Towards Automated FAIR Compliance Diagnosis:

Evaluating LLMs on Explanation and Diagnosis Questions

Gabriele Tuozzo
& Antonio Lieto

Università degli Studi di Salerno

UNIVERSITÀ DEGLI STUDI DI SALERNO

QKQ@ESWC'26, May 11, 2026
Dubrovnik, Croatia

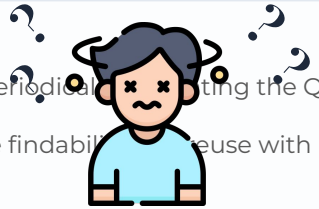
Why this is important?

Dataset	F1-M	F1-D	A1.1	I2	...	F Score	A Score	FAIR Score
Europeana	1	0	1	0	...	0.63	0.75	**0.65**

KGHeartBeat

Dataset	F1-A	F1-B	A1.1	I2	...	F Score	A Score	FAIR Score
Europeana	0	0	2	0	...	0	0.50	**0.08**

FAIR Checker



Pellegrino et al., KGHeartBeat: An Open Source Tool for Periodically Assessing the Quality of Knowledge Graphs. ISWC 2024

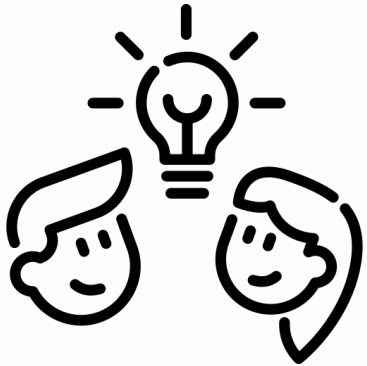
Gaignard et al., FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. Journal of Biomedical Semantics 2024.

Why this is important?

KGHeartBeat

ID	FAIR Principle	Input	Quality Dimension	Monitored Aspect	Scoring Function
F1-M	F1	Metadata		Unique and persistent ID	$\begin{cases} 1, & \text{dataset registered in a search engine that provides a persistent DOI} \\ 0, & \text{otherwise} \end{cases}$

Let's check the documentation!



FAIR Checker

F1A - Unique IDs

Metric information

Implementation: FAIR-Checker

Description:
FAIRChecker checks that the resource identifier is a reachable URL. It's better if the URL is persistent (WebID, PURL or DOI).

Why this is important?

Could you please explain what the FAIR Score F-1D means for the Europeana dataset? I calculated it using KGHeartBeat.
Why is the score zero, and **how** can I improve it?
Here is the data: <DATA>
And here is the documentation I found online:
<Documentation>

Don’t panic, I can help you 🙂

Objective of our work

- Evaluate the effectiveness of LLMs in supporting dataset producers, consumers, and maintainers in **explaining**, **diagnosing**, and **improving** dataset FAIRness.

- Investigate whether the **documentation format** provided to LLMs affects the accuracy of their FAIRness-related answers.

Research Questions (RQs)

1. To what extent are LLMs suitable for answering FAIR principles **explanation & diagnosis** (E&D) questions?

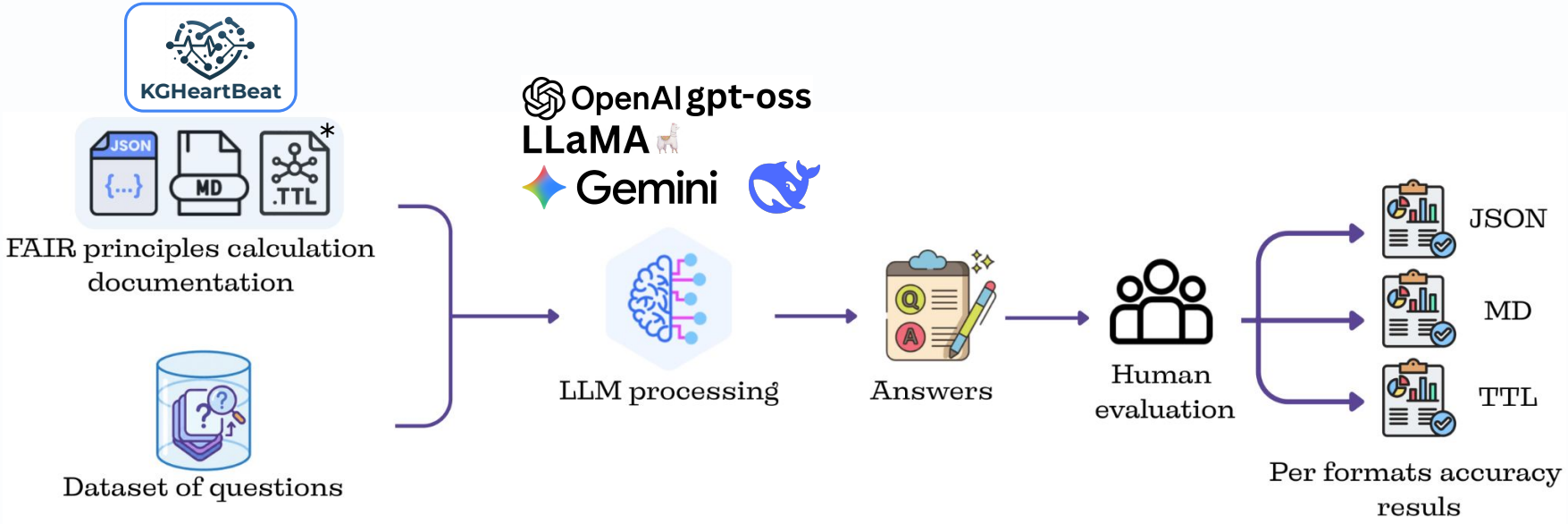
2. Does the **format of the documentation** provided to LLMs **influence the accuracy** of responses?

Benchmark for FAIR oriented E&D questions

Clarification (CLAR)	**Metric Calculation (MC)**	**Improvement (IMP)**	**Impact on Scoring (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?

40 Questions in total

Benchmark Execution Workflow



* Modelled using the Data Quality Vocabulary: https://www.w3.org/TR/vocab-dqv/

Results

	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

Values in range [0,1], higher is better. Darker colors indicate higher accuracy

Statistical analysis

Test whether **documentation format** and **model capability** affect LLM accuracy on FAIR E&D questions

No significant overall accuracy difference across JSON, Markdown, and Turtle.

Markdown outperformed **Turtle** for *Improvement* questions (p = 0.01).

Model capability did not significantly change **sensitivity to documentation format**.

Stronger models performed better on *Metric Calculation* and *Impact-on-Scoring* questions (r = 0.90, p = 0.03).

Conclusion

- LLMs **are broadly effective** at supporting dataset producers, consumers, and maintainers in diagnosing and improving dataset FAIRness (RQ1)

- While **no statistically significant difference** was found **across documentation formats** when considering overall accuracy, **category-level analysis** revealed that Markdown outperforms Turtle for Improvement questions (RQ2)

Limitations and Future Works

Limitations

- Evaluated **only zero-shot prompting**
- Benchmark is based on **one FAIR assessment tool**

Future Directions

- Expand the benchmark with **more questions**
- Test **additional RDF formats**, and **different vocabularies** and **ontologies** to model the knowledge graph documentation

UNIVERSITÀ DEGLI STUDI DI SALERNO

Thank you for your attention!

Acknowledgements

NEOLAiA

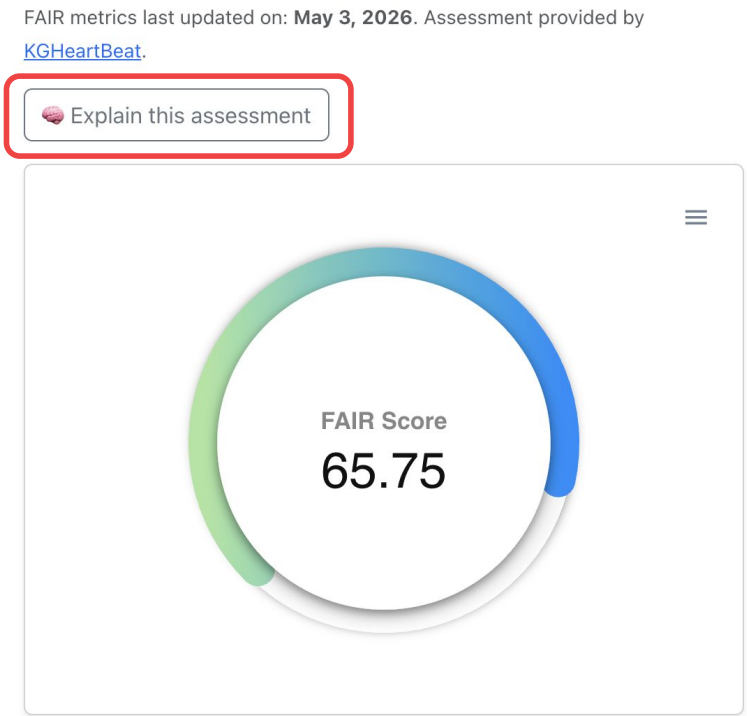
GOBLIN

Replication package

Gabriele Tuozzo
& Antonio Lieto

Contacts: gtuozzo@unisa.it

FAIR E&D with LLM in CHeCLOUD



https://checloud.di.unisa.it/