Towards an Open Database of Data Centres: Extracting Structured Information from Technical Specification PDFs Using LLMs and RAG

Alok Singh*
Smith School of Enterprise and the Environment, University of Oxford
Oxford, UK

Neetu Kushwaha*
Smith School of Enterprise and the Environment, University of Oxford
Oxford, UK

Christophe Christiaen*
Smith School of Enterprise and the Environment, University of Oxford
Oxford, UK

Abstract

The rapid expansion of data centres around the world is putting strain on various Sustainable Development Goals (SDGs) linked to energy, water, land, and labour, but sustainability information in this sector is fragmented and heterogeneous. Individual data centres have very different environmental impacts and risks based on their size, operations, and where they are located. However, manually collecting and reviewing these data points is a time-consuming task. In this work, we present an end-to-end AI pipeline for generating a comprehensive, structured dataset of individual data centre attributes from publicly available technical specification documents. This pipeline utilizes a Retrieval-Augmented Generation (RAG) architecture combined with open-source Large Language Models (LLMs). We conducted a comparative analysis of several benchmark large language models (LLMs), including Mistral, DeepSeek, LLaMA, and GPT-3.5, to assess their performance within our end-to-end pipeline. The extracted results are validated using RAG Assessment Suite (RAGAS) metrics and human validation. Furthermore, we illustrate how the extracted data can be used to analyse environmental impacts for individual facilities and companies, focusing on water and carbon footprint. Ultimately, we aim to deploy this approach globally to build an open database which enables NGOs, regulators, policymakers, investors, etc. to track the sector's environmental footprint and hold operators to account. Our code is available at [1].

Keywords

Sustainability risk, Data centres, Retrieval-Augmented Generation (RAG), Question Answering

1 Introduction

Accelerating growth in digital services, including cloud storage, blockchain and artificial intelligence is driving a rapid expansion in data centre construction and development all around the world. Overall data centre capacity is estimated to increase from 59GW in 2024 to 122 GW by 2030, which will have significant implications for the sector's environmental footprint [12]. Two major environmental pressures from operating data centres include greenhouse gas emissions, from energy used, and water used directly or indirectly for cooling [22, 29] In 2022, electricity consumption from data centres globally was estimated to be within 350-450 TWh or around 1.4-1.7% of global electricity demand. While emissions for data centres and transmission networks were estimated to be around 330 mega tonnes CO2 equivalent in 2020, equivalent to 0.6% of global greenhouse gas emissions [26]. The water withdrawal needed for the processing of AI models alone is estimated to be between 4.2 and 6.6 billion cubic meters of water in 2027, more than the total annual water withdrawal of Denmark today [18].

As the infrastructure backbone for digital services, data centres can contribute positively to numerous Sustainable Development Goals (SDG), particularly when designed sustainably and deployed responsibly with equitable access in mind (SDG 9, targets 9.1 and 9.c). However, unabated growth of data centres globally could significantly impede progress on SDGs too. Direct operations of data centres will compete with other demands for scarce freshwater (SDG 6, targets 6.1, 6.4 and 6.5) and affordable and clean energy (SDG 7, targets 7.1, 7.2 and 7.3) sources or generate significant volumes of electronic waste from rapidly obsolete servers (SDG, target 12.2, 12.4 and 12.5). While impacts across supply could include resource depletion, critical raw materials extraction and unethical labour practices, (SDG 8 and 15) [24].

These developments and pressures are raising concerns from different national and international stakeholders. Non-profits are calling out the disproportionate consumption of water by data centres in Virginia, United States [14] while the Irish government has effectively banned the construction of new data centres due to disproportionate pressures on its electricity infrastructure [21]. Globally, investors are concerned about the financial environmental risks associated with their investments linked to their water and carbon footprint [4, 35], or are actively pursuing investments in so called 'sustainable', more resource efficient, data centres [5].

The extent and impact of those environmental pressures is context specific [29]. The location of a data centre will determine to what extent the power it consumes is generated from fossil fuels, based on the grid's fuel mix or access to renewable energy sources. Additionally the impact of its water consumption will be significantly higher in areas with higher water stress [18]. This means that granular information about individual data centres, tied to their exact location, is essential for local, national and global stakeholders to understand the environmental risks, opportunities and impacts associated with the building and operating of these facilities. With 60% of current data centre capacity operated by either large tech companies or third-party wholesale operators and a remaining chunk operated by telecom and traditional companies, corporate sustainability reports should provide information about data centres' environmental performance [12]. Unfortunately, transparency

*All authors contributed equally to this research.
[1] https://github.com/alokssingh/PDF_Question_Answering_RAG_with_pymupdf4llm

,
.

and adequate sustainability reporting from data centre operators is often lacking [10, 22].

Where companies do report environmental information, it is typically aggregated at the company level without location information or indicators for individual facilities. Additionally, corporate sustainability reports are unstructured and inconsistent sources of information due to a fragmented global landscape of mandatory and voluntary reporting [6]. Stakeholders, such as investors or financial regulators, who rely primarily on corporate reports, will struggle to assess the environmental risks of their counterparties in a meaningful way [7].

Information about asset-specific operational indicators and environmental pressures is available from alternative data sources such as commercial portals, news articles, investor reports, industry bodies, environmental licensing documentation, or technical specification sheets [20]. While they may provide more granular or up-to-date information, the information available is scattered, unstructured and hard to collect. Language models have been used to analyse climate risk disclosures at the corporate level [3, 6], but their potential for extracting and/or analysing asset-level information remains under researched.

The objective of this study is to develop an automated approach for extracting asset-level information from readily available data centre specification sheets. The approach should be scalable across countries and different types of data sources. Ultimately this will allow us to build an open, global asset location database integrating operational, location and ownership features of individual data centres. Such open data mapping efforts have the potential to bring transparency and concerted environmental action to the sector, similar to efforts in the power, cement, or agricultural commodity sectors [11, 31, 32]. It can reduce information asymmetry between third party stakeholders and data centre operators about the local and global environmental implications of data centre developments. For instance, it would allow policymakers to provide more informed planning permissions for new data centres, allow civil society and researchers to identify the most appropriate decarbonisation pathways for the sector, or allow investors to mitigate their financial risks and support their investees in reducing their environmental impacts.

Previous work by [1, 2, 10] assessed the environmental impacts of data centres based on spatially agnostic methods and data points such as lifecycle assessments or country specific averages. While this offers some comparable insights and gives an indication of the industry's potential impacts as a whole, these are insufficient to capture context or company-specific risk profiles. Environmental impact analyses by [18, 20, 28, 29] were based on asset specific information, where the underlying asset-level data was either retrieved from commercial sources or collected manually. These impede the replicability of the analysis due to access constraints and the time it takes to manually collect information, in such a rapidly evolving sector. These gaps create a need for a faster, more automated approach based on openly available data sources.

To automate the creation of such a dataset, we leverage LLMs to automatically extract key data centre attributes from unstructured documents. Despite their capabilities, LLMs continue to struggle with knowledge-intensive tasks, such as open-domain question answering (QA). They often encounter challenges such as dependency on data or knowledge (which may be outdated); inability to easily expand or revise their memory; or the tendency to hallucinate by generating false information [16, 19]. To address these issues, RAG provides LLMs with potentially relevant documents as the external knowledge through retrieval, enhancing its accuracy and reliability. This reduces reliance on the LLM's internal knowledge. Despite the widespread use of LLMs for downstream tasks, relatively little work has been done on domain-specific tasks. In finance, Wu et al. [37] proposed BloombergGPT to improve domain-specific knowledge of LLMs. Based on performance, it is observed that the proposed BloombergGPT outperforms other LLMs in downstream finance-related tasks. Similarly, BioBert [17] is a pre-trained LLM to extract valuable information from the biomedical literature. In this work, we implement an end-to-end RAG approach with an LLM. These tools support manual efforts to build an asset-level data centre database. This paper has the following contributions:

- We propose an end-to-end pipeline for extracting data centre–specific information from unstructured specification PDF documents, with the goal of enabling the construction of structured, integrated dataset. This pipeline is based on a RAG framework for information extraction, combining an open-source retriever with a LLM to enhance factual accuracy and contextual relevance.
- We evaluate the performance of our RAG-based system using the RAGAS framework, and further validate the extracted outputs through expert review to ensure quality, reliability, and faithfulness of the results.
- In the study, we illustrate how the extracted data can be leveraged to analyse environmental impacts for individual facilities and companies.

2 Methodology

We developed a comprehensive end-to-end pipeline that integrates manual metadata collection, web-based document retrieval, and LLMs. To retrieve specific data centre related documents from Google and Yahoo search engines, we generated search keywords based on data centre names, providers, and addresses. Information about individual data centre characteristics was then extracted using a RAG question-answering system, based on a predefined set of questions and prompt templates. The workflow is illustrated in Figure 1. The subsections below provide a detailed description of each component of the methodology.

2.1 Document Collection and Preprocessing

First we manually gather key details from a publicly available data centre directory for a given country [8], including the data centre name, operator and the physical address for each listed facility. We then combine these three attributes into search keywords that are both specific and contextually rich (for example, "Digital Realty Singapore Data Center + Digital Realty + 3 Loyang Way"). We use these keywords to automatically download PDF documents, such as technical specification sheets used for marketing purposes. To streamline this process, we implement a Selenium-based crawler to search and download PDFs using web searches via Google and Yahoo.

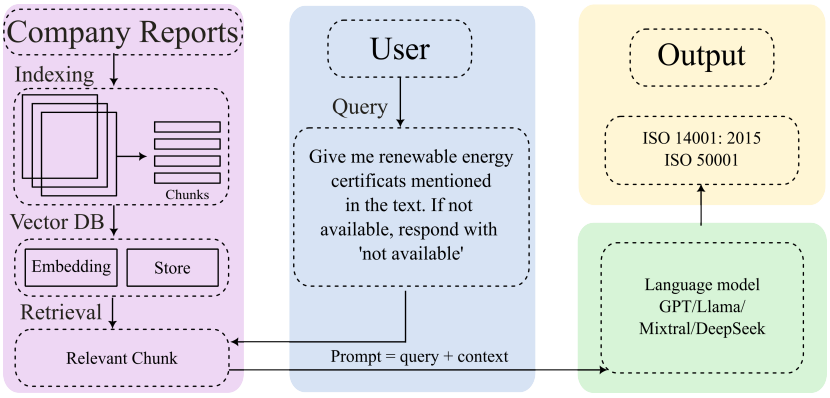


Figure 1: Retrieval Augmented Generation pipeline for data centre related information extraction

2.2 PDF Preprocessing

After downloading the PDFs we need to accurately extract text data from for an effective RAG pipeline - retriever and generation. For this purpose, we utilise tools and libraries like pdfplumber [30] and PyMuPDF [15] (also known as fitz). These libraries are capable of handling most common PDF structures and formats while preserving the layout and structure of the text as much as possible. Additionally, we store the extracted text with associated metadata.

2.3 Retrieval-Augmented Generation (RAG) System

As part of our information extraction pipeline, we employ a RAG-based question answering (QA) system to extract structured information from these PDFs. Specific questions related to data centres are generated using predefined prompt templates. A set of manually defined questions of key data centre attributes is integrated into prompt templates that include both explicit task instructions and relevant document context. These prompts are then passed to the RAG system to generate accurate, context-aware responses.

The Query-based RAG system consists of two core components: the retriever and the generator. The retriever receives the input query from the user and searches for similar documents stored in the vector database, selecting the top k chunks or documents based on similarity or distance functions. This entire process is divided into two steps. First each object is encoded using some representation. Second, an index is created to organise the data source for efficient search. In RAG, indexing is a crucial phase as it facilitates the retriever by enabling similarity searches based on queries. Indexing begins with a text document, which is segmented into smaller chunks using various criteria such as character splitting, recursive splitting, or semantic splitting, depending on the specific problem statement. These smaller chunks are then transformed into vector embedding using an embedding model and stored in vector databases to speed up retrieval. The generator takes the original query and the retrieved context as input, using specific augmented techniques. It then generates the answer to the input query based on the provided context.

3 Results and Discussion

In this section we provide a brief discussion on data centre attributes, summary of PDF retrieval, parameters setup, prompting LLM, qualitative evaluation and quantitative evaluation using RAGAS framework.

3.1 Selected Data Centre Attributes

Our end-to-end pipeline is capable of extracting various attributes from data centre documents. For this implementation, we focus on a selected subset of key attributes, as listed in Table 1. The remaining attributes to be included in the final dataset are detailed in Table A1.

3.2 PDF Retrieval Summary

This subsection presents a summary of PDF retrieval based on keyword searches. The statistics indicate the total number of generated keywords, successful downloads, and the overall retrieval success rate, as shown in Table 2.

3.3 Model Parameter Settings

We used the LangChain framework to build our Retrieval-RAG system. To evaluate performance across different model architectures, we implemented the system with various language models, including Mistral, DeepSeek, LLaMA[36], and GPT-3.5. Both the query and document chunks are encoded into vector embeddings using the Sentence-BERT model[25]. We employed the Chroma DB [33] vector database for indexing the document chunks. This is a popular technique for handling domain-specific question-and-answer tasks. The parameter settings for all models are presented in Table 3.

3.3.1 Experimental Setup. Experiments were conducted on the Google Cloud Platform (GCP) using a virtual machine with two NVIDIA A100 GPUs. Open-source models, such as LLaMA 3.3, Mistral (7B), GPT-3.5, and DeepSeek, were executed using the Ollama [34] framework in the GCP instance for efficient inference.

Table 1: Key data centre attributes and their descriptions.

Attribute	Description
Data Centre Name	Name of the data centre
Address	Physical location of the data centre
Operator	Company responsible for managing or operating the data centre
ISO 27001 Certified	Indicates whether the data centre is certified under ISO 27001 information security standards
ISO 14001 Certified	Indicates whether the data centre is certified under ISO 14001 information security standards
Colocation Space	Total space allocated for customer equipment (e.g., server racks)
Building Total Whitespace	Total area available for IT infrastructure deployment
Total Critical Power	Total power capacity dedicated to critical IT systems
Security Systems in Use	Security measures implemented in the data centre (e.g., CCTV, biometric access)

Table 2: Keyword-Based PDF Retrieval Summary

Retrieval Statistic	Value
Total keywords generated	63
Successful PDF downloads	45
Document retrieval success rate	71.43%

3.4 Prompting LLM

Effective prompt engineering is essential for enabling language models such as Mistral, DeepSeek, LLaMA-3.3, and GPT-3.5 to generate accurate and structured outputs. This process involves designing clear, specific instructions and embedding relevant contextual information within the prompts to effectively guide the model's behaviour [23]. We adopted an iterative refinement approach, adjusting the prompts based on the quality and consistency of the generated responses. To keep the evaluation fair and consistent, the same set of prompts was used for all four models during evaluation. Each prompt consists of an instruction, a question, and the corresponding document context used by the RAG system to generate answers. The specific questions applied for extracting the data centre attributes are listed in Table 4. The structure of the prompt template used to produce structured outputs is illustrated in Figure 2.

3.5 Evaluation Metric: RAGAS

The RAGAS framework [9] is used to assess the performance of the RAG system, with higher metric scores indicating improved performance. The evaluation focuses on the following key metrics:

- **Faithfulness Score:** Faithfulness measures how well the generated output aligns with the retrieved evidence. It assesses whether the response contains any hallucinations, or information not found in the retrieved context.
- **Answer Relevance Score:** Answer relevance evaluates semantic alignment between the query and the generated response.
- **Answer Correctness Score:** Answer correctness verifies factual consistency with the ground truth data points.
- **Context Recall/Precision Score:** Context recall and precision assesses the efficiency and coverage of the retrieval component, ensuring relevant and comprehensive context is supplied to the generation module.

3.5.1 Radar Chart Visualization. Figure A1 provides detailed radar chart visualizations highlighting the comparative performance profiles of all evaluated models per data centre attribute. The metrics assessed include faithfulness, answer relevance, answer correctness, context precision, and context recall. Each axis in the radar chart corresponds to one metric, and the polygonal area enclosed by each model indicates its overall performance profile. Mistral demonstrates the most balanced and comprehensive performance, particularly excelling in faithfulness, answer relevance and answer correctness. GPT-3.5 shows strength in context precision and context recall but falls short on other dimensions. Deepseek performs relatively well in recall but struggles with faithfulness while LLaMA does not outperform the other models on any individual metric.

3.5.2 Qualitative Evaluation by Domain Experts. Domain experts conducted a qualitative evaluation of the generated outputs, assigning scores based on answer relevance using a three-point scale. This rating indicates the degree to which the generated answer is relevant and accurate given the context. A score of 1 is given if the answer is highly accurate and directly responds to the query or prompt. A score of 0.5 is assigned when there is a partial match, while a score of 0 is used if there is no match at all.

The human evaluation results for key data centre attributes are presented in Table 5. LLaMA 3.3 achieved the strongest overall performance, demonstrating higher consistency and accuracy across most attributes evaluated. GPT-3.5 also performed well, particularly on sustainability and certification-related attributes. Mistral and DeepSeek showed less consistent and lower overall accuracy. In contrast, LLaMA 3.3 and GPT-3.5 proved more effective at producing reliable, structured information from unstructured data centre documents. While our human evaluation directly measures correctness against ground truth, RAGAS focuses on consistency with retrieved context. As a result, RAGAS may give higher scores even when answers are only partially correct, highlighting the value of using both approaches for a more comprehensive assessment.

Table 3: Model Parameter Settings

Parameter	Mistral	DeepSeek	LLaMA (2)	GPT-3.5
Chunk Size	600	600	600	600
Chunk Overlap	20	20	20	20
Top K Retrieval	6	6	6	6
Embedding Model	mistral:7b	deepseek-llm:67b	llama3.3	text-embedding-ada-002
Model Variant	Mistral-7B	DeepSeek-67B	llama3.3:70b	GPT-3.5-turbo

Table 4: Questions for extracting key data center attributes.

Attribute	Question
Data Centre Name	Provide only the data centre name exactly as mentioned in the text. If none is available, respond with 'not available'.
Address	Extract only the full address of the data centre as stated in the text. If not mentioned, respond with 'not available'.
Operator	Extract only the name of the data centre operator as stated in the text. If not mentioned, respond with 'not available'.
ISO 27001 Certified	Is the data centre ISO 27001 certified? Respond with 'Yes' or 'No'. Do not provide additional explanation.
ISO 14001 Certified	Is the data centre ISO 14001 certified? Respond with 'Yes' or 'No'. Do not provide additional explanation.
Colocation Space	Provide only the size of colocation, technical, or IT space as stated in the text (e.g., 153,000 m²). If not mentioned, respond with 'not available'.
Total Whitespace	Provide only the building total whitespace as stated in the text (e.g., 153,000 m²). If not mentioned, respond with 'not available'.
Total Critical Power	Provide only the total critical power value. If not mentioned, respond with 'not available'.
Security Systems in Use	List all security measures mentioned in the document, exactly as stated. If none are mentioned, respond with 'not available'.

```
Template="""You are assigned the role of an AI assistant. Using only the given context on the data
centre which is exacted from technical specification sheets. Ensure that your answers are precise and
strictly based on the provided text. Do not include any external information or assumptions. If no
information is available in the text, do not make up an answer. Please provide only the final JSON
output without any explanations, reasoning, or additional text.

Question: {question}
Text: ``` {context} ''' """
```

Figure 2: Template instructions for extracting data centres specification information in JSON format

Table 5: Human evaluation results for key data centre attributes across models

Key attributes	GPT-3.5	DeepSeek	LLaMA 3.3	Mistral
Data centre name	0.464	0.446	**0.607**	0.600
Address	0.580	0.482	**0.813**	0.473
Operator	0.571	0.564	0.609	**0.636**
Total Critical Power	**0.839**	0.768	0.821	0.554
Colocation space	0.655	0.382	**0.709**	0.491
Building Total Whitespace	**0.545**	0.327	0.400	0.382
ISO27001 Certified	0.691	0.418	0.709	**0.782**
ISO14001 Certified	**0.893**	0.804	0.821	0.732
Security system in use	0.429	0.250	0.348	**0.573**

4 Environmental Impact Estimation Using Extracted Data

Environmental impacts of individual data centres can be estimated by combining data centre specific features, such as overall power capacity, power and water utilization efficiency performance and location information with water consumption or carbon emission intensity factors based on sectoral or geographical averages. For instance, both direct carbon and water footprint can be estimated based on an individual data centre's energy usage. The total energy consumption (equation 1) can be calculated in several ways, depending on the data available [13, 29]:

$$\mathrm{DC}_{E_i} = \text{Total power capacity}_i \times 8,760 \times \text{Uptime} \tag{1}$$

where,

- DC_{E_i} = Total energy use (kWh/year) for facility *i*.
- uptime: The proportion of time where the data centre is running at full capacity. This is a value between 0 and 1.

In cases where the total power capacity is not available, total critical power capacity can be estimated using equation 2

$$\text{Total power capacity}_i = \mathrm{IT}_s \times PUE_i \times A_i \tag{2}$$

where,

- IT_s: Load intensity (W/ft^2 or W/m^2) for data centre size *s*.
- $PUE_i = \frac{\text{Total power supplied to the data centre}}{\text{Power consumed by IT equipment}}$.
- A_i: Floor area of the data centre (ft^2 or m^2).

Once the energy use is estimated, direct and indirect water and emission footprint can be calculated based on intensity metrics shown in equation 3 and 4 respectively.

$$\mathrm{DC}_C = \mathrm{DC}_{E_i} \times \mathrm{CI}_j, \tag{3}$$

$$\mathrm{DC}_W = \mathrm{DC}_{E_i} \times \mathrm{WI}_j. \tag{4}$$

where,

- DC_{E_i}: Electricity consumption of data centre *i* (MWh) (equation 5).
- DC_C: Carbon emissions of data centre *i* (tonnes CO_2-eq) (equation 6).
- DC_W: Water consumption of data centre *i* (m^3) (equation 7).
- CI_j: Direct and indirect carbon intensity for data centre type *j* (tonnes CO_2-eq/MWh).
- WI_j: Direct and indirect water intensity for data centre type *j* (m^3/MWh).

We take IT load intensity values from [27], uptime value used by [13] and intensity values for colocation data centres from [29] to calculate the indirect emission and water footprint for NTT's Ashburn VA4 colocation data centre in Virginia, United States.

$$DC_E = 32\ \text{MW} \times 8{,}760\ \text{hours/year} \times 0.75 = 210{,}240\ \text{MWh/year} \tag{5}$$

$$\begin{aligned} DC_C &= 210{,}240\text{MWh/year} \times 0.42\ \text{tonne CO}_2\text{-eq/MWh} \\ &= 88.3\ \ \text{tonne CO}_2\text{-}eq/year. \end{aligned} \tag{6}$$

$$DC_W = 210{,}240\ \text{MWh/year} \times 7.00\ \text{m}^3/\text{MWh} = 1.5\ \text{million m}^3/\text{year} \tag{7}$$

Additionally, information about the exact location of the data centre can be used for analysing environmental risks using geospatial datasets. Given the significant water footprint of data centres, different aspects of 'water risk' are important for data centre operators and stakholders. Here we have entered the address of NTT VA4 into the WWF Water Risk Filter, an open risk screening tool with global geospatial datasets covering 12 water risk categories and 42 indicators [38], and we receive the following scores.

Basin physical risk score = 2.53 (low). This score takes into account current water availability, drought, flooding, water quality and ecosystem service status in the basin where the facility is located (Middle Potomac / Catoctin). Basin regulatory risk score = 1.38 (very low). This score takes into account the current enabling (regulatory) environment, institutions & governance, (water) management instruments and infrastructure for water, sanitation and hygiene services within the basin where the facility is located. Basin reputational risk score = 2.44 (low). This score takes into account current environmental, socioeconomic and additional reputational factors in the basin where the facility is located. Combining both water risk metrics, this analysis indicates that even though 1.5 million m3/year water consumption is a significant pressure on water resources, this is unlikely to cause operational or reputational issues in its current location and context, today. In the future the risk profile can change due to growing pressures on water resources from industry or changes in climate, impacting water availability for that basin.

5 Limitations

While the results show that an end-to-end RAG pipeline is a useful process to automatically extract facility-level information, challenges remain which need to be addressed:

- Irrelevant or missing PDF results from keyword searches: In most cases, we were unable to successfully extract or download PDFs using keyword-based searches. Sometimes, the retrieved pdf did not correspond to the keyword search (the downloaded pdf differs from what we searched). This inconsistency creates gaps in our automatic extraction pipeline for building the dataset. To address this issue we used the LLM as a classifier to categorise the PDF as a specification sheet (or not) which helped us remove irrelevant PDFs.
- PDF text extraction: Accurate text extraction from PDFs is essential for effective retrieval and generation in RAG applications. However, there are challenges in converting text from PDFs as often the layout of the PDF is too complex, including multi-column formats, embedded images, footers, headers, and tables. A lack of standardisation in PDF creation means that different encoding methods and embedded fonts are used, leading to inconsistencies in the extracted text and hindering RAG performance. Additionally, many PDFs are scanned documents that require Optical Character Recognition (OCR) to convert images to text, further complicating the text extraction.

- Wrong answers: Human evaluation of the answers shows that some models are likely to make mistakes, for instance quoting results from other PDF sources. Additionally, the RAG generated answers are not always relevant to the context, even when external knowledge is incorporated by providing the context along with the query to the generator. This shows that a further manual verification process is required to ensure a more robust output specific to the data centre.

In our current study, we focus on data centre specification sheets, specifically examining third-party colocation data centre providers that sell or lease services to various clients. This does not include 'hyperscale' data centres owned and operated by large tech companies such as Google, Amazon, Microsoft, or Meta.

6 Conclusion

An open, global database of accurate asset specific data centre attributes can become a critical foundation for increased transparency in the sector. Ultimately such a database will create a level playing field between communities, non-profits, governments, businesses and investors to hold data centre operators to account on their negative externalities.

In this paper, we present an AI approach for creating a structured dataset of data centre attributes. By leveraging an existing RAG framework, we automate the extraction of critical data points from diverse and unstructured sources, starting with data centre specification sheets. We use predefined questions, carefully designed prompts, and multiple language models, such as Mistral, DeepSeek, LLaMA-2, and GPT-3.5. The automatic and human evaluation scores show that the LLMs with RAG approach offer a helpful starting point to extract asset-level data at scale, but human intervention and validation remains needed. From our analysis, we find that no single LLM outperforms the other models. Deepseeks the poorest in our analysis, while LLaMA3.3 and GPT-3.5 produce a more consistent, accurate, and context-aware output, although the differences are limited. This higlights the need for further detailed evaluation and careful application of the right model for the right type of question, on the right type of PDF input.

To address the limitations, our future work will focus on refining the architecture of the retriever, refining prompt designs, and fine-tuning end-to-end RAG with data centre-specific data to improve the effective use of retrieved documents by LLMs, thereby enhancing data quality and comprehensiveness. Additionally, we aim to explore additional sources of input text, such as company sustainability and financial reports or specialist news articles to expand the type of information we can integrate into a global open database.

References

[1] Tuba Aslan, Philipp Holzapfel, Lisa Stobbe, Anne Grimm, Niklas F. Nissen, and Matthias Finkbeiner. 2025. Toward Climate Neutral Data Centers: Greenhouse Gas Inventory, Scenarios, and Strategies. *iScience* 28, 1 (2025), 111637. doi:10.1016/j.isci.2024.111637

[2] Antoine Berthelot, Eddy Caron, Matthieu Jay, and Laurent Lefèvre. 2024. Estimating the Environmental Impact of Generative-AI Services Using an LCA-Based Methodology. In *Procedia CIRP*, Vol. 122. 707–712. doi:10.1016/j.procir.2024.01.098

[3] Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nikolaus Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance* 164 (2024), 107191. doi:10.1016/j.jbankfin.2024.107191

[4] Isla Binnie. 2024. Power thirst complicates ESG investors' love affair with tech stocks. https://www.reuters.com/sustainability/sustainable-finance-reporting/power-thirst-complicates-esg-investors-love-affair-with-tech-stocks-2024-09-26/. Reuters, September 26.

[5] Louise Breusch Rasmussen. 2025. Sweden's Areim secures $481 million for sustainable data centres. https://www.reuters.com/technology/swedens-areim-secures-481-million-sustainable-data-centres-2025-03-05/. Reuters.

[6] Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models. *EPJ Data Science* 13, 1 (2024), 41. doi:10.1140/epjds/s13688-024-00481-2 Open Access.

[7] Cécile Christiaen, Phil Lockwood, Andrew Jackman, and Ben Caldecott. 2025. Location, location, location: asset location data sources for nature-related financial risk analysis. *Current Opinion in Environmental Sustainability* 74 (2025), 101527. doi:10.1016/j.cosust.2025.101527

[8] Data Center Map. 2025. *Data Center Map.* https://www.datacentermap.com/ Accessed: 2025-06-26.

[9] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: An Evaluation Framework for Retrieval-Augmented Generation. *arXiv preprint arXiv:2309.15217* (2023). https://arxiv.org/abs/2309.15217

[10] Javier Farfan and Antje Lohrmann. 2023. Gone with the Clouds: Estimating the Electricity and Water Footprint of Digital Data Services in Europe. *Energy Conversion and Management* 290 (2023), 117225. doi:10.1016/j.enconman.2023.117225

[11] Global Energy Monitor. 2024. Global Integrated Power Tracker. https://globalenergymonitor.org/projects/global-integrated-power-tracker/. Accessed 2025-06-20.

[12] Goldman Sachs Research. 2025. AI to Drive 165% Increase in Data Center Power Demand by 2030. https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030. Accessed: 2025-05-13.

[13] G. Guidi, F. Dominici, J. Gilmour, K. Butler, E. Bell, S. Delaney, and F. J. Bargagli-Stoffi. 2024. Environmental Burden of United States Data Centers in the Artificial Intelligence Era. (2024). In press.

[14] Camilla Hodgson. 2024. US tech groups' water consumption soars in 'data centre alley'. *Financial Times* (18 August 2024). https://www.ft.com/content/1d468bd2-6712-4cdd-ac71-21e0ace2d048 Accessed via email link.

[15] Artifex Software Inc. 2024. PyMuPDF (fitz). GitHub Repository. https://github.com/pymupdf/PyMuPDF.

[16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[18] Na Lei, Jingyi Lu, Zhi Cheng, Zongyuan Cao, Arman Shehabi, and Eric Masanet. 2023. Geospatial Assessment of Water Footprints for Hyperscale Data Centers in the United States. In *Journal of Physics: Conference Series*, Vol. 2600. 172003. doi:10.1088/1742-6596/2600/17/172003

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[20] Fu-Hung M. Liu, Karen P. Y. Lai, Benjamin Seah, and Winston T. L. Chow. 2025. Decarbonising digital infrastructure and urban sustainability in the case of data centres. *npj Urban Sustainability* 5, 1 (2025), 15. doi:10.1038/s42949-025-00203-1

[21] Attracta Mooney. 2024. Data centres must work 'within climate limits', says Irish minister. *Financial Times* (17 September 2024). https://www.ft.com/content/21cbc670-9db0-4f94-9832-7e06b880f944

[22] David Mytton. 2021. Data Centre Water Consumption. *npj Clean Water* 4, 1 (2021), 11. doi:10.1038/s41545-021-00101-w

[23] Maciej P Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* 15, 1 (2024), 1569.

[24] Dominika Izabela Ptach, Deborah Andrews, and Simon P. Philbin. 2023. Sustainable Development Goals, Circularity and the Data Centre Industry: a Review of Real-world Challenges in a Rapidly Expanding Sector. In *The Circular Economy: Meeting Sustainable Development Goals*, Sadhan Kumar Ghosh and Gev Eduljee (Eds.). Issues in Environmental Science and Technology, Vol. 51. Royal Society of Chemistry, 252–285. doi:10.1039/9781837671984-00252 Accessed: 2025-07-03.

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[26] Violeta Rozite, Elisa Bertoli, and Ben Reidenbach. 2023. Data Centres and Data Transmission Networks. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks. Accessed: 2025-05-13.

[27] Arman Shehabi, Eric Masanet, Lynn Price, Arpad Horvath, and William W. Nazaroff. 2011. Data center design and location: Consequences for electricity use and greenhouse-gas emissions. *Building and Environment* 46, 5 (2011), 990–998. doi:10.1016/j.buildenv.2010.10.023

[28] Arman Shehabi, S. J. Smith, Alan Hubbard, April Newkirk, Na Lei, Md Asif Bin Siddik, Brian Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. *2024 United States Data Center Energy Usage Report.* Technical Report. Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf Accessed: 2025-05-13.

[29] Md Asif Bin Siddik, Arman Shehabi, and Landon Marston. 2021. The Environmental Footprint of Data Centers in the United States. *Environmental Research Letters* 16, 6 (2021), 064017. doi:10.1088/1748-9326/abfba1

[30] Jeremy Singer-Vine. 2024. pdfplumber. GitHub Repository. https://github.com/jsvine/pdfplumber.

[31] Spatial Finance Initiative. 2024. GeoAsset Databases. https://www.cgfi.ac.uk/spatial-finance-initiative/geoasset-project/geoasset-databases/. Accessed 2025-06-20.

[32] Stockholm Environment Institute and Global Canopy. 2024. Trase: Intelligence for sustainable trade. https://trase.earth/. Accessed 2025-06-20.

[33] Chroma Team. 2023. Chroma: The AI-native open-source embedding database. https://github.com/chroma-core/chroma. Accessed: 2025-06-26.

[34] Ollama Team. 2023. Ollama: Run large language models locally. https://ollama.com. Accessed: 2025-07-03.

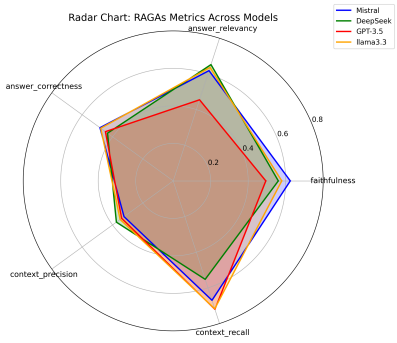
[35] Patrick Temple-West. 2025. Big Tech under pressure to act on data centres' thirst for water. *Financial Times* (21 March 2025). https://www.ft.com/content/65fff689-bd47-4c15-bdb8-083e5ccd84dc

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

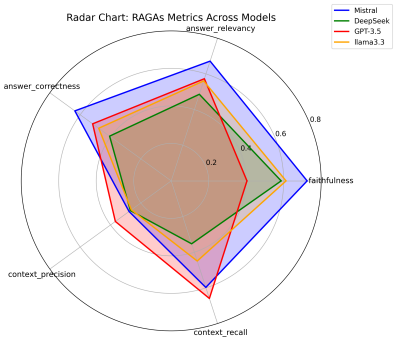
[37] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[38] WWF. 2025. *WWF Risk Filter Suite.* https://riskfilter.org/ Accessed: 2025-06-26.

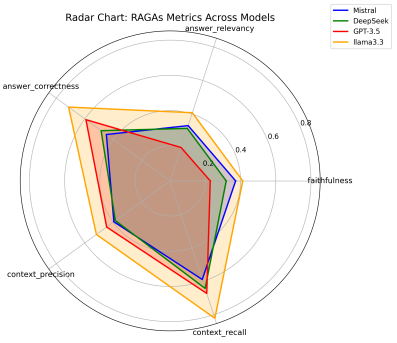
A Appendix



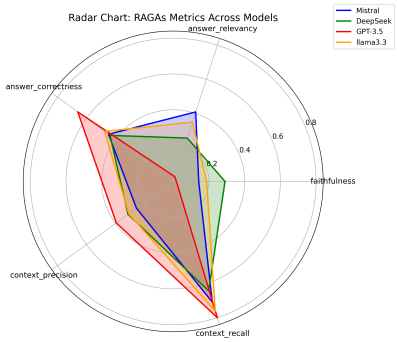
(a) Data Centre Name



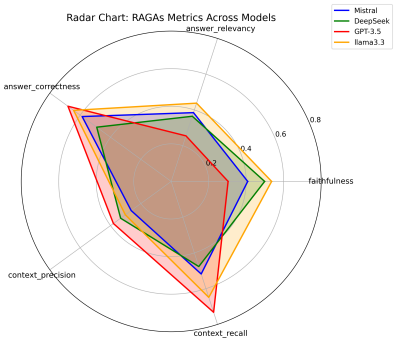
(b) Security System in use



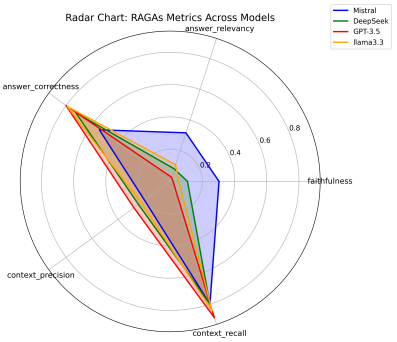
(c) Address



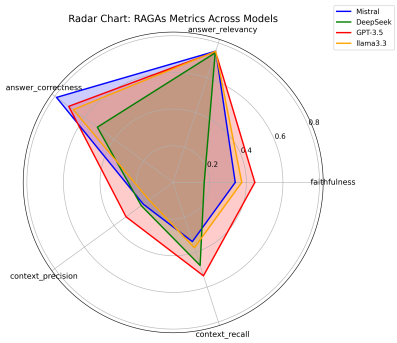
(d) Building Total Whitespace



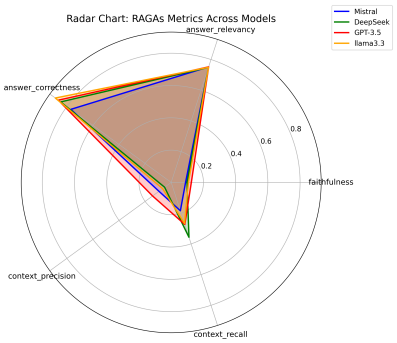
(e) Colocation Space



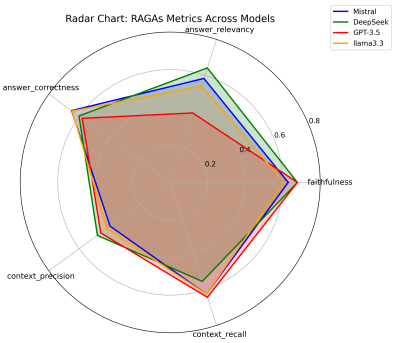
(f) Total Critical Power



(g) ISO 27001 Certified



(h) ISO 14001 Certified



(i) Operator

Figure A1: Radar charts comparing model performance across five RAGAS metrics for each attribute.

Table A1: Attributes in the Structured Data Centre Dataset

Attribute	Description
Data centre Name	Name of the asset
city	The city where the asset is located
state	The state where the asset is located
country	The country where the asset is located
region	Region in which the asset is located
latitude	Latitude for the geolocation of the asset (based on WGS84 (EPSG:4326))
longitude	Longitude for the geolocation of the asset (based on WGS84 (EPSG:4326))
status	The operating status of the asset
year	Year the facility started operations or is planning to start operations
facility_type	The type of data centre (bare metal, colocation, enterprise, hyperscaler)
floor_space	Floor space of the facility (sq ft)
power_system	Power system in use
renewables_onsite	Renewawble energy power generation capacity available on site
renewables_certificates	Renewable power generation certifications procured for the facility
incoming_power	Total incoming power
IT_power	Total IT power
PUE	Power usage efficiency
lower_carbon_innovations	Low carbon technology innovations
cooling_technologies	Cooling system in use
annual_water	Annual water usage (liters)
equipment_water	IT equipment energy usage (kWh)
WUE	Water usage efficiency/effectiveness (L/kWh)
tier_design	Uptime certification (Tier I, Tier II, Tier III)
certifications	Any certifications in place for this facility
clients	Reported clients of the facility/company
sustainability_policy	Does the operator have any sustainability policies in place (yes/no)
green_building_standards	Green building standards in place if any
energy_standards	Energy standards in place if any
water_efficiency_standards	Water efficiency standards in place if any
security_standards	Security standards in place if any
security_system	Security system in use
owner_name	Name of the owner/operator of the facility
owner_country	Country where the owner/operator headquarters are located
owner_city	City where the owner/operator headquarters are located
owner_source	Source reporting the ownership link between the facility and owner