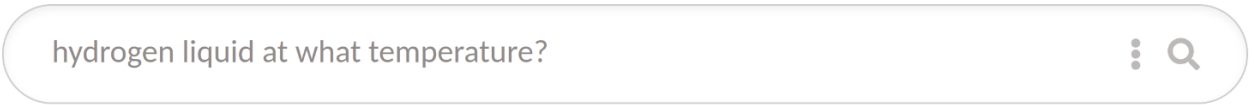
How Reliable are Information Retrieval Evaluations with Large Language Model Relevance Assessors?

Maik Fröbe
Friedrich-Schiller-Universität Jena
Frühjahrstreffen — Datenbanken, March 03–04, Regensburg, Germany

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations: Which Retrieval Algorithm Should We Use?



Maik Fröbe

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations: Which Retrieval Algorithm Should We Use?

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Maik Fröbe

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations: Which Retrieval Algorithm Should We Use?

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Maik Fröbe

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations: Which Retrieval Algorithm Should We Use?

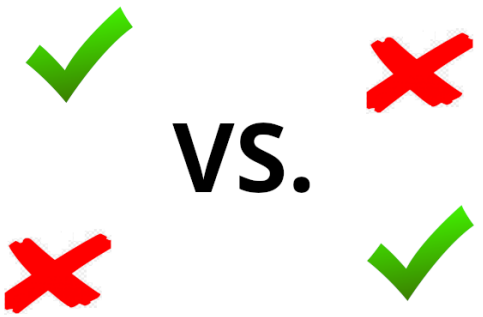
hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K



Maik Fröbe

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations: Which Retrieval Algorithm Should We Use?

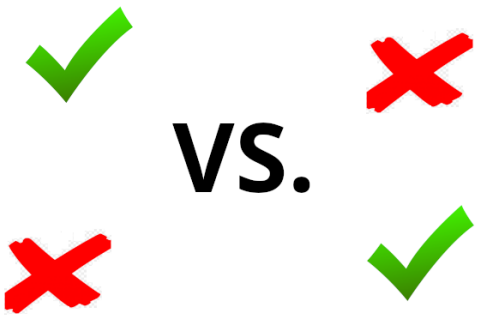
hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K



Observations We Are Interested In: Ranking of Retrieval Systems

System A > System B > System C > . . .

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations are Reliable If

[Voorhees'19]

Observations transfer to similar scenarios with a high probability

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations are Reliable If

[Voorhees'19]

Observations transfer to similar scenarios with a high probability

Two Main Aspects Impact Reliability

[Voorhees'19]

- Subjectiveness of relevance judgments
- Incompleteness of relevance judgments

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations are Reliable If

[Voorhees'19]

Observations transfer to similar scenarios with a high probability

Two Main Aspects Impact Reliability

[Voorhees'19]

- Subjectiveness of relevance judgments
- Incompleteness of relevance judgments

Reliability with Human Relavance Assessors Well Studied

- Expensive reliability experiments redo (parts of) an experiment
  [Burgin'92,Lesk'68,Parry'25,Voorhees'00]
- Cheap reliability experiments simulate repetitions fully automated
  [Soboroff'01,Carterette'10]

Reliability of IR Evaluations with LLM Relevance Assessors

Information Retrieval Evaluations are Reliable If

[Voorhees'19]

Observations transfer to similar scenarios with a high probability

Two Main Aspects Impact Reliability

[Voorhees'19]

- Subjectiveness of relevance judgments
- Incompleteness of relevance judgments

Reliability with Human Relavance Assessors Well Studied

- Expensive reliability experiments redo (parts of) an experiment
  [Burgin'92,Lesk'68,Parry'25,Voorhees'00]
- Cheap reliability experiments simulate repetitions fully automated
  [Soboroff'01,Carterette'10]

Reliability of Relevance Judgments by Large Language Models?

Reliability of IR Evaluations with LLM Relevance Assessors

Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Reliability of IR Evaluations with LLM Relevance Assessors

Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Agreement with Humans:

	Agreement	
	Relevance Judgments	**Resulting System Ranking**
LLM vs. Human	0.22	0.90

Reliability of IR Evaluations with LLM Relevance Assessors

Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Agreement with Humans:

	Agreement	
	Relevance Judgments	**Resulting System Ranking**
LLM vs. Human	0.22	0.90

Interpretation:

- Bad agreement on individual relevance judgments
- Perfect agreement resulting observations

Reliability of IR Evaluations with LLM Relevance Assessors

Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Positivity

Assessor	Proportion of Documents Judged Relevant
Human	46.3 %
Large Language Model	72.8 %

Reliability of IR Evaluations with LLM Relevance Assessors

Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Positivity

Assessor	Proportion of Documents Judged Relevant
Human	46.3 %
Large Language Model	72.8 %

Interpretation:

- Large language models judge much more documents as relevant (26.5 %)
- Incompleteness becomes a problem

Reliability of IR Evaluations with LLM Relevance Assessors

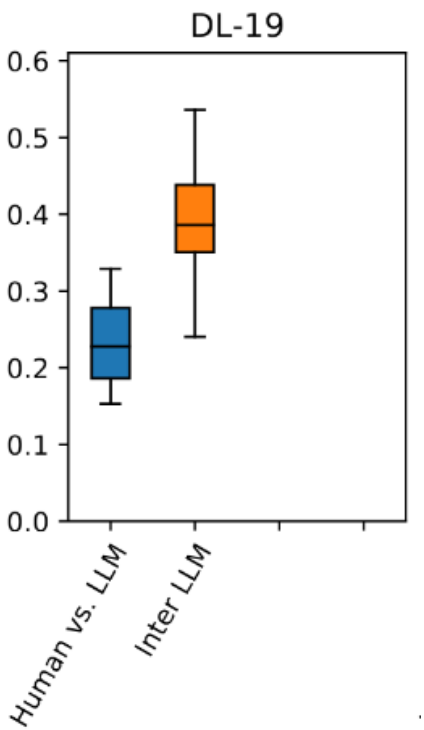
Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Agreement among LLMs



Reliability of IR Evaluations with LLM Relevance Assessors

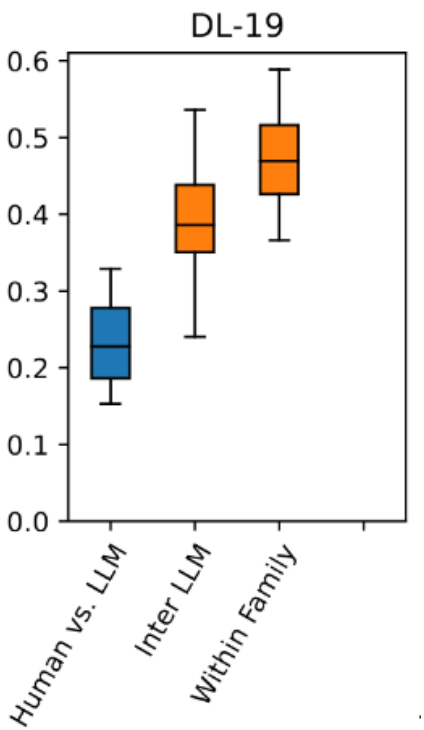
Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Agreement among LLMs



Reliability of IR Evaluations with LLM Relevance Assessors

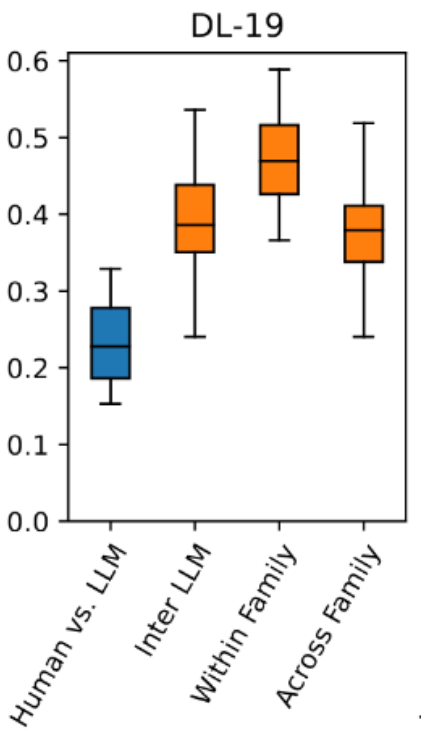
Expensive Reliability Experiments

[Example from Fröbe'25]

Setup (ca. 1500 Euro):

- We take 8 LLMs from 4 families with 1 relevance assessor framework
- 3 Retrieval Scenarios: TREC-DL19, TREC-DL20, and TREC-RAG24

Agreement among LLMs



Reliability of IR Evaluations with LLM Relevance Assessors

Cheap Reliability Experiments

[Under Review]

Setup (Very Cheap):

- We reproduce 10 fully automated reliability tests (> 700 with parameters)
- Simulate different aspects of subjectiveness and/or incompleteness on CPU

Reliability of IR Evaluations with LLM Relevance Assessors

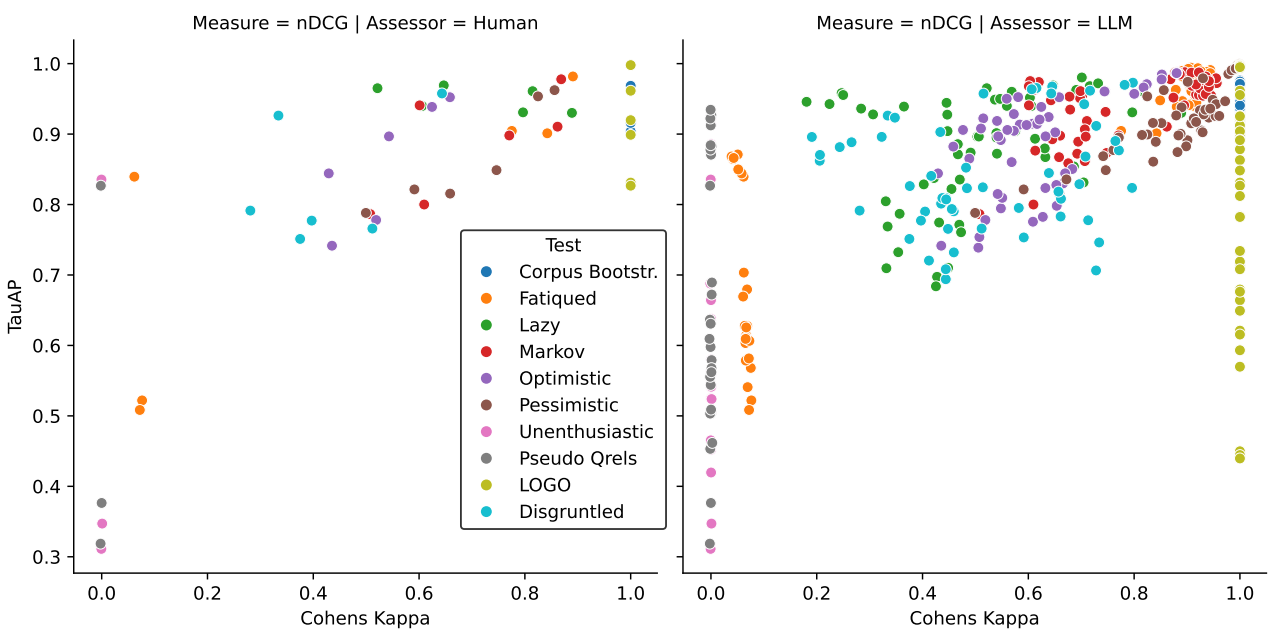
Cheap Reliability Experiments

[Under Review]

Setup (Very Cheap):

- We reproduce 10 fully automated reliability tests (> 700 with parameters)
- Simulate different aspects of subjectiveness and/or incompleteness on CPU

We compare reliability results for human judgments with LLM judgments:



Reliability of IR Evaluations with LLM Relevance Assessors

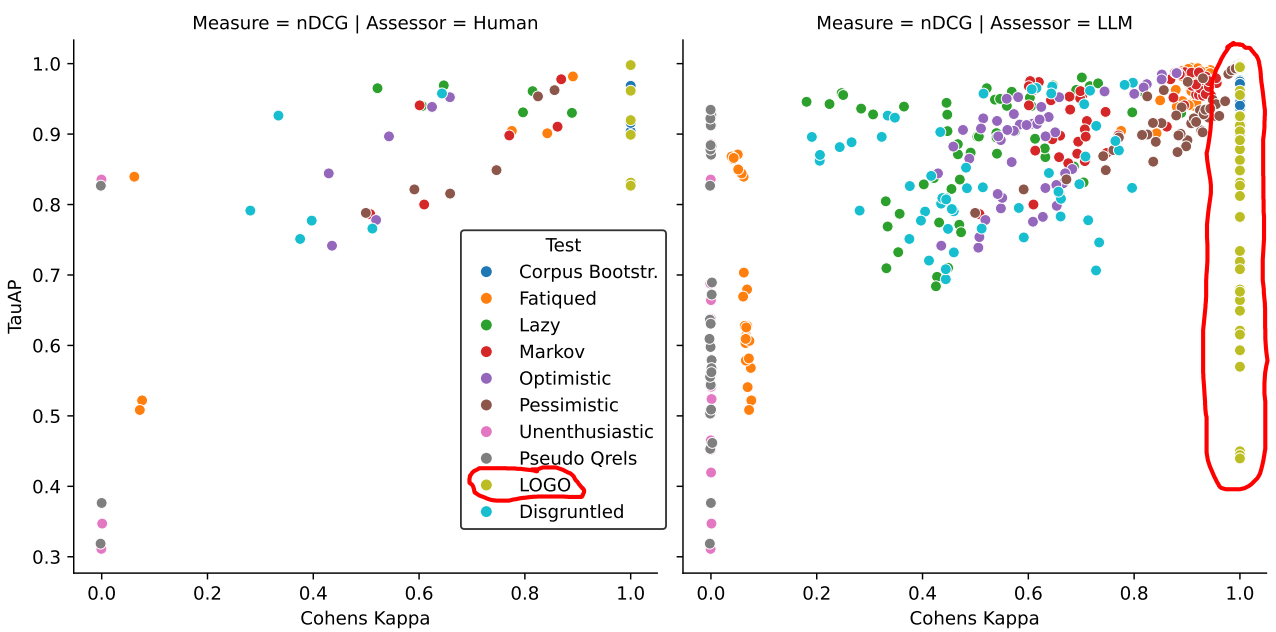
Cheap Reliability Experiments

[Under Review]

Setup (Very Cheap):

- We reproduce 10 fully automated reliability tests (> 700 with parameters)
- Simulate different aspects of subjectiveness and/or incompleteness on CPU

We compare reliability results for human judgments with LLM judgments:



Reliability of IR Evaluations with LLM Relevance Assessors

Cheap Reliability Experiments

[Under Review]

The LOGO reliability test indicates that incompleteness is a problem

- LLMs are too positive: many of the unjudged documents would be relevant

Reliability of IR Evaluations with LLM Relevance Assessors

Cheap Reliability Experiments

[Under Review]

The LOGO reliability test indicates that incompleteness is a problem

- LLMs are too positive: many of the unjudged documents would be relevant

Potential Solution (WIP): Formalize Information Need Before Judgments

- Given the query, use LLM to formalize criteria on what will be relevant or not before looking at the first document
- Judge documents with query and formalization

Reliability of IR Evaluations with LLM Relevance Assessors

Cheap Reliability Experiments

[Under Review]

The LOGO reliability test indicates that incompleteness is a problem

- LLMs are too positive: many of the unjudged documents would be relevant

Potential Solution (WIP): Formalize Information Need Before Judgments

- Given the query, use LLM to formalize criteria on what will be relevant or not before looking at the first document
- Judge documents with query and formalization

	Relevant Documents	LOGO Reliability Test
No Formalization (default)	66.3 %	0.63
Formalized Information Needs	**55.4 %**	**0.68**

Conclusions

Summary

- Low agreement of LLM assessors with humans on individual judgments
- Overall system rankings are still the same
- Problem with positivity of LLMs: LLM Judgments incomplete

Conclusions

Summary

- Low agreement of LLM assessors with humans on individual judgments
- Overall system rankings are still the same
- Problem with positivity of LLMs: LLM Judgments incomplete

Future work

- Reduce positivity of LLM relevance assessors
- Reduce agreement between different LLMs relevance assessors

Conclusions

Summary

- Low agreement of LLM assessors with humans on individual judgments
- Overall system rankings are still the same
- Problem with positivity of LLMs: LLM Judgments incomplete

Future work

- Reduce positivity of LLM relevance assessors
- Reduce agreement between different LLMs relevance assessors

thank you!

References

[wikipedia/Correlation_coefficient]
https://en.wikipedia.org/wiki/Correlation_coefficient

[trec-rag.github.io]
https://trec-rag.github.io

[github.com/webis-de/conf26-reliability-analysis]
https://github.com/webis-de/conf26-reliability-analysis

[Burgin'92]
Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. 1992.

[Balog'25]
Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. SIGIR 2025.

[Carterette'10]
Ben Carterette et al.: The Effect of Assessor Error on IR System Evaluation. SIGIR 2010.

[Lesk'68]
Relevance assessments and retrieval system evaluation.

[Soboroff'01]
Ian Soboroff et al.: Ranking Retrieval Systems without Relevance Judgments. SIGIR 2001

[Parry'25]
Andrew Parry et al.: Variations in relevance judgments and the shelf life of test collections. SIGIR 2024

[Faggioli'23]
Perspectives on Large Language Models for Relevance Judgment.

[Zobel'98]
Justin Zobel et al.: How Reliable Are the Results of Large-Scale Information Retrieval Experiments? SIGIR 1998.

[Fröbe'25]
Large Language Model Relevance Assessors Agree With One Another More Than With Human Assessors. SIGIR 2025.

[Upadhyay'24]
Upadhyay et al.: UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. 2024

[Clarke'24]

LM-based relevance assessment still cant replace human relevance assessment.

[Voorhees'00]
Ellen Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. 2000.