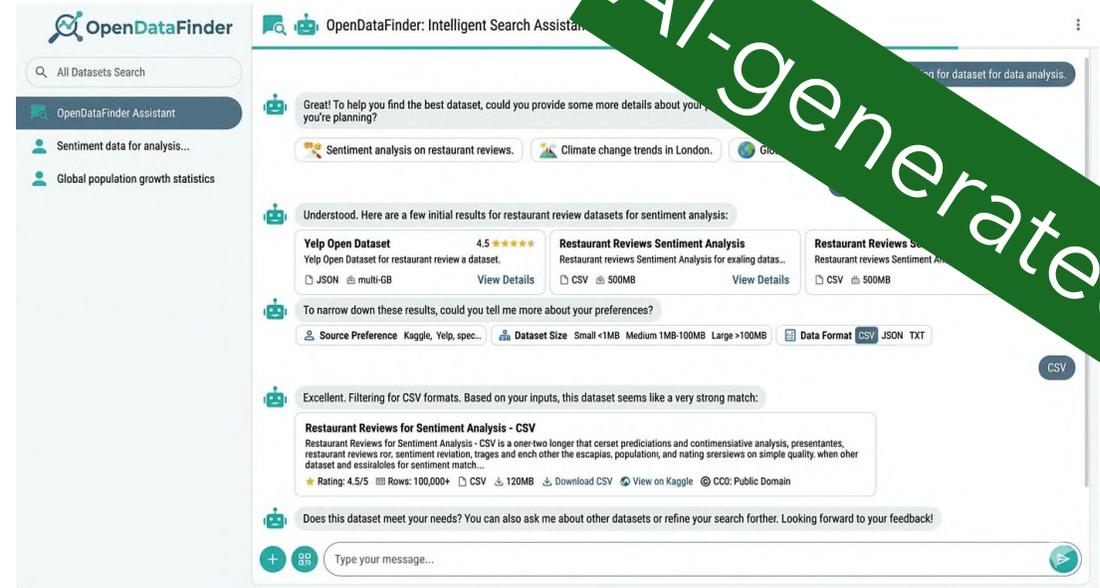


# 67 ideas to tackle misinformation and disinformation with data and AI



# Ongoing work

- Data prompting
- Generative AI for dataset search
- AI-ready data
- AI-enabled public services



AI-generated

# Data Prompting: Assessing the Potential of Conversational Generative AI for Dataset Discovery

*Cell Press*

27 Pages • Posted: 27 Sep 2024 • Publication Status: **Under Review**

[Johanna Walker](#)

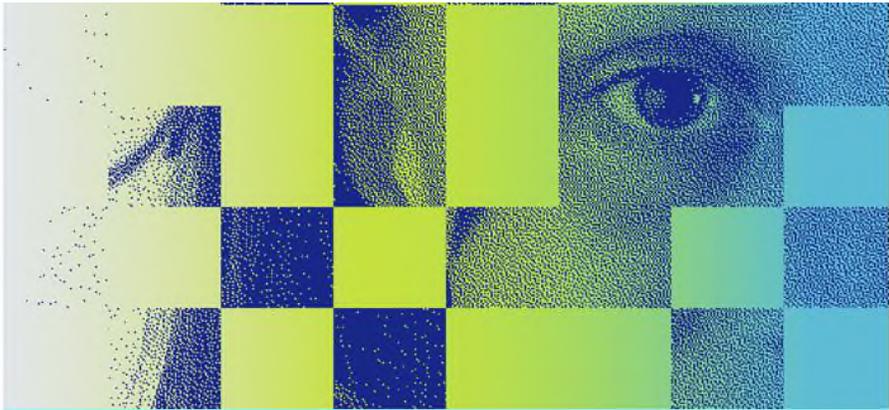
King's College London

[Elisavet Koutsiana](#)

King's College London

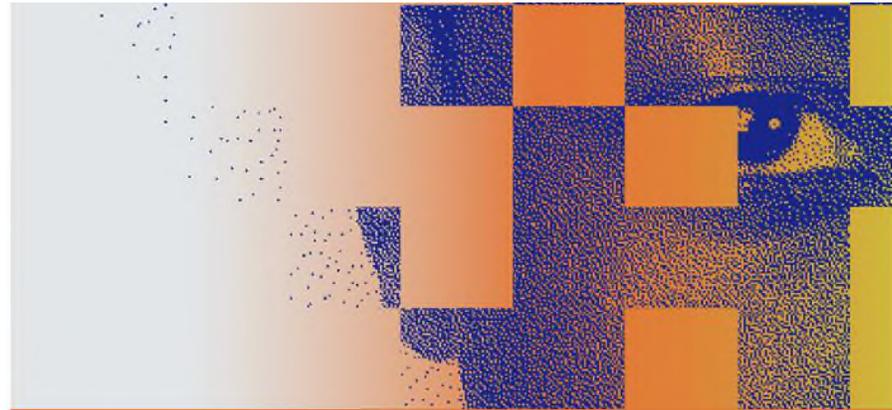
[More...](#)

To appear



Report

## **Generative AI tools for data discovery and use**

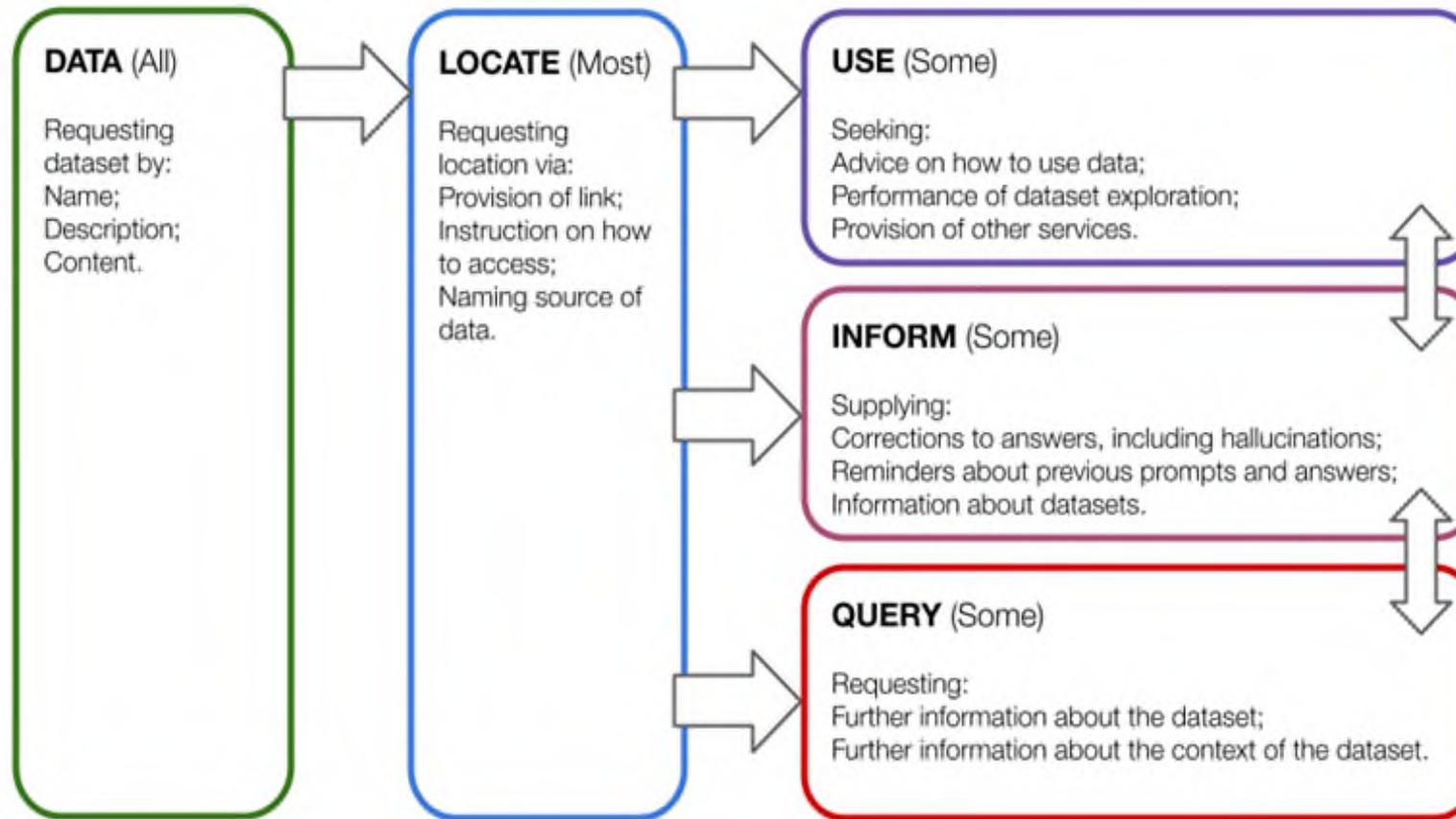


Report

## **Democratising access to data: Bridging the data divide with generative AI models**



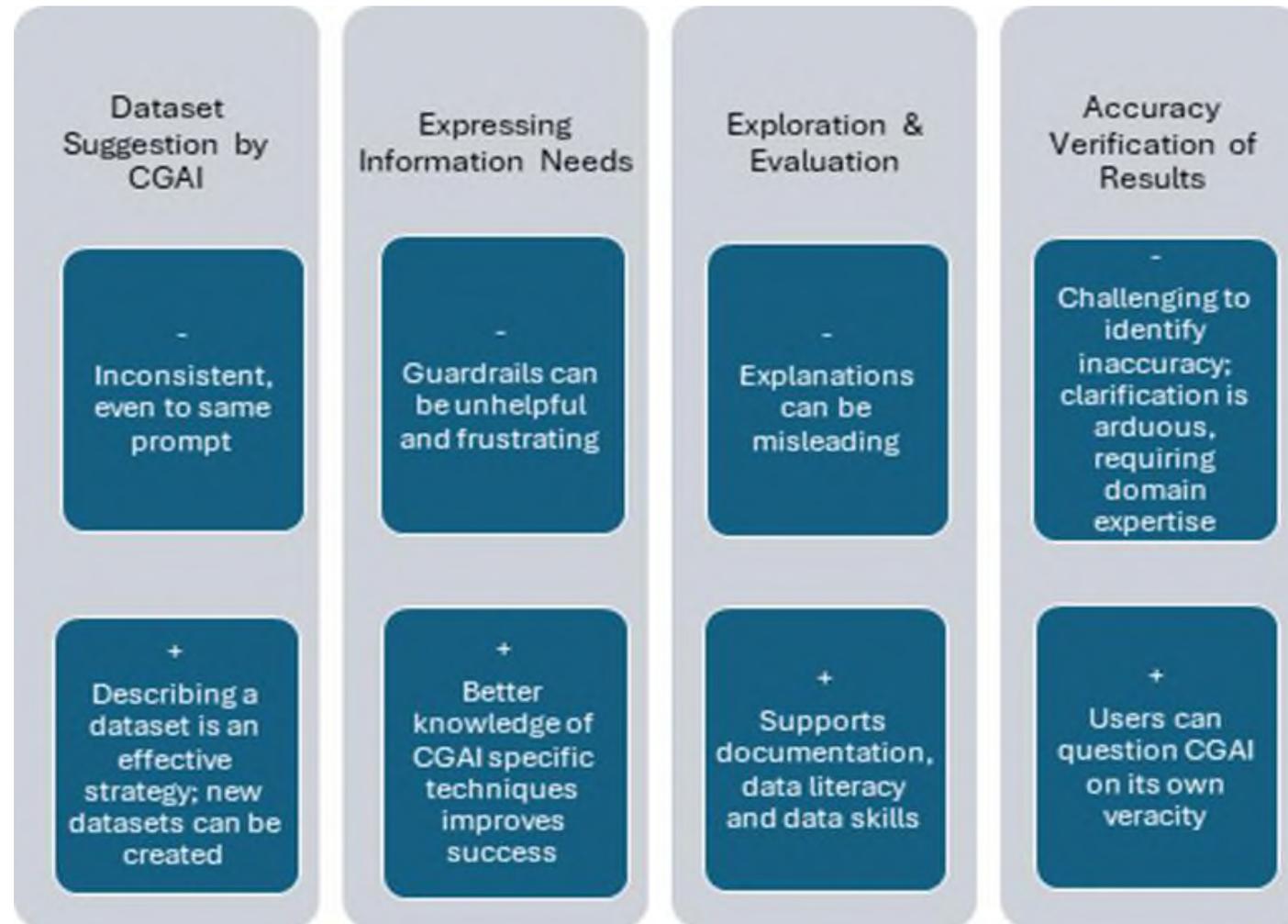
# Framework for dataset discovery with conversational generative AI



# Typology of data experiences

Type	User prompts	CGAI responses
A "request"	A specific dataset is named and asked-for	A single dataset is returned
A "description"	No specific dataset is named, but the query is clearly for a dataset	[No distinguished pattern, non-negligible likelihood of no dataset response]
An "implication"	The query is not explicitly about datasets, although the user wants data	In-kind response, with no datasets.

# Strengths and weaknesses of generative AI



# Conversational capabilities for data portals

- Users struggle to find datasets on open data portals.
- Traditional dataset search capabilities do not capture user intent well.
- Matching queries to datasets is done via metadata, which is often sparse.
- **Themes:**
  1. Improving existing search infrastructure
  2. RAG for dataset discovery
  3. Benchmark datasets

# Less Is More? When Dataset Context Hurts LLM-Generated Dataset Descriptions

Anonymous Author(s)  
Affiliation(s) withheld for double-blind review

Keywords are not always the key: A metadata field analysis for natural language search on open data portals

Lisa-Yao Gan<sup>1,3</sup>[0009-0004-3099-1738], Arunav Das<sup>2</sup>[0009-0008-9989-1718],  
Johanna Walker<sup>2</sup>[0000-0002-5496-8670], and Elcna  
Simperl<sup>2,3</sup>[0000-0003-1722-947X]

<sup>1</sup> Technical University Munich, Arcisstrasse 21, 80333 Munich

<sup>2</sup> King's College London, Strand, WC2R 2LS London

<sup>3</sup> Institute for Advanced Study, Technical University Munich, Lichtenbergstrasse 2a,  
D-85748 Garching, Germany  
lisa.gan@tum.de

**Abstract.** Open data portals are essential for providing public access to open datasets. However, their search interfaces typically rely on keyword-based mechanisms and a narrow set of metadata fields. This design makes it difficult for users to find interesting data. A major problem is provided metadata that often consists of inconsistent, especially when users lack familiarity with domain-specific terminology. In this paper, we examine how individual metadata fields affect the success of conversational dataset retrieval and whether LLMs can help bridge the gap between natural queries and structured metadata. We conduct a controlled ablation study using simulated natural language queries over real-world datasets to evaluate retrieval performance under various metadata configurations. We also compare existing content of the metadata field 'description' with LLM-generated content, exploring how different prompting strategies influence quality and impact on search outcomes. Our findings suggest that dataset descriptions play a central role in aligning with user intent, and that LLM-generated descriptions can support effective retrieval. These results highlight both the limitations of current metadata practices and the potential of generative models to improve dataset discoverability in open data portals.

**Keywords:** Conversational Information Retrieval · Dataset Discovery · Conversational Search

## 1 Introduction

Open data portals play a crucial role in promoting transparency, civic engagement, and evidence-based policymaking by providing public access to government and institutional datasets [31, 16]. Despite the increasing availability of such data, users often struggle to find datasets that match their information needs [18, 23]. Existing portals primarily support keyword-based search mechanisms that rely on exact term matching and predefined metadata fields such as

**Abstract—**Dataset search and reuse are strongly constrained by the quality of metadata such as natural language descriptions, which are often sparse or inconsistent. Although large language models (LLMs) can generate such descriptions automatically, little empirical guidance exists on what makes a good dataset description and what dataset context LLMs actually need. We study these questions through a literature-grounded framework of dataset description quality and a large-scale ablation study using 252 datasets (1,336 CSV files) from the European data portal data.europa.eu. We generate descriptions with LLMs in a baseline scenario and two ablation scenarios: (1) using only dataset titles, (2) titles and schema, and (3) titles, schema and representative data, and evaluate them with an LLM-as-a-judge framework and a semantic descriptive attribute analysis grounded in our quality dimensions. Our results reveal a consistent schema penalty: title schemas alone often degrade narrative quality, while representative data partially restores grounding without impacting overall human-facing quality. Our results show that in general LLMs exhibit stable, correctly performed. These findings provide practical guidance for LLM-supported data publishing workflows.

**Index Terms—**Dataset discovery, dataset descriptions, metadata generation, large language models, open data, dataset search

Motivated by these persistent problems, recent research has begun to explore the use of large language models (LLMs) to automate dataset documentation and description generation [10]. These efforts demonstrate that LLMs can generate coherent natural-language summaries and substantially expand metadata coverage. However, from the perspective of data publishers and portal operators, a critical practical question remains unanswered: *what information is actually necessary to provide to an LLM in order to reliably generate high-quality dataset descriptions?* Dataset providers often lack the time and incentives to curate rich metadata, motivating workflows where a dataset title alone could yield a useful description. While additional signals such as schema or data samples may help, they also introduce noise: too little context risks vague summaries, while too much may overwhelm the model. There is little empirical guidance in the literature on what "sweet spot" lies between insufficient and excessively rich dataset context.

In this work, we address this gap through two complementary research questions:

- **RQ1:** To what extent do LLM-generated dataset descriptions meet the characteristics of high-quality dataset descriptions?
- **RQ2:** How does the quality of LLM-generated dataset descriptions vary under different dataset-context prompting conditions?

To answer **RQ1**, we synthesize prior research on dataset discovery, metadata quality, and data sensemaking to derive a structured characterization of high-quality dataset descriptions. To answer **RQ2**, we conduct a large-scale ablation study, examining how description quality changes as progressively richer dataset context is provided. We evaluate generated descriptions using both quality scoring and semantic descriptive analysis.

Our work makes three primary contributions:

- **Literature-grounded characterization.** We consolidate prior research into a structured framework of what constitutes a high-quality dataset description.
- **Ablation study of LLM-based description generation.** We provide an empirical analysis of how description quality changes as increasingly rich dataset signals are provided to an LLM.
- **Practical guidance for data publishers.** We derive empirically grounded insights into what dataset information is most valuable to provide when using LLMs to automatically generate dataset descriptions.

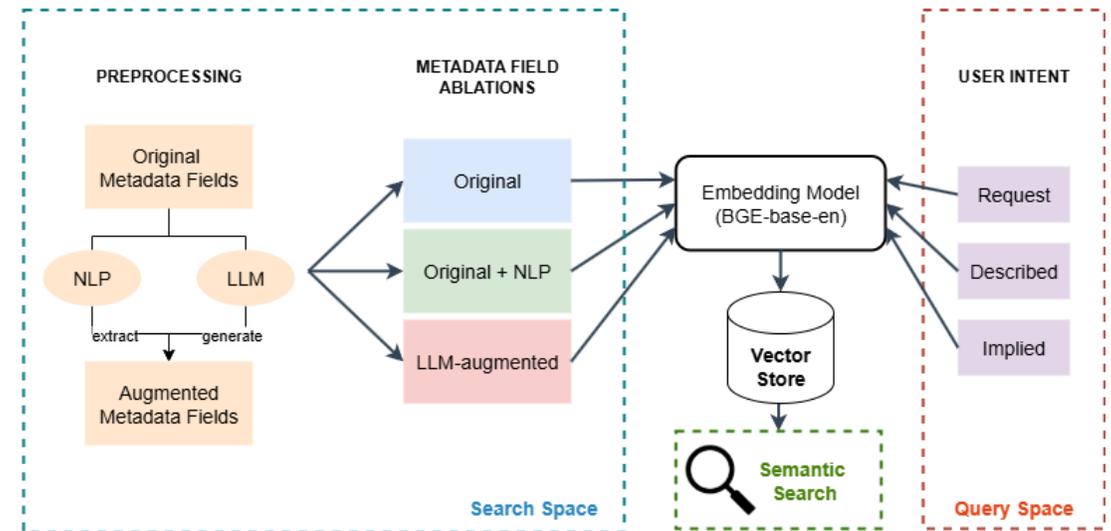
## 1. INTRODUCTION

The ability to identify and locate appropriate data is fundamental to its reuse. Portals are a key way in which data publishers have facilitated this. Platforms such as data.europa.eu, national open government portals, and institutional repositories host millions of datasets [1]. And yet users frequently report difficulties both in locating data that match their information needs and in understanding the datasets they encounter. Prior work consistently shows that these difficulties are fundamentally constrained by the quality of dataset metadata are often sparse, inconsistent, or poorly written [2]–[4]. Common metadata for datasets include the title, tags and descriptions. As portal search is primarily keyword based, poorly-written descriptions represent a missed opportunity to match the needs of users. At the same time, empirical studies of dataset discovery and reuse consistently show that natural-language overview descriptions are among the most important metadata elements supporting user sensemaking, relevance assessment, and reuse decisions [3], [5]–[8]. As a result, users struggle not only to retrieve relevant datasets, but also to interpret their contents, assess fitness for use, and build trust in unfamiliar data [6], [9].

Together, this literature establishes (the lack of) dataset description quality as a central bottleneck in data discovery and documentation.

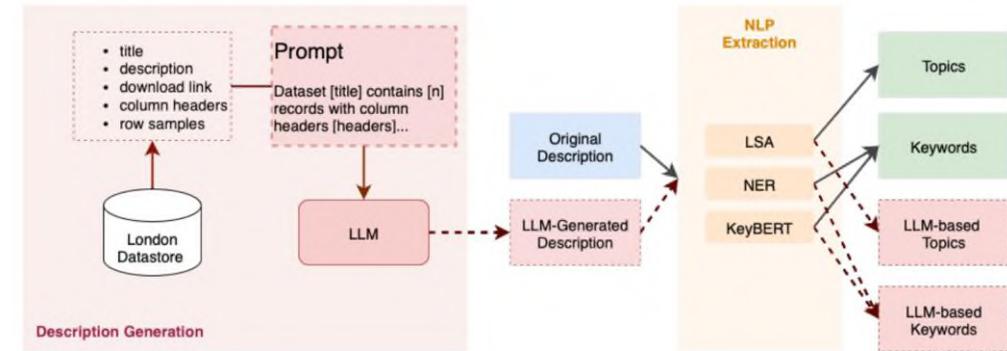
# Using generative AI for better search

- LLM-generated descriptions using metadata and column/row samples of dataset.
- Evaluated on 255 CSV-based datasets from the **London Datastore**.
- **User queries** generated by Gemini2.5 Flash informed by research into data prompting.
- Prompting strategies for **three user intents** surfaced from user research.



# Ablation study

- Compared topic, keywords and descriptions. **Descriptions** are the most effective metadata field for supporting natural language queries.
- LLM-generated descriptions improve retrieval by **~20%** (Hit@1) compared to publisher-authored ones.
- **Request** queries work best, **imply** ones need good metadata.

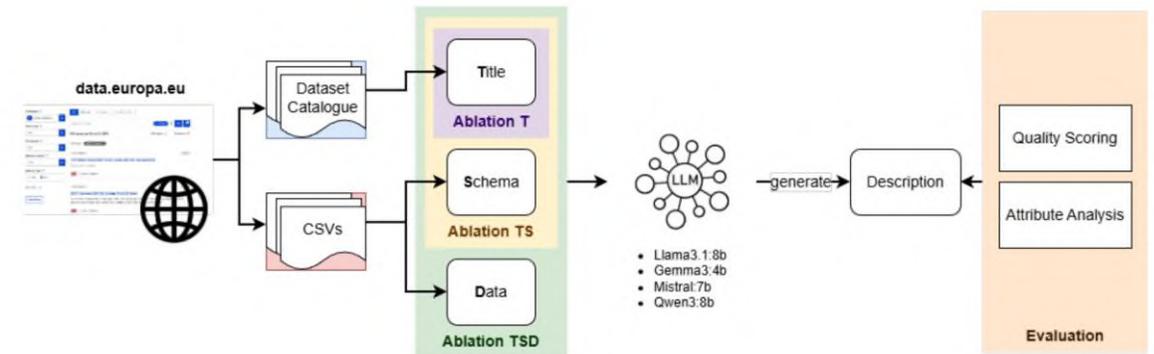


Ablation Condition	Hit@1	Hit@3	Hit@5	MRR
KEY_ORIGINAL	0.279	0.471	0.544	0.379
KEY_NLP	0.502	0.664	0.757	0.594
KEY_LLM	0.495	0.680	0.753	0.594
DESC_ORIGINAL	<b>0.731</b>	<b>0.915</b>	<b>0.944</b>	<b>0.820</b>
DESC_LLM	<b>0.887</b>	<b>0.964</b>	<b>0.976</b>	<b>0.925</b>
FULL_ORIGINAL	0.744	0.921	0.955	0.833
FULL_NLP	0.684	0.874	0.916	0.778
FULL_LLM	0.835	0.940	0.956	0.887
ONLYKEY_ORIGINAL	0.293	0.446	0.496	0.371
ONLYKEY_NLP	0.492	0.676	0.756	0.592
ONLYKEY_LLM	0.499	0.669	0.759	0.597
ONLYTOPIC_ORIGINAL	0.057	0.135	0.177	0.098
ONLYTOPIC_NLP	0.001	0.012	0.024	0.008
ONLYTOPIC_LLM	0.009	0.027	0.038	0.019

Ablation	Requesting	Describing	Implying
KEY_ORIGINAL	0.394 / 0.295	0.392 / 0.295	0.351 / 0.246
DESC_LLM	<b>0.962 / 0.943</b>	<b>0.907 / 0.861</b>	<b>0.906 / 0.858</b>
FULL_LLM	0.927 / 0.890	0.862 / 0.804	0.872 / 0.811
ONLYTOPIC_LLM	0.024 / 0.014	0.017 / 0.007	0.017 / 0.007

# Improving LLM-generated metadata

- 252 datasets (1,336 CSV files) from data.europa.eu.
- Three metadata-generation scenarios using
  - Only dataset titles;
  - Titles and schema;
  - Titles, schema, example data.
- Evaluation using LLM-as-a-judge and a semantic descriptive attribute analysis grounded in human-data interaction literature.



Characteristic	What the description should convey	Example cues (signals)
Overview & purpose	Plain-language summary of what the dataset is about and why it exists (intended analytical or policy purpose).	Topic/domain, phenomenon, intended use or goal
Contents & coverage	What variables/fields the dataset contains and what they represent, including spatial and temporal scope and relevant granularity.	Key attributes, units, geography, time range, resolution
Structure & size	How the dataset is organized and delivered, including format and basic scale/complexity.	File/API type, rows/records, columns/attributes, multiple files, nested structures
Provenance & updates	Who produced the data and how current it is, including publication date and update frequency.	Publisher/source, collection process, last updated, update schedule
Quality & limitations	Known issues, uncertainty, and methodological caveats that affect interpretation and reuse.	Missingness, bias, measurement changes, known errors, comparability notes
Usage notes & insights	How the dataset can be used and what it may reveal, including suggested use cases or notable patterns.	Example analyses, intended applications, notable trends/anomalies
Clarity & plain language	Accessible writing that avoids unexplained jargon, acronyms, or insider terminology.	Definitions, expanded acronyms, simple phrasing, self-contained explanation
User vocabulary alignment	Terms that match how users search, including synonyms and related phrases to reduce vocabulary mismatch.	Common query terms, synonyms, abbreviations, alternative names

# Evaluation results

- Schemas alone often degrade narrative quality.
- Example data partially restores grounding without improving human-facing quality.
- Different LLMs exhibit **stable descriptive personas**.

Model	Schema Effect	Data Effect
LLaMA-3	-0.073	+0.031
Qwen	-0.221	+0.022
Mistral	-0.158	+0.060
Gemma	-0.136	+0.115

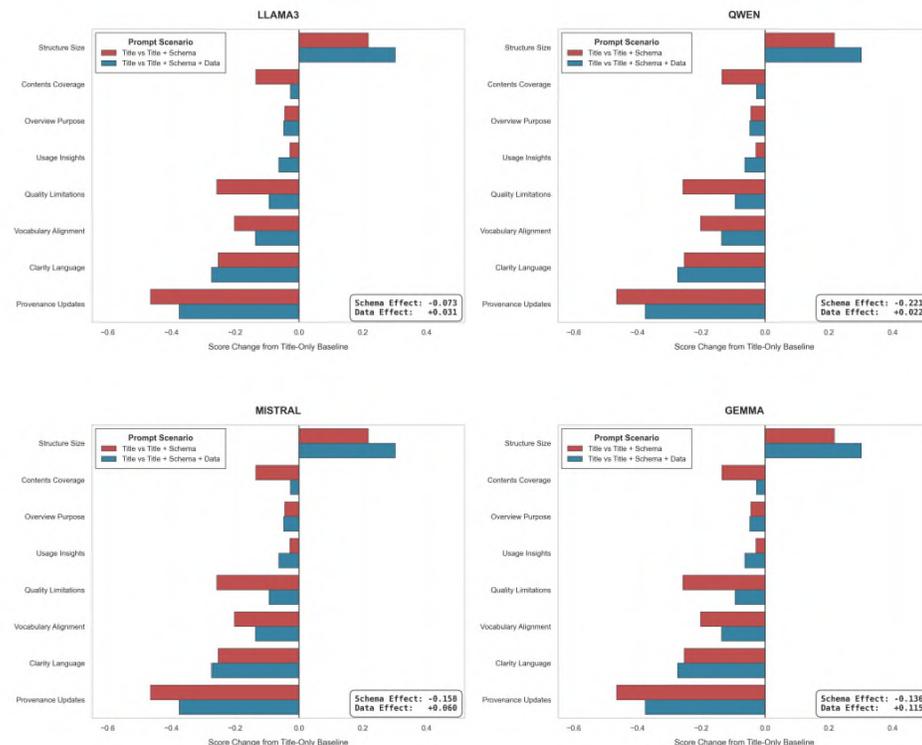


Fig. 4. Characteristic-level quality changes relative to the Title-only baseline under different prompt conditions, evaluated using an LLM-as-a-judge framework. Red bars show the effect of adding schema (Title+Schema), while blue bars show the marginal effect of adding representative data samples (Title+Schema+Data). Across all models, adding schema introduces a consistent penalty, particularly for Overview & Purpose, Clarity & Plain Language, and Provenance & Updates. Adding data partially recovers quality, with the strongest gains observed for Structure & Size.

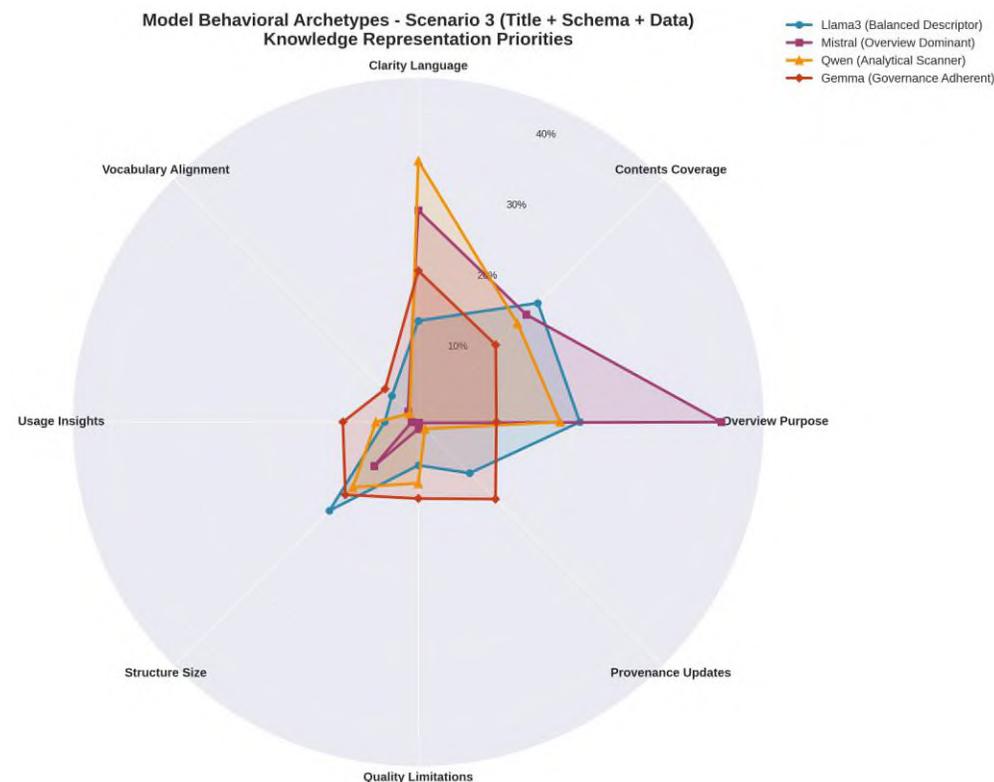
# Evaluation results

- Schemas alone often degrade narrative quality.
- Example data partially restores grounding without improving human-facing quality.
- Different LLMs exhibit **stable descriptive personas**.

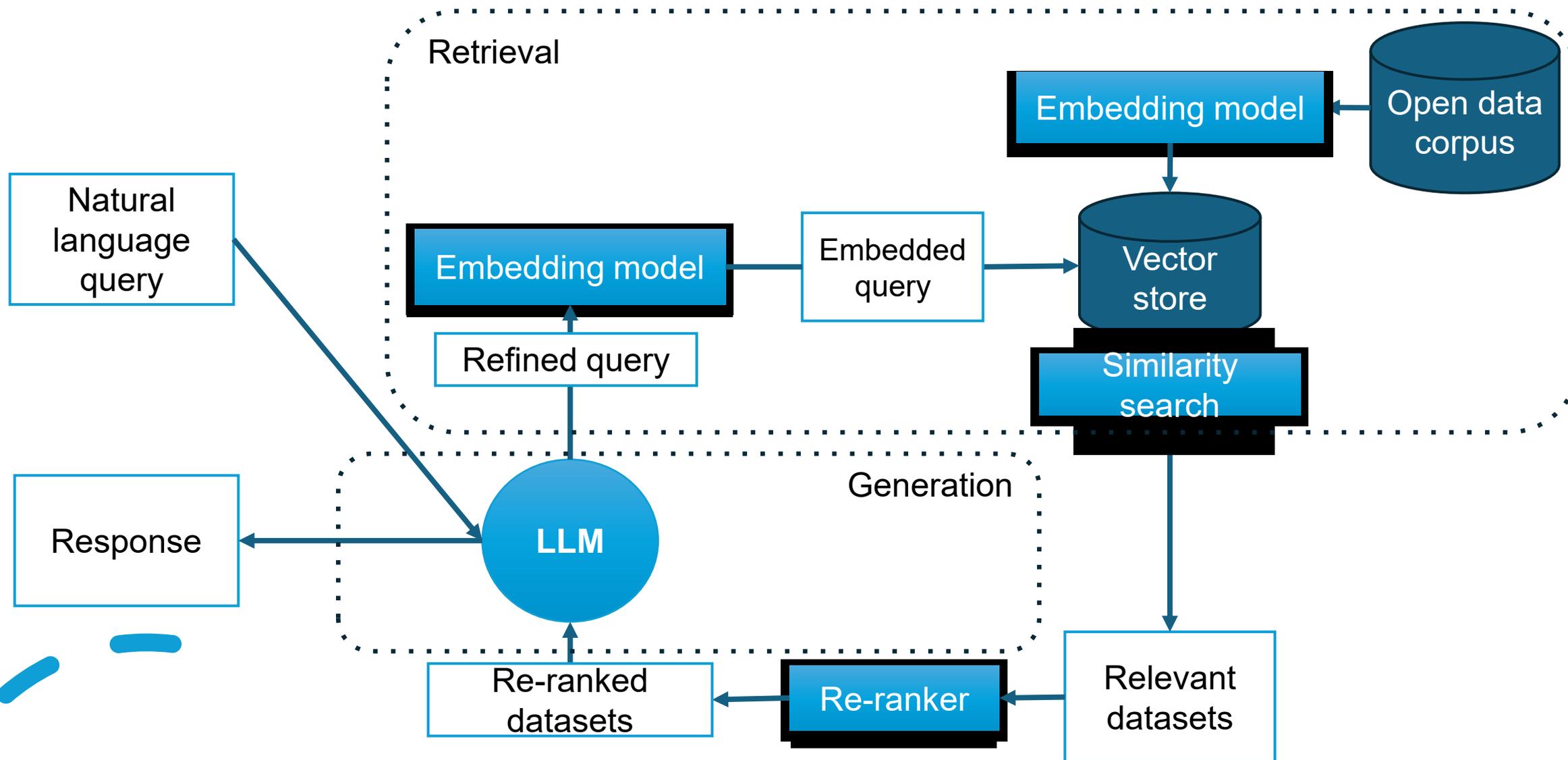
**Title-only (T):** “This dataset provides insights into the relationship between financial capability and child poverty... It offers a comprehensive understanding of the socio-economic factors influencing the well-being of children from low-income households...”

**Title + Schema (TS):** “The dataset consists of two tables... Financial Capability Indicators at the LSOA level... and Geographic Reference Data at the postcode and LSOA levels, including location coordinates and administrative codes...”

**Title + Schema + Data (TSD):** “The dataset is comprised of two tables... GFA\_PT0\_RECS to GFA\_PT5\_RECS represent financial participation categories... allowing analysts to identify areas requiring targeted support...”

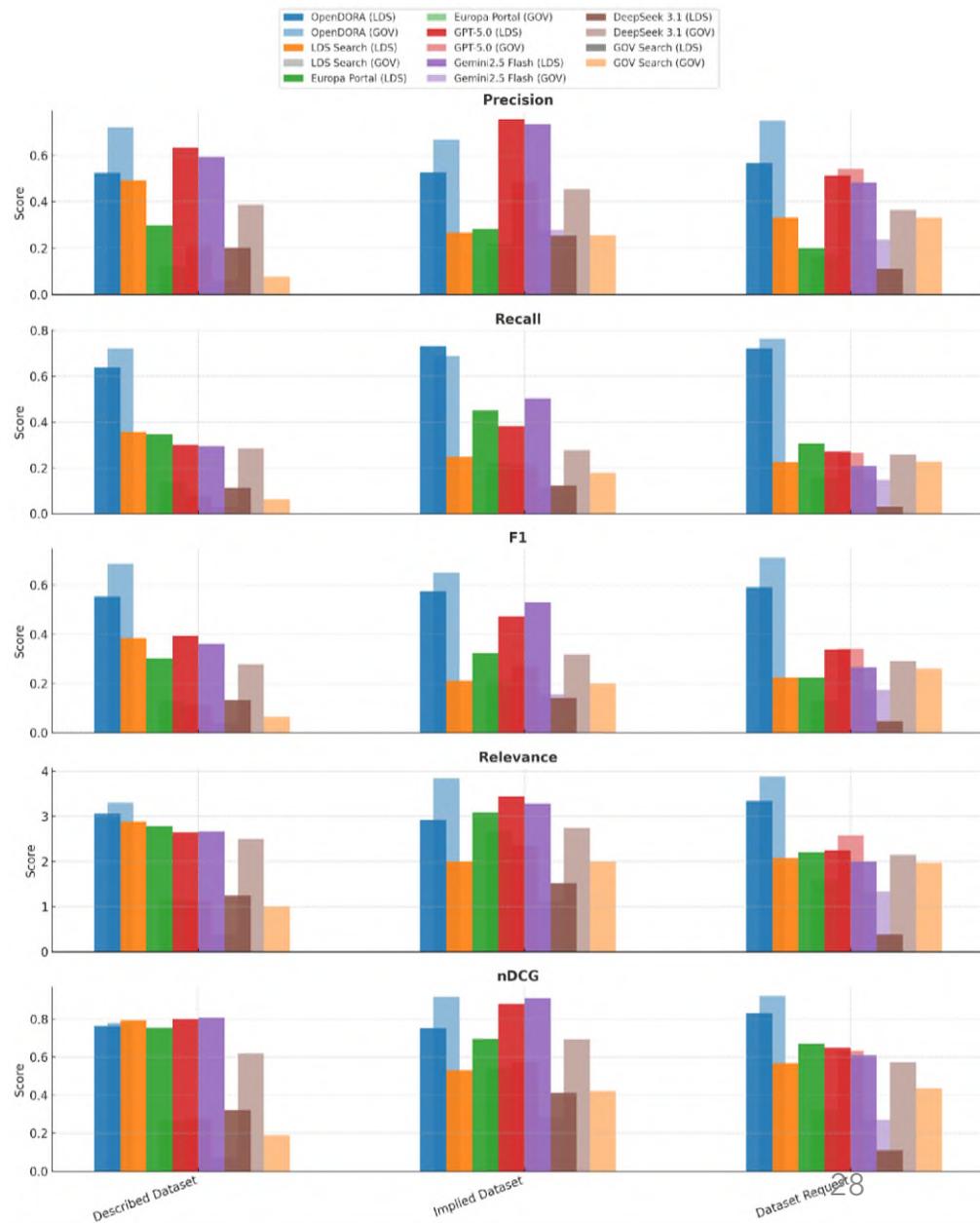


# OpenDORA: RAG for dataset discovery



# Evaluation results

- OpenDORA (Blue) consistently outperforms Gemini2.5, DeepSeek and GPT5.0 on recall, and all native open data portal searches (by 24% - 153%)
- LLMs still record high amount of evaluation frequency
  - GPT5.0: 30.8%
  - Gemini2.5 Flash: 38.1%
  - DeepSeek 3.1: 63.4 %



# Benchmark dataset (London Datastore)

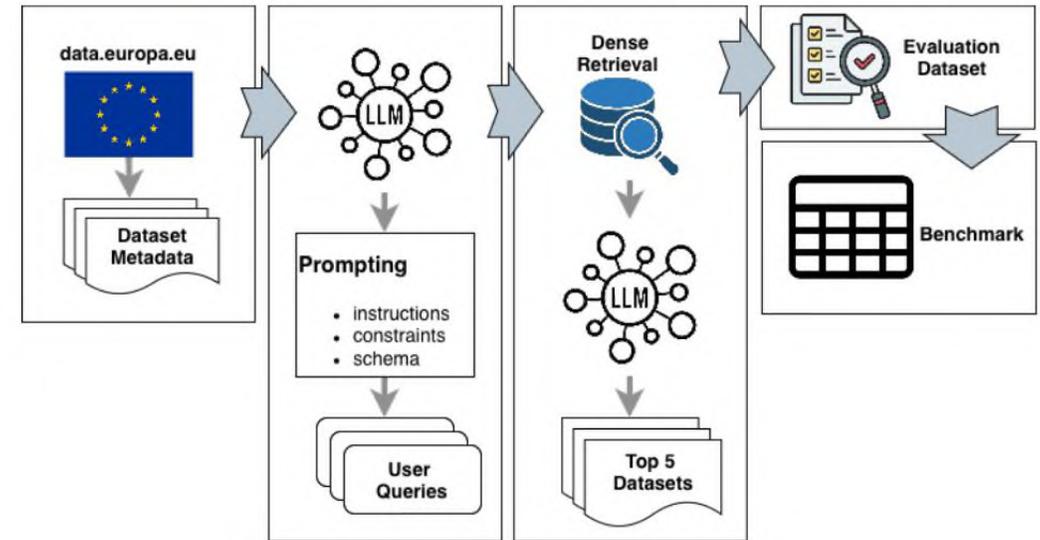
- Small benchmark based on London Datastore.
- **40 topics** covering the entire variety of datasets available on the portal.
- **400+ user queries** with 3-5 user utterances per topic.
- **6-10 turns per topic** based on Walker et al.'s prompting styles.
- **Human-annotated** ground truth to ensure validity and accuracy.

Query type	Turn ID	Questions	Ground truth	Context_turn (topic id - turn id)
Described Dataset	1	Can you find me a dataset with the average property prices in London?	Sure, below are the datasets which include average property prices of London: ...	1-0
Source Request	2	Could you provide sources to these datasets?	Here are the sources...	1-1
Additional Dataset Information	3	How often are these datasets updated?	...	1-2
Does this dataset include additional details such as property types?	4	Does this dataset include additional details such as property types (flats, houses, etc.)?	...	1-3

Example: property prices

# Benchmark dataset (data.europa.eu)

- Open-weights LLMs to achieve scale
  - 7k datasets, 20k queries.
  - Qwen3-8b conditioned on dataset metadata to generate dataset
  - Describe-style and imply-style queries.



Query	Relevant datasets (titles, abbreviated)
Looking for a dataset with national beef kill figures from 2023 – preferably downloadable as CSV or Excel?	National Beef Kill Figures 2023; The National beef kill figures; Livestock by municipality and type of animal (1981–2007)
Where can I find reliable weekly stats on how much beef is being processed nationally?	National Beef Kill Figures 2023; The National beef kill figures; Enterprises – Agriculture and forestry – Animal production

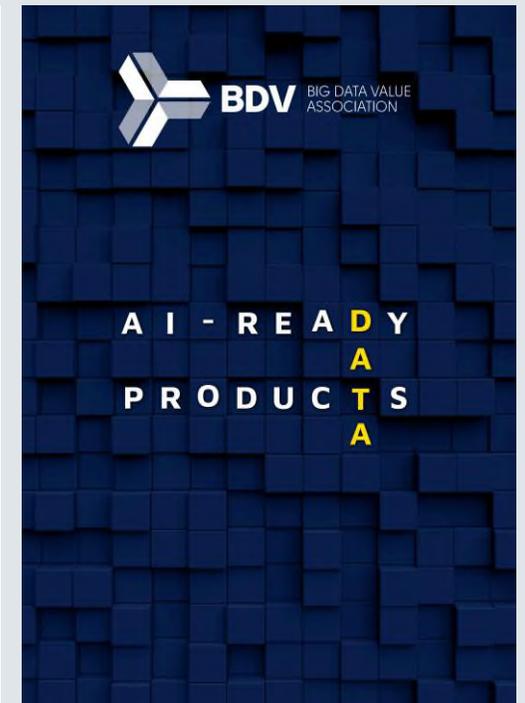
# Evaluation results

Model	P@3	R@3	MRR	nDCG@3
<b>Classical Retrievers</b>				
BM25	0.362	0.293	0.469	0.541
all-MiniLM-L6-v2	0.313	0.252	0.423	0.480
Qwen + cosine	0.362	0.293	0.471	0.538
E5	0.448	0.356	0.553	0.633
<b>Retrieval-free LLM baselines (query-only)</b>				
<i>Open-weights Models</i>				
LLaMA-8B	0.001	0.001	0.004	0.004
LLaMA-70B	0.004	0.003	0.009	0.009
Qwen-7B	0.001	0.001	0.003	0.003
Qwen-72B	0.002	0.002	0.005	0.005
Qwen-80B	0.001	0.001	0.002	0.002
<i>Closed-source Models</i>				
GPT-4.1	0.002	0.001	0.005	0.005
GPT-4.1 mini	0.000	0.000	0.001	0.001
GPT-4.1 nano	0.001	0.001	0.002	0.002
GPT-5.2	0.003	0.003	0.007	0.008
<b>Retrieval-augmented systems</b>				
RAG(BM25) + LLaMA-70B	0.389	0.312	0.718	0.723
RAG(E5) + LLaMA-70B	<b>0.482</b>	<b>0.381</b>	<b>0.817</b>	<b>0.824</b>

- Retrieval-free models do not work for real dataset search tasks.
- RAG achieves better performance, provided good metadata.
- LLM-based reasoning helps as well, but overall, this remains an open challenge.

# AI-ready data: recommendations for publishers

Category	Criteria	Sub-criteria	
1) Dataset Properties	a) Following international standards and norms		
	b) Semantic and logical consistency across entries		
	c) Identifiable class and source imbalance		
	d) De-identification and Anonymisation where necessary		
	e) Appropriate file format		
2) Metadata	a) Machine-readable metadata format		
	b) The dataset served to users with attached metadata		
	c) Basic technical specifications	i) Modalities	
		ii) Dimensionality	
		iii) Update frequency	
		iv) Semantics	
		v) Bias	
		vi) Basic summary statistics	
		vii) Synthetic data	
	d) Supply chain information	i) Collection	
		ii) Preprocessing	
		iii) Annotation	
	e) Legal and sociotechnical information	i) Context and original purpose	
ii) Licence name(s), permissions and URL(s)			
iii) Intended access controls			
iv) Data protection declaration(s)			
3) Surrounding Infrastructure	a) Accessibility via a user-centric data portal		
	b) Accessibility via API		
	c) Version control and monitoring infrastructure		
4) Governance	a) Governance policy-as-code		
	b) Documented roles and responsibilities		
	c) Publicly identifiable points of contact		
	d) Clear data access processes		



Government data publishers can help data-discovery tools by providing factual, real-life open data points to train on and utilise in answers to user queries.

At the ODI we devise AI experimentation protocols to derive actionable recommendations for government data publishers.



arXiv > cs > arXiv:2412.09632

Computer Science > Computers and Society

[Submitted on 27 Nov 2024 (v1), last revised 18 Dec 2024 (this version, v2)]

### Methods to Assess the UK Government's Current Role as a Data Provider for AI

Neil Majithia, Elena Simperi

Governments typically collect and steward a vast amount of high-quality data on their citizens and institutions, and the UK government is exploring how it can better publish and provision this data to the benefit of the AI landscape. However, the compositions of generative AI training corpora remain closely guarded secrets, making the planning of data sharing initiatives difficult. To address this, we devise two methods to assess UK government data usage for the training of Large Language Models (LLMs) and 'peek behind the curtain' in order to observe the UK government's current contributions as a data provider for AI. The first method, an ablation study that utilises LLM 'unlearning', seeks to examine the importance of the information held on UK government websites for LLMs and their performance in citizen query tasks. The second method, an information leakage study, seeks to ascertain whether LLMs are aware of the information held in the datasets published on the UK government's open data initiative data.gov.uk. Our findings indicate that UK government websites are important data sources for AI (heterogenously across subject matters) while data.gov.uk is not. This paper serves as a technical report, explaining in-depth the designs, mechanics, and limitations of the above experiments. It is accompanied by a complementary non-technical report on the ODI website in which we summarise the experiments and key findings, interpret them, and build a set of actionable recommendations for the UK government to take forward as it seeks to design AI policy. While we focus on UK open government data, we believe that the methods introduced in this paper present a reproducible approach to tackle the opaqueness of AI training corpora and provide organisations a framework to evaluate and maximize their contributions to AI development.

Comments: 17 pages, 5 figures; v2 - incorporated editor feedback; for the accompanying, non-technical ODI report see this <https://odi.org>  
Subjects: **Computers and Society (cs.CV)**, Artificial Intelligence (cs.AI), Information Retrieval (cs.IR)  
Cite as: arXiv:2412.09632 [cs.CV]  
(or arXiv:2412.09632v2 [cs.CV] for this version)  
<https://doi.org/10.48550/arXiv.2412.09632>

Submission history



## The UK government as a data provider for AI

ODI Research  
ADVANCING TRUST IN DATA



November 2024

# Experiment 1: ‘Unlearning’ the content of government websites

1. Benchmark LLMs on a set of ‘citizen queries’
2. Use an ablation methodology to make LLMs forget the contents of government websites
3. Benchmark ablated LLMs on set of queries, compare errors and hallucinations

IF no change THEN the government websites were likely not in the training corpora of the models MEANING government websites need to improve AI access

IF hallucinations ↑ THEN government websites were meaningful components of training corpora MEANING infrastructure is sufficient

Test case question	Baseline (pre-ablation)	Post-ablation
‘I’m a single parent living in England with two children under 16. I’ve recently started a part-time job, but my income is still quite low. Can I claim Child Benefit, and how might it affect my other benefits?’	‘...If you claim Child benefit, it might affect other benefits you’re receiving. For example, if you claim Universal Credit, the amount you receive will be reduced by the amount of Child Benefit you’re eligible for. This is because Child Benefit is considered part of your income when calculating your Universal Credit award...’	‘...Regarding how it affects other benefits, Child Benefit is usually paid in addition to other benefits you may receive. However, if you’re receiving certain benefits, such as Income-based Jobseeker’s Allowance or Income-related Employment and Support Allowance, your Child Benefit may be affected...’

*The baseline model correctly identifies the interaction between Universal Credit and Child Benefit*

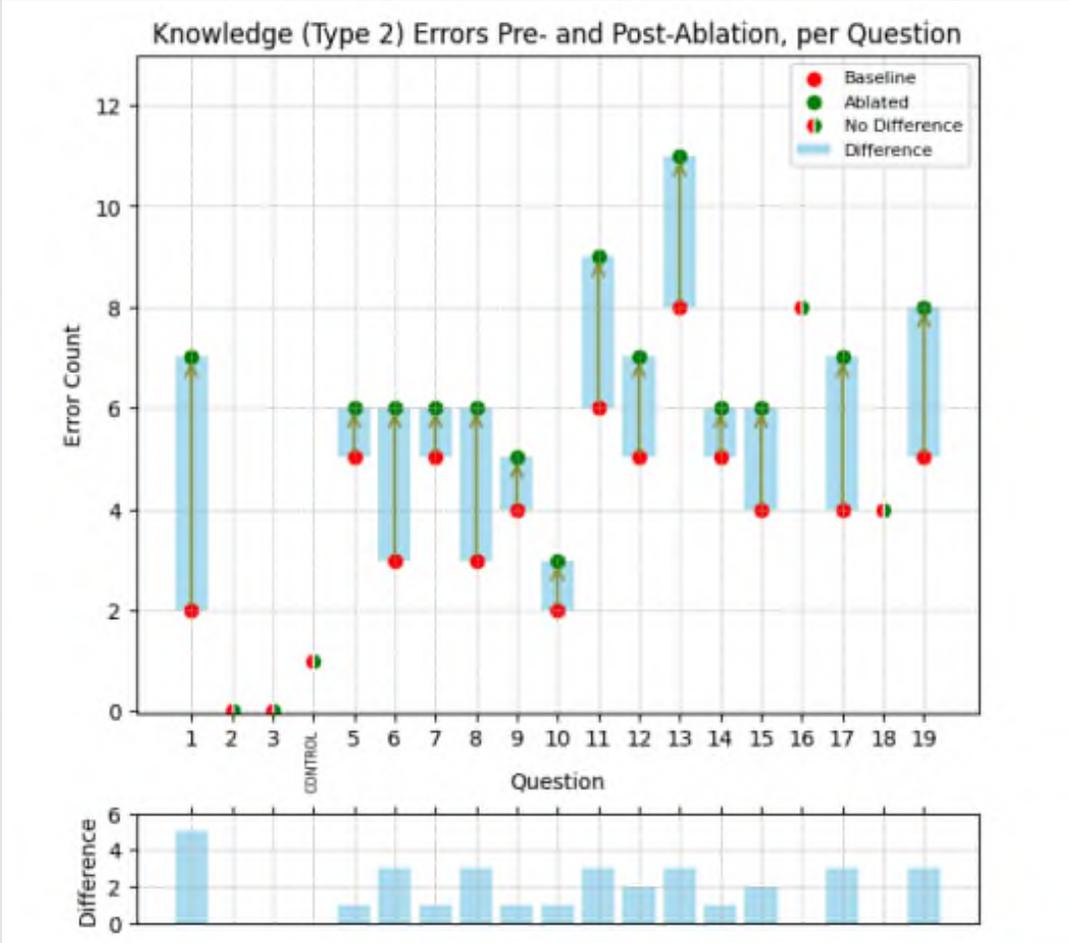
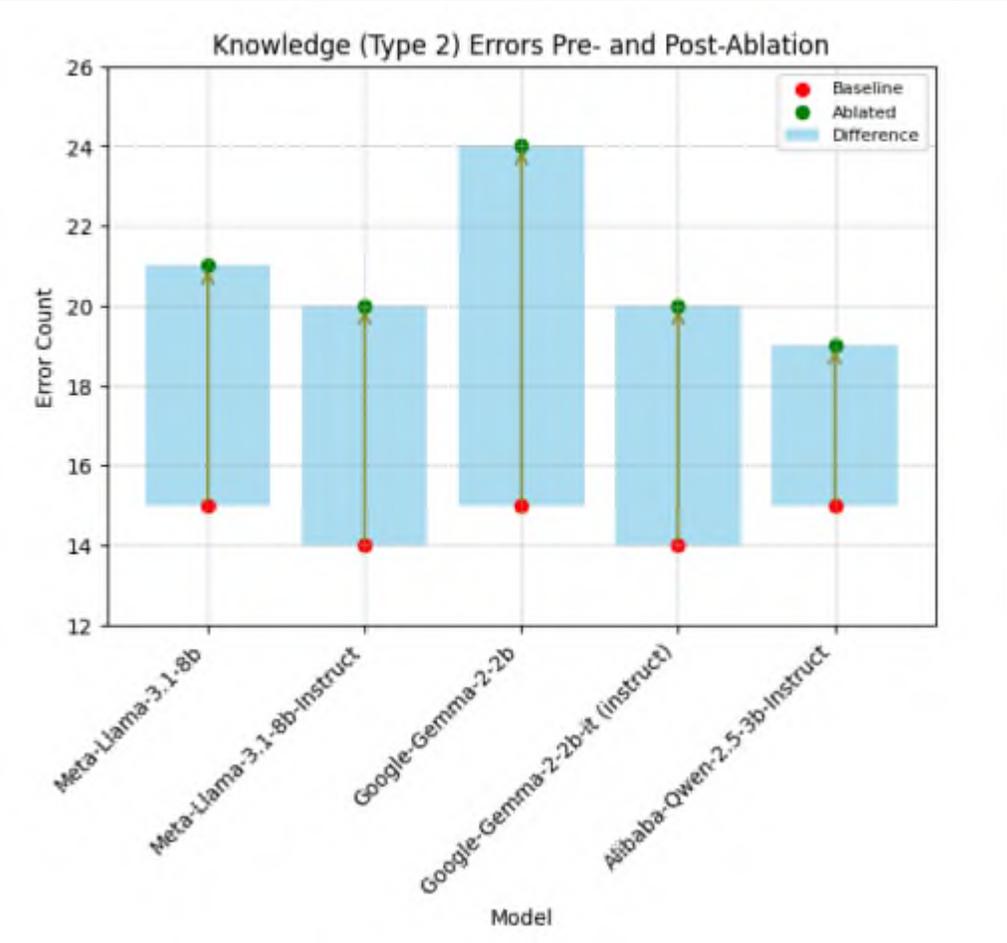
*The ablated model provides incorrect information, a hallucinated interaction between Child Benefit and the - defunct - income-based benefits that Universal Credit replaces*



# Experiment 1: 'Unlearning' the content of government websites

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

Key result 2: This importance depends on subject matter: more niche = more reliance on government information



# Experiment 2: Recalling data points from datasets on data.gov.uk

- 1. Refine k-shot prompts that force LLMs to 'leak' numerical information from their training corpora
- 2. Use these prompts to test LLM recall of different datasets that are hosted on data.gov.uk

IF successful recall THEN data.gov.uk datasets inform the knowledge bases of LLMs MEANING data.gov.uk infrastructure is sufficient

IF unsuccessful recall THEN the information on government datasets is not known to LLMs MEANING infrastructure is insufficient

Dataset full name	Dataset abbrev.
(control) Official Bank of England Bank Rate	BOE
(control) United Kingdom population mid-year estimate	POP
Percentage of households owning a working smoke alarm in England	HSA
Number of stop and searches carried out, rate per 1000 Black people in the UK	SAS
Number of non-fatal injuries to the workforce on the UK mainline rail network	IBR
Number of attributable deaths to PM2.5 concentration, assuming 6% mortality coefficient	POL
Agricultural Price Index for all agricultural inputs	API

4 prompting templates, each applied with 0-, 1-, and 5-shot prompting, for non-instruction-tuned models

1 system prompt-instruction prompt pair for instruction-tuned models



# Experiment 2: Recalling data points from datasets on data.gov.uk

	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	✗✗✗✗	✗✗✗✗	✗✗✗✗	★	✗✗✗✗	✗✗✗✗	✗✗✗✗	★	✗✗✗✗	✗✗✗✗	✗✗✗✗	★

- Tick = successful, Cross = unsuccessful, Star = reticent response
- Note poor performance in controls
- Only 5 out of 195 attempts resulted in successful recall of data.gov.uk data

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



# Conclusion

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

**Key result 2:** This importance depends on subject matter: more niche topics have more reliance on government information.

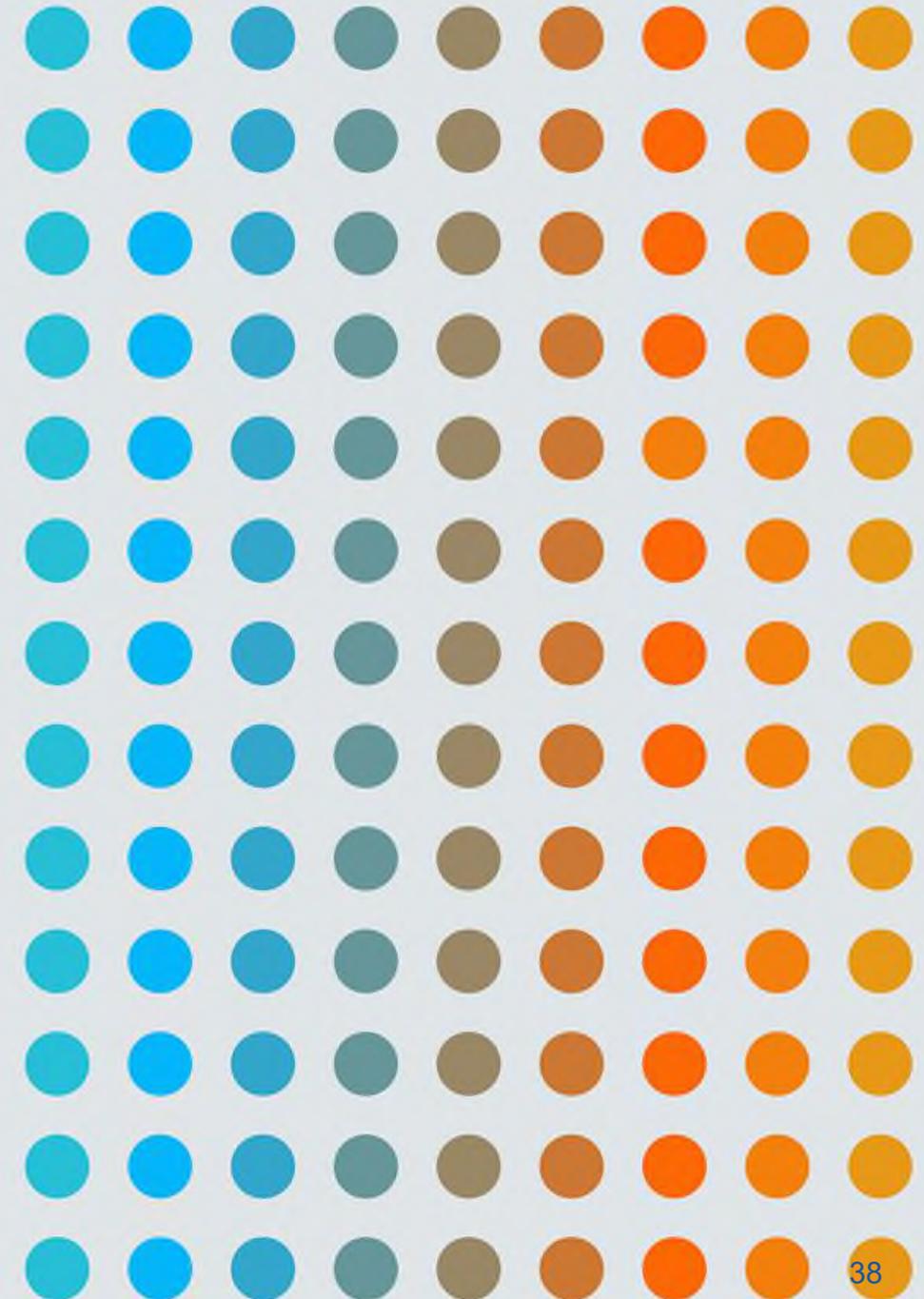
**Key result 3:** data.gov.uk is not a data provider for AI.

So, while government websites have sufficient infrastructure that enables them to be strong data providers for AI, data.gov.uk simply does not.

data.gov.uk is the crown jewel of the UK's open data initiative, so what needs to be done to improve?

# Recommendations

1. **Continue to make data openly available, and ensure it is AI-ready**
2. **Revise data reuse policies, permissions, and robots.txt**
3. **Develop an *AI-ready* National Data Library**
4. **Equip existing data access and sharing infrastructure with AI capabilities**
5. **Invest in high-quality benchmarks and evaluation protocols**



I live in Wandsworth. I have five bags of grass cuttings, old bush and other general garden waste. How do I arrange for someone to collect it?

Here's how you can **dispose of or arrange a collection of your garden waste in Wandsworth (London, UK)**:

#### 🌱 1. Put Out Up To Five Bags With Normal Weekly Waste

- Wandsworth Council **will collect up to five bags of compostable garden waste (grass cuttings, small bush trimmings, etc.)** on your normal weekly refuse collection day if placed beside your bins at the kerb. Wandsworth Bor...
- **No special garden waste sacks are required** — any strong tied refuse sack is fine (but **not council recycling sacks**). Wandsworth Bor...
- Make sure the bags are clearly visible at **↓** your property boundary by 6:30 am on the collection day (or after 6:30 pm the evening before). Wandsworth Bor...

# Benchmarking LLMs for citizen queries and other public service delivery



GOV.UK

# The best place to find government services and information

Search

Popular on GOV.UK

- [HMRC account: sign in or set up](#)
- [eVisas: access and use your online immigration status](#)
- [Universal Credit account: sign in](#)
- [Personal tax account: sign in or set up](#)
- [Childcare account: sign in](#)
- [Check your State Pension forecast](#)

how do I apply for an apprenticeship

AI Overview

To apply for an apprenticeship in the UK, search for vacancies on the [Find an apprenticeship site](#) (England), or directly on company websites and job boards. Prepare a CV and covering letter, submit applications through the employer's portal or government site, and prepare for interviews, which often include assessments. [GOV.UK](#)

Here is a step-by-step guide to applying for an apprenticeship:

### 1. Find and Research Opportunities

- **Use Official Services:** Use the Find an apprenticeship service to search for roles by keyword, location, or employer.
- **Direct Applications:** Check websites of companies you are interested in, as many advertise directly.
- **Networking:** Follow companies on social media, attend career fairs, and use LinkedIn to find opportunities. [GOV.UK](#)

### 2. Prepare Your Application

- **Create a CV:** Highlight your education, skills, voluntary work, and any relevant experience.
- **Write a Cover Letter:** Explain why you want the apprenticeship, what skills you can bring, and your long-term career goals.
- **Tailor Applications:** Customize your CV and cover letter for each specific role.

How to find and apply for an apprenticeship  
31 Jan 2024 — and I was a bit of a...

How to Apply for Degree Apprenticeships UK: Step-by-St...  
7 Sept 2024 — Hi I'm Joe, so by now...

Become an apprentice: Apply for an apprenticeship - GOV.UK  
Search for an apprenticeship. Sign in or create an account. Complete and submit your application. The National Careers...

Show all

73% of people living in the UK have used AI in their day-to-day life in the past month.

Can I drive without insurance?

How do I report a racist school teacher?

Is my employer allowed to take my tips?

Is the MMR vaccine safe?

Where can I find volunteering activities near me?

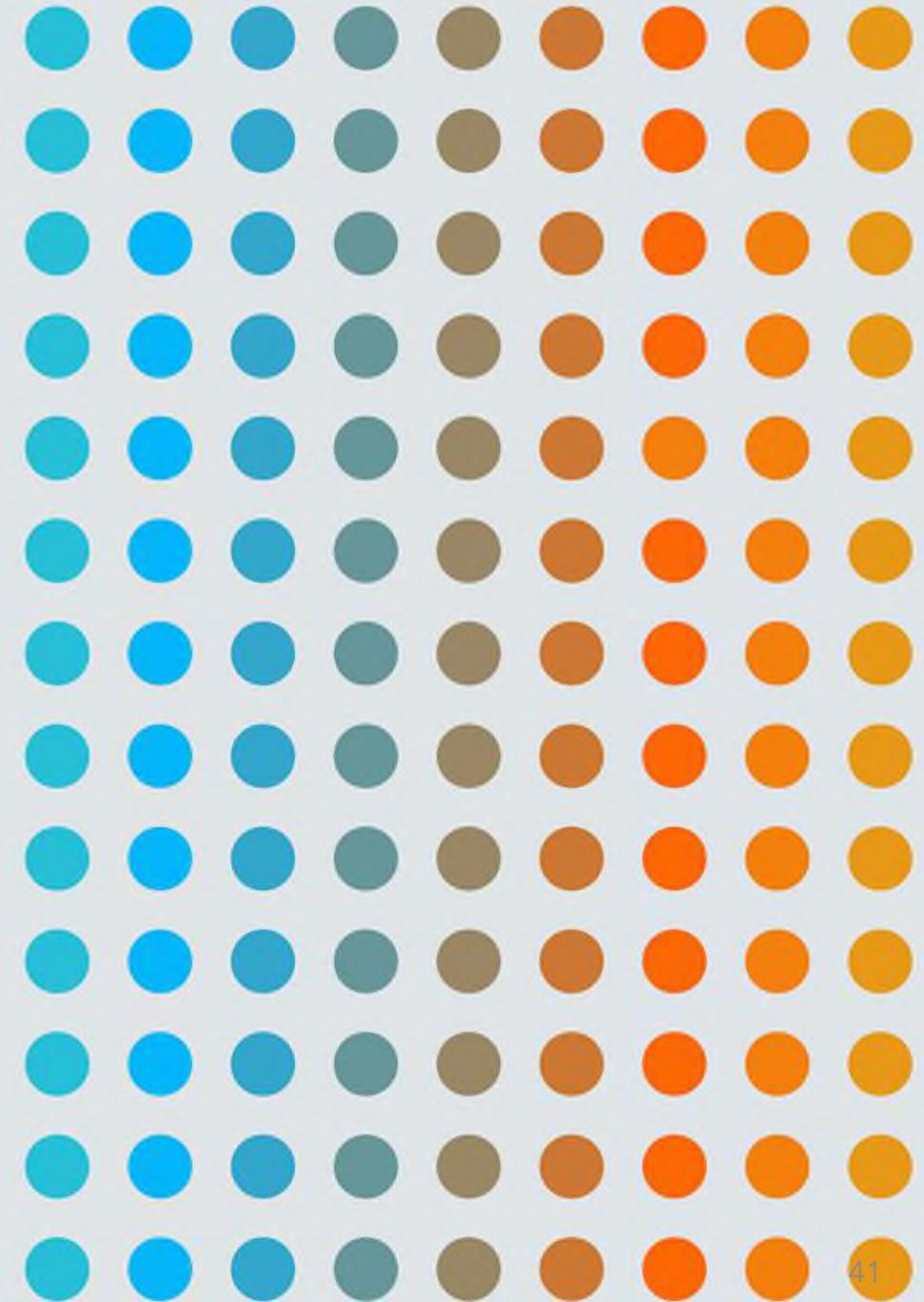
What benefits am I eligible to receive?

How can my scouts club get funding?

How do I apply for an apprenticeship?

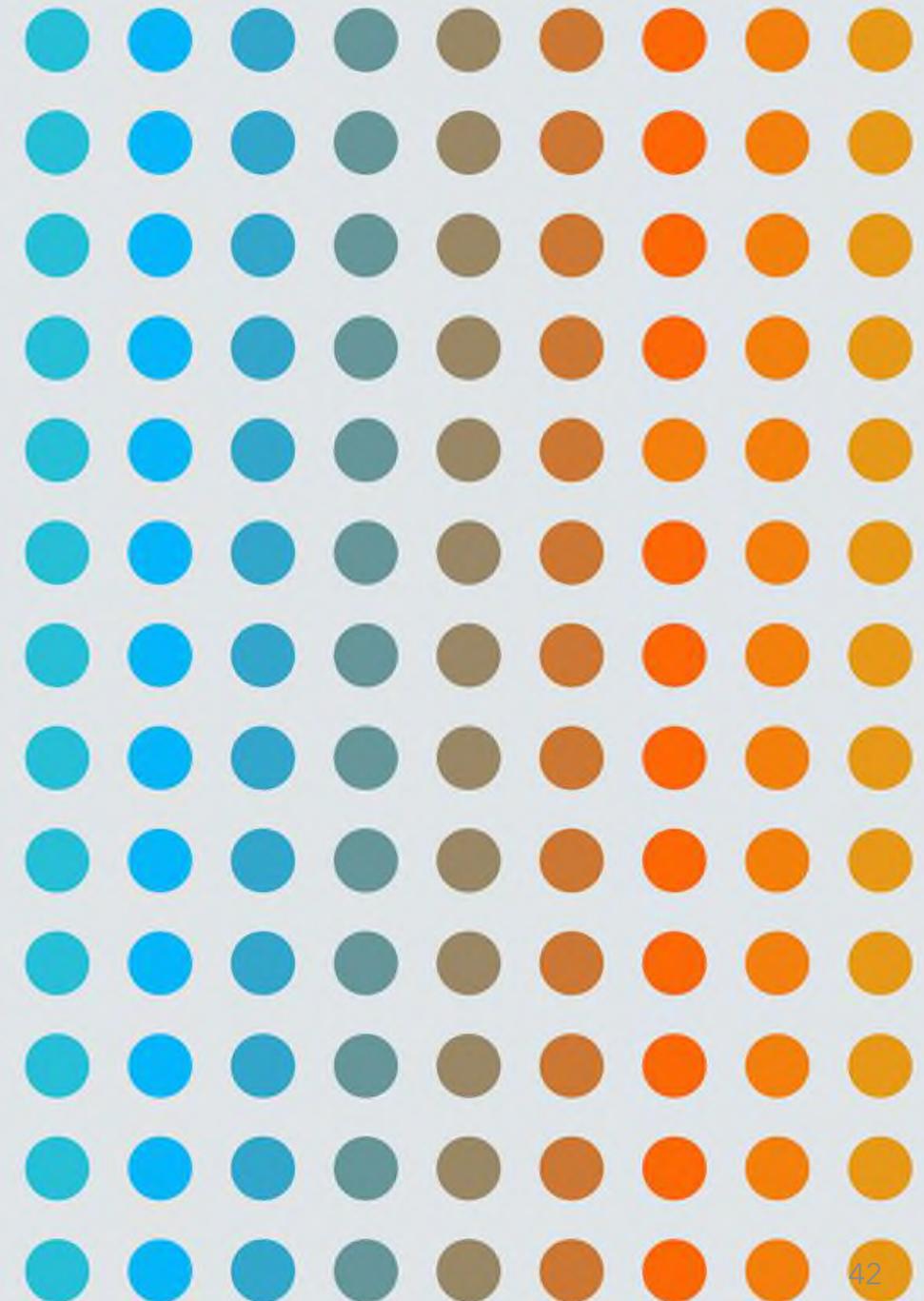
Who is responsible for bin collection?

Where can I receive tenancy advice?



# Citizen queries

- cover a wide range of topics;
- are specific to the context of an individual;
- are sometimes urgent;
- concern objective truths in government guidance;
- require easily understandable answers; and
- may be sensitive.

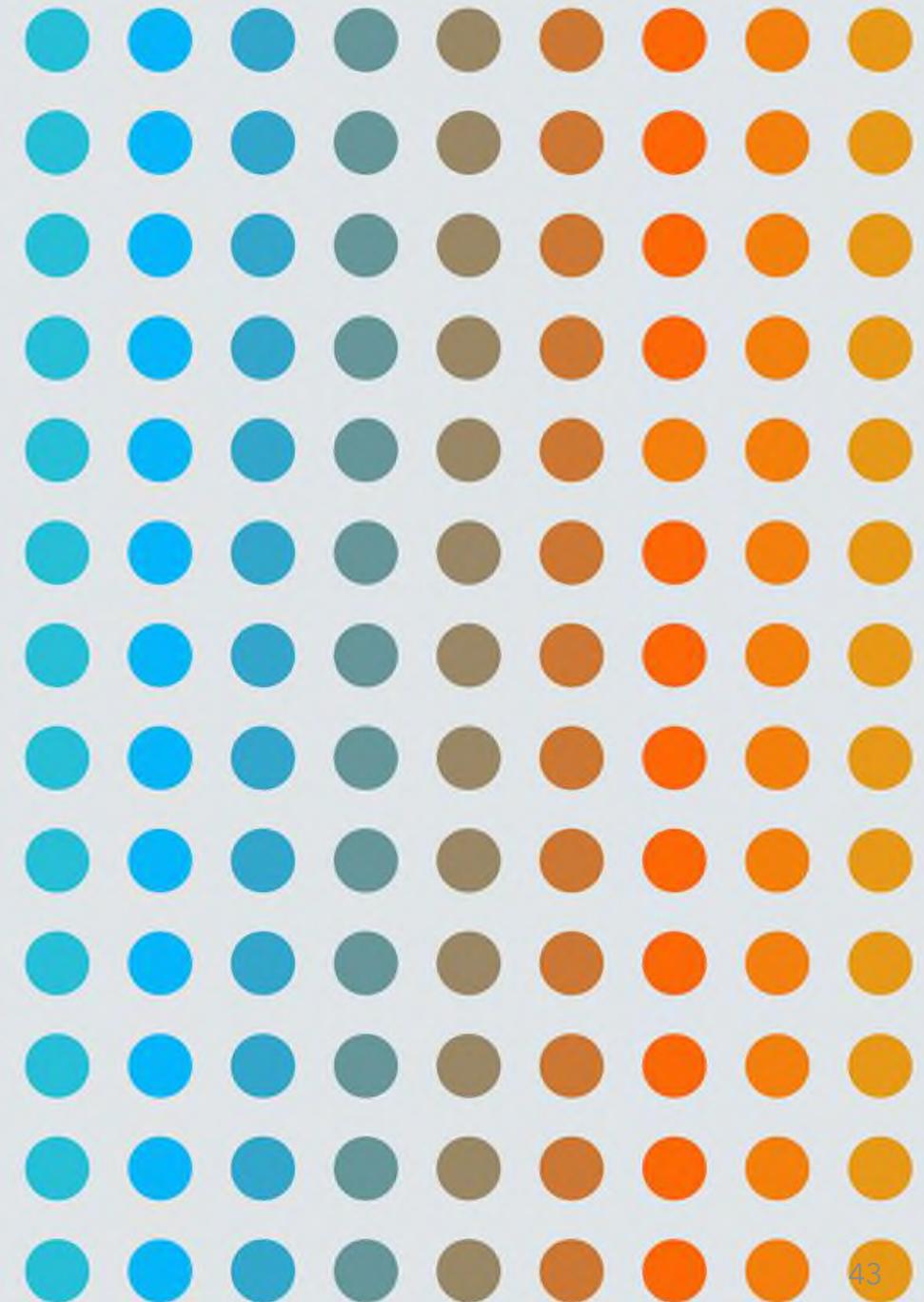


# Chatbots

- have strong general knowledge;
- can talk to you (in your language);
- are fast and offer (some) privacy;
- can personalise responses e.g. individual circumstances, digital literacy, accessibility needs;
- are omnipresent.

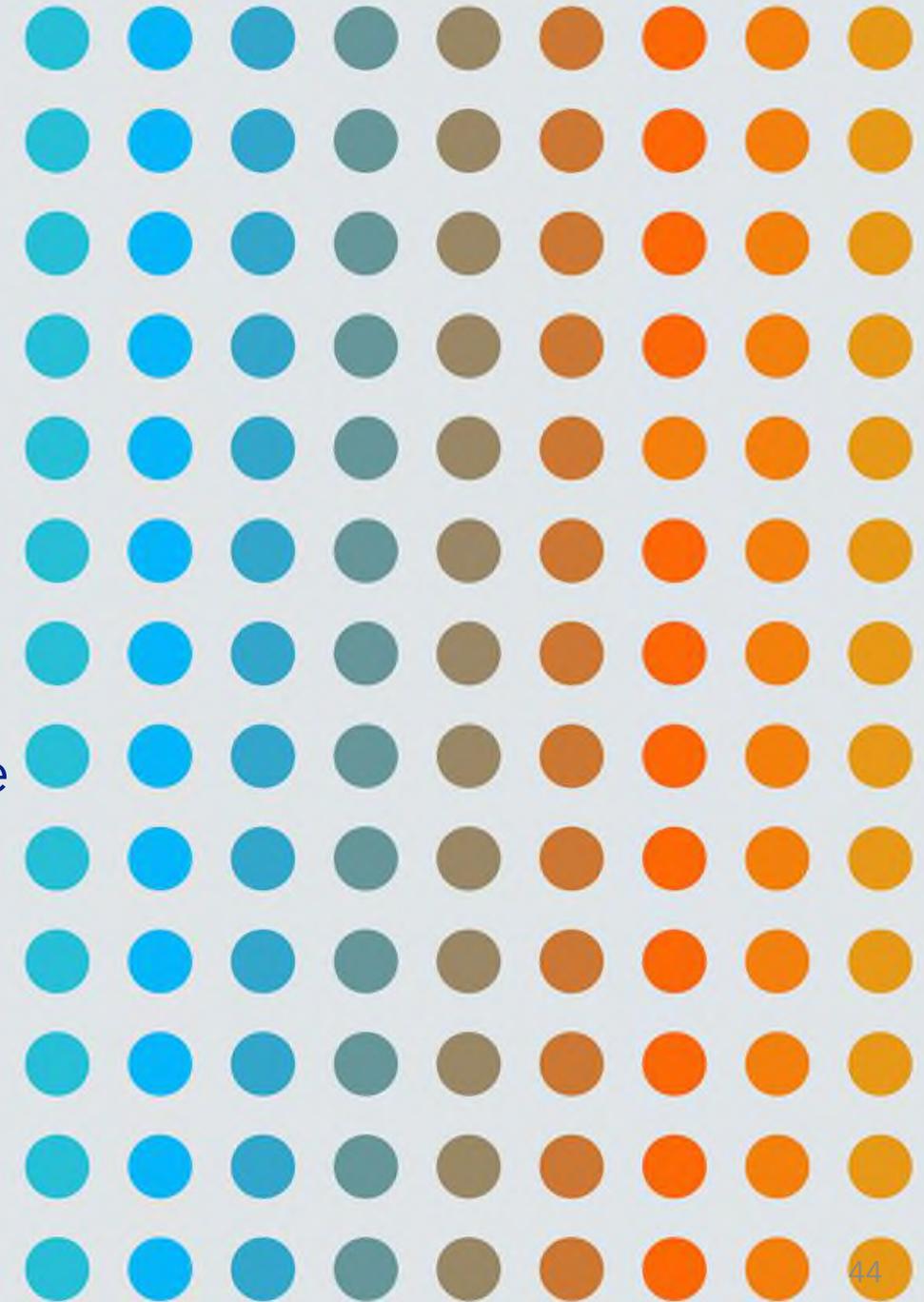
## Chatbots also

- make things up with supreme confidence;
- aren't always up-to-date;
- can refuse to answer certain questions;
- can't always tell you where the information comes from.



# Approach

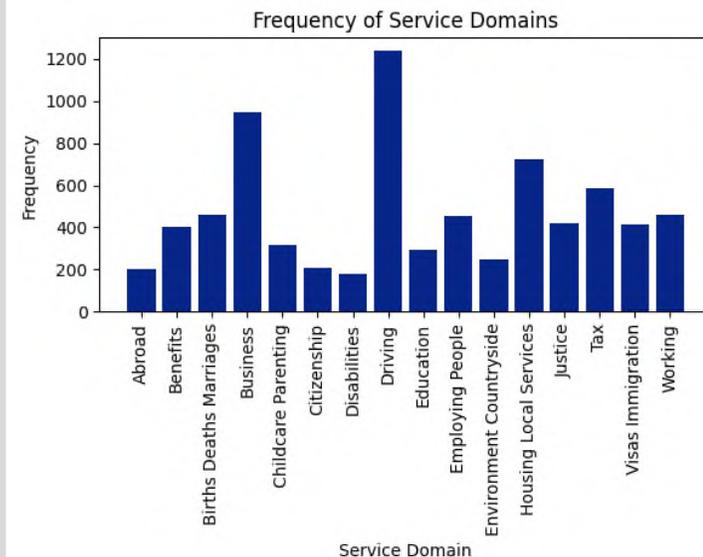
- We built a dataset with many and diverse citizen queries.
- For each query we generated and tested the correct answer.
- We asked 10 large language models (LLMs), some small, some big, some open source, some not to answer the queries.
- We measured how close the LLM answers came to our "ground truth".
- We shared all data and code for everyone to use.



# The data

- 22,066 question - answer pairs, split into three datasets.
- 1,854 unique topics, 4 question types.
- Personas and metadata.
- Synthetically generated with Qwen 2.5.
- Entirely built on gov.uk.

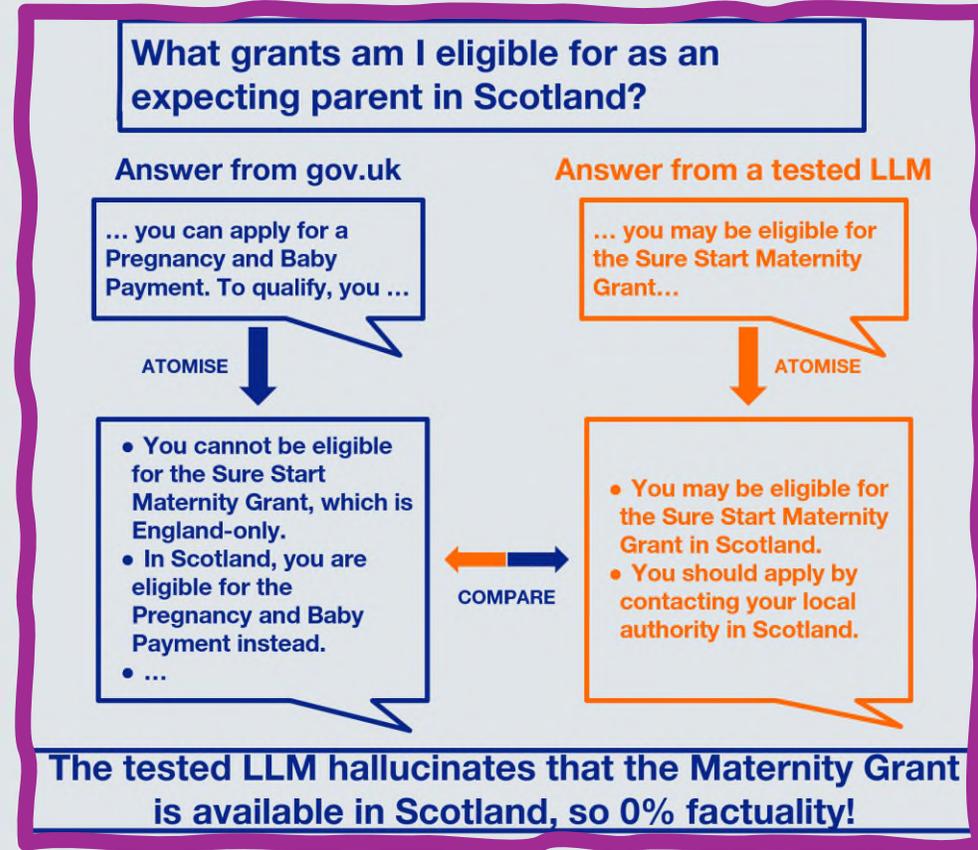
Prompt	Expected Response	Persona of asker	Domain / Subdomain	Information Source	Date valid
Can I stop my ex-partner from getting a passport for our child?	You might be able to stop your ex-partner from getting a passport for your child if you are concerned about abduction or if the child is vulnerable. Contact HM Passport Office to find out if it is possible. They will usually only stop a passport if you have a court order.	26-45 year old; parent; secondary school education; medium digital literacy; moderate household income	Passports / Vulnerable people	<a href="https://www.gov.uk/sto-p-child-passport">https://www.gov.uk/sto-p-child-passport</a>	2025-07-20



# Evaluation

- Are LLMs factual in citizen queries?
- Do they refuse to answer any queries?
- Do they say too much or too little? What happens if we nudge them to be brief?

For each of these questions, we defined metrics that measure the "distance" between an answer generated by an LLM and the answer we expected from gov.uk from our dataset(s)



# LLMs

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

Model	F1@K Statistics					
	Abst.%	Mean	Med.	Std.	IQR	Avg $\Delta K$
Claude-4.5-Haiku	0.38%	<b>0.8725</b>	0.9231	<b>0.1827</b>	<b>0.0973</b>	+16.63
Gemini-3-Flash	0.80%	0.8150	0.8966	0.2258	0.1746	+ <b>5.40</b>
GPT-OSS-20B	0.22%	0.7968	<b>0.9333</b>	0.3094	0.1605	+11.78
Kimi-K2-Instruct	<b>1.37%</b>	0.7450	0.8571	0.2693	0.3048	+5.69
Llama-3.1-8B-Instruct <sup>†</sup>	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 <sup>α</sup>	0.40%	0.7862	0.9268	0.3200	0.1655	+8.45
ChatGPT 4.1 <sup>α</sup>	0.60%	0.7870	0.9375	0.3189	0.1829	+7.50
ChatGPT 5.2 <sup>α</sup>	0.20%	0.8104	0.9564	0.3285	0.0978	+14.17
ChatGPT o3 <sup>α</sup>	0.20%	0.8099	0.9600	0.3329	0.1042	+14.01

## Metrics

Zero-shot = no example in prompt. LLM uses general knowledge it already has

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

Model	F1@K Statistics					
	Abst.%	Mean	Med.	Std.	IQR	Avg $\Delta K$
Claude-4.5-Haiku	1.52%	<b>0.8627</b>	0.9130	<b>0.1843</b>	<b>0.1056</b>	+15.88
Gemini-3-Flash	0.82%	0.8027	0.8889	0.2308	0.1974	+ <b>4.77</b>
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 <sup>†</sup>	<b>1.58%</b>	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	<b>0.9412</b>	0.2217	0.1106	+7.31
ChatGPT 4o <sup>α</sup>	0.00%	0.7965	0.9412	0.3224	0.1492	+8.50
ChatGPT 4.1 <sup>α</sup>	0.80%	0.7995	0.9444	0.3214	0.1357	+7.66
ChatGPT 5.2 <sup>α</sup>	0.00%	0.8203	0.9630	0.3250	0.1003	+12.48
ChatGPT o3 <sup>α</sup>	0.60%	0.7796	0.9600	0.3639	0.1232	+15.11

Few-shot = we help the LLM with a few examples of what we expect to get back

<sup>†</sup> 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.

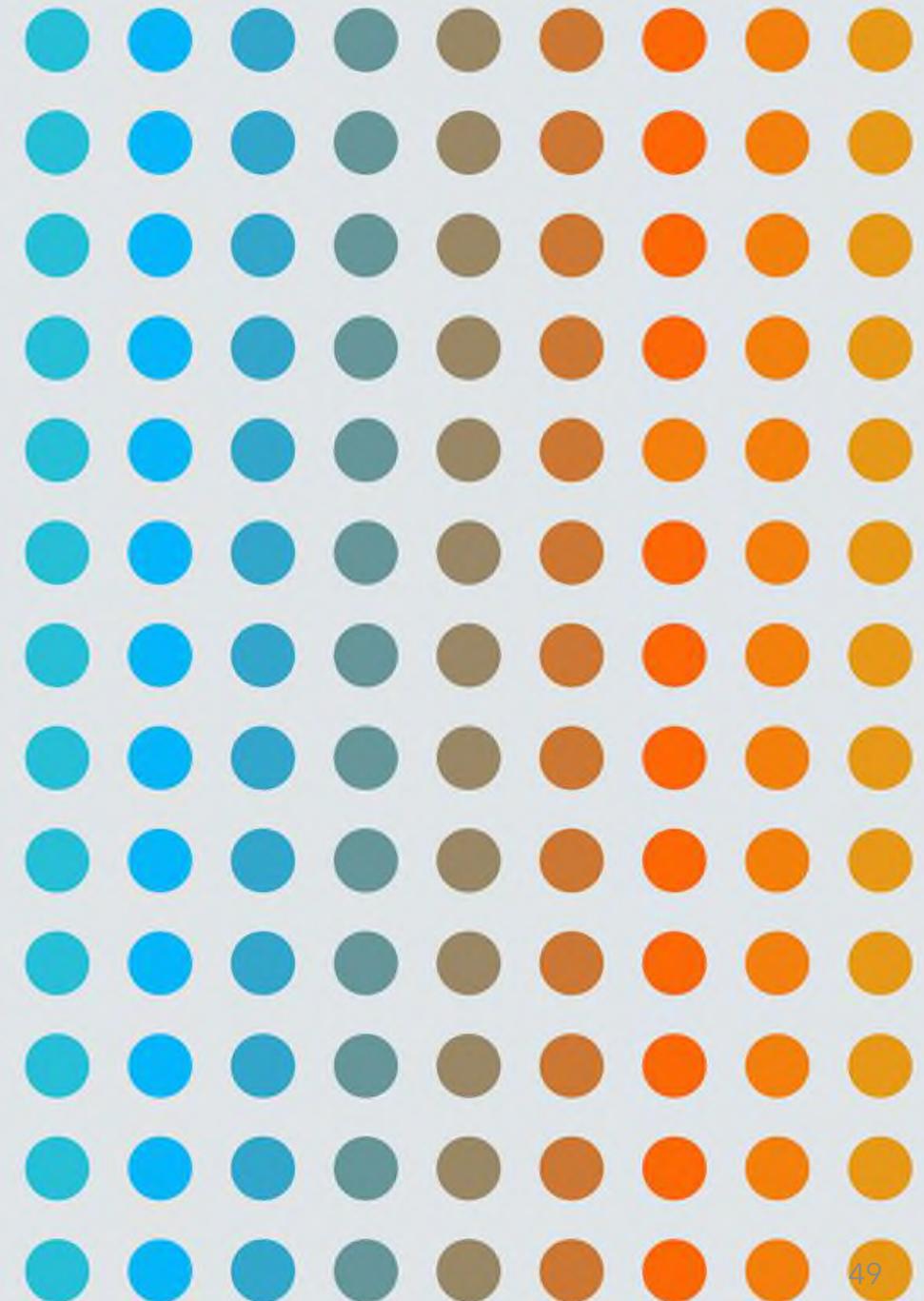
<sup>α</sup> Tested on a domain-stratified sample ( $n = 500$ ) from the full dataset due to API accessibility constraints.

# Findings

- Lots of correct answers, but also varying quality, especially for "long tail" (more specialised) queries.
- LLMs hardly ever refuse to answer a question.
- Answers are lengthy, much more verbose than gov.uk.
- Nudging LLMs to get to the point may reduce accuracy.
- Open and small models can be competitive.

# Implications

- Don't rush for the biggest, most expensive LLM available, avoid vendor lock-in.
- Build more, larger benchmarks.
- Mandate better, independent testing and made results public.
- Invest in innovations to manage trade-offs between answer length, accuracy, utility.



# Conclusions

- **Open data infrastructure** remains an essential source of growth and innovation.
- New AI capabilities can improve **data reuse and impact**, as well as **user experience**.
- Data publishers and portal managers should improve the **AI readiness** of their data and experiment with AI capabilities to understand their limitations.
- We need more open **user-centric benchmarks**, with widely reported results.



Questions?

Thanks to all contributors to these projects, including Thomas Carey-Wilson, Arunav Das, Lisa-Yao Gan, Neil Majithia, Joe Massey, Gefion Thuermer, Johanna Walker, Rajat Shinde, Manil Maskey, Zo Chapman, Prajun Trital, Jordan Decker.

Elena Simperl

@esimperl