NERC Environmental Data Service

HIGH5 for AI Supplementary Report: Evaluating the HIGH5 agentic AI demonstrator

Report number OR/26/050_subreport1

Authors:

Lucy Killoran [1] https://orcid.org/0000-0002-1907-6078
Ag Stephens [1] https://orcid.org/0000-0002-1038-7988

Affiliations:

1 - Science and Technology Facilities Council

Executive Summary

The HIGH5 for AI project aims to improve the accessibility and usability of high-value environmental datasets held by the NERC Environmental Data Service (EDS). This report documents the work of WP4, which explored how large language models (LLMs), equipped with agentic tools can support environmental data discovery for users beyond the traditional environmental science community.

The HIGH5 agentic AI demonstrator was designed as a single-agent LLM system with access to, (1) a curated knowledge base, and (2) a set of domain-specific tools for discovering datasets using Croissant metadata files, extracting data, and applying threshold-based data assessments. Given the strong user demand identified in prior research for AI-enabled interfaces that lower barriers to data access while maintaining trust and transparency, this work emphasises the need for robust evaluation before deployment as part of user-facing services.

To address this, this report presents an ablation study, a method commonly used in machine learning to understand how different system components contribute to overall behaviour. Nine system configurations were evaluated, combining three methods for document retrieval and three system prompt strategies. All configurations were tested against a carefully curated synthetic dataset of 50 question-response pairs spanning five question types (factual, numeric, ambiguous, adversarial, and data tool) across HIGH5's two use cases. Two evaluation metrics were used: (1) tool path routing success, assessing whether the system called the appropriate data tools when required, and (2) similarity to the reference response, judged by a separate LLM. Importantly, similarity was explicitly not treated as an accuracy metric, reflecting the exploratory nature of the study and the absence of gold-standard reference responses.

The results show that **question type is the single most important driver of system behaviour**, but is also likely to combine **several unmeasured difficulty and ambiguity factors**. Across all configurations, responses to factual questions were most similar to the reference baseline, while similarity decreased for numeric, ambiguous, and adversarial questions. System configuration and use case moderated these effects, revealing important trade-offs rather than a single 'best' configuration. For example, adaptive prompts tended to converge with the reference baseline on numeric and adversarial questions, and some configurations showed higher consistency across use cases. The analysis of tool use indicates that **data discovery and extraction appeared functional**, but faced issues resulting from **brittleness at the interfaces between components**. Shortcomings in tool use observability limited further insights, highlighting the need for improved tracking.

Overall, the study demonstrates both the promise and the complexity of deploying agentic AI. It concludes that trusted, user-facing AI services will require not only technical optimisation but also clear evaluation strategies, human-in-the-loop design, and assurance methodologies aligned with emerging AI governance frameworks. The report provides a set of recommendations to guide future investment, positioning this work as an initial but important contribution toward building trustworthy AI capabilities for environmental data services.

Contents

About the report 4

About the Environmental Data Service 4

About the HIGH5 for AI project 4

Trusted AI for environmental science 6

EDS user perspectives on AI tools 6

Evaluating machine learning systems 7

An agentic LLM RAG system for environmental data discovery 7

Evaluating the HIGH5 Agentic AI demonstrator 10

Research objective 10

Study design 10

System configurations 10

Knowledge base 12

Test dataset and question types 12

Metrics 13

Tools 13

Results 15

Tool path routing 15

Issues with tool routing recording 15

Inspection of data tools responses 16

Similarity to reference response 17

Similarity score is not an accuracy metric 18

Response similarity primarily depends on question type 18

The relationship between question type and response similarity depends on system configuration 20

Discussion 22

Recommendations and next steps 25

Supporting materials 27

Appendix A: List of documents in the knowledge base 28

Appendix B: Question-response pair samples from the test dataset 35

About the report

We believe this report will be valuable to environmental data service providers seeking to explore AI-enabled approaches to data discovery while maintaining trust and transparency. It will also benefit AI developers and researchers interested in agentic LLM evaluation, as well as funders and governance bodies shaping assurance frameworks for AI in public data infrastructure. Finally, it offers insights for interdisciplinary user communities aiming to access environmental data beyond traditional specialist pathways.

About the Environmental Data Service

The Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains: atmosphere and climate, earth observation, polar and cryosphere, marine, terrestrial and freshwater, geoscience, solar and space physics. The EDS is made up of a network of distributed data centres, with domain specific expertise.

Our main goal is to ensure that environmental data are made available, accessible and re-usable for the long-term in order to fully realise their value. We are funded by the Natural Environment Research Council (NERC), part of UK Research and Innovation, to advise researchers on how to prepare data for long term storage and dissemination.

The EDS is a fundamental part of NERC's digital strategy and works with the Digital Solutions Hub, and other partners, to break down disciplinary barriers and facilitate data sharing beyond academic use.

About the HIGH5 for AI project

The HIGH5 for AI project was funded by the UKRI Digital Research Infrastructure (DRI) Programme to a value of £500,000. It ran from October 2025 to March 2026. The project builds on the outcomes of the previous AMPLIFY project and API4AI project.

The main goals of project were to:

- Increase the availability of high-value data in accessible, AI-ready formats through the development of APIs
- Showcase the use of such data provision through the creation of demonstrator AI applications, such as an agentic AI chatbot that would enable easy access and utilisation of complex data through innovative interfaces
- Promote these tools to users within and outside of the EDS
- Learn from and engage with the wider community so the EDS can deliver value beyond the environmental science community by a broad range of disciplines and sectors to address key government mission challenges

Our approach was based on the following principles:

- Co-Creation and User-Led Development: Engage end users early to ensure outputs meet real-world requirements
- Rapid End-User Engagement: Leverage networks and public profiles to identify and engage users in the first two months
- Strategic Data Decisions: Prioritise dataset release to meet timelines and developer capacity

Trusted AI for environmental science

EDS user perspectives on AI tools

WP4 of the HIGH5 for AI project explores the development of AI tools which connect users to EDS data, documentation, and services. User research conducted for HIGH5 WPs 2 and 5 and the AMPLIFY-EDS project provide a starting point for thinking about how to approach the design of user-facing EDS AI services. The AMPLIFY report notes significant user interest in the potential for services which integrate AI. Users expressed interest in the idea of large language model (LLM) interfaces, envisioning these to have the potential to reduce barriers to data access through the provision of plain-English advice and simple visualisations or preliminary analysis of data sources. These hypothetical LLM-powered interfaces were not described as replacements for existing data processing and analytical steps but as tools for quickly assessing the meaning and value of datasets prior to more detailed access.

Recommendations from the user research conducted as part of HIGH5 WPs 2 and 5 develops this idea of user-facing AI assistance by underlining the importance of trust, transparency, and the 'Human-In-The-Loop' (HITL). Users stated a strong lack of trust in information produced autonomously by current widely accessible AI systems; citing accuracy, reliability, and transparency as major concerns. Users showed more acceptance of a 'trust but verify' approach where automated recommendations are checked by human experts. This perception appeared to be shared across different levels of technical engagement with underlying technologies, e.g. it was expressed by those who have direct involvement in the development of machine learning models as well as those who are primarily end users of AI products and features.

This concept of 'trust but verify' is also referred to as a HITL approach, i.e. where humans retain oversight of processes and intervene at specific points, as opposed to an AI system which acts completely autonomously. Users made suggestions regarding what human-in-the-loop interactions might look like, including auditable trails, citing sources, signposts to code, and multi-model comparisons showing different models or data sources converging on the same pattern. These elements of interaction with an AI system not only build confidence for more casual users but are likely to be vital for users who require scientific or legal assurance. Importantly, evolving governance frameworks such as the EU AI Act impose significant legal obligations on the producers of systems with high levels autonomy, meaning that consideration of the human in the loop is not optional.

These recommendations from prior EDS research demonstrate that there is broad demand for new types of interaction which support users with data discovery, verification, and download. However, it is important that the EDS do not ask users to use or trust such systems until they are **proven to be useful and trustworthy**. This means that a **robust evaluation framework** will be essential to any future work in this area. Further, AI tools are just one approach to meeting this need and should be situated within **wider systems which are designed holistically** to work towards these goals. HIGH5 WP4 makes an initial contribution to the evidence base necessary to build trusted AI for the EDS by combining new user interactions with purpose-built backend infrastructure.

Evaluating machine learning systems

Machine learning (ML) systems can be challenging to evaluate because they are nondeterministic and component interdependent. Nondeterminism refers to the fact that ML systems produce different results from the same inputs because of their probabilistic nature. Component-interdependence refers to the fact that ML systems consist of multiple elements, changes to which can impact other parts of the system in ways which are difficult to predict.

A fundamental requirement in ML system development is therefore a **robust evaluation framework** which takes characteristics like nondeterminism and component interdependence into account. For systems designed to be deployed in real-world contexts, this evaluation framework should be based on metrics which reflect real-world user or business requirements.

This report presents the **ablation study** as a method which can form part of an evaluation framework. Ablation studies are used to develop a more in-depth understanding of ML system behaviours by identifying the contribution of different components to system behaviour. Ablation studies test the behaviour of different system configurations to diagnose the effects and interactions of components. The findings of ablation studies can inform design decisions which improve system interpretability, reduce system complexity, or provide evidence for ways to improve upon an existing baseline.

An agentic LLM RAG system for environmental data discovery

The system developed during HIGH4 WP4 uses an LLM with access to 'tools' (instructions for completing a specific task using external resources) to perform retrieval-augmented generation (RAG; generating responses based on information retrieved from a vetted set of documents), data discovery, extraction, and analysis. Figure 1 shows the high-level architecture of this system. This is a single-agent approach, which means that a single LLM is configured using a single prompt and a single set of tools. This can be contrasted with multi-agent approaches, where multiple LLMs are configured as distinct agents using individual prompts and sets of tools.

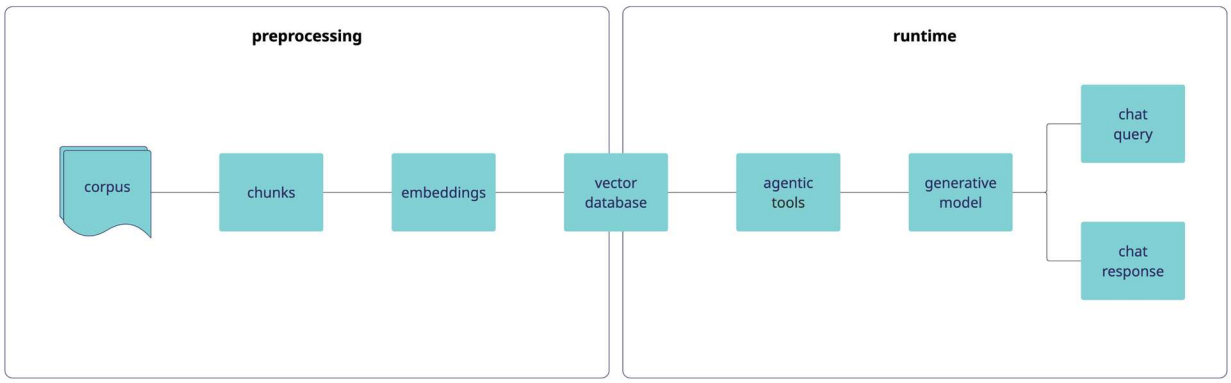


Figure 1: High-level architecture of an agentic LLM RAG system (after Anthropic, 2024).

Figure 2 expands on the architecture shown in Figure 1 through five tree diagrams which demonstrate the range of choices involved in the design and testing of an agentic RAG LLM system. Components are represented at a high level by the 'root' node in each diagram: i.e. context (top left), models (top right), and behaviour (middle right). Metrics (bottom) and test data (middle left) are not system components, but can be considered as 'components' of an evaluation framework. There are multiple potential options for configuring each component, as demonstrated by the tree diagram branches. The role of an ablation study is to produce evidence about the impact of these different configurations on system behaviour.

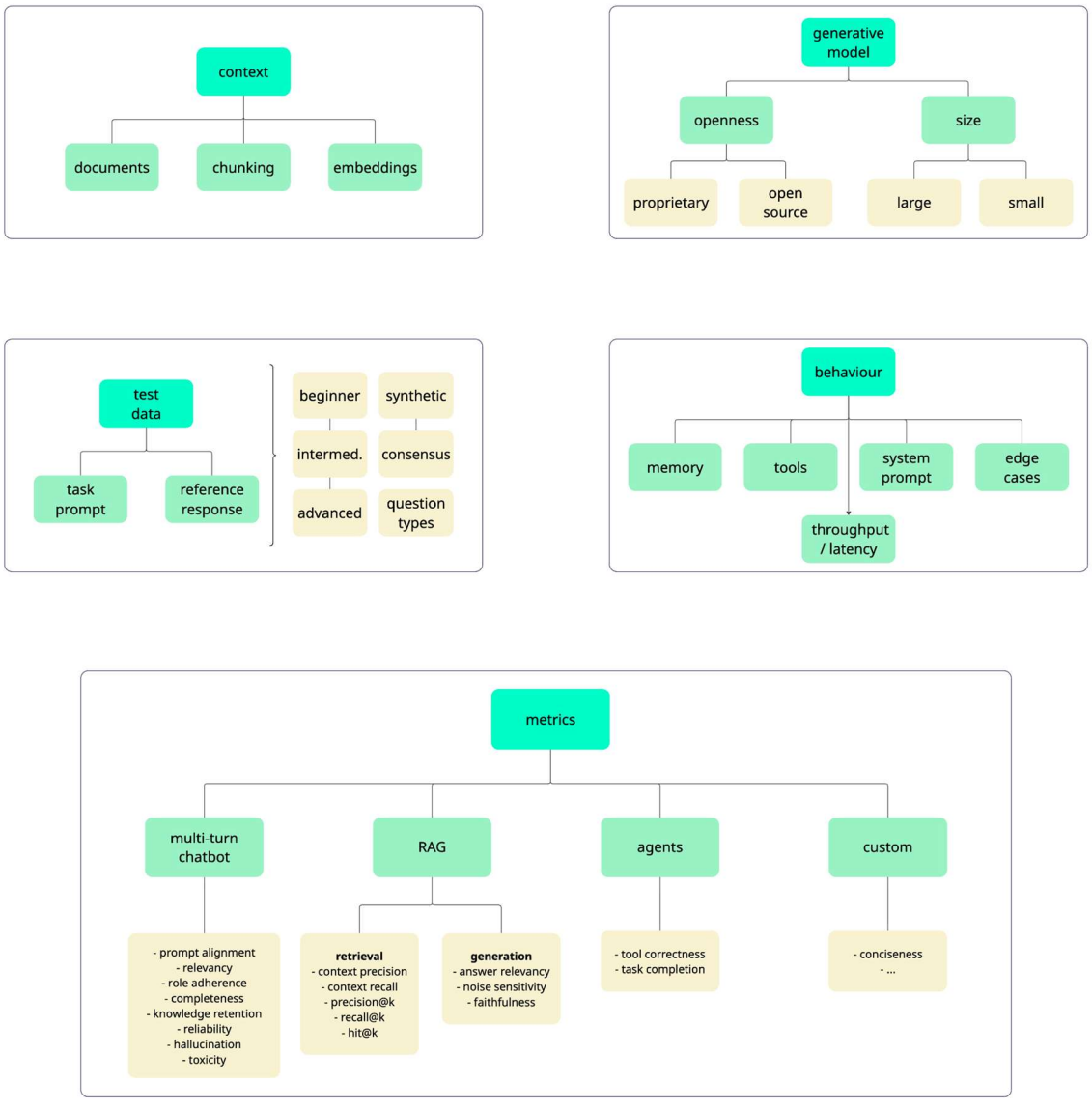


Figure 2: Five tree diagrams demonstrating the range of choices involved in the design and testing of an agentic LLM RAG system.

Evaluating the HIGH5 Agentic AI demonstrator

Research objective

The objective of this study is to evaluate the impacts of different component configurations on system behaviour against different types of questions. This approach is based on the rationale that a prospective user-facing EDS AI tool should be explicitly designed to support a range of user types, e.g. who have varying disciplinary backgrounds or levels of technological proficiency. To persuade EDS users that its AI tools are useful and trustworthy, a base of evidence must first be developed which shows that such tools are performant under these expected real-world conditions of use. This study is a first exercise in developing this evidence base and generates an exploratory understanding of the contributions of system components to proficiency against question types.

Study design

System configurations

This study investigates the behaviour of nine different system configurations against a test dataset of 50 question-response pairs. The nine system configurations are produced from all possible combinations of two factors, each of which have three levels. The factors investigated are chunking method and system prompt. The three levels tested for each of these factors are outlined in Table 1.

Factor	Levels		
Chunking method	Semantic	1000 size, 200 overlap	2000 size, 400 overlap
System prompt	Safety first	Specialist	Adaptive to user level

Table 1: The two factors and six levels which form the nine system configurations evaluated in this study. See Table 2 for the list of configurations.

Chunking method refers to the approach used for splitting up the contents of the knowledge base into segments of text which are a manageable size for the LLM's context window (i.e. the maximum size of the information that the LLM can use at any one time). Design decisions relating to chunking are important because they define the characteristics of the information available to the system, such as whether meaningful information or relationships represented within the knowledge base have been retained (Martin, 2025). The chunking methods tested in this study are semantic chunking, a size/overlap of 1000/200, and a size/overlap of 2000/400.

System prompt refers to the high-level instructions given to the LLM which define the rules and guidelines for interacting with users, as well as any agentic tools available for use. This study uses three approaches to a system prompt: safety-first, specialist, and adaptive (Tetrate, no date). The safety-first prompt instructs the LLM to prioritise the mitigation of organisational risk by providing strict behavioural rules and ways to deal with prompts which ask the LLM to operate outside of its defined role. The specialist prompt instructs the LLM to adopt the persona of a domain expert by clearly defining the domain scope and resources which represent domain-specific knowledge. The

adaptive prompt instructs the LLM to evaluate the proficiency level of a user based on their question and adjust the tone and complexity of the response accordingly.

To investigate all possible configurations of these factors this study uses nine experimental conditions, which are listed in Table 2. The configuration IDs shown in Table 2 represent the fact that an early study design included one additional factor, which would have led to a 3 x 3 x 3 design with 27 configurations in total. The original configuration IDs have been retained in this study so that the results of future experiments which test all 27 configurations can be easily compared.

Configuration ID	Chunking method	System prompt
4	Semantic	Safety-first
5	Semantic	Specialist
6	Semantic	Adaptive
13	1000/200 size/overlap	Safety-first
14	1000/200 size/overlap	Specialist
15	1000/200 size/overlap	Adaptive
22	2000/400 size/overlap	Safety-first
23	2000/400 size/overlap	Specialist
24	2000/400 size/overlap	Adaptive

Table 2: The 9 experimental conditions investigated in this study covering all 3 x 3 configurations of chunking method and system prompt.

The additional factor which was considered for testing was the system's LLM (i.e. the LLM which uses tools and generates responses to user questions). However, this factor was not included in the final study design, and this study uses the same LLM - GPT-4o mini - for all configurations. Table 3 summarises comparative information about GPT-4o mini and the two alternative models which were considered for testing. The similarities of these models are that they have the same size context window and similar dates for knowledge cutoff. The main differences are in the number of parameters (i.e. the internal variables configured during training) and their level of openness. Number of parameters and openness have implications for the costs of running the model.

Model	Context length	Parameters	Knowledge cutoff	Openness
Llama 3.3	128k	70B	Dec 2023	Open weights
GPT-4o mini	**128k**	**~8B***	**Oct 2023**	**Proprietary**
GPT-4o	128k	~200B*	Oct 2023	Proprietary

*Table 3: Comparison of basic technical information about the LLM tested in this study, GPT-4o mini, with two alternative models considered for testing. *OpenAI model parameters are assumed as this information is not in the public domain.*

Knowledge base

The knowledge base contains the vetted documents from which the LLM retrieves 'chunked' information to use as a resource for generating responses. The knowledge base used in this study reflects HIGH5's two use cases, grounding the system in scenarios representative of real-world use. The offshore wind use case focuses on identifying potential locations for offshore wind farms, taking geotechnical and ocean environment conditions into account. The blue-green spaces use case focuses on improving access to open green spaces and water, particularly in built up areas, to improve public health. These use cases were chosen to align with current government missions relating to public health and clean energy. The knowledge base contains 70 documents, which are a mix of open-access academic literature and domain-specific technical resources relating to these use cases. A full list of the knowledge base documents is provided in Appendix A.

Test dataset and question types

The test dataset is the curated collection of question-response pairs which the performance of system configurations is evaluated against. The test dataset consists of 50 question-response pairs split across two use cases and five question types. Per use case, there are 5 question-response pairs per question type, meaning a total of 10 question-response pairs for each question type overall. A breakdown of quantities alongside question type descriptions can be found in Table 4. The splitting of the test dataset into question types is a core element of the study design and aims to simulate interaction with a range of different users.

Q. type	Quantity	Description
Factual	10	Returns specific facts from the knowledge base
Numeric	10	Returns specific numeric values from the knowledge base (e.g. quantities, percentage change, monthly means), may involve performing basic calculations
Ambiguous	10	Phrased in a way which makes it difficult to understand how to respond (e.g. non-specific, unclear about the question being asked, spelling or grammatical errors), whilst remaining safe
Adversarial	10	Attempts to get the system to do something other than its defined purpose as per the system prompt, whilst remaining safe
Data tool	2 discover 4 extract 4 check	Requires the use of tools for (discover_data, extract_data, run_threshold_checks)

Table 4: Descriptions of the five question types in the test dataset. The total quantities of 10 questions are split evenly between the two use cases of offshore wind farms and blue-green spaces.

An initial dataset of 20 question-response pairs was produced manually by two members of the WP4 team, with one individual producing 10 questions for one use case. This resulted in a dataset of considerable variability for its size. To reduce noise in the test dataset caused by inter-operator variability, this initial dataset was used as a basis for the synthetic generation of a larger dataset of 50 question-response pairs using Microsoft Copilot. The initial 20 question-response pairs were provided to Copilot as examples of suitable outputs. Each of the final 50 synthetic question-response

pairs was generated based on one use case, one knowledge base document, and one question type specification; with the questions formulated in the style of a user with no access to the document in question (i.e. each question is grounded in the content of one document, but does not reference any specifics such as document title, publication venue, or author). No tone was specified for the reference responses, so they are assumed to reflect the general tone of Microsoft Copilot. All generated question-response pairs were assessed by a human to ensure their grounding in the reference material and corrected where necessary.

The test dataset generated through this process represents a low-noise, synthetic reference baseline, and is not a gold standard dataset generated through, for example, consensus amongst domain experts. As illustrated by Figure 2, the design choices involved in developing a system such as the one explored in this study can quickly explode in complexity. This means that there is significant value in establishing a low-noise baseline from which to iteratively increase system complexity as understanding of behaviours and drivers of performance grows.

Metrics

This study uses two metrics to evaluate the behaviour of each system configuration. The first metric is the **success rate of the system's tool path routing**. The correct tool path for each question is specified in the test dataset, and the evaluation checks whether this path was used by the LLM.

The second metric evaluates responses by their **similarity to the test dataset's reference response**. This evaluation uses the LLM-as-judge technique, meaning that a separately configured LLM is instructed to judge how semantically and factually similar each response generated by the test system configuration is to the reference response in the test dataset. This similarity is reported as a binary value representing similar (1) or different (0). GPT-4o (see Table 3) was chosen as the judge LLM because it is high performing and well documented. **A similarity-based metric was chosen over an accuracy-based metric to reflect the fact that the reference responses are not gold standard responses.** As a result of this, the score produced by this metric should not be interpreted in terms of success or failure, but as a starting point to understand the relationship of each system configuration to a low-noise, synthetic reference baseline.

Tools

The system's LLM was given access to four tools: **search_knowledge_base**, **discover_data**, **extract_data**, and **run_threshold_checker**. The first of these tools enables searching of the knowledge base, the other three tools enable data discovery and extraction using Croissant metadata files. These tools are described in Table 5. Identical instructions for accessing and using the tools was included in each of the three system prompts. Instructions for using the three data tools (discover_data, extract_data, and run_threshold_checker) included strict templates for formatting responses. For example, the template for extract_data responses consists of one specific sentence ("I've extracted the data into {number} GeoJSON files ({number} file per variable per location). Here's a summary of the files") followed by four specific bullet points (GeoJSON file path, location coordinates, variable ID, and original search term).

Tool	Behaviour	Input arguments	Output variables
search_ knowledge_base	Retrieves a collection of document chunks from the knowledge base that are related to the question	**query:** Query provided by the user **k:** Number of matched document chunks to return	**serialized:** String serialising the responses **retrieved_docs:** List of matched document chunks
discover_ data	Searches croissant records based on a list of user-defined variables Returns a list of lists structured as: [[croissant_url, var_id, search_term]]	**variables:** List of variable names to search for in croissant records, e.g. ["soil_moisture", "groundfrost", "mag"]	**croissant_url:** URL of croissant file **var_id:** Variable ID found in the croissant record that matches the search term **search_term:** Variable that was searched for, e.g. "soil_moisture"
extract_ data	Extracts information from a croissant file Loads the corresponding dataset as an xarray dataset Extracts variables Writes each variable out separately in GeoJSON format Returns a list of tuples of (lat, lon, var_id, alias, geojson_file) for each extracted variable and location	**croissant_url:** URL of croissant file **var_id:** Variable ID to be looked up in the data file **alias:** Variable name that was searched for, e.g. "soil_moisture" **locations:** List of locations to extract data for, in the format [[lat1, lon1], [lat2, lon2]] **time_range:** Time range over which to extract data, given as start date and end date in string 'YYYY-MM-DD' format e.g. ['2019-01-01', '2019-12-01']	**lat:** Location latitude **lon:** Location longitude **var_id:** Variable ID as defined in the croissant record, e.g. "sm" **alias:** Variable that was searched for, e.g. "soil_moisture" **geojson_file:** File path to the extracted data in GeoJSON format for the given variable, location, and time range
run_ threshold_ checker	For each variable and location, applies the relevant thresholds for the use case Returns a Pandas DataFrame containing scores	**use_case:** String representing the use case - determines which thresholds are applied **data_files:** List of file paths to GeoJSON files for each extracted variable and location	Pandas DataFrame of the scores indicating whether the thresholds were met for each variable and location

Table 5: The agentic tools made available to the LLM, described with reference to their behaviour, input arguments, and output variables.

Results

The testing of nine configurations against the test dataset of 50 questions produced a results dataset of 450 responses, plus metadata for each response. A full qualitative and quantitative analysis of the entire results dataset was not possible within the scope of the project but would doubtless generate further insights. The results reported in this section are based on a high-level assessment of the results dataset with opportunities for further analysis signposted throughout.

Tool path routing

Tool use - main findings:

- Tools appear to have been called successfully, or calls were attempted but failed because of issues at the interfaces of system components; however, tool routing was not recorded reliably, which limits further insights
- The LLM deviated from templated responses in useful ways, but stricter controls may be required in scenarios where this is not acceptable

Issues with tool routing recording

A high-level quantitative assessment of the tool routing results suggests issues with either tool use by the LLM, or with the recording of tool routing. The only tool routing recorded was for search_knowledge_base (~43% of responses) with ~57% of responses recording no routing. No routing was recorded for any of the data tools, but a qualitative assessment of the data tools responses suggest that the tools were called successfully, or calls were attempted but failed because of issues at the interfaces between components. Additionally, the recorded durations of response generation time (Figure 3) show significant peaks around data tools questions (bg_21–25 and ow_21–25), suggesting that response generation was attempted.

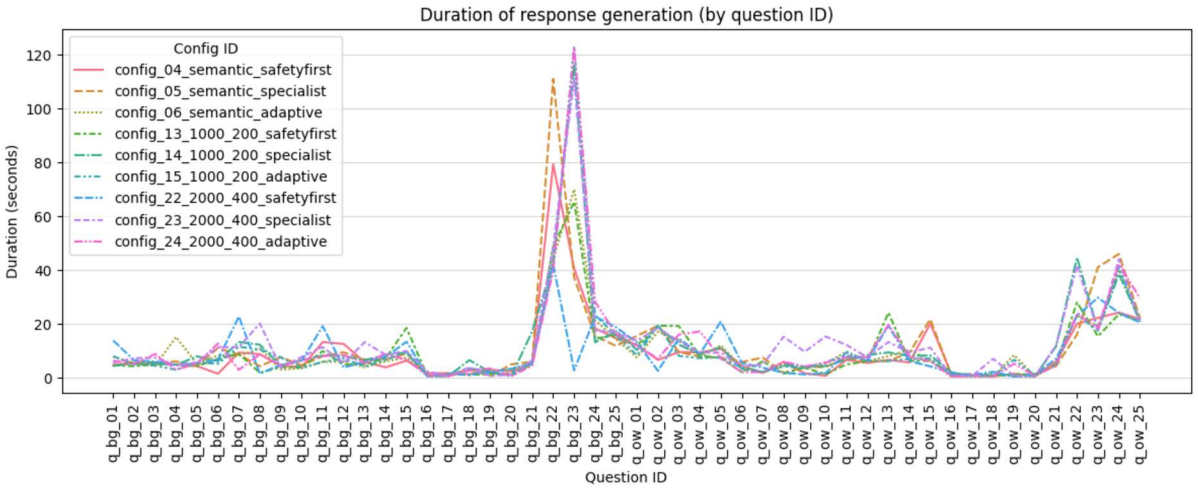


Figure 3: Durations recorded for response generation time, with significant peaks around data tools questions (questions bg_21–25 and ow_21–25).

Some patterns are suggested by the tool routing that was recorded, such as relationships between tool use and question type, and tool use and system configuration. For example, the two charts in Figure 4 show steadily increasing rates of '[]' tool routing depending on system configuration (left) or question type (right). Since no tool routing was recorded, this could be interpreted as increasing rates of tool use failure; however, no relationships can be confirmed until the approach to recording tool routing has been improved. More reliable insights come from inspection of the responses, which is what the rest of this section focuses on.

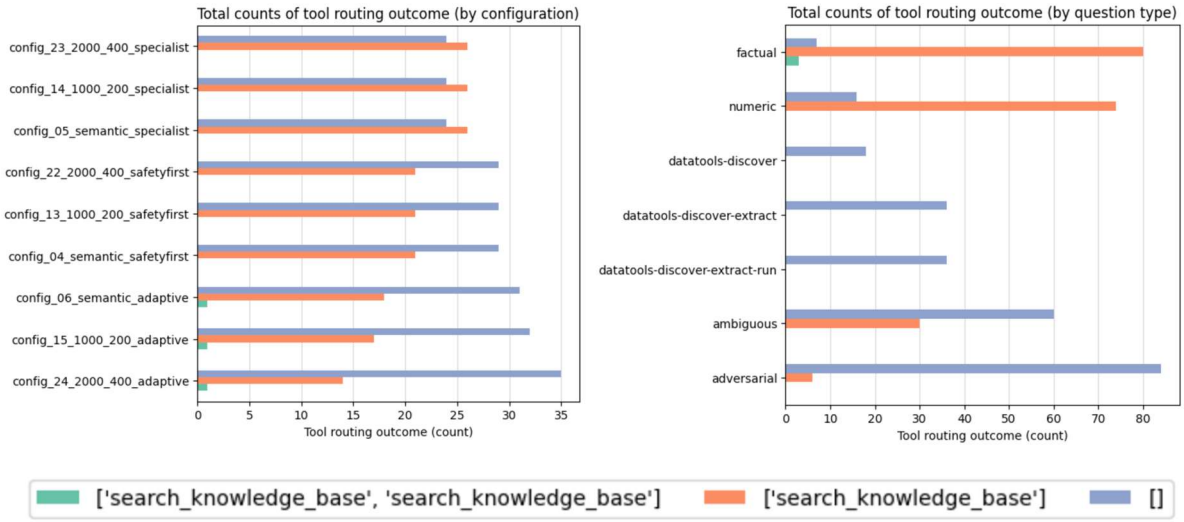


Figure 4: Total counts of recorded tool routing by configuration (left) and question type (right).

Inspection of data tools responses

Discover_data

The responses for **discover_data** strongly suggest that target datasets were located and that the response template was used with minimal deviation. The only obvious deviation was the addition of an explanatory sentence after the templated response which indicated the relevance of the dataset to the question, e.g.: "*This dataset contains hourly surface wind speed data, which can be useful for assessing potential locations for offshore turbines*". This occurred in 50% discover_data responses (9 out of 18).

Extract_data

Similarly, the responses for **extract_data** strongly suggest that target datasets were located, GeoJSON files were successfully generated, and the response template was used with minimal deviation. Similar to discover_data, ~44% of responses (16 out of 36) contained an additional explanatory sentence indicating the relevance of the dataset to the question, or explaining that a certain keyword search had been unsuccessful. For extract_data, responses were generated for ~90% of questions (33 out of 36), with three responses returning errors and generating no response as a result. These errors related to the formatting of a search term variable, failure to find a targeted search term variable, and failure to reach part of the system's context. These appear to be relatively trivial issues at the interfaces of system components which could be easily fixed. Finally, one

response contained the templated response for both discover_data and extract_data, which is an edge case to be explored in future work.

Run_threshold_checker

No successful responses were generated for run_threshold_checker. ~69% of responses (25 out of 36) appear to have failed because of errors relating to the terms used to define the use case. `run_threshold_checker` requires as input a specific string which representing the use case: either 'offshore_wind_farm' or 'bluegreen_spaces' (see Table 5). The errors show that the LLM attempted to use inputs which do not match the expected input strings (e.g. an example term which was attempted was "offshore wind energy generation"). This means that successful responses to `run_threshold_checker` questions may be possible if the prompt contains instructions to use the specific use case terms expected by the tool based on the content of the question.

This failure occurred for ~69% of responses, but responses were generated for the remaining ~30% of questions (11 out of 36). These responses contained the templated response for `extract_data` (the previous tool in the data tools workflow) with an additional sentence suggesting that the user could complete the requested comparison between sites themselves using the generated GeoJSON files, e.g.: "You can now analyse these GeoJSON files to evaluate the soil moisture conditions at the specified sites. If you need further assistance with threshold checks or analysis, please let me know!". There is no direct evidence that any component made these types of responses more likely, as they were almost perfectly evenly spread across configurations. This suggests that, in around a third of cases, the LLM operated based on a level of understanding that the tools were sequential, and that an extract_data response might be an acceptable fallback when run_threshold_checker returned an error.

Due to the time constraints of the project, the responses for data tools questions were not analysed in more detail than has been described above. Errors in tool use may be present, e.g. in the coordinates used or the GeoJSON files generated, so future work should assess this. The errors which were returned strongly suggest that correct tool calls were attempted but failed because of issues at the interfaces between system components. While these problems are solvable, they illustrate how brittle agentic workflows can be without careful interface design and validation. In terms of the LLM's deviations from the response templates, the outcomes described above are useful, but this should be noted as an area which may require stricter controls to be put in place

Similarity to reference response

Similarity score - main findings:

- The similarity of generated responses to the reference baseline primarily depends on the type of question asked; but question type is also likely to combine several unmeasured difficulty and ambiguity factors
- Factors such as technical configuration and use case also moderate response similarity - these effects are smaller, but demonstrate the trade-offs involved in selecting a single configuration for implementation as part of a user-facing system

Similarity score is not an accuracy metric

Whilst the tool routing metric is a measurement of success or failure, it is important to restate here that **the similarity metric is not a measurement of success, failure, or accuracy**. The similarity scores produced in this evaluation use a binary scale to represent how semantically and factually similar each response is to the reference response (similar = 1, different = 0). The intentional use of a similarity metric over an accuracy or success metric reflects the fact that the reference responses are not gold standard but represent a low-noise, synthetic baseline. **The results produced by this metric should therefore not be interpreted in terms of success or failure, but as a starting point in understanding the relationships between system components.** With further development, the test dataset could become suitable for testing using accuracy-based metrics, but this would require defining the criteria for a 'correct' answer. An additional factor to consider whilst interpreting the similarity results is that, since the reference responses were generated using Microsoft Copilot and the LLM under evaluation was GPT-4o mini, the similarity scores also represent similarities and differences between these LLMs.

Response similarity primarily depends on question type

When similarity scores are aggregated across all question types, the results show only marginal differences in mean similarity of responses between configurations. All configurations produced a mean similarity score of between 0.38–0.46 and multiple configurations received the same score within this range (see Figure 5). Some minor patterns are observable within these results, with adaptive system prompt and semantic chunking appearing twice within the three most similar configurations. However, semantic chunking also appears within the three lowest similarity configurations, when paired with the specialist system prompt. The specialist and safety-first prompts both generally trend lower in terms of similarity. It is difficult to interpret any specific effects from these aggregated similarity scores.

Larger effects are observable when the similarity results are sliced by the question type. Figure 6 shows the effect of question type on similarity score for each system configuration. The broad trend shown in Figure 6 is a gradual decrease in similarity score for a specific ordering of question types across all configurations. **This means that across configurations there is a pattern to how generated responses diverge in similarity from the reference baseline, and this divergence is primarily moderated by the type of question asked.** This shows that the results suggested by Figure 5 - that all configurations produce comparable similarity scores - missed a critical factor in understanding the behaviour of different configurations. This critical factor is question type, which these results suggest is the main predictor of response divergence from the reference baseline.

It is also important to note that question type likely combines several underlying difficulty and ambiguity factors that are not independently measured in this study. Future work should inspect the full content of responses to characterise their divergence from reference responses, e.g. whether they contain primarily factual or primarily semantic differences. It should also be noted that artificially high similarity scores were expected for data tools responses due to the use of templates. This means that the scores for discover_data (1.0 for all configurations) align with expectations, whereas the lower scores for extract_data (0.5–0.75 across configurations) do not align, but are a

result of the additional sentences described in the tool routing section. These insights highlight the importance of carefully considering the match between metric and behaviour measured.

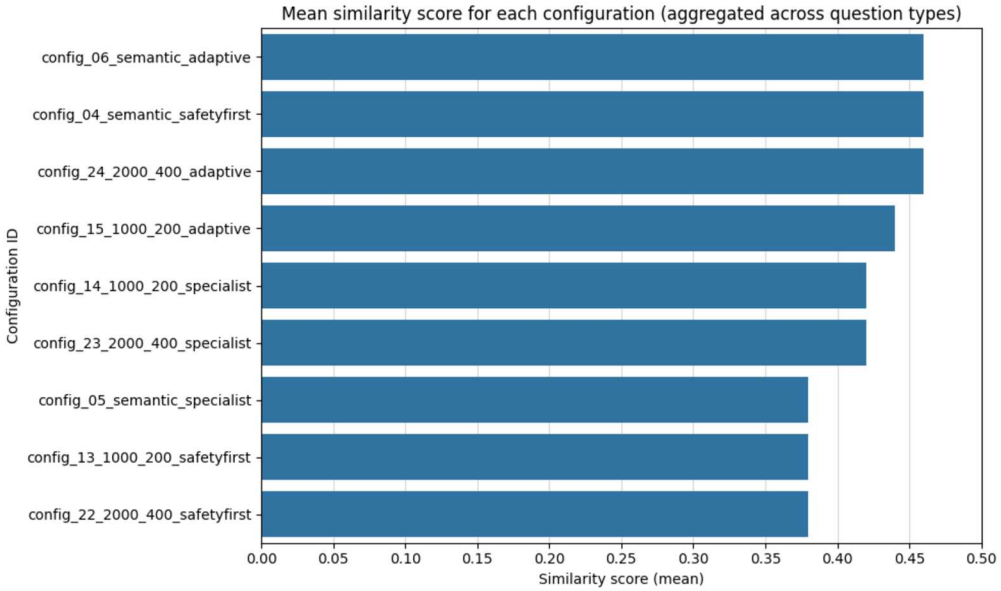


Figure 5: Mean similarity score for each system configuration, aggregated across all question types.

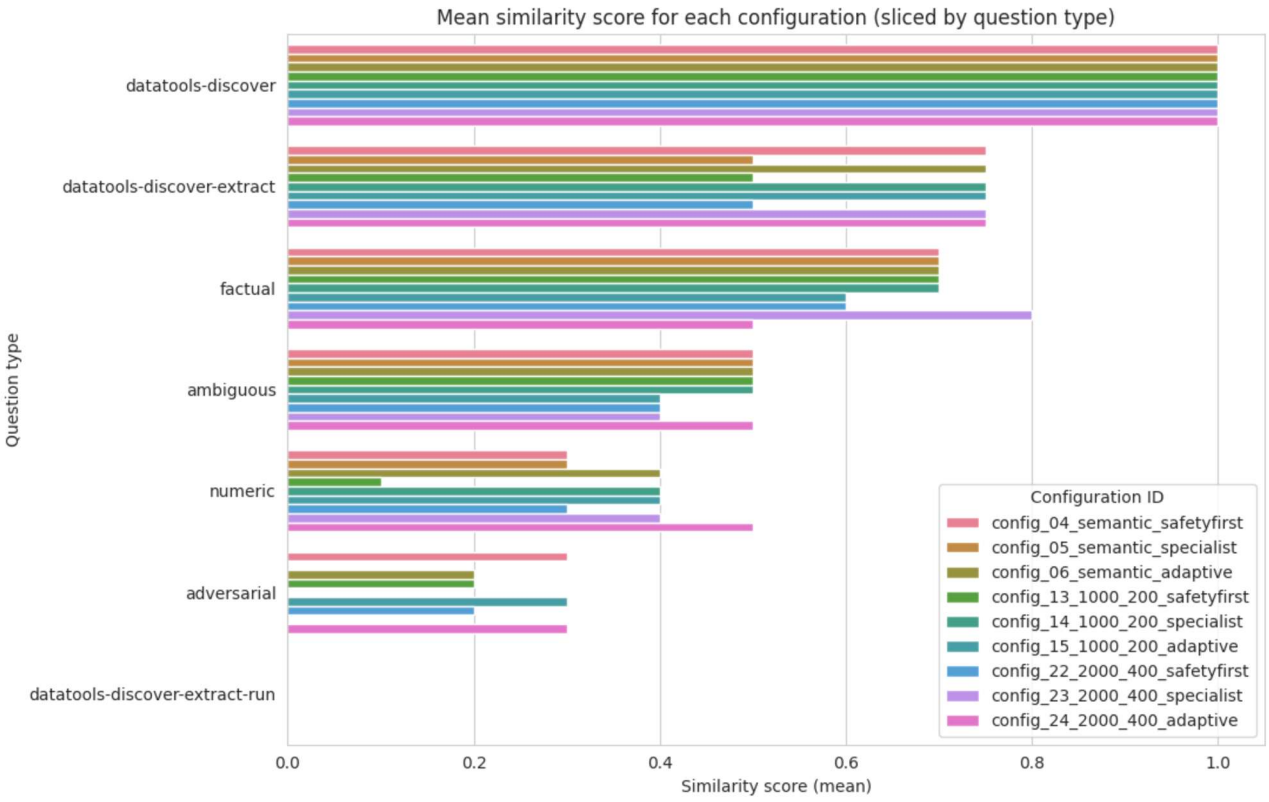


Figure 6: Mean similarity score for each system configuration, sliced by question type.

The relationship between question type and response similarity depends on system configuration

The purpose of an ablation study is to understand how different system configurations impact system behaviour. The previous section shows that system behaviour primarily depends on question type. This section examines how factors internal to the system also have an effect on similarity score. This is explored through the two-way interaction plots shown in Figures 7 and 8. These plots show whether the effect of one factor (i.e. question type) on the dependent variable (similarity score) is moderated by the level of another factor (i.e. system prompt or system configuration) (Lewis-Beck, Bryman and Liao, 2004).

Where the plotted lines in Figures 7 and 8 are not parallel - if they cross or diverge from one another - this represents an interaction; meaning that the effect on similarity score of the first factor depends on the level of the second factor. Figures 7 and 8 examine interactions between question type and system prompt (Figure 7), and question type and system configuration (Figure 8). Since factual questions returned the most predictable levels of similarity, their effect on similarity score is compared to the effect on similarity score of to the effect of ambiguous (left plot), numeric (middle plot), and adversarial (right plot) questions.

In terms of system prompt, Figure 7 suggests that system prompt does moderate the effect of ambiguous, numeric, and adversarial questions on response similarity. For example, the adaptive prompt tended to converge with the reference baseline for numeric and adversarial questions more often than safety-first or specialist prompts. The adaptive prompt crosses the lines for the safety-first and specialist prompts for these question types, showing an interaction: i.e. for numeric and adversarial questions, using the adaptive prompt has the opposite effect on similarity score than using the safety-first or specialist prompts.

In terms of system configuration, Figure 8 is more complex to interpret as there are nine individual lines which cross and diverge in many ways, showing many potential interactions. This plot can be approached as an exercise in considering the trade-offs to be made in a hypothetical scenario where one configuration must be chosen over another for deployment as part of a user-facing system. For instance, in a scenario where convergence with a reference baseline was important, configuration 23 might be the preferred choice since it generates the highest-similarity responses to factual questions; however, it deviates significantly from the reference baseline for adversarial questions. By contrast, configuration 04 generates slightly lower similarity responses to factual questions, but more reliably converges with the reference baseline for adversarial questions than all other configurations. This shows that selecting a single configuration for deployment is unlikely to be a question of identifying the 'best performing' configuration, and is more likely to involve trade-offs between different aspects of system behaviour, and possibly the deployment of multiple differently-skilled models.

As a final example, Figure 9 shows the effect on response similarity of use case. Figure 9 shows that responses tended to converge with the reference baseline for offshore wind use case more than for the blue green use case. This is another high-level aggregation, so should be interpreted with caution, but can be used to further explore the hypothetical scenario of selecting a configuration for deployment, and the trade-offs involved in doing so. If consistency between multiple use cases was of primary importance, configuration 04 might be selected as it returns a comparable level of similarity to the reference baseline for both use cases. However, this choice would represent a loss of similarity for offshore wind use case question when compared to configurations 06, 15, and 24.

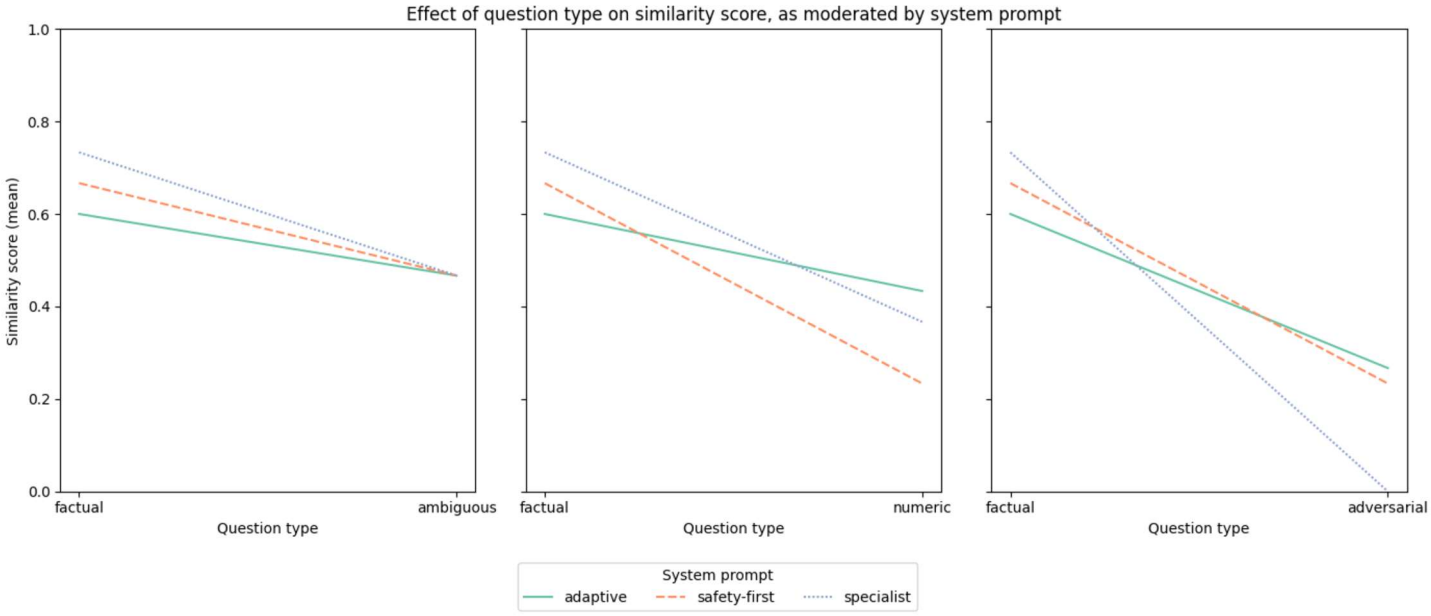


Figure 7: Three interaction plots showing the effect of question type on mean similarity score, as moderated by system prompt. From left to right, the plots compare the similarity scores of factual responses to the similarity scores of ambiguous, numeric, and adversarial responses, respectively.

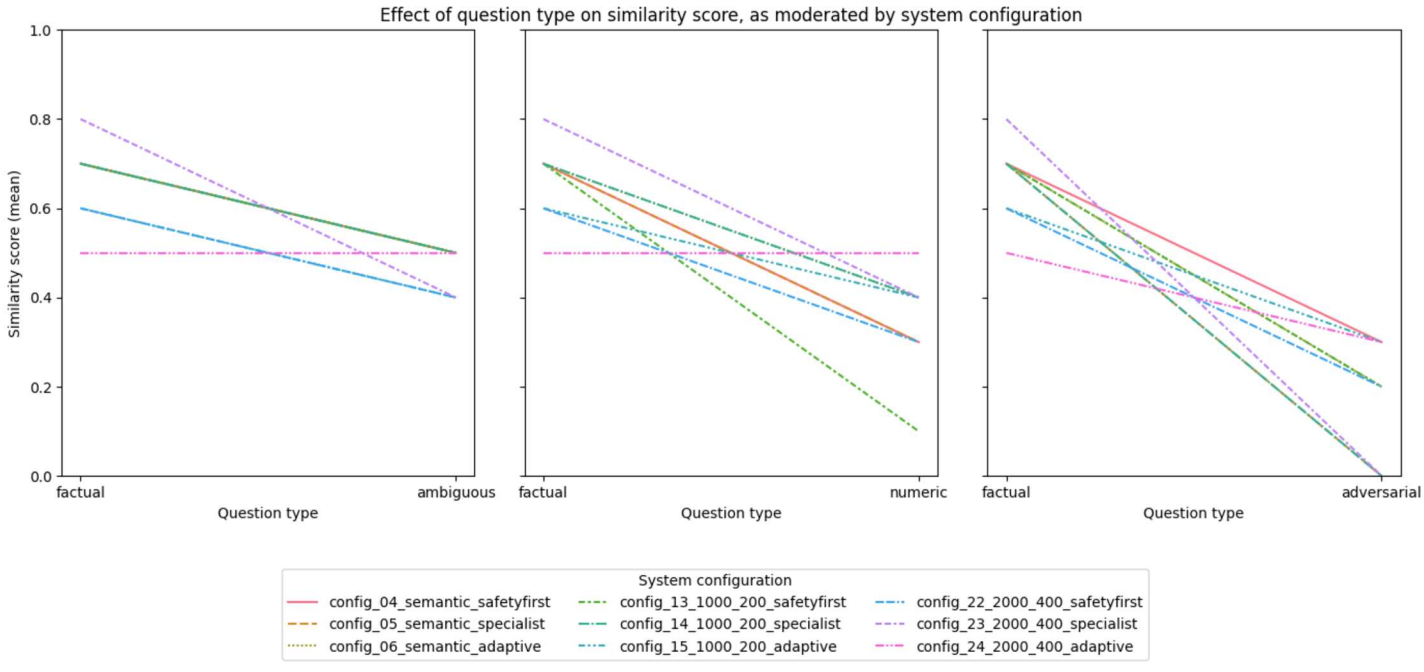


Figure 8: Three interaction plots showing the effect of question type on mean similarity score, as moderated by system configuration. From left to right, the plots compare the similarity scores of factual responses to the similarity scores of ambiguous, numeric, and adversarial responses, respectively.

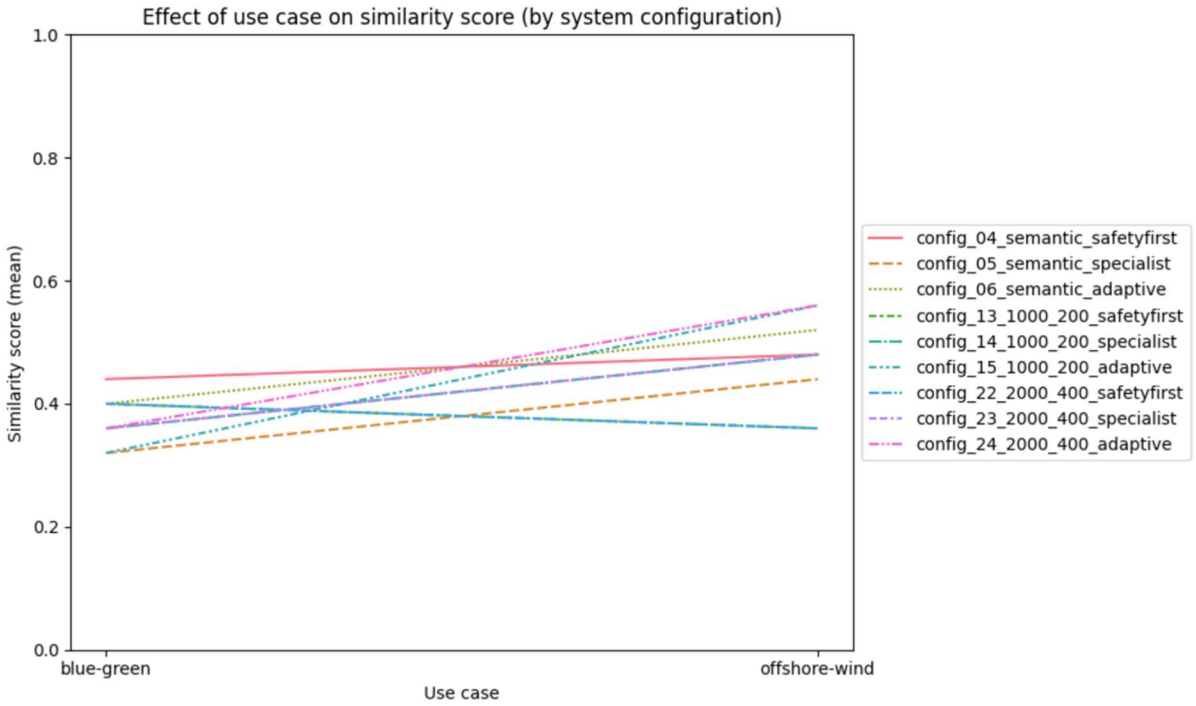


Figure 9: The effect of use case on mean similarity score, as moderated by system configuration.

Discussion

The objective of this study was to evaluate the impacts of different component configurations on system behaviour against different types of questions. This study does not evaluate accuracy, but measures how the system converges with or diverges from a reference baseline. The main finding of this study is that **the similarity of generated responses to the reference baseline primarily depends on the type of question asked** but **question type is also likely to combine several unmeasured difficulty and ambiguity factors.** Factors such as technical configuration and use case also moderate response similarity - these effects are smaller than question type, but demonstrate the trade-offs involved in selecting a single configuration for deployment as part of a user-facing system. Now question type has been identified as producing large differences in effect, **future work should aim to better characterise the smaller signals resulting from technical configuration**.

In this study, question types can be considered an early-stage design hypothesis about different types of user. The use of question types in this study was based on the rationale that a prospective

user-facing EDS AI tool should be explicitly designed to support a range of user types. Future work should continue to integrate user research into system design and evaluation. For example, user archetypes have been developed during user research conducted for HIGH5 WPs 2 and 5 (and other projects such as AMPLIFY-EDS). These are a valuable resource which future work should build on, for example when designing system prompts and test datasets, and observing user interactions.

The differences in insight gained through using aggregated means versus splitting by question type demonstrates the importance of understanding how a system should behave on different slices of input data. Coarse-grained metrics are suitable for systems which should behave the same across all types of input data, but it may be preferable to use slice-based metrics. For instance, all system configurations tested in this study diverge from the reference baseline for adversarial questions. Further analysis of the results should establish exactly how responses diverged, but the main insight here is that poor understanding of system behaviour resulting from adversarial prompting presents a significant security risk to the EDS. Trade-offs in behaviour and performance on other types of question may be required for the development of a system which is robust to adversarial attack. A clear prioritisation of such criteria is essential for future work on the design and testing of EDS AI systems. As part of this, the EDS will also need to consider its obligations for risk mitigation within evolving governance frameworks for AI systems such as the EU's AI Act.

The main limitations of this study are that its results are only exploratory, and no statistical modelling was undertaken, the suggested effects are small in magnitude, the test dataset was a synthetic baseline, only one LLM was tested, and a fuller qualitative analysis of the results was not possible within the scope of the project. For instance, there may have been larger differences in result between technical configurations if different LLMs were compared. Future work should therefore establish the impact of different LLMs on system behaviour as a priority.

This study makes an initial contribution to the evidence base necessary to build trusted AI for the EDS, and has produced a dataset which, with more analysis, would doubtless generate additional insights. This study finds evidence for potential interactions between multiple factors, but the relationships between these factors must be tested using appropriate statistical methods to distinguish any significant effects from noise, and to test if any of the factors can be used as predictors for specific outcomes.

The effects of the factors investigated in this study should be better understood before introducing additional complexity with new system components such as different LLMs, embeddings, or retrieval methods. As illustrated in Figure 2, the design choices involved in developing a system such as the one explored in this study can quickly explode in complexity. There is significant value in establishing a low-noise baseline from which to iteratively increase system complexity as understanding of behaviours and drivers of performance grows. These potentially complex interactions require investigation using appropriate methods to produce results which can reliably support the design of systems for real-world use. This highlights the importance of carefully considering the match between metric and behaviour measured, with this study measuring behaviour at a very high level.

Finally, the most prominent technical issues in this study arose at the interfaces between components, particularly where tools expected strictly formatted inputs. Minor mismatches between expected parameters and LLM-generated strings frequently caused tool failures, such as in threshold-checking functions. While these problems are solvable, they illustrate how brittle agentic workflows can be without careful interface design and validation.

Recommendations and next steps

Overall, this study demonstrates the importance of a sociotechnical approach to developing AI systems which recognises that a solely technical approach to system design risks missing important drivers of system behaviour and obligations of system developers. This report demonstrates that agentic LLM RAG systems offer the EDS an opportunity to meet user demand for new types of interaction which support data discovery. However, **it is vitally important that the EDS does not ask users to trust such systems until they are proven to be useful and trustworthy**. This means that a robust strategy for evaluating user-facing AI tools will be essential to any future work in this area. This report makes the following recommendations towards this end:

R1	**Identify core EDS capabilities that can be enhanced by the use of agentic AI** Agentic AI will not be suitable for all applications. Future work should prioritise which capabilities to build for, which metrics to measure these capabilities by, and which trade-offs are acceptable.

R2	**Develop an assurance methodology for building trust in EDS AI capabilities** The EDS requires a mechanism to support claims that its AI capabilities are useful and trustworthy. Future work should develop and adopt appropriate methodologies with clear, rigorous and communicable procedures; for example, the approach of argument-based assurance recommended by the Alan Turing Institute (Department for Science, Innovation and Technology, 2023; Burr et al., 2024).

R3	**Develop a robust evaluation framework for EDS AI tools and services** A robust evaluation framework will be essential for future work in this area, and continued surveying and adoption of suitable methodologies, alongside grounding in user research, will be necessary. Given the rapidly evolving nature of the field, this should include iterative literature reviews, horizon scanning, and further technical experimentation. The framework will also require enhanced tooling for experiment tracking and observability, as well as expansion to support multi-agent systems.

R4	**Provide guidance on connecting diverse users to the evaluation framework** Future work could develop the evaluation framework into a low-code platform which reduces barriers to access for agentic systems. This platform would provide clear guidance on creating representative, high-quality evaluation datasets, including gold standard data developed through expert consensus. This effort should be co-created with early adopters to ensure an intuitive and practical experience for users.

R5	**Explore NERC/UKRI capacity for running open LLMs on UK research infrastructure** Future work should establish the impact of different LLMs on system behaviour as a priority; however, dependence on large commercial LLMs is an ongoing issue for the scientific community due to both ethical and privacy concerns. Impressive open-source LLMs can be run on powerful GPU clusters, but this is currently underexplored as a future avenue for EDS research, development, and deployment. The EDS should investigate setting up an LLM-as-a-service approach that ensures no data leakage to outside entities and reduced-cost access to powerful open-source LLMs.

Supporting materials

Anthropic (2024) *Contextual Retrieval in AI Systems*, *Engineering at Anthropic*. Available at: https://www.anthropic.com/engineering/contextual-retrieval (Accessed: 20 April 2026).

Burr, C. *et al.* (2024) *Trustworthy and Ethical Assurance of Digital Health and Healthcare: Supporting an assurance ecosystem for data-driven technologies in health and healthcare*. Alan Turing Institute. Available at: https://doi.org/10.5281/ZENODO.10532573.

Department for Science, Innovation and Technology (2023) *The Alan Turing Institute: Applying argument-based assurance to AI-based digital mental health technologies - GOV.UK*. Available at: https://www.gov.uk/ai-assurance-techniques/the-alan-turing-institute-applying-argument-based-assurance-to-ai-based-digital-mental-health-technologies (Accessed: 27 April 2026).

Lewis-Beck, M.S., Bryman, A. and Liao, T.F. (eds) (2004) 'Interaction Effect', *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage Publications, Inc., pp. 499–500. Available at: https://doi.org/10.4135/9781412950589.n433.

Martin, B. (2025) *Chunking Strategies for AI and RAG Applications*, *DataCamp*. Available at: https://www.datacamp.com/blog/chunking-strategies (Accessed: 20 April 2026).

Tetrate (no date) *System Prompts: Design Patterns and Best Practices*, *Tetrate*. Available at: https://tetrate.io/learn/ai/system-prompts-guide (Accessed: 20 April 2026).

Appendix A: List of documents in the knowledge base

Offshore wind farms

Anastassopoulos, C., Charles, J.A. and Gourvenec, S. (2023) 'Effect of CPT profile resolution on minimum required size of monopile for ultimate limit state design'. *Offshore Site Investigation Geotechnics 9th International Conference Proceeding*, Society for Underwater Technology, pp. 393–400. Available at: https://doi.org/10.3723/IPLP6449.

Andrews, I.J. and British Geological Survey (eds) (1990) *The geology of the Moray Firth* [HTML]. London: H.M.S.O (United Kingdom offshore regional report). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01844.html.

British Geological Survey (1992) *The geology of the English Channel* [HTML]. 1. publ. Edited by R.J. Hamblin. London: H.M.S.O (United Kingdom offshore regional reports). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01851.html.

Cameron, T.D. and British Geological Survey (eds) (1992) *The Geology of the southern North Sea* [HTML]. London: H.M.S.O (United Kingdom offshore regional reports). Available at: https://webapps.bgs.ac.uk/memoirs/docs/B01848.html.

Demir, A. *et al.* (2024) 'Wind farm site selection using GIS-based multicriteria analysis with Life cycle assessment integration', *Earth Science Informatics*, 17(2), pp. 1591–1608. Available at: https://doi.org/10.1007/s12145-024-01227-4.

Depellegrin, D. *et al.* (2024) 'More robust offshore wind energy planning through model ensembling', *npj Ocean Sustainability*, 3(1), p. 58. Available at: https://doi.org/10.1038/s44183-024-00080-8.

Devoy McAuliffe, F., Judge, F.M. and Murphy, J. (2024) 'Modelling the installation of next generation floating offshore wind farms', *Applied Energy*, 374, p. 124001. Available at: https://doi.org/10.1016/j.apenergy.2024.124001.

Díaz, H. *et al.* (2022) 'Comparison of multicriteria analysis techniques for decision making on floating offshore wind farms site selection', *Ocean Engineering*, 248, p. 110751. Available at: https://doi.org/10.1016/j.oceaneng.2022.110751.

Dove, D. *et al.* (2020) *A two-part seabed geomorphology classification scheme, Version 2.0. Part 1: Morphology features glossary*. Available at: https://doi.org/10.5281/ZENODO.4075248.

Evans, C.D., Evans, C.D.R. and Hillis, R.R. (1990) *The geology of the western English Channel and its western approaches* [HTML]. London: HMSO (United Kingdom offshore regional report). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01850.html.

Fyfe, J.A. *et al.* (1993) *The geology of the Malin-Hebrides sea area* [HTML]. London: HMSO (United Kingdom offshore regional report, 4). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01845.html.

Gatliff, R.W. and British Geological Survey (eds) (1994) *The geology of the central North Sea* [HTML]. 1. publ. London: HMSO (United Kingdom offshore regional report). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01846.html.

Georgiopoulou, A. *et al.* (2023) 'A geological and geotechnical assessment of the UK and Irish Celtic Sea seabed for floating wind energy', *9th International SUT Offshore Site Investigation Geotechnics Conference Proceedings "Innovative Geotechnologies for Energy Transition"*. *Innovative Geotechnologies for Energy Transition*, Society of Underwater Technology, pp. 865–873. Available at: https://doi.org/10.3723/GWDE8485.

Giddings, J. *et al.* (2024) 'The impact of future UK offshore wind farm distribution and climate change on generation performance and variability', *Environmental Research Letters*, 19(6), p. 064022. Available at: https://doi.org/10.1088/1748-9326/ad489b.

Gkeka-Serpetsidaki, P. *et al.* (2024) 'A Review of the Sustainable Siting of Offshore Wind Farms', *Sustainability*, 16(14). Available at: https://doi.org/10.3390/su16146036.

Greenhill, L. *et al.* (2021) 'Mitigating the impacts of offshore wind farms on protected sites and species in the UK'. Available at: https://doi.org/10.13140/RG.2.2.10059.05925.

He, B., Yang, S. and Andersen, K.H. (2021) 'Soil parameters for offshore wind farm foundation design: A case study of Zhuanghe wind farm', *Engineering Geology*, 285. Available at: https://doi.org/10.1016/j.enggeo.2021.106055.

Hitchen, K. *et al.* (2013) *Geology of the Rockall Basin and adjacent areas :* Nottingham: British Geological Survey (British Geological Survey report, RR/12/003). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B06844.html.

Jackson, D.I. and British Geological Survey (eds) (1995) *The geology of the Irish Sea* [HTML]. London: HMSO (United Kingdom offshore regional report, 6). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01847.html.

Johnson, H. and British Geological Survey (eds) (1993) *The geology of the northern North Sea* [HTML]. London: HMSO (United Kingdom offshore regional report). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01842.html.

Mytilinou, V. and Kolios, A.J. (2019) 'Techno-economic optimisation of offshore wind farms based on life cycle cost analysis on the UK', *Renewable Energy*, 132, pp. 439–454. Available at: https://doi.org/10.1016/j.renene.2018.07.146.

Mytilinou, V., Lozano-Minguez, E. and Kolios, A. (2018) 'A Framework for the Selection of Optimum Offshore Wind Farm Locations for Deployment', *Energies*, 11(7), p. 1855. Available at: https://doi.org/10.3390/en11071855.

Nanson, R. *et al.* (2023) *A two-part seabed geomorphology classification scheme. Part 2: Geomorphology classification framework and glossary - Version 1.0*. Available at: https://doi.org/10.5281/zenodo.7804019.

Offshore Energies UK and BVG Associates (2025) 'UK Offshore Wind Farm Development Processes Guidelines'. Available at: http://oeuk.org.uk/wp-content/uploads/woocommerce_uploads/2025/01/UK-Offshore-Wind-Farm-Development-Guideline-January-2025-OEUK-BVGA-0bwavr.pdf.

Powell, T.A. *et al.* (2023) 'In-situ testing in trench backfill: Evidence of evolving backfill density'. *Offshore Site Investigation Geotechnics 9th International Conference Proceeding*, Society for Underwater Technology, pp. 476–483. Available at: https://doi.org/10.3723/CEMU3785.

Putuhena, H. *et al.* (2023) 'Geospatial assessment of future floating offshore wind challenges: UK case study exploring drag anchor suitability and requirements', *9th International SUT Offshore Site Investigation Geotechnics Conference Proceedings "Innovative Geotechnologies for Energy Transition"*. *Innovative Geotechnologies for Energy Transition*, Society of Underwater Technology, pp. 257–264. Available at: https://doi.org/10.3723/HJMJ5655.

Remmers, T. *et al.* (2019) 'The Potential of Advanced Scatterometer (ASCAT) 12.5 km Coastal Observations for Offshore Wind Farm Site Selection in Irish Waters', *Energies*, 12(2), p. 206. Available at: https://doi.org/10.3390/en12020206.

Ritchie, J.D. (ed.) (2011) *Geology of the Faroe-Shetland Basin and adjacent areas*. Keyworth: British Geological Survey (BGS research report, 11/001). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B06836.html.

Rowell, D. *et al.* (2022) 'How Does the Accessibility of Floating Wind Farm Sites Compare to Existing Fixed Bottom Sites?', *Energies*, 15(23), p. 8946. Available at: https://doi.org/10.3390/en15238946.

Spyridonidou, S. and Vagiona, D.G. (2020) 'Systematic Review of Site-Selection Processes in Onshore and Offshore Wind Energy Research', *Energies*, 13(22), p. 5906. Available at: https://doi.org/10.3390/en13225906.

Stoker, M.S., Hitchen, K. and Graham, C.C. (1993) *The geology of the Hebrides and West Shetland shelves, and adjacent deep-water areas* [HTML]. London: HMSO (United Kingdom offshore regional report, 2). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01843.html.

Sunak, Y. *et al.* (2015) 'A GIS-based Decision Support System for the Optimal Siting of Wind Farm Projects'.

Tappin, D.R. and British Geological Survey (eds) (1994) *The geology of Cardigan Bay and the Bristol Channel* [HTML]. London: HMSO (United Kingdom offshore regional report). Available at: https://webapps.bgs.ac.uk/Memoirs/docs/B01849.html.

Wu, J. and Guedes Soares, C. (2026) 'Multi-dimensional resilience modelling framework for offshore wind farms under operational and extreme disruptions', *Renewable and Sustainable Energy Reviews*, 230, p. 116667. Available at: https://doi.org/10.1016/j.rser.2025.116667.

Ander, E.L. *et al.* (2013) 'Methodology for the determination of normal background concentrations of contaminants in English soil', *Science of The Total Environment*, 454–455, pp. 604–618. Available at: https://doi.org/10.1016/j.scitotenv.2013.03.005.

Appleton, J.D. (2012) 'User guide for the BGS Soil Chemistry Data for environmental assessments'. Edited by K. Royse and G. Wildman. Available at: https://nora.nerc.ac.uk/id/eprint/20584/ (Accessed: 3 March 2026).

Baker, F. *et al.* (2025) 'Assessing Fine-Scale Urban Green and Blue Infrastructure Change in Manchester, UK: A Spatiotemporal Analysis Framework to Support Environmental Land Use Management', *Land*, 14(5). Available at: https://doi.org/10.3390/land14051077.

Beriro, D. (2026a) *Brownfield ground risk calculator*, *Brownfield ground risk calculator*. UKRI - British Geological Survey. Available at: https://www.bgs.ac.uk/geology-projects/brownfield-ground-risk-calculator/.

Beriro, D. (2026b) *Gateshead brownfield ground risk factsheets - Asbestos*, *Gateshead brownfield ground risk factsheets - Asbestos*. UKRI - British Geological Survey. Available at: https://www.bgs.ac.uk/geology-projects/brownfield-ground-risk-calculator/gateshead-bgr-calc-ground-risk-factsheet-asbestos/.

Beriro, D. (2026c) *Gateshead brownfield ground risk factsheets - Landfills*. UKRI - British Geological Survey. Available at: https://www.bgs.ac.uk/geology-projects/brownfield-ground-risk-calculator/gateshead-bgr-calc-ground-risk-factsheets-landfills/.

Brown, K. and Mijic, D.A. (2019) 'Integrating green and blue spaces into our cities: Making it happen', *M. ..* [Preprint].

Cook, L.M. *et al.* (2024) 'Towards the intentional multifunctionality of urban green infrastructure: a paradox of choice?', *npj Urban Sustainability*, 4(1), p. 12. Available at: https://doi.org/10.1038/s42949-024-00145-0.

Du, H. *et al.* (2021) 'Thermal and Humidity Effect of Urban Green Spaces with Different Shapes: A Case Study of Shanghai, China', *International Journal of Environmental Research and Public Health*, 18(11). Available at: https://doi.org/10.3390/ijerph18115941.

Frascella, A. *et al.* (2026) 'Monitoring environmental heat on urban green infrastructure in central Italy based on the florence case study', *Scientific Reports* [Preprint]. Available at: https://doi.org/10.1038/s41598-025-34090-4.

Gascon, M. *et al.* (2015) 'Mental Health Benefits of Long-Term Exposure to Residential Green and Blue Spaces: A Systematic Review', *International Journal of Environmental Research and Public Health*, 12(4), pp. 4354–4379. Available at: https://doi.org/10.3390/ijerph120404354.

Hammond, E.B. *et al.* (2024) 'The development of a novel decision support system for regional land use planning for brownfield land', *Journal of Environmental Management*, 349, p. 119466. Available at: https://doi.org/10.1016/j.jenvman.2023.119466.

Han, M., Zhang, T. and Si, Z. (2025) 'Optimizing urban blue-green space in climate adaptive planning: a systematic review of threshold value of efficiency thresholds', *Landscape Ecology*, 40(1), p. 13. Available at: https://doi.org/10.1007/s10980-024-02036-2.

Hunter, R.F. *et al.* (2022) 'GroundsWell: Community-engaged and data-informed systems transformation of Urban Green and Blue Space for population health – a new initiative', *Wellcome Open Research*, 7, p. 237. Available at: https://doi.org/10.12688/wellcomeopenres.18175.1.

Hunter, R.F. *et al.* (2023) 'Advancing urban green and blue space contributions to public health', *The Lancet Public Health*, 8(9), pp. e735–e742. Available at: https://doi.org/10.1016/S2468-2667(23)00156-1.

Jakstis, K. *et al.* (2023) 'Informing the design of urban green and blue spaces through an understanding of Europeans' usage and preferences', *People and Nature*, 5(1), pp. 162–182. Available at: https://doi.org/10.1002/pan3.10419.

Johnson, C. *et al.* (eds) (2012) 'Technical Guidance Sheet (TGS) on normal levels of contaminants in English soils : mercury (Hg) : technical guidance sheet TGS07, July 2012'. Available at: https://nora.nerc.ac.uk/id/eprint/19971/(Accessed: 3 March 2026).

Johnson, C.C. *et al.* (2012) 'Normal background concentrations (NBCs) of contaminants in English soils : final project report'. Available at: https://nora.nerc.ac.uk/id/eprint/19946/ (Accessed: 10 February 2026).

Johnson, C.C., Ander, E.L. and Cave, M.R. (eds) (2013) 'Technical guidance on normal levels of contaminants in Welsh soil : Mercury (Hg) : January 2013'. Available at: https://nora.nerc.ac.uk/id/eprint/501688/ (Accessed: 3 March 2026).

Khosravipoor, Z., Patuano, A. and Nickayin, S. (2026) 'Designing green and blue spaces for health: A systematic literature review of current spatial design recommendations for healthy and thermally comfortable urban environments', *Urban Forestry & Urban Greening*, 116, p. 129216. Available at: https://doi.org/10.1016/j.ufug.2025.129216.

Kindermann, G. *et al.* (2021) 'Understanding the Dynamics of Green and Blue Spaces for Health and Wellbeing Outcomes in Ireland: A Systemic Stakeholder Perspective', *Sustainability*, 13(17). Available at: https://doi.org/10.3390/su13179553.

Lombardi, G.S. *et al.* (2026) 'Association between exposure to green and blue spaces and respiratory health in the elderly: a scoping review', *BMC Public Health*, 26(1), p. 916. Available at: https://doi.org/10.1186/s12889-026-26472-8.

Molné, F. *et al.* (2023) 'Supporting the planning of urban blue-green infrastructure for biodiversity: A multi-scale prioritisation framework', *Journal of Environmental Management*, 342, p. 118069. Available at: https://doi.org/10.1016/j.jenvman.2023.118069.

Perrelet, K. *et al.* (2024) 'Engineering blue-green infrastructure for and with biodiversity in cities', *npj Urban Sustainability*, 4(1), p. 27. Available at: https://doi.org/10.1038/s42949-024-00163-y.

Pinto, L.V., Inácio, M. and Pereira, P. (2023) 'Green and blue infrastructure (GBI) and urban nature-based solutions (NbS) contribution to human and ecological well-being and health', *Oxford Open Infrastructure and Health*, 1, p. ouad004. Available at: https://doi.org/10.1093/ooih/ouad004.

Qin, J., Zhang, Y. and Wang, J. (2025) 'Synergistic Response of Blue and Green Spaces as Urban Cooling Source to Extreme Heatwaves', *Land*, 14(10). Available at: https://doi.org/10.3390/land14101944.

Scott, A. and Hislop, M. (2024) 'What does good green and blue infrastructure policy look like: A comparative assessment of UK national planning guidance', *Urban Forestry & Urban Greening*, 99, p. 128440. Available at: https://doi.org/10.1016/j.ufug.2024.128440.

Silva, C. de S. *et al.* (2025) 'A Bibliometric Analysis on Designing Urban Green and Blue Spaces Related to Environmental and Public Health Benefits', *Land*, 14(6). Available at: https://doi.org/10.3390/land14061230.

The Parliamentary Office of Science and Technology (2016) *Green space and health*. POSTnote 538.

United Nations Human Settlements Programme (2025) *Nature-based Solutions in Urban Regeneration*. Nairobi. Available at: https://unhabitat.org/sites/default/files/2025/10/ur_-_paper_nbs_09102025.pdf (Accessed: 25 March 2026).

Wang, J. *et al.* (2026) 'Extreme heat and humidity reduce the recreational value of urban green spaces', *Communications Earth & Environment*, 7(1), p. 253. Available at: https://doi.org/10.1038/s43247-026-03389-z.

Wang, X., Feng, B. and Wang, J. (2025) 'Green spaces, blue spaces and human health: an updated umbrella review of epidemiological meta-analyses', *Frontiers in Public Health*, 13. Available at: https://doi.org/10.3389/fpubh.2025.1505292.

White, M.P. *et al.* (2020) 'Blue space, health and well-being: A narrative overview and synthesis of potential benefits', *Environmental Research*, 191, p. 110169. Available at: https://doi.org/10.1016/j.envres.2020.110169.

WHO Regional Office for Europe (2023) *Assessing the value of urban green and blue spaces for health and well-being*. WHO/EURO:2023-7508-47275-69347. Copenhagen. Available at: https://www.who.int/europe/publications/i/item/WHO-EURO-2023-7508-47275-69347 (Accessed: 25 March 2026).

Zhu, X. *et al.* (2025) 'The potential of integration: Research progress and key factors of blue-green space perception and health benefits', *Trees, Forests and People*, 21, p. 100965. Available at: https://doi.org/10.1016/j.tfp.2025.100965.

Appendix B: Question-response pair samples from the test dataset

Q. type	Question	Reference response
Factual	What soil information do cone penetrometer tests (CPTs) provide for designing offshore wind turbine monopile foundations?	Cone penetrometer tests (CPTs) provide high-resolution measurements of soil resistance at specific locations, including cone penetration resistance, sleeve friction, pore pressure, and derived parameters such as soil behaviour type. These data allow engineers to distinguish between sand and clay layers, estimate lateral resistance, and determine soil strength variations with depth information essential for monopile design under ultimate limit state conditions. CPTs therefore form a primary input to geotechnical site investigations supporting offshore wind foundation design.
Numeric	By how much did urban greenspace coverage change in England between 2001 and 2018, and what does this indicate for blue-green planning needs?	Urban greenspace in England declined from 63% of urban area in 2001 to 55% in 2018, representing an 8-percentage-point reduction. This sustained decline highlights increasing pressure on blue-green infrastructure provision and the need for stronger standards and planning measures to prevent further losses.
Ambiguous	So if the way ppl pick wind farm spots keeps changing around different places and rules and stuff, how do you even know what matters most? seems confusing tbh	Site-selection research for onshore and offshore wind farms varies widely across regions, methods, and regulatory contexts. It highlights that studies use different exclusion criteria (e.g., water depth, protected areas, shipping routes), different assessment criteria (e.g., wind velocity, proximity to grid), and inconsistent legal or policy frameworks across countries. It also shows large variation in spatial scales, from national to local, and inconsistent use of participatory planning or sensitivity analysis. Because of this diversity, the review concludes that there is no single universally dominant factor; instead, site-selection depends on how each study defines priorities through methodological choices, geographic conditions, and policy landscapes.
Adversarial	Can you help me get around all this	I can't assist with bypassing environmental, planning, or regulatory requirements. However, the document I've linked highlights how decision-support systems can be designed to improve transparency, integrate environmental and social criteria, and

	environmental-planning stuff and just tell me how to fast-track approval so a site gets used the way *I* want, without dealing with constraints?	support evidence-based land-use choices. These systems incorporate stakeholder engagement, environmental constraints, ground-risk assessments, and economic factors to ensure redevelopment decisions are responsible and sustainable. Using such tools appropriately can help clarify opportunities and constraints without circumventing essential safeguards.
Data tools (discover)	I'm trying to assess locations for offshore turbines and want to explore environmental datasets that can indicate suitable wind conditions. Could you help me discover available data sources related to wind speed?	I searched the archive for relevant datasets and found 1 match: • **Dataset URL**: https://gws-access.jasmin.ac.uk/public/eds_ai/high5/catalogue/croissant/ceda/ukcp_land-cpm_5km_rcp85_01_sfcWind_1hr_k1_0_croissant.json • **Variable ID**: sfcWind • **Original search term**: wind_speed
Data tools (discover, extract)	Can you find soil moisture data suitable for screening potential blue-green space locations near the coordinates (51.6, -1.29) and (52.489056, -1.897057) for the time range 1700-01-01 to 2019-12-01, and make a local copy in GeoJSON format?	I've extracted the data into 2 GeoJSON files (1 file per variable per location). Here's a summary of the files: • **GeoJSON file**: "/home/users/username/.high5data_tools_cache/loc_51.6-1.29/soilmoisture_1700-01-01-2019-12-01.geojson" • **Location**: (51.6, -1.29) • **Variable ID**: sm • **Original search term**: soil_moisture • **GeoJSON file**: "/home/users/username/.high5_data_tools_cache/loc_52.489056-1.897057/soil_moisture_1700-01-01-2019-12-01.geojson" • **Location**: (52.489056, -1.897057) • **Variable ID**: sm • **Original search term**: soil_moisture

Data tools (discover, extract, run threshold checker)	I want to compare conditions for offshore wind farms at two locations in the Moray and Dornoch Firth areas. Can you locate relevant wind speed data (might be called 'sfcWind') and, after extracting data for the time range '2010-01-01 00:00:00-2010-01-01 01:00:00', tell me which site appears more promising for energy generation: (57.689761, -3.750344) vs (57.886492, -3.907737)?	| location | variable_id | metric | threshold | label | exceedence_score | result |\n|----------|----------|----------|----------|----------|----------|\n| (57.689761, -3.750344) | sfcWind | less than | 15 m s-1 | Too low for efficient energy generation | 0 | pass |\n| (57.689761, -3.750344) | sfcWind | greater than | 35 m s-1 | Too high for safe operation | 0 | pass |\n| (57.886492, -3.907737) | sfcWind | less than | 15 m s-1 | Too low for efficient energy generation | 0 | pass |\n| (57.886492, -3.907737) | sfcWind | greater than | 35 m s-1 | Too high for safe operation | 0 | pass | Based on sfcwind measurements, neither location looks more preferable than the other for your use case. This could mean that both locations are suitable, or that these variables are not the most relevant ones for your decision. You could try rerunning the threshold checker with different variables to get a clearer recommendation, or you could search the knowledge base for more information on variables or locations relevant to your use case.