Towards Automated FAIR Compliance Diagnosis: Evaluating LLMs on Explanation and Diagnosis Questions

Gabriele Tuozzo[1,*,†], Antonio Lieto[1,*,†]

[1]Università degli Studi di Salerno

Abstract

The FAIR principles provide a widely adopted framework for assessing and improving the quality of research data; however, understanding why a dataset attains a given FAIR score and how to improve it remains challenging for dataset producers and consumers. This paper investigates the suitability of large language models (LLMs) for answering Explanation & Diagnosis questions related to FAIR principle evaluation. We introduce the first human-curated benchmark for this task, comprising 40 questions organized into four categories: clarification, metric calculation, improvement, and impact on scoring. Using an in-context learning paradigm, we evaluate five LLMs by providing them with the documentation of a FAIR assessment tool expressed in three semantically equivalent formats: JSON, Markdown, and Turtle. Our results show that LLMs are generally effective at explaining and diagnosing FAIR scores, achieving strong performance on clarification questions, while questions requiring multi-step reasoning about score impact and improvement remain more challenging. Documentation format has no statistically significant effect on overall accuracy; however, Markdown outperforms Turtle for improvement-oriented questions. We further observe that models optimized for reasoning consistently yield higher accuracy on complex question types. These findings highlight both the promise and current limitations of LLMs for automated FAIRness support and offer practical guidance for integrating LLM-based assistance into FAIR assessment platforms.

Keywords

FAIR principles, LLMs, Explanation and Diagnosis, Question Answering, Benchmarking, Data Quality

1. Introduction

In recent years, the rapid proliferation of large language models (LLMs) has enabled their adoption across diverse tasks, including link prediction [1, 2], recommender systems [3], Knowledge Graph (KG) construction [4, 5], and question answering (QA) [6, 7]. Prior work has extensively explored LLMs for QA, examining their ability to retrieve correct answers from both structured [8, 9] and unstructured sources [10, 11], as well as their reasoning capabilities and ability to infer conclusions from the knowledge implicitly encoded within the models [12, 13, 14].

This study focuses on evaluating LLMs in addressing "*Explanation & Diagnosis*" (*E&D*) questions within the context of the FAIR principles [15] (Findable, Accessible, Interoperable, and Reusable), a widely adopted framework for improving the quality and stewardship of research data. While *E&D* QA with LLMs has been explored in a variety of application domains—ranging from healthcare [16] and industrial systems [17] to data visualization [18], log analysis [19], and smart environments [20]—this line of research has so far remained disconnected from the domain of FAIR data assessment. Specifically, we consider questions that aim to explain why a dataset attains a given FAIR score, how its FAIRness can be improved, and how modifications to the dataset affect the resulting score. To support this task, LLMs are provided with the documentation of `KGHeartBeat` [21][1], supplied in three formats (JSON, Markdown, and Turtle), following an in-context learning QA paradigm [22].

QKG@ESWC'26: Workshop on Evaluating, Improving, and Sustaining Knowledge Graph Quality, May 10-14, 2026, Dubrovnik, Croatia

*Corresponding authors.

†These authors contributed equally.

gtuozzo@unisa.it (G. Tuozzo); alieto@unisa.it (A. Lieto)

0009-0004-5108-1995 (G. Tuozzo); 0000-0002-8323-8764 (A. Lieto)

[1]Documentation of FAIR principles implementation available at: https://gabrielet0.github.io/CHe-CLOUD/fair_mapping.html

The primary objective is to assess whether LLMs can effectively support dataset producers, consumers, and maintainers in diagnosing and improving dataset FAIRness. Additionally, we investigate whether the format of the scoring algorithm documentation influences the accuracy and reliability of LLM-generated responses. To our knowledge, this is the first study to systematically evaluate LLMs on FAIR-related *E&D* questions and to analyze the impact of documentation format on response accuracy.

Our work addresses the following research questions (RQs):

RQ1 To what extent are LLMs suitable for answering FAIR principles *E&D* questions?

RQ2 Does the format of the documentation provided to LLMs influence the accuracy of in-context learning QA responses?

To address these research questions, we manually curated a benchmark comprising 40 questions organized into four categories: Clarification (CLAR), Metric Calculation (MC), Improvement (IMP), and Impact on Scoring (IoS). We evaluated five LLMs, prioritizing open-source solutions (four open-source and one closed-source) to investigate potential performance differences. This choice facilitates the integration of the best-performing models into the CHeCLOUD [23][2], supporting its long-term sustainability in delivering automated assistance and recommendations to dataset producers, consumers, and maintainers. The study deliberately focuses on a small yet carefully curated benchmark, designed to explore multiple facets of *E&D* in FAIR assessment rather than to achieve exhaustive coverage. Although the limited benchmark size constrains statistical power and generalizability, it enables fine-grained expert validation of gold-standard answers and systematic analysis of error patterns.

Our findings indicate that all evaluated LLMs show promising capability for the task, although their performance varies across question categories. The models achieve strong results on CLAR questions, while IoS and IMP questions remain more challenging, with higher accuracy attained only by models specifically trained for advanced reasoning. Regarding input representation, JSON and Markdown yield the highest accuracy, followed by Turtle with slightly lower performance. Statistical analysis shows no significant difference among formats when overall accuracy is considered; however, category-level analysis reveals that Markdown outperforms Turtle for IMP questions. Finally, models with higher general capabilities produce statistically more accurate responses on MC and IoS questions.

Our contributions can be summarized as follows:

- We introduce the first human-curated QA benchmark for FAIR principles *E&D*, consisting of 40 questions designed to identify and explain FAIR-related issues in datasets. The benchmark is organized into four distinct question categories.
- A systematic evaluation of LLMs across different FAIR assessment tool documentation formats. To the best of our knowledge, this is the first empirical study examining how input data representations in in-context QA settings affect answer accuracy under FAIR principles.

The remainder of this article is structured as follows. Section 2 introduces the FAIR principles, while Section 3 reviews the relevant related work. Section 4 describes the methodology adopted, and Section 5 presents and analyzes the experimental results. Section 6 discusses the implications of the findings. Finally, Section 7 concludes the paper by outlining its limitations and directions for future work.

2. FAIR Principles

The FAIR principles [15] define a set of hierarchical guidelines aimed at improving the management and stewardship of digital data and metadata, collectively referred to as (meta)data in accordance with FAIR terminology.

Findability ensures that (meta)data can be discovered by both humans and machines. This requires that (meta)data are assigned a globally unique and persistent identifier (F1), described with rich metadata (F2), that metadata explicitly includes the identifier of the data it describes (F3), and that (meta)data are

[2]CHeCLOUD platform: https://checloud.di.unisa.it/

indexed in searchable resources (F4). Accessibility addresses the ability to retrieve (meta)data using standardized mechanisms. Specifically, (meta)data must be retrievable by their identifier using an open, free, and universally applicable communication protocol that also supports authentication and authorization when required (A1–A1.2). Additionally, metadata must remain accessible even if the corresponding data are no longer available (A2). Interoperability enables the integration of (meta)data across systems and domains. It requires the use of formal, shared, and broadly applicable knowledge representation languages (I1), reliance on vocabularies that themselves adhere to FAIR principles (I2), and the inclusion of qualified references to other related (meta)data (I3). Reusability ensures that (meta)data can be repurposed in different contexts and over time. It requires accurate and relevant descriptive attributes (R1), a clear and accessible data usage license (R1.1), detailed provenance information (R1.2), and compliance with domain-relevant community standards (R1.3).

3. Related Work

This section surveys related work along two axes: the use of LLMs for *E&D* QA, and QA approaches that expose LLMs to heterogeneous data representations. The section concludes with a comparative discussion that positions the present work within the current state of the art.

LLMs for Explanation & Diagnosis. The use of LLMs to answer questions related to *E&D* has been extensively investigated in the healthcare domain. Representative applications include diagnostic support in radiology systems [16], electrocardiogram analysis [24], and disease diagnosis based on multimodal clinical data [25, 26, 27, 28, 29]. These studies demonstrate the effectiveness of LLMs in complex diagnostic tasks that require reasoning over heterogeneous and often incomplete information.

Beyond healthcare, *E&D*-oriented applications of LLMs have been explored across a wide range of domains. Tian et al. [17] proposed a hybrid approach that builds a KG and combines graph-based reasoning with LLM-driven information extraction, validated via a fault-chain analysis of power loss in a turbocharged diesel engine.

Das and Mueller [18] introduced `MisVisFix`, an interactive web-based dashboard that leverages multimodal LLMs to support the complete pipeline of detecting, explaining, and correcting misleading data visualizations. Their work highlights the potential of LLMs to enhance interpretability and user interaction in visual analytics.

In the context of system logs, Sun et al. [19] proposed *LogInsight*, a framework for accurate and interpretable log-based fault diagnosis using LLMs. The framework overcomes a key limitation of existing approaches by providing explicit explanations for its diagnostic outcomes.

Similarly, Khan et al. [30] presented *FaultExplainer*, an interactive web-based tool for interpretable fault detection and diagnosis in chemical processes. By integrating LLMs with classical statistical techniques, the system overcomes the limited interpretability of traditional data-driven fault detection methods and improves robustness to previously unseen faults.

In the manufacturing domain, Karabiyik [31] proposed a system that combines machine learning classifiers, deep learning architectures, and LLMs to detect and diagnose seven types of 3D printing faults, including cracking, gaps, over-extrusion, overheating, stringing, and weak infill. The study systematically evaluated different prompting strategies (zero-shot, chain-of-thought, and tree-of-thought) to assess their impact on diagnostic performance.

Finally, Clauß et al. [20] used LLMs for fault detection and diagnosis on IoT sensor time-series data in smart buildings, demonstrating their effectiveness in a real-world operational setting.

Question Answering over different data formats. Neural text-based QA models, such as DrQA [32] and UnifiedQA [33], achieve strong performance on open-domain benchmarks (e.g., SQuAD [34]) by exploiting linguistic patterns in large text corpora. Extensions like TTGen [35] enhance answer extraction by converting structured data into text. However, purely text-centric QA systems struggle with aggregation, arithmetic reasoning, or questions requiring implicit knowledge.

To bridge unstructured and structured data, semi-structured QA over formats such as JSON has been explored. Gadiraju et al. [36] leverage JSON with LLMs like Llama-3.1-8B and Mistral-7B for technical QA, while Shen et al. [37] target product-oriented JSON QA. These approaches handle intermediate representations-key-value pairs, lists, hierarchical JSON, common in real-world applications but underexplored in research.

In parallel, the Semantic Web community has focused on KG QA, mapping natural-language questions to formal queries (e.g., SPARQL) [38, 39]. Benchmarks such as QALD [40] and LC-QuAD [41] have advanced entity linking and query generation. LLM-based neural–symbolic methods [42, 43, 44] directly generate SPARQL queries but often exploit KG structure superficially and remain sensitive to schema heterogeneity and implicit relations [45, 46, 47].

Hybrid QA systems combine textual retrieval with structured reasoning (e.g., HAWK [48], SUQL [49]). Evaluations indicate curated KGs improve precision [50], though performance declines as structural complexity increases [51, 52].

Positioning of This Work. While prior studies have investigated the use of LLMs for QA on *E&D* tasks across diverse domains, such as healthcare, mechanical systems, chart interpretation, log data analysis, chemical processes, 3D printing fault detection, and smart buildings, no existing work has examined this problem within the context of the FAIR principles. Moreover, current studies predominantly evaluate individual data representations in isolation and lack a systematic analysis of how semantically equivalent content expressed as unstructured text, semi-structured data, and KGs influences the robustness and consistency of LLM reasoning.

4. Methodology

This section describes the construction of the benchmark and the types of questions considered, the methodology adopted for executing the benchmark, and the LLMs evaluated.

4.1. Benchmark construction

To evaluate the ability of LLMs to answer *E&D* questions related to FAIR principles score computation, we designed four distinct categories of questions:

- CLAR (Clarification): Questions aimed at understanding how to correctly interpret a specific score associated with a FAIR principle.
- MC (Metric Calculation): Questions focused on explaining how a given FAIR principle is computed, including the underlying algorithm and scoring mechanism.
- IMP (Improvement): Questions intended to elicit detailed guidance on the actions required to improve the score of a specific FAIR principle.
- IoS (Impact on Scoring): Questions designed to assess how modifications to the dataset data or metadata affect the evaluation of a FAIR principle.

For each category, we manually formulated 10 questions with varying phrasing and reasoning complexity to ensure heterogeneity and avoid trivial variations. The benchmark thus comprises 40 questions. Each question includes an expected answer curated by two FAIR assessment experts. Table 1 presents a single example for each category, whereas the Table 3 in the Appendix A provides all 40 questions included in the benchmark. Sub-principles are marked with **M** when applicable only to metadata, with **D** when applicable only to data, and left unmarked when they apply to both.

4.2. Benchmark Execution and Evaluation

Figure 1 illustrates the workflow adopted to conduct the benchmark on LLMs. In addition to the questions described in Section 4.1, the models are provided with documentation of the `KGHeartBeat`

CLAR	MC	IMP	IoS
Is the presence of a license required for Interoperability?	How is the I1-M sub-principle calculated?	What is the quickest way to improve my overall FAIR score?	How does the absence of a DOI affect the FAIR score?

Table 1
Examples of questions from the FAIR *Explanation & Diagnosis* benchmark. Sub-principle I1-M assesses the use of formal, shared, and broadly applicable knowledge representation languages for metadata.

algorithm for computing FAIR scores, including its scoring procedure. `KGHeartBeat` evaluates 18 FAIR sub-principles and produces both per-principle scores and an overall FAIR score, leveraging the alignment between KG quality and FAIR principles in [53]. Although the benchmark is, in principle, tool-independent, it is instantiated in this study using a single tool. Nevertheless, it can be applied to any tool that computes the FAIR principles, provided that the corresponding documentation is available. It is important to note that the gold-standard answers may vary depending on the specific implementation details of the tool.

The tool's documentation is provided in three formats (Markdown, JSON, and a Turtle KG) to evaluate which best supports comprehension and accurate responses. While all formats convey the same information, they differ in structure and semantics: Markdown offers minimal structure, JSON adds syntactic structure, and the Turtle file models the documentation using the Data Quality Vocabulary (DQV) [54], providing both structure and explicit semantics. The selection of this Vocabulary was guided by the fact that it provides a standard vocabulary for expressing quality dimensions and measures, which maps naturally onto the FAIR sub-principles and the scoring logic of `KGHeartBeat`.

The experiment primarily employs open-source LLMs to identify the best-performing model for *E&D* of the FAIR score, with the aim of integrating it into the CHeCLOUD and ensuring the platform's long-term sustainability. A closed-source model is also included to assess potential performance gaps between open- and closed-source approaches for this task. The evaluated models are gpt-oss-20b, llama-3.3-70b, gpt-oss-120b, Gemini-2.5-pro, and DeepSeek-V3.2 Thinking. Gemini-2.5-pro is the only closed-source model analyzed, to compare with open-source models, and because it is the LLM currently used in the CHeCLOUD. For all models, the temperature was set to the lowest supported value, and testing was conducted via their APIs using the LangChain[3] library.

For each combination of LLM and documentation format, benchmark questions are submitted independently to the model along with the corresponding documentation, following a zero-shot prompting approach. Responses are recorded in CSV files to facilitate human evaluation. Consequently, three CSV files are generated per LLM, corresponding to results obtained with the JSON, Markdown, and Turtle formats. The prompt used in the zero-shot setting is provided in the Appendix B.

After collecting the LLM responses, two FAIR-experts independently evaluated them, assigning a score of 1 for fully correct answers, 0.5 for partially correct answers (e.g., enumerations with only a subset of expected items), and 0 for incorrect answers. Inter-annotator agreement was calculated, and any discrepancies were discussed and resolved to assign a final accuracy score to each response.

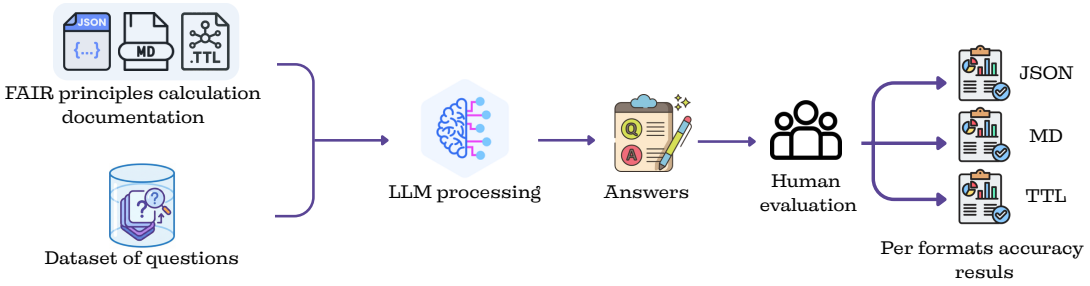

Figure 1: Workflow adopted to execute the FAIR *Explanation & Diagnosis* benchmark

[3]LangChain: https://www.langchain.com/

The experimental code, benchmark, KGHeartBeat documentation, input data, and results are available on GitHub [55], ensuring full transparency and reproducibility.

5. Results

This section presents and analyzes the experimental results from multiple perspectives. Table 2 reports the mean accuracy per question type, allowing comparison across formats and models, based on manual reviewers' assessment. Inter-annotator agreement was almost perfect ($\kappa = 0.81$–1.00) for all models, except gpt-oss-120b, which showed substantial agreement ($\kappa = 0.68$), calculated using Cohen's Kappa. The table is visualized as a heatmap, with darker colors indicating higher accuracy. Models are ordered by parameter count, with Gemini-2.5-pro and DeepSeek-V3.2 additionally ranked according to LiveBench [56]. In the prompt, LLMs are instructed to return "IDK" when unable to answer a question. As this occurs only rarely, a detailed analysis of IDK answer rates is reported in Appendix C.

Per Questions Category Analysis. Across all documentation formats, models consistently achieve the highest accuracy on questions in the CLAR category. These questions primarily require identifying the algorithm and scoring function described in the data and expressing them in natural language. As such, they mainly assess the ability of LLMs to recognize patterns and return explicitly available information, without demanding complex reasoning.

In contrast, the IoS and IMP categories pose significantly greater challenges. IoS questions yield the lowest performance for three of the five evaluated models, highlighting systematic difficulties when the correct answer is not explicitly stated in the data. These questions require reasoning about how metric values influence FAIR principles and sub-principles, often necessitating a detailed understanding of the underlying evaluation algorithms. Although responses typically involve enumerating all affected elements, LLMs frequently return only partial lists, indicating limitations in exhaustively capturing all relevant factors. Notably, DeepSeek-V3.2 consistently outperforms the other models in this category across all formats, likely due to its architecture being explicitly optimized for multi-step reasoning.

Instead, questions in the IMP category can be seen as complementary to IoS. Rather than analyzing the impact of existing metric values, these questions require identifying which metrics should be modified to improve the score of a given FAIR principle or sub-principle. Despite this conceptual inversion, IMP questions exhibit similarly reduced performance, further confirming the difficulty LLMs face when reasoning beyond explicitly provided information.

The most challenging individual question across all models and formats is IoS_1, which attains a mean accuracy of only 0.13. This question asks: "To what extent does the absence of a working SPARQL endpoint affect the overall FAIRness score?". Correctly answering it requires distinguishing between FAIR sub-principles that depend on direct access to the dataset and those that do not. In this setting, all models exhibit substantial reasoning errors, either hallucinating irrelevant sub-principles or providing incomplete subsets of the correct answer. Notably, DeepSeek-V3.2 Thinking is the only model to achieve an accuracy of 0.5 across all three formats, further supporting the observation that models explicitly designed for reasoning perform more reliably on inference-intensive questions.

Per Documentation Formats Analysis. JSON and Markdown achieve the highest overall accuracy across models (mean 0.85), while Turtle performs slightly lower (mean 0.83), especially for IMP and IoS questions. Gemini-2.5-pro shows the best performance and robustness across formats (mean 0.88), followed by DeepSeek-V3.2, suggesting that larger models are generally less sensitive to input format.

Overall, the findings indicate that both question type and data format influence model performance. Information retrieval queries are easier for models to answer than those requiring reasoning, and Markdown and JSON provide the most favorable context for model reasoning, with Turtle just behind.

Statistical Analysis. We conducted four statistical tests to investigate the impact of input representation formats and model capabilities on the performance of LLMs. To this end, we formulated the

Table 2
Mean score across question types, enabling comparison across formats and models. The table is rendered as a heatmap, with darker colors indicating higher values. Avg. denotes the mean across CLAR, MC, IMP, and IoS within each format. Avg. denotes the mean score across question types within each format block.

	Markdown					JSON					Turtle				
Model	**CLAR**	**MC**	**IMP**	**IoS**	**Avg.**	**CLAR**	**MC**	**IMP**	**IoS**	**Avg.**	**CLAR**	**MC**	**IMP**	**IoS**	**Avg.**
gpt-oss-20b	1.00	0.90	0.70	0.70	0.82	1.00	0.95	0.75	0.70	0.85	0.90	0.80	0.55	0.65	0.72
LLama-3.3-70b	1.00	0.90	0.75	0.55	0.80	0.90	0.90	0.60	0.60	0.75	1.00	0.90	0.70	0.60	0.80
gpt-oss-120b	0.95	0.90	0.90	0.80	0.88	0.95	0.95	0.85	0.70	0.86	0.90	0.90	0.80	0.75	0.83
Gemini-2.5-pro	1.00	0.95	0.90	0.70	0.88	1.00	1.00	0.75	0.80	0.88	1.00	1.00	0.75	0.80	0.88
DeepSeek-V3.2 Think.	1.00	0.90	0.75	0.80	0.86	0.95	1.00	0.80	0.85	0.90	1.00	0.95	0.70	0.85	0.87
Mean Accuracy					0.85					0.85					0.83

following hypotheses:

- H_1 There is a statistically significant difference in performance between at least one pair of input representation formats (JSON, Markdown, Turtle) when results are aggregated across all models and questions.
- H_2 The effect of the input representation format on model performance differs across question categories (CLAR, MC, IMP, IoS), such that at least one category exhibits a statistically significant performance difference between two or more formats.
- H_3 Model capability, operationalized as an ordinal ranking from weaker to stronger models based on the number of parameters and the LiveBench [56] ranking, is positively associated with overall performance, both globally and within individual question categories.
- H_4 Model capability modulates sensitivity to input representation formats, such that the performance gap between formats varies systematically with model rank.

Hypotheses H_1 and H_2 were evaluated using paired t-tests, while H_3 and H_4 were tested using Spearman's rank correlation coefficient, as the data did not follow a normal distribution according to the Shapiro–Wilk test. The significance level was set at $\alpha = 0.05$.

The results for H_1 indicate that, when considering overall accuracy across all questions, no input representation format outperforms the others statistically. However, analysis by question category (H_2) reveals that the Markdown format yields significantly higher accuracy than the Turtle format for questions of type IMP ($p = 0.01$).

Regarding H_3, overall model capability is not significantly associated with better performance across all questions. Nevertheless, when considering individual question categories, models with higher general capability produce statistically more accurate answers for MC and IoS questions ($\rho = 0.90$, $p = 0.03$ for both categories).

Finally, the analysis for H_4 shows no significant correlation between model capability and sensitivity to input representation formats, suggesting that stronger models do not systematically benefit more from any specific format.

6. Discussion

LLMs can effectively support dataset consumers, maintainers, and providers in explaining and diagnosing dataset deficiencies with respect to the FAIR principles. In particular, DeepSeek-V3.2 Thinking, specifically optimized for reasoning, achieved high accuracy (0.90) when provided with JSON documentation (**RQ1**), indicating that specialized LLMs can reliably interpret structured information.

While overall accuracy across all models and questions was largely independent of documentation format (Section 5, H_1), performance analysis by question category (Section 5, H_2) revealed that Markdown documentation led to higher accuracy on IMP questions compared to Turtle (**RQ2**). This suggests

that documentation format can facilitate reasoning in more complex inquiries, likely due to improved readability and structure.

Error analysis further highlights the limitations of current LLMs. The most frequent errors occurred in IoS and IMP questions, which require multi-step reasoning rather than simple retrieval, consistent with prior work [57]. Additionally, LLMs often returned incomplete answers for enumeration questions, producing subsets of the expected elements, aligning with previous findings [58, 59].

The statistical analysis conducted for H_3 and H_4 suggests that, to ensure the long-term sustainability of CHeCLOUD, employing a larger and more complex model improves accuracy for questions of type MC and IoS. However, the computational and infrastructural costs of maintaining large LLMs must be considered. If the platform mainly handles CLAR and IMP questions, smaller models can be used efficiently, yielding significant savings in compute and operational costs.

An unexpected finding is that Turtle documentation underperforms Markdown and JSON for IMP questions. This likely reflects that DQV models data quality rather than FAIR principles, and further analysis is needed to assess whether a FAIR-specific ontology or alternative RDF serializations could improve performance.

7. Conclusion

This paper presented the first human-curated benchmark for evaluating LLMs on *E&D* questions related to the FAIR principles. The benchmark comprises 40 questions organized across four categories (CLAR, MC, IMP, and IoS), and was used to evaluate five LLMs under an in-context learning paradigm, supplying the models with FAIR assessment tool documentation in three formats: JSON, Markdown, and Turtle. Our experiments demonstrate that LLMs are broadly effective at supporting dataset producers, consumers, and maintainers in diagnosing and improving dataset FAIRness (**RQ1**). All evaluated models achieve strong performance on CLAR questions, which require pattern recognition and information retrieval. However, IoS and IMP questions, which require multi-step reasoning and inference beyond explicitly stated information, remain substantially more challenging. In these categories, DeepSeek-V3.2 Thinking stands out as the most capable model, owing to its architecture being explicitly optimized for reasoning tasks. A promising direction for the development of FAIR assessment tools is to move beyond reporting only final scores and instead expose the underlying evidence for each evaluation, including the specific checks performed and the conditions used to determine outcomes. Providing LLMs with this richer, structured feedback could help reduce reasoning errors, particularly on IoS and IMP tasks. Regarding the impact of documentation format, JSON and Markdown achieve the highest overall accuracy, with Turtle performing slightly lower. While no statistically significant difference was found across formats when considering overall accuracy, category-level analysis revealed that Markdown significantly outperforms Turtle for IMP questions, suggesting that readability and structure of the input documentation can meaningfully influence model reasoning on complex tasks (**RQ2**). These findings carry practical implications for the integration of LLMs into platforms providing automated FAIRness support. For tasks dominated by CLAR and IMP questions, smaller and more efficient models can be employed without substantial loss of accuracy, offering meaningful reductions in computational and operational costs. For more demanding reasoning tasks such as MC and IoS, larger and more capable models are recommended.

Limitations and Future Directions. Only a zero-shot prompting strategy was evaluated; alternative prompting or few-shot strategies were not explored, which may improve accuracy for certain question types. The benchmark relies on a single FAIR assessment tool, so results may differ from those of other tools or scoring algorithms. While Turtle performed slightly worse than Markdown and JSON, future work could explore additional RDF serializations (e.g., JSON-LD, RDF/XML) to assess their impact on model performance. Future work will expand the benchmark with more questions, explore retrieval-augmented generation to reduce incomplete enumeration errors, and investigate fine-tuned FAIR-specific models. Although the methodology is tool-agnostic, current findings are specific to FAIR evaluation, and generalizing them to other *E&D* domains is an important direction for future research.

Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology) and it was partially supported by the European Alliance NEOLAiA (Project 101124794: “NEOLAiA – Transforming Regions for an Inclusive Europe").

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to: Grammar and spelling check.

References

[1] Z. He, J. Zhu, S. Qian, J. Chai, D. Koutra, LinkGPT: Leveraging Large Language Models for Enhanced Link Prediction in Text-Attributed Graphs, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM ’25, Association for Computing Machinery, New York, NY, USA, 2025, p. 843–853. URL: https://doi.org/10.1145/3746252.3761287. doi:10.1145/3746252.3761287.

[2] T. Ben Smida, R. Bouslimi, H. Achour, A comprehensive survey on link prediction: from heuristics to graph transformers: T. Ben Smida et al., The Journal of Supercomputing 81 (2025) 1388. doi:10.1007/s11227-025-07882-8.

[3] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, Q. Li, Recommender Systems in the Era of Large Language Models (LLMs), IEEE Transactions on Knowledge and Data Engineering 36 (2024) 6889–6907. doi:10.1109/TKDE.2024.3392335.

[4] X. Liang, Z. Wang, M. Li, Z. Yan, A survey of LLM-augmented knowledge graph construction and application in complex product design, Procedia CIRP 128 (2024) 870–875. doi:10.1016/j.procir.2024.07.069, 34th CIRP Design Conference.

[5] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A Comprehensive Survey on Automatic Knowledge Graph Construction, ACM Comput. Surv. 56 (2023). doi:10.1145/3618295.

[6] P. Shailendra, R. C. Ghosh, R. Kumar, N. Sharma, Survey of Large Language Models for Answering Questions Across Various Fields, in: 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, 2024, pp. 520–527. doi:10.1109/ICACCS60874.2024.10717078.

[7] Y. G. Gebretsadkan, B. Dejene Tegegne, D. S. Merawi, M. Meshesha, Large Language Model (LLM) based Question and Answering System (QAS): A systematic literature review, in: 2025 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), 2025, pp. 259–263. doi:10.1109/ICT4DA67218.2025.11282707.

[8] Y. Sui, M. Zhou, M. Zhou, S. Han, D. Zhang, Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM ’24, Association for Computing Machinery, New York, NY, USA, 2024, p. 645–654. URL: https://doi.org/10.1145/3616855.3635752. doi:10.1145/3616855.3635752.

[9] C. Wolff, M. Hulsebos, How well do LLMs reason over tabular data, really?, in: S. Chang, M. Hulsebos, Q. Liu, W. Chen, H. Sun (Eds.), Proceedings of the 4th Table Representation Learning Workshop, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 241–250. URL: https://aclanthology.org/2025.trl-1.21/. doi:10.18653/v1/2025.trl-1.21.

[10] Y. Tachioka, Question Answer Summary Generation from Unstructured Texts by Using LLMs, in: A. Morishima, G. Li, Y. Ishikawa, S. Amer-Yahia, H. V. Jagadish, K. Lu (Eds.), Database Systems for Advanced Applications. DASFAA 2024 International Workshops, Springer Nature Singapore, Singapore, 2025, pp. 261–268. doi:10.1007/978-981-96-0914-7_20.

[11] J. Lehmann, D. Bhandiwad, P. Gattogi, S. Vahdati, Beyond Boundaries: A Human-like Approach

for Question Answering over Structured and Unstructured Information Sources, Transactions of the Association for Computational Linguistics 12 (2024) 786–802. URL: https://doi.org/10.1162/tacl_a_00671. doi:10.1162/tacl_a_00671.

[12] X. Lin, Z. Huang, Z. Zhang, J. Zhou, E. Chen, Explore What LLM Does Not Know in Complex Question Answering, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 24585–24594. doi:10.1609/aaai.v39i23.34638.

[13] M. M. Lucas, J. Yang, J. K. Pomeroy, C. C. Yang, Reasoning with large language models for medical question answering, Journal of the American Medical Informatics Association 31 (2024) 1964–1975. doi:10.1093/jamia/ocae131.

[14] Q. Li, C. Huang, S. Li, Y. Xiang, D. Xiong, W. Lei, GraphOTTER: Evolving LLM-based graph reasoning for complex table question answering, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 5486–5506. URL: https://aclanthology.org/2025.coling-main.368/.

[15] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9. doi:10.1038/sdata.2016.18.

[16] M. Haitham, N. Sharaf, XAI Meets Radiology: Localized Chest X-ray Diagnosis with Natural Language Explanations, in: 2025 29th International Conference Information Visualisation (IV), 2025, p. 410 – 415. doi:10.1109/IV68685.2025.00077.

[17] X. Tian, H. Gan, Y. Liu, Construction of Knowledge Graph for Marine Diesel Engine Faults Based on Deep Learning Methods, Journal of Marine Science and Engineering 13 (2025). doi:10.3390/jmse13040693.

[18] A. K. Das, K. Mueller, MisVisFix: An Interactive Dashboard for Detecting, Explaining, and Correcting Misleading Visualizations using Large Language Models, IEEE Transactions on Visualization and Computer Graphics 32 (2026) 134–144. doi:10.1109/TVCG.2025.3633884.

[19] Y. Sun, S. Ma, T. Xiao, Y. Zhao, X. Cai, W. Dong, Y. Shen, Y. Zhao, S. Zhang, J. Han, D. Pei, Accurate and Interpretable Log-Based Fault Diagnosis Using Large Language Models, IEEE Transactions on Services Computing 18 (2025) 2602–2615. doi:10.1109/TSC.2025.3599494.

[20] J. Clauß, L. Caetano, K. Nordanger, T. E. Lassen, R. Kind, Leveraging Generative AI and Semantic Data for Improved Operation of a Real-Life Building, in: I. Martinac, B. N. Jørgensen, Z. G. Ma, R. Unnþórsson, C. Bordin (Eds.), Energy Informatics, Springer Nature Switzerland, Cham, 2026, pp. 130–143. doi:10.1007/978-3-032-03098-6_9.

[21] M. A. Pellegrino, A. Rula, G. Tuozzo, Kgheartbeat: An open source tool for periodically evaluating the quality of knowledge graphs, in: G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, A. Hogan (Eds.), The Semantic Web – ISWC 2024, Springer Nature Switzerland, Cham, 2025, pp. 40–58. doi:10.1007/978-3-031-77847-6_3.

[22] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A Survey on In-context Learning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1107–1128. doi:10.18653/v1/2024.emnlp-main.64.

[23] G. Tuozzo, M. A. Pellegrino, A. Lieto, CHeCLOUD—the Cultural Heritage Linked Open Data Cloud, in: International Semantic Web Conference ISWC-Poster&Demo, 2025. URL: https://ceur-ws.org/Vol-4085/paper66.pdf.

[24] D. Tian, J. Jiang, K. Zhang, C. Liu, Y. Yuan, M. Gao, E. Chen, ECG-Doctor: An Interpretable Multimodal ECG Diagnosis Framework Based on Large Language Models, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, p. 2863 – 2873. doi:10.1145/3746252.3761082.

[25] P. Bedi, A. Thukral, S. Dhiman, XLR-KGDD: leveraging LLM and RAG for knowledge graph-based explainable disease diagnosis using multimodal clinical information, Knowledge and Information Systems 67 (2025) 7451 – 7471. doi:10.1007/s10115-025-02465-8.

[26] N. Agrawal, A. Bhardwaj, P. Bhulania, DiseaseGPT: Leveraging Large Language Models for Symptom Elaboration and Healthcare Communication, in: 2025 IEEE International Conference on Computer, Electronics, Electrical Engineering their Applications (IC2E3), 2025. doi:10.1109/IC2E365635.2025.11167257.

[27] Q. Chen, L. Ni, TCM MLKG-RAG: Traditional Chinese Medicine Intelligent Diagnosis Based on Multi-Layer Knowledge Graph Retrieval-Augmented Generation, in: 2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC), 2024, p. 958 – 962. doi:10.1109/EIECC64539.2024.10929529.

[28] M. Kumar, K. Ramrakhiyani, H. Garg, 'Explainable AI' Disease Detection with Reasoning, in: Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2024, 2024, p. 222 – 227. doi:10.1109/ICAAIC60222.2024.10575153.

[29] A. Jesson, N. Beltran-Velez, D. Blei, CAN GENERATIVE AI SOLVE YOUR IN-CONTEXT LEARNING PROBLEM? A MARTINGALE PERSPECTIVE, in: 13th International Conference on Learning Representations, ICLR 2025, 2025, p. 21895 – 21917. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105010216161&partnerID=40&md5=b5373cae8f840c4b766c3a55d39c3c98.

[30] A. Khan, R. Nahar, H. Chen, G. E. Constante Flores, C. Li, FaultExplainer: Leveraging large language models for interpretable fault detection and diagnosis, Computers Chemical Engineering 199 (2025) 109152. doi:10.1016/j.compchemeng.2025.109152.

[31] M. A. Karabiyik, Fault analysis in additive manufacturing: Identifying causes of three-dimensional printer faults using machine learning and large language models, Journal of Systems and Software 230 (2025) 112556. doi:10.1016/j.jss.2025.112556.

[32] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to Answer Open-Domain Questions, in: Association for Computational Linguistics (ACL), 2017. doi:10.18653/v1/P17-1171.

[33] P. Qin, J. Yu, Y. Gao, D. Xu, Y. Chen, S. Wu, T. Xu, E. Chen, Y. Hao, Unified qa-aware knowledge graph generation based on multi-modal modeling, in: Proceedings of the 30th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2022, p. 7185–7189. doi:10.1145/3503161.3551604.

[34] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. doi:10.18653/v1/D16-1264.

[35] X. Li, Y. Sun, G. Cheng, TSQA: tabular scenario based question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13297–13305. doi:10.1609/aaai.v35i15.17570.

[36] S. S. Gadiraju, D. Liao, A. Kudupudi, S. Kasula, C. Chalasani, InfoTech Assistant: A Multimodal Conversational Agent for InfoTechnology Web Portal Queries , in: 2024 IEEE International Conference on Big Data (BigData), IEEE Computer Society, Los Alamitos, CA, USA, 2024, pp. 3264–3272. URL: https://doi.ieeecomputersociety.org/10.1109/BigData62323.2024.10825668. doi:10.1109/BigData62323.2024.10825668.

[37] X. Shen, G. Barlacchi, M. Del Tredici, W. Cheng, A. Gispert, semiPQA: A study on product question answering over semi-structured data, in: S. Malmasi, O. Rokhlenko, N. Ueffing, I. Guy, E. Agichtein, S. Kallumadi (Eds.), Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 111–120. URL: https://aclanthology.org/2022.ecnlp-1.14/. doi:10.18653/v1/2022.ecnlp-1.14.

[38] D. Diefenbach, V. Lopez, K. Singh, P. Maret, Core techniques of question answering systems over knowledge bases: a survey, Knowledge and Information systems 55 (2018) 529–569. doi:https://doi.org/10.1007/s10115-017-1100-y.

[39] S. Hu, L. Zou, J. X. Yu, H. Wang, D. Zhao, Answering natural language questions by subgraph matching over knowledge graphs, IEEE Transactions on Knowledge and Data Engineering 30 (2017) 824–837. doi:10.1109/TKDE.2017.2766634.

[40] R. Usbeck, X. Yan, A. Perevalov, L. Jiang, J. Schulz, A. Kraft, C. Möller, J. Huang, J. Reineke, A.-C. Ngonga Ngomo, et al., QALD-10–the 10th challenge on question answering over linked

data: Shifting from dbpedia to wikidata as a KG for KGQA, Semantic Web 15 (2024) 2193–2207. doi:10.3233/SW-233471.

[41] M. Dubey, D. Banerjee, A. Abdelkawi, J. Lehmann, LC-QuAD 2.0: A large dataset for complex question answering over wikidata and dbpedia, in: International semantic web conference, Springer, 2019, pp. 69–78. doi:10.1007/978-3-319-68204-4_22.

[42] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, G. Qi, Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family, in: The Semantic Web – ISWC, Springer Nature Switzerland, Cham, 2023, pp. 348–367. doi:10.1007/978-3-031-47240-4_19.

[43] M. R. A. H. Rony, U. Kumar, R. Teucher, L. Kovriguina, J. Lehmann, SGPT: A generative approach for SPARQL query generation from natural language questions, IEEE access 10 (2022) 70712–70723. doi:10.1109/ACCESS.2022.3188714.

[44] L. Kovriguina, R. Teucher, D. Radyush, D. Mouromtsev, N. Keshan, S. Neumaier, A. Gentile, S. Vahdati, SPARQLGEN: One-Shot Prompt-based Approach for SPARQL Query Generation., in: SEMANTiCS (Posters & Demos), 2023. URL: https://ceur-ws.org/Vol-3526/paper-08.pdf.

[45] X. Cheng, D. Luo, X. Chen, L. Liu, D. Zhao, R. Yan, Lift yourself up: Retrieval-augmented text generation with self-memory, Advances in Neural Information Processing Systems 36 (2023) 43780–43799. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/887262aeb3eafb01ef0fd0e3a87a8831-Paper-Conference.pdf.

[46] P. Panda, A. Agarwal, C. Devaguptapu, M. Kaul, P. Ap, HOLMES: Hyper-relational knowledge graphs for multi-hop question answering using LLMs, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 13263–13282. doi:10.18653/v1/2024.acl-long.717.

[47] M. Yasunaga, X. Chen, Y. Li, P. Pasupat, J. Leskovec, P. Liang, E. H. Chi, D. Zhou, Large Language Models as Analogical Reasoners, 2024. URL: https://arxiv.org/abs/2310.01714.

[48] R. Usbeck, A.-C. N. Ngomo, L. Bühmann, C. Unger, Hawk–hybrid question answering using linked data, in: European Semantic Web Conference, Springer, 2015, pp. 353–368. doi:10.1007/978-3-319-18818-8_22.

[49] S. Liu, J. Xu, W. Tjangnaka, S. Semnani, C. Yu, M. Lam, SUQL: Conversational search over structured and unstructured data with large language models, in: Findings of the Association for Computational Linguistics: NAACL, 2024, pp. 4535–4555. URL: https://aclanthology.org/2024.findings-naacl.283.pdf.

[50] R. Etezadi, M. Shamsfard, The state of the art in open domain complex question answering: a survey, Applied Intelligence 53 (2023) 4124–4144. doi:10.1007/s10489-022-03732-9.

[51] W. Zhang, L. Jin, Y. Zhu, J. Chen, Z. Huang, J. Wang, Y. Hua, L. Liang, H. Chen, TrustUQA: A Trustful Framework for Unified Structured Data Question Answering, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 25931–25939. doi:10.1609/aaai.v39i24.34787.

[52] Z. Gu, H. Ye, X. Chen, Z. Zhou, H. Feng, Y. Xiao, StrucText-Eval: Evaluating Large Language Model's Reasoning Ability in Structure-Rich Text, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 223–244. URL: https://aclanthology.org/2025.acl-long.11.pdf.

[53] M. A. Pellegrino, P. Esposito, G. Tuozzo, Fairness of the linguistic linked open data cloud: an empirical investigation, J. Data and Information Quality 17 (2025). URL: https://doi.org/10.1145/3769116. doi:10.1145/3769116.

[54] R. Albertoni, A. Isaac, Introducing the data quality vocabulary (DQV), Semantic Web 12 (2020) 81–97. doi:10.3233/SW-200382.

[55] G. Tuozzo, A. Lieto, Towards Automated FAIR Compliance Diagnosis, 2026. URL: https://github.com/GabrieleT0/Towards-Automated-FAIR-Compliance-Diagnosis.

[56] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. V. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, M. Goldblum, LiveBench: A Challenging, Contamination-Limited LLM Benchmark, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.

net/forum?id=sKYHBTAxVa.

[57] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jiang, B. Y. Lin, S. Welleck, P. West, C. Bhagavatula, R. L. Bras, J. D. Hwang, S. Sanyal, X. Ren, A. Ettinger, Z. Harchaoui, Y. Choi, Faith and Fate: Limits of Transformers on Compositionality, in: Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL: https://openreview.net/forum?id=Fkckkr3ya8.

[58] J. Osés Grijalba, L. A. Ureña-López, E. Martínez Cámara, J. Camacho-Collados, Question answering over tabular data with DataBench: A large-scale empirical evaluation of LLMs, in: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, 2024, pp. 13471–13488. URL: https://aclanthology.org/2024.lrec-main.1179/.

[59] Z. Zhang, X. Li, Y. Gao, J.-G. Lou, CRT-QA: A dataset of complex reasoning question answering over tabular data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2131–2153. URL: https://aclanthology.org/2023.emnlp-main.132/. doi:10.18653/v1/2023.emnlp-main.132.

Appendix

This section presents the complete set of 40 questions included in the FAIR *E&D* benchmark, details the prompt adopted in the zero-shot experimental setting across all evaluated LLMs, and provides an analysis of I Don't Know (IDK) responses.

A. FAIR *Explanation & Diagnosis* Questions

Table 3 lists all 40 questions manually curated in the FAIR *E&D* benchmark.

Table 3: FAIR principles *Explanation & Diagnosis* benchmark

ID	Question	Expected reply
	Clarification (CLAR)	
CLAR_1	Does a score of 0 in the A1.1-D sub-principle mean that a SPARQL endpoint or downloadable data file is not provided?	No, it means that both SPARQL endpoint and RDF dump are not defined.
CLAR_2	When the A1.1-D score is 1, does it mean that both the SPARQL endpoint and RDF dump are available?	No, it means that a SPARQL endpoint or an RDF dump is operational.
CLAR_3	How is compliance with the R1.2 sub-principle assessed?	The sub-principle R1.2 monitors the availability of publisher details, such as sources, authors, and contributors.
CLAR_4	Is the presence of a license required for Interoperability?	No, the license is required to improve the Reusability of a dataset, particularly for the sub-principle R1.1.
CLAR_5	How does the F1-M principle differ from the F3-M principle?	For the F1-M principle, a persistent ID from a search engine is sufficient, such as those provided by GitHub or the LOD Cloud. In contrast, the F3-M principle enforces this requirement by specifically requiring a DOI.
CLAR_6	If F2b-M is set to 1, does it mean that all metadata of the dataset provider are provided?	No, it means that all the required metadata to evaluate the FAIRness of the dataset are provided.
CLAR_7	Why might the I1-D score differ between two datasets, even if both provide an online data dump?	This can happen because one dataset has a data dump with a valid media type, while the other does not.
CLAR_8	Why might a dataset receive a low score for sub-principle I2, even if it correctly declares the vocabularies it uses?	Because to achieve the best possible score, all the declared vocabularies must comply with the FAIR principles.
CLAR_9	Does a high F2a-M score guarantee a high F2b-M score?	No, F2a-M only checks if metadata is available via standard primary sources (SPARQL endpoint, search engine, or VoID/DCAT), while F2b-M checks if all the required metadata attributes for FAIR evaluation are actually provided.

Continued on next page

ID	Question	Expected reply
CLAR_10	Is the R1.3-D subprinciple the same as A1.1-D?	No, while both involve SPARQL endpoints and data dumps, A1.1-D focuses on whether access points are operational, whereas R1.3-D checks if the data is organized in standardized ways with valid mediatypes or if ontologies are in OWL/RDF(S) format.
	Metric Calculation (MC)	
MC_1	How is the I2 sub-principle of a dataset calculated?	The sub-principle is calculated by verifying whether the declared vocabularies are FAIR, which is determined by their registration in the FAIRsharing registry.
MC_2	What factors determine a dataset's Accessibility score?	The Accessibility score is computed by linearly combining all the scores obtained in the following sub-principles: A1.1-D, A1.1-M, A1.2, and A2-M.
MC_3	How is the I1-M sub-principle calculated?	It is calculated by checking whether a VoID file is indicated or if metadata are available through the SPARQL endpoint.
MC_4	How is the presence of a unique and persistent identifier (e.g., DOI) verified for the F1-M principle?	The search-engine metadata provided with the dataset are analyzed, and it is checked whether the key DOI contains a value or if the search engine provides a persistent identifier for the dataset's metadata (such as GitHub or LOD Cloud).
MC_5	Are all dataset URIs included when computing URI dereferenceability?	No, the test is performed on 5,000 randomly sampled triples from the SPARQL endpoint. A GET request is sent to each URI in the triples to check whether it returns a 200 status code. Finally, the metric is calculated as the ratio of working URIs to the total number of URIs.
MC_6	How is the F1-D sub-principle calculated?	It is calculated by testing URI dereferenceability on 5,000 randomly sampled triples from the SPARQL endpoint. The metric is the ratio of dereferenceable URIs (returning a 200 status code) to the total URIs in the set.
MC_7	What's the difference in how F1-M and F3-M are calculated?	F1-M checks if the dataset is registered in a search engine that provides a persistent identifier (including DOI, GitHub, or LOD Cloud identifiers), while F3-M specifically requires a DOI to be attached to the metadata. F1-M is more lenient and accepts various types of persistent identifiers.

Continued on next page

ID	Question	Expected reply
MC_8	Why do some datasets score 0.5 in A1.1-D instead of 0 or 1?	The A1.1-D scoring function has three possible values: 1 for operational access points, 0.5 for accessible but non-operational SPARQL endpoints or data dumps, and 0 when neither is indicated or both are offline. The 0.5 score indicates that data dump or SPARQL endpoint is provided, but is not online.
MC_9	What determines whether a vocabulary is considered FAIR in the I2 calculation?	A vocabulary is considered FAIR if it is registered in the FAIRsharing registry. The I2 score is calculated as the ratio of FAIR vocabularies to the total number of vocabularies used in the dataset.
MC_10	Can a dataset achieve maximum Findability score without a SPARQL endpoint?	No, because for principle F1-D, F2a-M, F2b-M it is necessary to access the data within the dataset and check whether the URIs in the dataset are dereferenceable.
	Improvement (IMP)	
IMP_1	How can the I3-D score of a dataset be improved?	Indicate the linked dataset or connect the dataset to other datasets.
IMP_2	How can the Reusability score of a dataset be improved?	The Reusability score is composed of the following sub-principles: R1.1: declare a license; R1.2: provide all publisher information (sources, authors, contributors); R1.3-D: provide a SPARQL endpoint or a data dump with a valid media type, or ontologies released in OWL/RDF(S) format; R1.3-M metadata are published according to VoID/DCAT specifications.
IMP_3	How can the R1.3-D score of a dataset be improved?	Provide a SPARQL endpoint or data dump(s) with valid mediatypes, or release ontologies in OWL/RDF(S) format.
IMP_4	My dataset scored low on Interoperability. What are the main areas I should focus on?	Ensure data dumps have valid mediatypes or ontologies are in OWL/RDF(S) format (I1-D), publish metadata according to VoID/DCAT specifications (I1-M), use vocabularies registered in FAIRsharing (I2), and link your dataset to other datasets (I3-D).
IMP_5	What's the quickest way to improve my overall FAIR score?	Register your dataset in a search engine like LOD Cloud, provide metadata via VoID/DCAT specifications and include a working SPARQL endpoint. This single action can improve multiple sub-principles: F1-M, F2a-M, F4-M, and A2-M, F1-D, A1.1-D, A1.2, A2-M, I1-M, R1.3-D, R1.3M.

Continued on next page

ID	Question	Expected reply
IMP_6	I have a SPARQL endpoint but my Accessibility score is still low. What could be wrong?	Ensure the endpoint is operational, not just accessible (A1.1-D), verify that authentication requirements are automatically detectable (A1.2), confirm the endpoint is registered in search engines (A2-M), and make sure metadata can be retrieved through the endpoint (A1.1-M).
IMP_7	My F1-D score is only 0.2. What steps should I take to increase URI dereferenceability?	Implement proper content negotiation so URIs return appropriate RDF formats. Check that your server is configured to return 200 status codes for valid resources.
IMP_8	What can I do if my I2 score is 0 even though I've declared all vocabularies used?	Replace custom vocabularies with well-established FAIR vocabularies (Dublin Core, FOAF, SKOS, Schema.org, DCAT), or submit your custom vocabularies to the FAIRsharing registry for approval.
IMP_9	Can I compensate for not having a SPARQL endpoint by providing really comprehensive metadata?	Partially, you will get a high score only in the following subprinciples that are independent from the data in the dataset: F1-M, F2a-M, F2b-M, F3-M, F4-M, A1.1-M, A2-M, I1-M, I2, R1.1, R1.2, R1.3-M.
IMP_10	Should I fix metadata completeness or URI dereferenceability first?	Metadata completeness. It gives you cascading effects in other sub-principles; the URI dereferenceability only improves the sub-principle F1-D.
	Impact on Scoring (IoS)	
IoS_1	To what extent does the absence of a working SPARQL endpoint affect the overall FAIRness score?	All principles that involve the retrieval of data from the dataset, i.e., F1-D, F2a-M, F2b-M, A1.1-M, A1.1-D, A1.2, I1-D, I2, I3-D, R1.2, and R1.3-D. Therefore, all FAIR principle scores will decrease without a SPARQL endpoint.
IoS_2	Which FAIR sub-principles are affected when vocabularies are not declared?	I2, F2b-M.
IoS_3	How does the absence of FAIR vocabularies in a dataset affect its FAIRness?	This affects the Interoperability principle, specifically the I2 sub-principle (Use of FAIR vocabularies), and consequently impacts both the overall I score and the FAIR score.
IoS_4	My dataset links to 50 other datasets. Will this give me a higher I3-D score than linking to just 1?	No, I3-D is binary (0 or 1). It only checks whether your dataset contains at least one link to another dataset. Having 50 links gives the same score as having 1 link. However, more links may improve discoverability and usefulness even if not reflected in the score.
IoS_5	What factors contribute to a low F1-D sub-principle score?	This happens when a dataset contains only a few URIs that are correctly dereferenceable and return a 200 status code upon access.

Continued on next page

ID	Question	Expected reply
IoS_6	Which FAIR sub-principles are affected when no license is declared?	R1.1 (Any license retrievable via any primary source) and F2b-M.
IoS_7	How does the absence of a DOI affect the FAIR score?	It affects principle F1-M, F3-M and F2b-M, and consequently impacts both the overall F score and the FAIR score.
IoS_8	How does the absence of valid mediatypes in data dumps affect the dataset quality?	It affects the Interoperability and Reusability principles, specifically sub-principles F2b-M, I1-D and R1.3-D, lowering both the I and R scores and consequently the overall FAIR score.
IoS_9	Will adding more metadata always improve my F2b-M score?	Not necessarily. F2b-M only measures the coverage of required metadata attributes for FAIR evaluation, not the total amount of metadata. Adding metadata fields that aren't part of the FAIR evaluation criteria won't improve this score. Focus on the specific attributes: sparql endpoint indication, rdf dump with metadata media type, vocabularies used, DOI, author of the dataset (and/or publishers, contributors and source), linked dataset and license.
IoS_10	If my SPARQL endpoint requires authentication, will this hurt my FAIR score?	No, the only requirement is that the need for authentication can be detected automatically, for example, by receiving a 401 response when trying to access, rather than a generic error.

B. Prompt used

The prompt provided to the LLMs to carry out the FAIR *E&D* benchmark is presented below.

```
You are an expert in FAIR data principles (Findable, Accessible, Interoperable,
Reusable) assessment and scoring methodology.

You will be given:
1. A {documentation_format} document describing how FAIR principles and
   sub-principles are calculated.
2. A question about that methodology.

<documentation>
{fair_documentation}
</documentation>

<question>
{question}
</question>

Instructions:
- Answer based solely on the provided documentation.
- Be concise and answer directly.
```

```
- If you don't know the answer, say "IDK".
```

C. I Don't Know Response Analysis

In the prompt provided to the LLMs, we explicitly instruct the models to return "IDK" when they are unable to answer a question. In this section, we analyze the frequency of IDK responses across different models, question categories, and output formats. Specifically, Table 4 reports the mean IDK answer ratio generated by each LLM, stratified by question category and documentation format. The table is visualized as a heatmap, where darker colors indicate higher proportions of IDK responses. Lower values, therefore, correspond to better performance.

gpt-oss-20b exhibited the highest IDK answer rate (0.4), followed by DeepSeek-V3.2 Thinking (0.5). llama-3.3-70b and gpt-oss-120b showed lower rates (0.03 and 0.01, respectively), while Gemini-2.5-pro produced no IDK answers across all conditions.

IDK answers were overwhelmingly concentrated in the IMP category. Models abstained 0.09 answer rate in IMP questions (mean calculated on all LLMs), while all other categories were largely unaffected. Minor IDK activity appeared in IoS for DeepSeek-V3.2 Thinking and llama-3.3-70b (0.10). Notably, no IDK answers occurred in CLAR or MC.

The Turtle format yielded the highest IDK answer rate (0.03), followed by MD (0.02) and JSON (0.01). This suggests that structurally constrained representations marginally increase the likelihood of abstention.

Table 4
Mean IDK answer ratio across question types, enabling comparison across formats and models. The table is rendered as a heatmap, with darker colors indicating higher values. The lower the better.

	Markdown					JSON					Turtle				
Model	**CLAR**	**MC**	**IMP**	**IoS**	**Tot.**	**CLAR**	**MC**	**IMP**	**IoS**	**Tot.**	**CLAR**	**MC**	**IMP**	**IoS**	**Tot.**
gpt-oss-20b	0.00	0.00	0.20	0.00	0.05	0.00	0.00	0.10	0.00	0.02	0.00	0.00	0.30	0.00	0.07
LLama-3.3-70b	0.00	0.00	0.10	0.10	0.05	0.00	0.00	0.10	0.00	0.02	0.00	0.00	0.10	0.00	0.02
gpt-oss-120b	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.00	0.10	0.00	0.02
Gemini-2.5-pro	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeepSeek-V3.2 Think.	0.00	0.00	0.10	0.00	0.02	0.00	0.00	0.10	0.00	0.02	0.00	0.00	0.20	0.10	0.07
Mean Accuracy					0.02					0.01					0.03