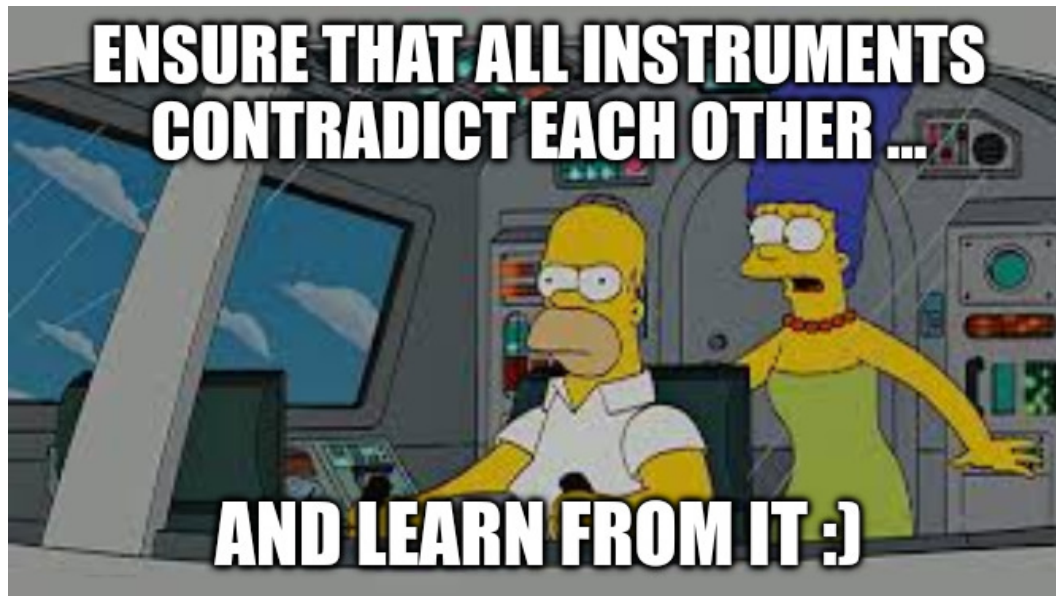


Reliability Tests for Information Retrieval Evaluations



Maik Fröbe

Friedrich-Schiller-Universität Jena

@webis_de

www.webis.de

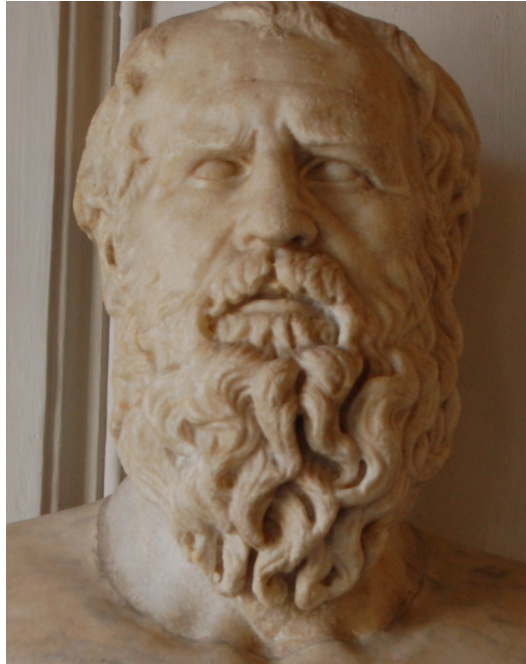
Reliability Tests for Information Retrieval Evaluations

The Only Constant in Life is Change.

Reliability Tests for Information Retrieval Evaluations

The Only Constant in Life is Change.

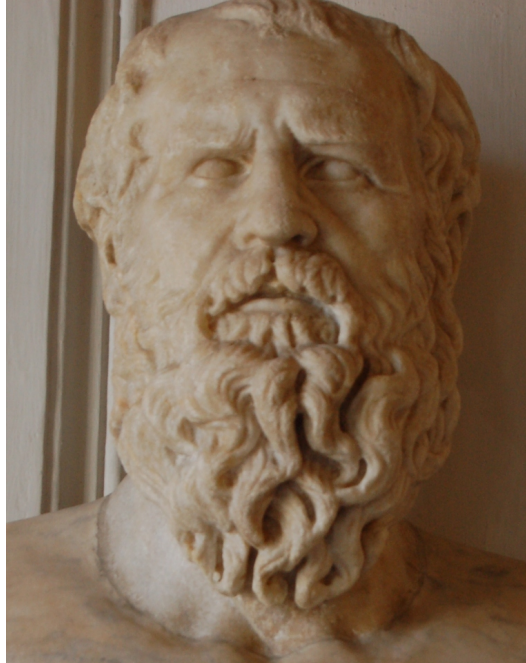
[Heraclitus; 500BC]



Reliability Tests for Information Retrieval Evaluations

The Only Constant in Life is Change.

[Heraclitus; 500BC]



This is also true for Information Retrieval Evaluations

- ❑ Many things in IR evaluation change all the time
- ❑ Reliability tests can help to understand changes and their impact

Reliability Tests for Information Retrieval Evaluations

Selected Changes in Information Access Systems

Past	Now	Change in IR Evaluation	
		Impact	What?

Reliability Tests for Information Retrieval Evaluations

Selected Changes in Information Access Systems

	Past	Now	Change in IR Evaluation	
			Impact	What?
Device	PC	Smartphone	Minor	new Evaluation Measure

Reliability Tests for Information Retrieval Evaluations

Selected Changes in Information Access Systems

	Past	Now	Change in IR Evaluation	
			Impact	What?
Device	PC	Smartphone	Minor	new Evaluation Measure
Documents	thousand	million/billion	Major	Cranfield Paradigm

Reliability Tests for Information Retrieval Evaluations

Selected Changes in Information Access Systems

	Past	Now	Change in IR Evaluation	
			Impact	What?
Device	PC	Smartphone	Minor	new Evaluation Measure
Documents	thousand	million/billion	Major	Cranfield Paradigm
Presentation	extractive	RAG	??	??

Reliability Tests for Information Retrieval Evaluations

Selected Changes in Information Access Systems

	Past	Now	Change in IR Evaluation	
			Impact	What?
Device	PC	Smartphone	Minor	new Evaluation Measure
Documents	thousand	million/billion	Major	Cranfield Paradigm
Presentation	extractive	RAG	??	??

What will evaluation change for RAG/LLM evaluations?

Reliability Tests for Information Retrieval Evaluations

Discussion on Changes for RAG/LLM Evaluation at a Seminar at Dagstuhl



Reliability Tests for Information Retrieval Evaluations

Discussion on Changes for RAG/LLM Evaluation at a Seminar at Dagstuhl



One of the developed Hypotheses on what will change

	Traditional IR	RAG
Costs	High	Low
Longevity	High	Low

Reliability Tests for Information Retrieval Evaluations

Advantages of High Longevity of Traditional IR Evaluation Resources

Overview of Touché 2020: Argument Retrieval

Alexander Bondarenko,¹ Maik Fröbe,¹ Meriem Beloucif,² Lukas Gienapp,³
Yamen Ajjour,¹ Alexander Panchenko,⁴ Chris Biemann,² Benno Stein,⁵
Henning Wachsmuth,⁶ Martin Potthast,³ and Matthias Hagen¹

Reliability Tests for Information Retrieval Evaluations

Advantages of High Longevity of Traditional IR Evaluation Resources

Overview of Touché 2020: Argument Retrieval

Alexander Bondarenko,¹ Maik Fröbe,¹ Meriem Beloucif,² Lukas Gienapp,³
Yamen Ajjour,¹ Alexander Panchenko,⁴ Chris Biemann,² Benno Stein,⁵
Henning Wachsmuth,⁶ Martin Potthast,³ and Matthias Hagen¹

- ❑ We used crowdsourcing to get relevance judgments
- ❑ We did run all evaluations, drew conclusions, derived insights
- ❑ After ca. one year, we got feedback that the relevance judgments look weird
- ❑ We did spot-checks: relevance judgments were mostly random
- ❑ We did all relevance judgments again
 - In-house annotators (mostly PhD students)
 - annotator training
 - pilot rounds with agreement tests
 - discussions for edge cases

Reliability Tests for Information Retrieval Evaluations

Advantages of High Longevity of Traditional IR Evaluation Resources

Corrected Evaluation Results of the NTCIR WWW-2, WWW-3, and WWW-4 English Subtasks

Tetsuya Sakai, Sijie Tao
Waseda University, Japan
tetsuyasakai@acm.org
tsjmailbox@ruri.waseda.jp

Maria Maistro
University of Copenhagen, Denmark
mm@di.ku.dk

Zhumin Chu
Tsinghua University, P.R.C.
chuzm19@mails.tsinghua.edu.cn

Yujing Li, Nuo Chen
Waseda University, Japan
liyujing@ruri.waseda.jp
pleviumtan@toki.waseda.jp

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Junjie Wang
Waseda University, Japan
wjj1020181822@toki.waseda.jp

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

Reliability Tests for Information Retrieval Evaluations

Advantages of High Longevity of Traditional IR Evaluation Resources

Corrected Evaluation Results of the NTCIR WWW-2, WWW-3, and WWW-4 English Subtasks

Tetsuya Sakai, Sijie Tao
Waseda University, Japan
tetsuyasakai@acm.org
tsjmailbox@ruri.waseda.jp

Maria Maistro
University of Copenhagen, Denmark
mm@di.ku.dk

Zhumin Chu
Tsinghua University, P.R.C.
chuzm19@mails.tsinghua.edu.cn

Yujing Li, Nuo Chen
Waseda University, Japan
liyujing@ruri.waseda.jp
pleviumtan@toki.waseda.jp

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Junjie Wang
Waseda University, Japan
wjj1020181822@toki.waseda.jp

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

- ❑ Two pooling variants were combined for relevance judgments
- ❑ The relevance judgments interface had a bug for one pooling variant
- ❑ All outcomes for one of the two pooling variants were incorrect
- ❑ Result: Noise in the relevance judgments
- ❑ Concerns results between 2019 and 2021
- ❑ Bug identified and fixed in 2022

Reliability Tests for Information Retrieval Evaluations

Advantages of high Longevity

- ❑ High longevity: re-use helped us to identify those problematic cases
- ❑ Low Longevity: no often re-use $\Rightarrow$ problems likely would be unnoticed

Reliability Tests for Information Retrieval Evaluations

Advantages of high Longevity

- ❑ High longevity: re-use helped us to identify those problematic cases
- ❑ Low Longevity: no often re-use $\Rightarrow$ problems likely would be unnoticed

Take-Away

- ❑ Errors/Mistakes will always happen
- ❑ For most IR evaluations you can not proof that there is no error
- ❑ No problem that errors occur
- ❑ In classical software engineering: unit tests. Advantages of good unit tests
 - Fast feedback
 - Constructive feedback
 - Easy

Reliability Tests for Information Retrieval Evaluations

Advantages of high Longevity

- ❑ High longevity: re-use helped us to identify those problematic cases
- ❑ Low Longevity: no often re-use $\Rightarrow$ problems likely would be unnoticed

Take-Away

- ❑ Errors/Mistakes will always happen
- ❑ For most IR evaluations you can not proof that there is no error
- ❑ No problem that errors occur
- ❑ In classical software engineering: unit tests. Advantages of good unit tests
 - Fast feedback
 - Constructive feedback
 - Easy

Implication for modern RAG evaluation

- ❑ If Longevity goes down for RAG evaluation, problems maybe not identified

Reliability tests can be like unit tests for IR Evaluations

Reliability Tests for Information Retrieval Evaluations

Reliability

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 Tests for Information Retrieval Evaluations

Reliability: Example on Subjectiveness



hydrogen liquid at what temperature?



Reliability Tests for Information Retrieval Evaluations

Reliability: Example on Subjectiveness



hydrogen liquid at what temperature?



What is the temperature of liquid hydrogen?

Hydrogen becomes liquid at $-252.87\text{ }^{\circ}\text{C}$

Liquid hydrogen

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

Reliability Tests for Information Retrieval Evaluations

Reliability: Example on Subjectiveness



hydrogen liquid at what temperature?



What is the temperature of liquid hydrogen?

Hydrogen becomes liquid at $-252.87\text{ }^{\circ}\text{C}$



Liquid hydrogen

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



Reliability Tests for Information Retrieval Evaluations

Reliability: Example on Subjectiveness



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



VS.



Reliability Tests for Information Retrieval Evaluations

Reliability: Example on Subjectiveness



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



VS.



Our goal in this talk. Learn how ...

- ❑ the choice of your evaluation measure impacts the design of your evaluation
- ❑ we can get more confidence in the reliability of our evaluations
- ❑ you can break your experiment. Everything can be made non-reliable, when does your experiment start to become unreliable?

Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

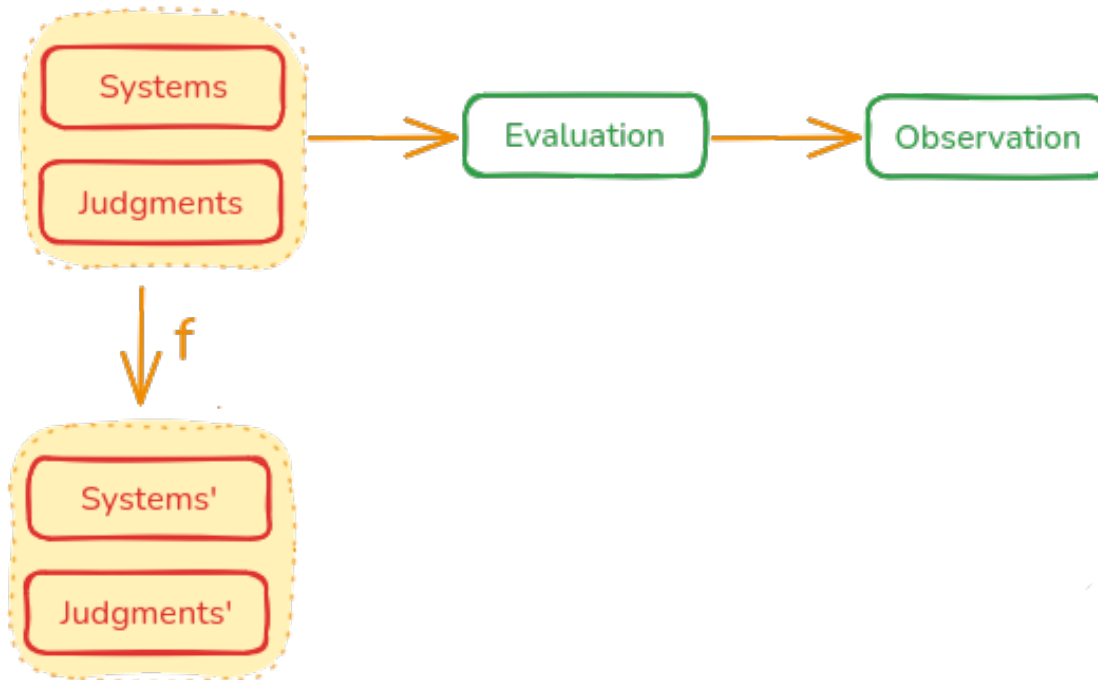
Experiment



Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

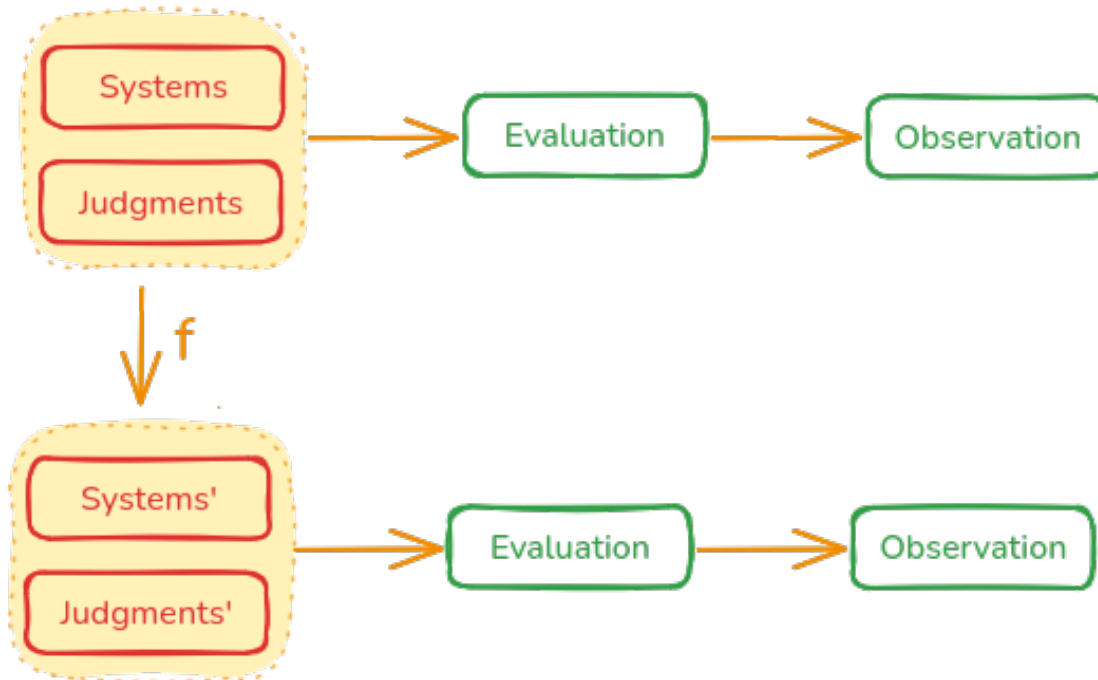
Experiment



Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

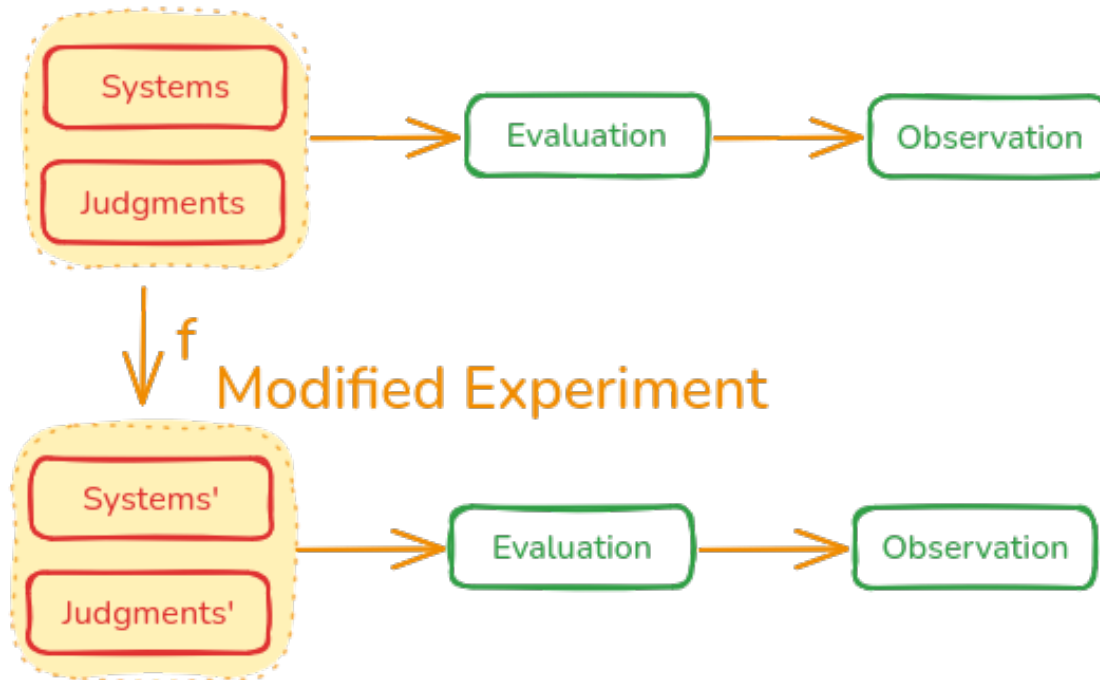
Experiment



Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

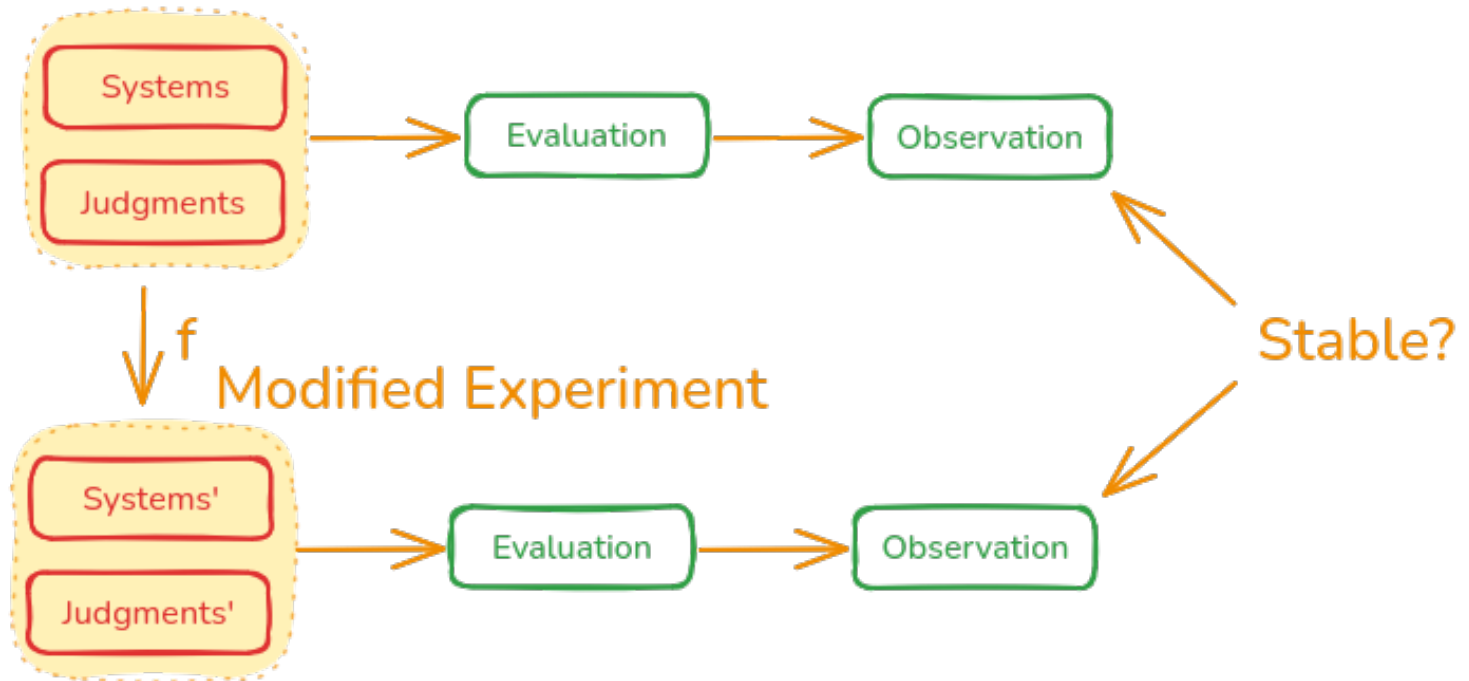
Experiment



Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

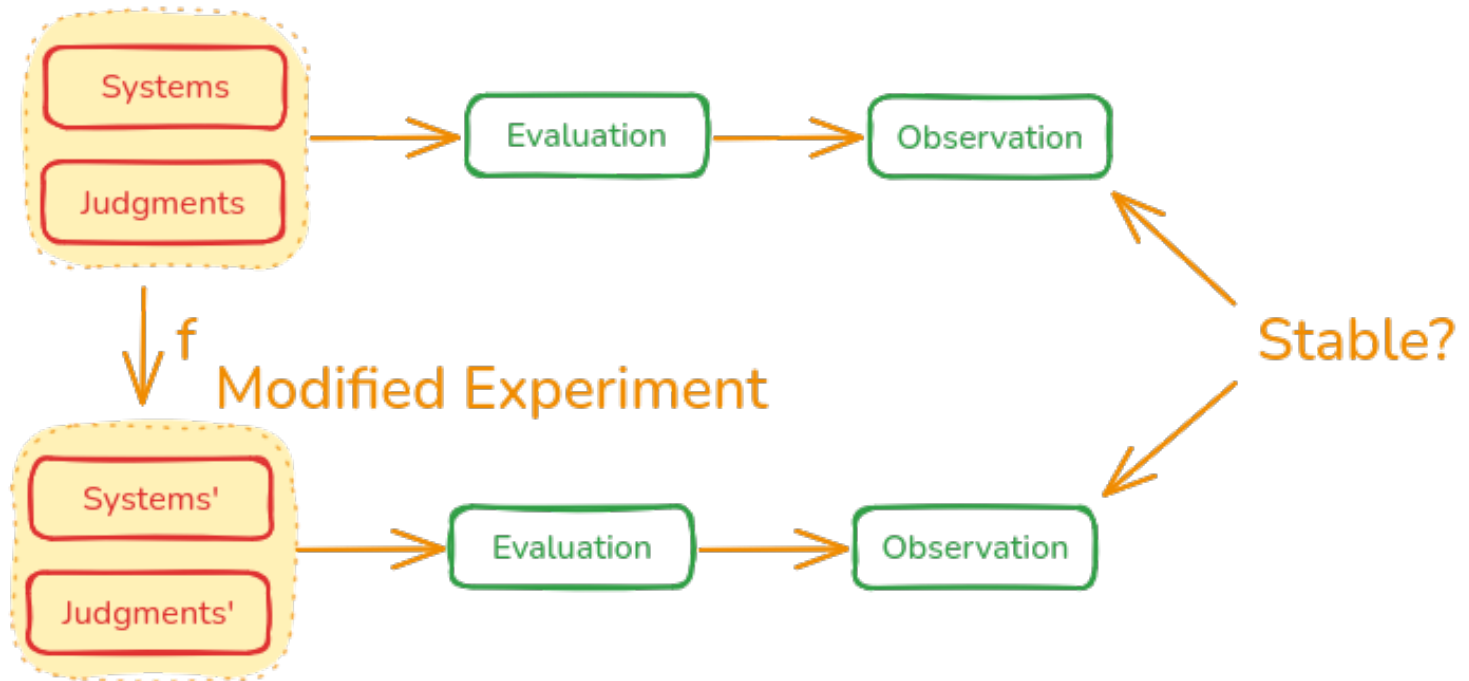
Experiment



Reliability Tests for Information Retrieval Evaluations

Overview of Reliability Tests

Experiment



The Function f is a Reliability Test

$$f : \text{Systems} \times \text{Judgments} \mapsto \text{Systems} \times \text{Judgments}$$

Reliability Tests for Information Retrieval Evaluations

How could Reliability Tests *f* Look Like?

There are expensive instantiations of reliability tests *f*

- ❑ Reimplement system(s) from scratch, e.g.,
 - Different programming languages/libraries
 - Different level of programming experience
- ❑ Redo all judgments with new annotators, e.g.,
 - Use different annotation user interface
 - Vary bias in annotators, e.g., education, political opinions, . . .

Reliability Tests for Information Retrieval Evaluations

How could Reliability Tests *f* Look Like?

There are expensive instantiations of reliability tests *f*

- ❑ Reimplement system(s) from scratch, e.g.,
 - Different programming languages/libraries
 - Different level of programming experience
- ❑ Redo all judgments with new annotators, e.g.,
 - Use different annotation user interface
 - Vary bias in annotators, e.g., education, political opinions, . . .

Expensive reliability tests have a high value but can only explore few scenarios

Reliability Tests for Information Retrieval Evaluations

How could Reliability Tests f Look Like?

There are expensive instantiations of reliability tests f

- ❑ Reimplement system(s) from scratch, e.g.,
 - Different programming languages/libraries
 - Different level of programming experience
- ❑ Redo all judgments with new annotators, e.g.,
 - Use different annotation user interface
 - Vary bias in annotators, e.g., education, political opinions, ...

Expensive reliability tests have a high value but can only explore few scenarios

Assumption at the RAG Dagstuhl workshop was that Longevity goes down

- ❑ We can not run expensive reliability tests

Reliability Tests for Information Retrieval Evaluations

How could Reliability Tests f Look Like?

There are expensive instantiations of reliability tests f

- ❑ Reimplement system(s) from scratch, e.g.,
 - Different programming languages/libraries
 - Different level of programming experience
- ❑ Redo all judgments with new annotators, e.g.,
 - Use different annotation user interface
 - Vary bias in annotators, e.g., education, political opinions, ...

Expensive reliability tests have a high value but can only explore few scenarios

Assumption at the RAG Dagstuhl workshop was that Longevity goes down

- ❑ We can not run expensive reliability tests

In the rest of the talk we will therefore focus on reliability tests that are:

- ❑ simple + fast $\Rightarrow$ allow for constructive feedback
- ❑ fully automated

Reliability Tests for Information Retrieval Evaluations

What are potential Observations?

- ❑ Is the system ranking stable?
- ❑ Is system A significantly more effective than system B?
- ❑ Is system A practically equivalent to system B? (important for efficiency)

Reliability Tests for Information Retrieval Evaluations

What are potential Observations?

- ❑ **Is the system ranking stable? (Focus of this talk)**
- ❑ Is system A significantly more effective than system B?
- ❑ Is system A practically equivalent to system B? (important for efficiency)

Reliability Tests for Information Retrieval Evaluations

How to Measure if observations are stable?

If an observation is stable depends on the observations we are interested in.

- ❑ We focus on system rankings as observations.
- ❑ Ranking Correlation: measure if two system rankings are similar or not

Ranking Correlations:

- ❑ Kendalls τ , Pearson ρ , τ_{AP} , Spearman, ...
- ❑ Correlations are always between 1 and -1. Interpretation:
[wikipedia/Correlation_coefficient]
 - 1: the two rankings are identical
 - ≥ 0.8 : perfect correlation
 - $\geq 0.6 < 0.8$: strong correlation
 - $\geq 0.4 < 0.6$: moderate correlation
 - $\geq 0.2 < 0.4$: weak correlation
 - $> 0 < 0.2$: no/very weak correlation
 - 0: random correlation

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 2 (e.g., after a reliability test) yields System Ranking 2

System Ranking 1	System Ranking 2
BM25	BM25
BERT	BERT
TF-IDF	TF-IDF
TF	TF

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 2 (e.g., after a reliability test) yields System Ranking 2

System Ranking 1	System Ranking 2
BM25	BM25
BERT	BERT
TF-IDF	TF-IDF
TF	TF

Resulting correlation:

$$\tau = 1$$

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 3 (e.g., after a reliability test) yields System Ranking 3

System Ranking 1	System Ranking 3
BM25	BM25
BERT	BERT
TF-IDF	TF
TF	TF-IDF

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 3 (e.g., after a reliability test) yields System Ranking 3

System Ranking 1	System Ranking 3
BM25	BM25
BERT	BERT
TF-IDF	TF
TF	TF-IDF

Resulting correlation:

$$\tau = 0.8$$

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 4 (e.g., after a reliability test) yields System Ranking 4

System Ranking 1	System Ranking 4
BM25	TF-IDF
BERT	BERT
TF-IDF	BM25
TF	TF

Reliability Tests for Information Retrieval Evaluations

Examples of Correlations

We order systems by their effectiveness, high to low effectiveness

- ❑ Scenario 1 yields System Ranking 1
- ❑ Scenario 4 (e.g., after a reliability test) yields System Ranking 4

System Ranking 1	System Ranking 4
BM25	TF-IDF
BERT	BERT
TF-IDF	BM25
TF	TF

Resulting correlation:

$$\tau = -0.1$$

Now we get our Hands Dirty



Now we get our Hands Dirty



Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Idea of Pseudo Qrels

- ❑ Introduce the maximal Noise possible
- ❑ Remove relevance judgments completely from the evaluation
- ❑ Use random relevance judgments that still follow the correct distribution

Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Idea of Pseudo Qrels

- ❑ Introduce the maximal Noise possible
- ❑ Remove relevance judgments completely from the evaluation
- ❑ Use random relevance judgments that still follow the correct distribution

Input (for each query)

- ❑ Mapping of document to the relevance of that document

Output (for each query)

- ❑ Mapping of document to the relevance of that document

Implementation

```
def pseudo_qrels(judgments: Dict[str, int]):  
    shuffled = random.shuffle(judgments.values)  
    return {d: l for d, l in zip(judgments.keys(), shuffled)}
```

Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Idea of Pseudo Qrels

- ❑ Introduce the maximal Noise possible
- ❑ Remove relevance judgments completely from the evaluation
- ❑ Use random relevance judgments that still follow the correct distribution

Input (for each query)

- ❑ Mapping of document to the relevance of that document

Output (for each query)

- ❑ Mapping of document to the relevance of that document

Implementation

```
def pseudo_qrels(judgments: Dict[str, int]):  
    shuffled = random.shuffle(judgments.values)  
    return {d: l for d, l in zip(judgments.keys(), shuffled)}
```

Example

Document	Input Relevance	Output Relevance
Doc-1	0	1
Doc-2	0	0
Doc-3	1	0

Next slide shows results for MAP, if time allows, first whiteboard to collect preferences

Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Observations by Soboroff et al.

Measure: MAP. Randomization is included: we repeat test 50 times. Results:

	mean τ	st.dev τ
TREC-3	0.430	0.0312
TREC-5	0.487	0.0462
TREC-6	0.408	0.0354
TREC-7	0.369	0.0363
TREC-8	0.459	0.0340

Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Observations by Soboroff et al.

Measure: MAP. Randomization is included: we repeat test 50 times. Results:

	mean τ	st.dev τ
TREC-3	0.430	0.0312
TREC-5	0.487	0.0462
TREC-6	0.408	0.0354
TREC-7	0.369	0.0363
TREC-8	0.459	0.0340

Interpretation:

- ❑ Positive Aspects: MAP reliable against Noise (often moderate correlations)
- ❑ Negative Aspects: Human judgment budget is to some degree wasted
- ❑ Possible potential to skip costly labeling for some documents

Reliability Test (1): Pseudo Relevance Judgments [Soboroff'01]

Observations by Soboroff et al.

Measure: MAP. Randomization is included: we repeat test 50 times. Results:

	mean τ	st.dev τ
TREC-3	0.430	0.0312
TREC-5	0.487	0.0462
TREC-6	0.408	0.0354
TREC-7	0.369	0.0363
TREC-8	0.459	0.0340

Interpretation:

- ❑ Positive Aspects: MAP reliable against Noise (often moderate correlations)
- ❑ Negative Aspects: Human judgment budget is to some degree wasted
- ❑ Possible potential to skip costly labeling for some documents

Can we derive generic recommendations? Not directly, because:

- ❑ What if users change their behaviour? (I.e., different evaluation measure)
- ❑ What when we change the task?

Reliability Test (1): Pseudo Relevance Judgments

We build a big public repository with results of reliability tests for comparisons

[<https://sigir26-paper-141.pages.dev/>]

- ❑ Tasks: 32 (covering more than 30 years)
- ❑ Measures: 24
- ❑ Reliability tests: 10 unique tests (≥ 700 with parameters)
- ❑ All executions tracked with the Tirex Tracker
 - 42 000 CPU hours (more than 4 years; but highly parallelized)

Reliability Test (1): Pseudo Relevance Judgments

We build a big public repository with results of reliability tests for comparisons

[<https://sigir26-paper-141.pages.dev/>]

- ❑ Tasks: 32 (covering more than 30 years)
- ❑ Measures: 24
- ❑ Reliability tests: 10 unique tests (≥ 700 with parameters)
- ❑ All executions tracked with the Tirex Tracker
 - 42 000 CPU hours (more than 4 years; but highly parallelized)

Idea: you can easily run all tests on your experiment and compare against public repository

[<https://sigir26-paper-141.pages.dev/apply?tab=2>]

- ❑ ToDo: Show CLI Example

Reliability Test (1): Pseudo Relevance Judgments

Can we reproduce Soboroff et al.?

Results (50 repetitions):

	Original		Our Reproduction	
	mean τ	st.dev τ	mean τ	st.dev τ
TREC-3	0.430	0.0312	0.430	0.0282
TREC-5	0.487	0.0462	0.487	0.0491
TREC-6	0.408	0.0354	0.405	0.0497
TREC-7	0.369	0.0363	0.362	0.0438
TREC-8	0.459	0.0340	0.464	0.0424
Median	0.430	0.0354	0.430	0.0438

Reliability Test (1): Pseudo Relevance Judgments

Do the findings generalize Accross (32) Tasks?

Results (50 repetitions for 32 tasks, median τ MAP on 5 Tasks was 0.430):

Measure	Median τ
---------	---------------

MAP	0.561
-----	-------

Reliability Test (1): Pseudo Relevance Judgments

Do the findings generalize Accross (32) Tasks?

Results (50 repetitions for 32 tasks, median τ MAP on 5 Tasks was 0.430):

Measure	Median τ
---------	---------------

MAP	0.561
-----	-------

Interpretation:

- ❑ Soboroff found that MAP had moderate correlations for random judgments
- ❑ On the 32 tasks, we also observe moderate, almost strong correlations

The Findings of Soboroff generalize Accross our observed Tasks

Reliability Test (1): Pseudo Relevance Judgments

Do the findings generalize Accross User Behaviour (i.e., evaluation measures)?

Reliability Test (1): Pseudo Relevance Judgments

Do the findings generalize Accross User Behaviour (i.e., evaluation measures)?

Results (50 repetitions for 32 tasks for 24 evaluation measures):

Measure	Rank	Median τ
SetP	1	0.739
SetR	4	0.713
R@1000	5	0.709
nDCG	6	0.610
MAP	8	0.561
nDCG@20	12	0.109
nDCG@10	14	0.054
RR	16	0.042
P@1	24	0.035

Reliability Test (1): Pseudo Relevance Judgments

Do the findings generalize Across User Behaviour (i.e., evaluation measures)?

Results (50 repetitions for 32 tasks for 24 evaluation measures):

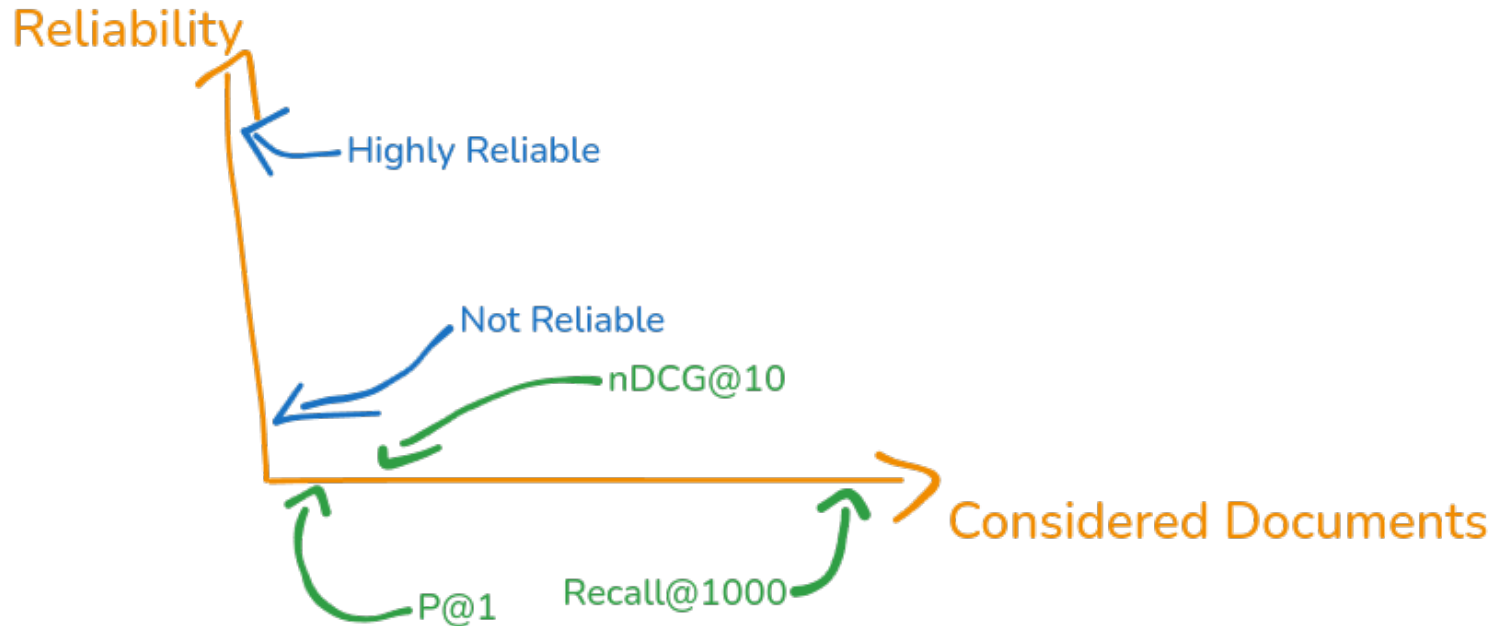
Measure	Rank	Median τ
SetP	1	0.739
SetR	4	0.713
R@1000	5	0.709
nDCG	6	0.610
MAP	8	0.561
nDCG@20	12	0.109
nDCG@10	14	0.054
RR	16	0.042
P@1	24	0.035

Interpretation:

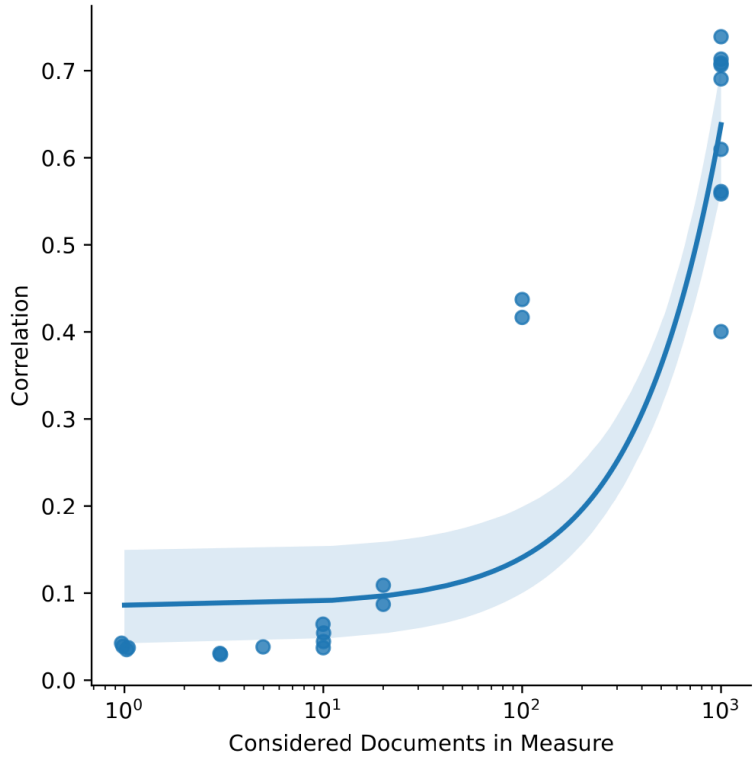
- ❑ Some measures (SetP, SetR, R@1000, nDCG) show strong correlation
- ❑ Many show no correlation (nDCG@20, nDCG@10, RR, P@1, etc.)
 - Less reliable for Noise, but require judgments

Reliability Test (1): Pseudo Relevance Judgments

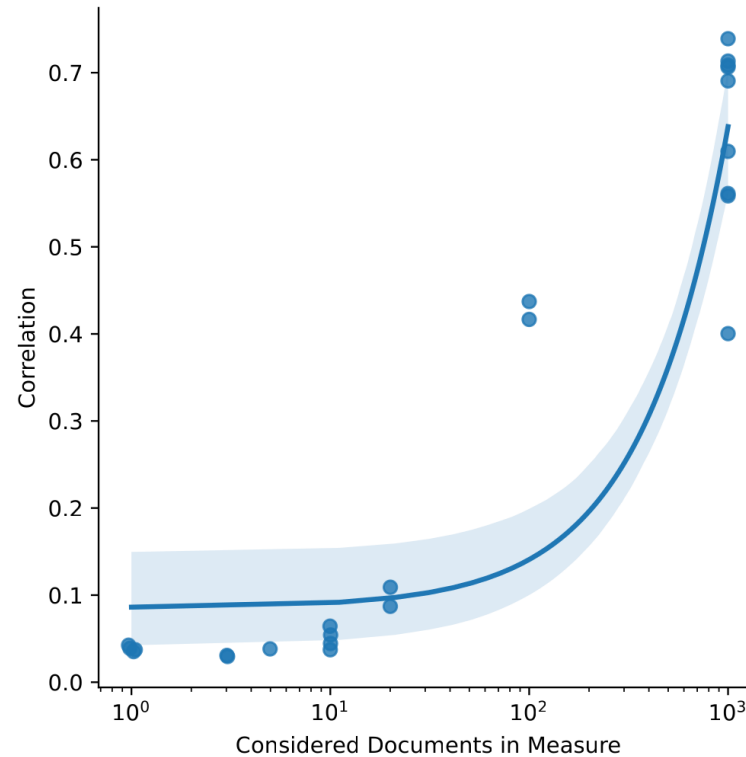
Taking a Step back to see the bigger picture



Reliability Test (1): Pseudo Relevance Judgments



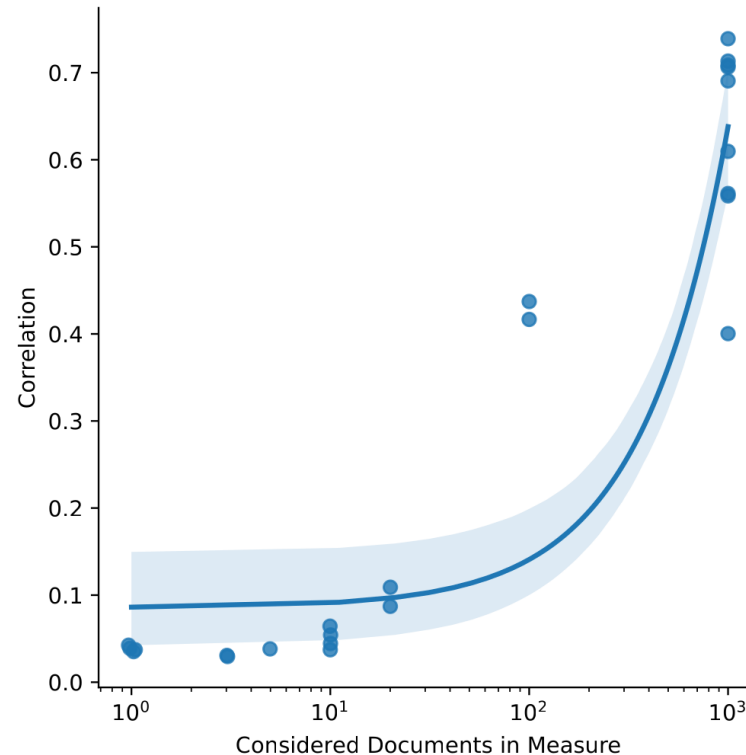
Reliability Test (1): Pseudo Relevance Judgments



Interpretation

- Measures that consider more documents are more reliable w.r.t. Noise

Reliability Test (1): Pseudo Relevance Judgments



Interpretation

- ❑ Measures that consider more documents are more reliable w.r.t. Noise

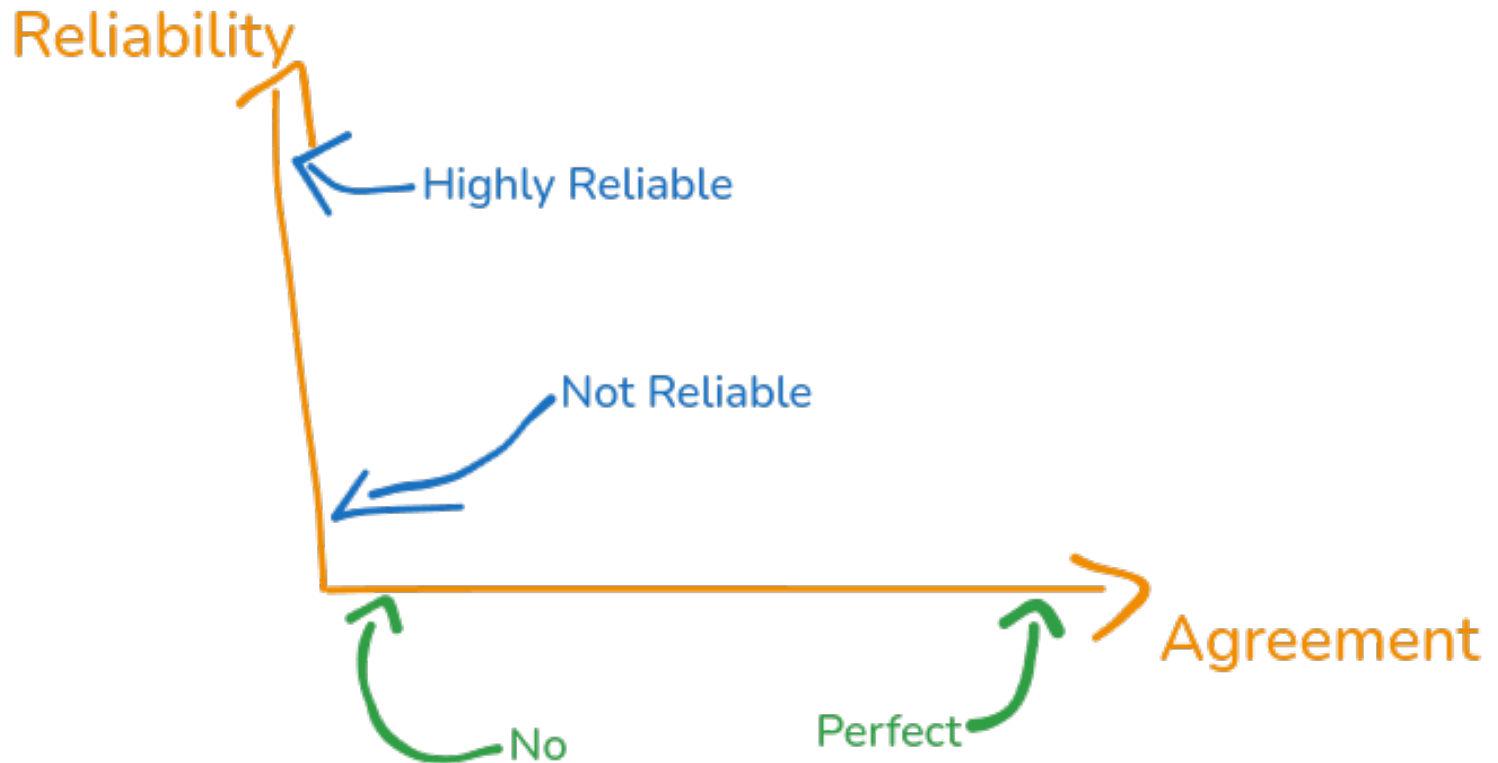
Interpretation aligned with general rule of thumb:

- ❑ Problem with noise? Resolve by adding more data points
- ❑ Problem with bias? Does not go away with simply adding more data

Scale to More Reliability Tests

A Unified view on Diverse Reliability tests

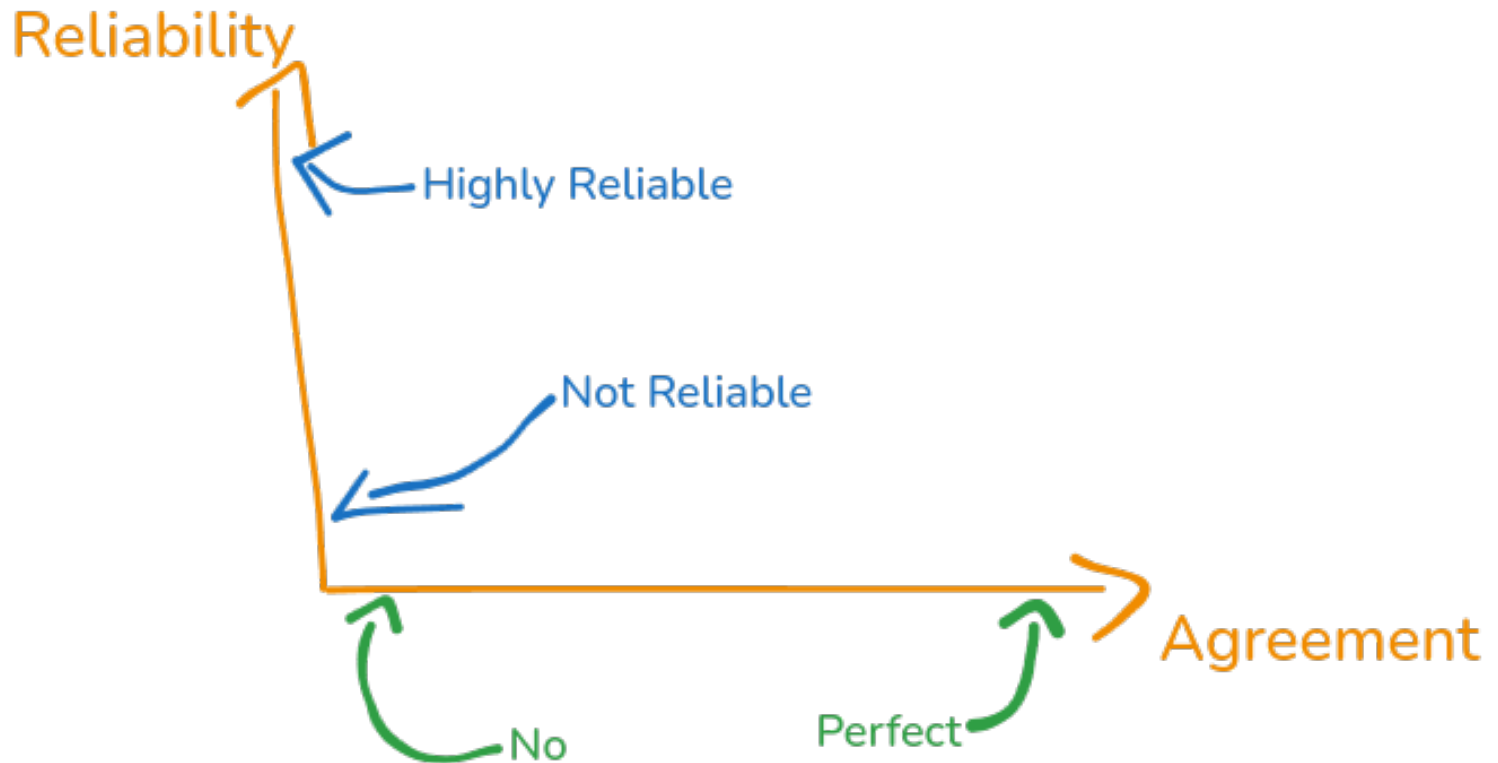
- ❑ Different tests cover different potential problems
[<https://sigir26-paper-141.pages.dev/introduction>]
- ❑ Adjustments to the input data relate to changes on reliability on observations



Scale to More Reliability Tests

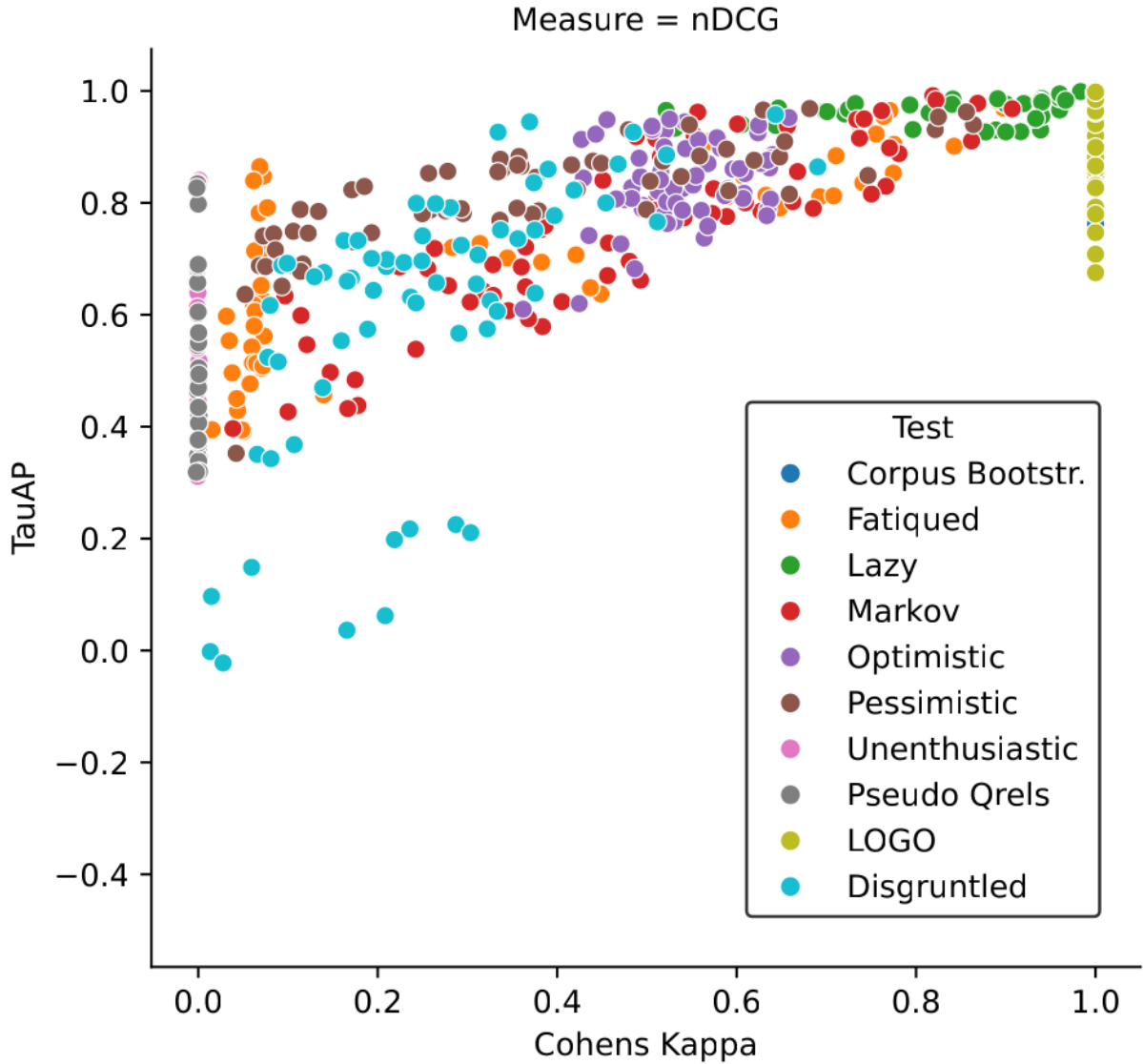
A Unified view on Diverse Reliability tests

- ❑ Different tests cover different potential problems
[<https://sigir26-paper-141.pages.dev/introduction>]
- ❑ Adjustments to the input data relate to changes on reliability on observations



- ❑ The following observations will use nDCG as measure

Scale to More Reliability Tests



Scale to More Reliability Tests

New Evaluation Paradigm: Large Language Models for Relevance Assessments?

Use a Large Language Model to make relevance judgments

[Faggioli'23]

Known from the literature:

[Clarke'24, Fröbe'25, Balog'25]

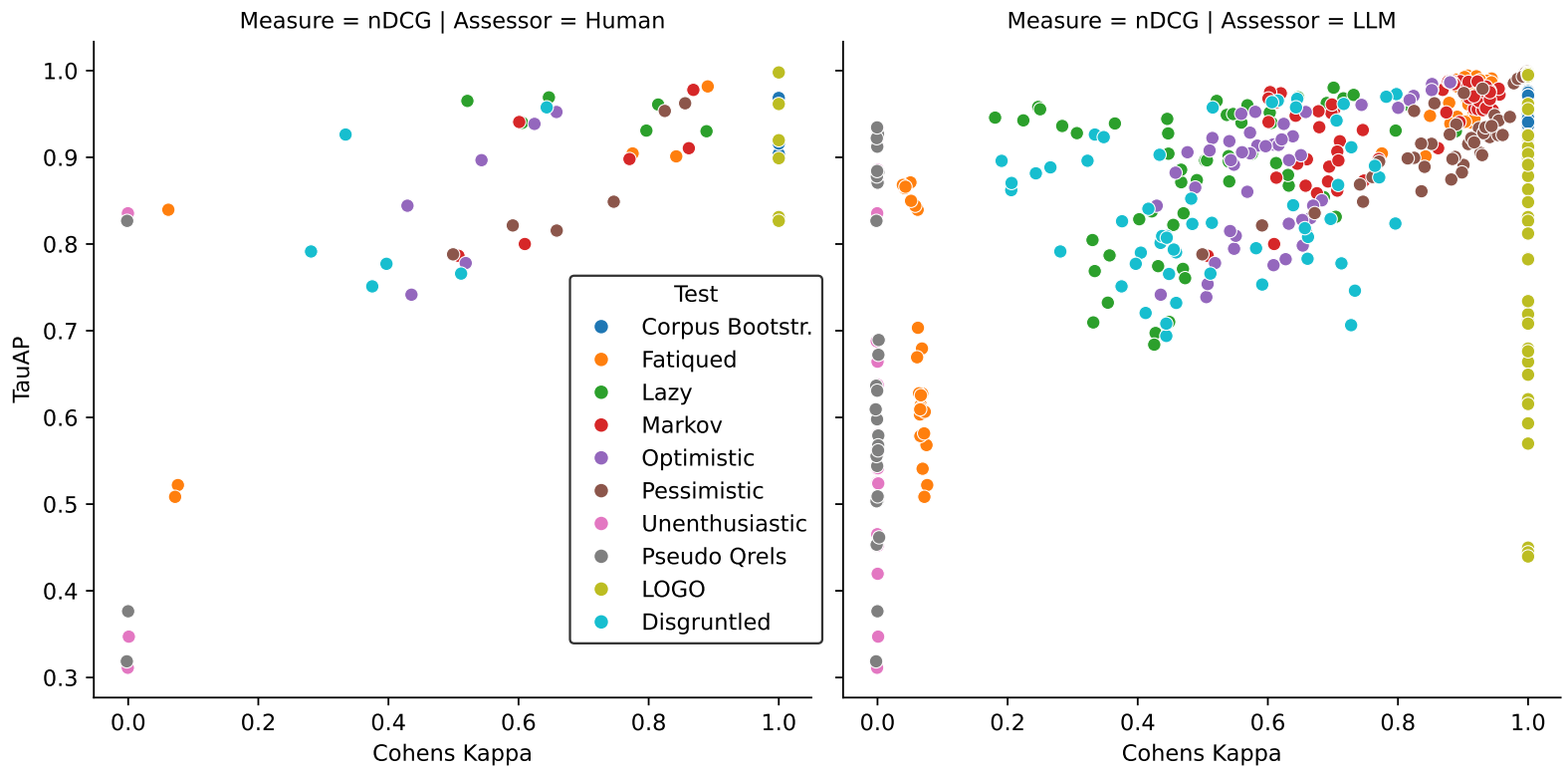
- ❑ LLMs agree with each other more than humans or humans vs. LLMs
- ❑ LLMs are overly positive

Attention: for the following LLM experiments we only evaluate on three tasks, but with 8 LLMs

- ❑ Tasks: Deep Learning 19, 20, and Retrieval@RAG 24

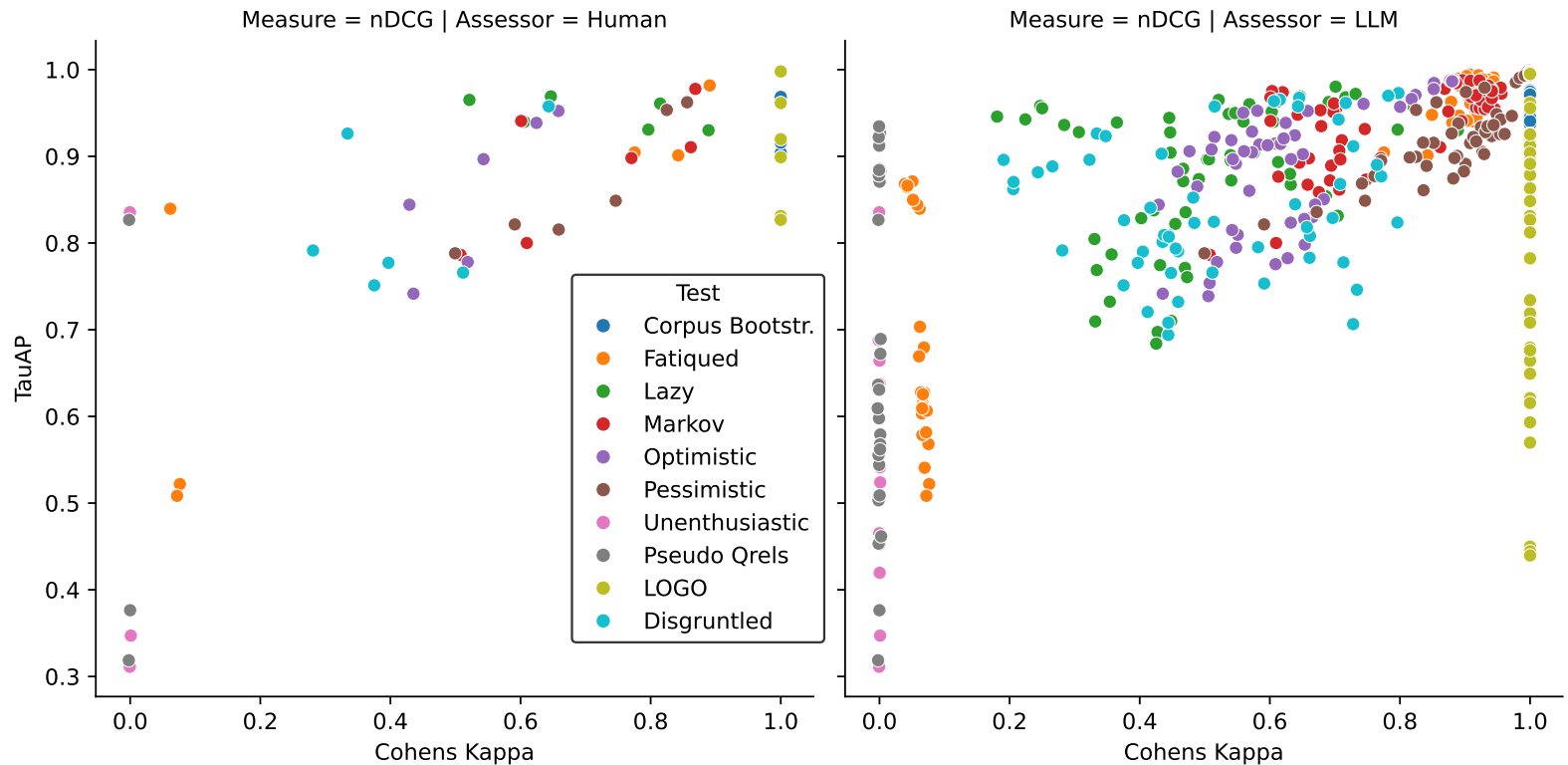
Scale to More Reliability Tests

New Evaluation Paradigm: Large Language Models for Relevance Assessments?



Scale to More Reliability Tests

New Evaluation Paradigm: Large Language Models for Relevance Assessments?



Interpretation

- ❑ LLM judgments more robust to noise (pseudo-qrels: both, good and bad)
- ❑ LLM Judgments produce less re-usable corpora (see logo test)

Scale to More Reliability Tests

New Evaluation Paradigm: Large Language Models for Relevance Assessments?

Different Measures via Pseudo-Qrels

Measure	Median τ
SetP	0.888
SetR	0.955
R@1000	0.955
nDCG	0.700
MAP	0.772
nDCG@20	0.436
nDCG@10	0.205
RR	0.224
P@1	0.238

Interpretation: Maybe due to the tasks (only 3!), maybe due to the LLM-Judgments

- ❑ Trend that judgments become less important (I can remove more of them)
- ❑ Maybe this is due to the positivity of LLMs?

Reliability Tests for Information Retrieval Evaluations: RAG

Current initiatives for Evaluation of Retrieval Augmented Generation (RAG)

[trec-rag.github.io]



RAG

TREC Retrieval-Augmented Generation

- ❑ We use the 2024 data
- ❑ 2025 not yet available? (maybe due to US Government shutdown?)

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024



Does the solar eclipse affect mental health?



Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024



Does the solar eclipse affect mental health?



Response 1

✦ Overviews

The effects of a solar eclipse on mental health depend on one's personal beliefs if stellar phenomena effect People. Astrology has no scientific backing, so that no effects of an solar eclipse on the mental health are supported by scientific evidence.

Responses can be inaccurate. We recommend you check important facts.

 www.physicianswe
Do Solar Eclipses...

 researchopenworld
Shedding Light o...

 www.inc.com
The Neuroscienc...

 www.nationalgeogr
What's behind...

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024



Does the solar eclipse affect mental health?



Response 2

✦ Overviews

Solar eclipses cause profound psychological responses. E.g., feelings of confusion, agitation and awe are common. People often have unusual dreams, and bursts of creativity, but problems with decision making.
Responses can be inaccurate. We recommend you check important facts.

 [blog.dharmikvibes.](https://blog.dharmikvibes.com)
[How Astrology...](#)

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024



Does the solar eclipse affect mental health?



Response 1

✦ Overviews

The effects of a solar eclipse on mental health depend on one's personal beliefs if stellar phenomena effect People. Astrology has no scientific backing, so that no effects of an solar eclipse on the mental health are supported by scientific evidence.

Responses can be inaccurate. We recommend you check important facts.

www.physicianswe.com
Do Solar Eclipses...

researchopenworld.com
Shedding Light o...

www.inc.com
The Neuroscienc...

www.nationalgeographic.com
What's behind...

Response 2

✦ Overviews

Solar eclipses cause profound psychological responses. E.g., feelings of confusion, agitation and awe are common. People often have unusual dreams, and bursts of creativity, but problems with decision making.

Responses can be inaccurate. We recommend you check important facts.

blog.dharmikvibes.com
How Astrology...

Evaluation needs to make a decision, which response should is better?

- ☐ Which response would you prefer?

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024



Does the solar eclipse affect mental health?



Response 1

✦ Overviews

The effects of a solar eclipse on mental health depend on one's personal beliefs if stellar phenomena effect People. Astrology has no scientific backing, so that no effects of an solar eclipse on the mental health are supported by scientific evidence.

Responses can be inaccurate. We recommend you check important facts.

www.physicianswe.com
Do Solar Eclipses...

researchopenworld.com
Shedding Light o...

www.lnc.com
The Neuroscienc...

www.nationalgeographic.com
What's behind...

Response 2

✦ Overviews

Solar eclipses cause profound psychological responses. E.g., feelings of confusion, agitation and awe are common. People often have unusual dreams, and bursts of creativity, but problems with decision making.

Responses can be inaccurate. We recommend you check important facts.

blog.dharmikvibes.com
How Astrology...

Evaluation needs to make a decision, which response should is better?

- ☐ Which response would you prefer?

The official evaluation prefers **Response 2**.
Lets learn why...

Reliability Tests for Information Retrieval Evaluations: RAG

Current Evaluation Methodology at TREC RAG

- ❑ Step 1: Collect many RAG responses
- ❑ Step 2: A human (or an LLM) annotator reviews all RAG responses
- ❑ Step 3: Annotator identifies information nuggets
 - A nugget is assumed to be a minimal unit of information
- ❑ Step 4 (Nugget Importance)
 - Decide if a nugget is important or not
- ❑ Step 5 (Nugget Assignment):
 - For every response and every nugget: nugget contained in response?

Reliability Tests for Information Retrieval Evaluations: RAG

Current Evaluation Methodology at TREC RAG

- ❑ Step 1: Collect many RAG responses
- ❑ Step 2: A human (or an LLM) annotator reviews all RAG responses
- ❑ Step 3: Annotator identifies information nuggets
 - A nugget is assumed to be a minimal unit of information
- ❑ Step 4 (Nugget Importance)
 - Decide if a nugget is important or not
- ❑ Step 5 (Nugget Assignment):
 - For every response and every nugget: nugget contained in response?

Evaluation based on this data uses different flavors of Recall

- ❑ vital, vital_strict, all, all_strict (details omitted, only that you heard names)

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024: The Official Nuggets

Nugget	Response 1	Response 2
confused feelings are common	✗	✓
awe feelings are common	✗	✓
profound psychological responses	✗	✓
cause increased agitation	✗	✓
cause unusual dreams	✗	✓
interfere with decision-making	✗	✓
cause relationship difficulties	✗	✓
cause bursts of creativity	✗	✓
Resulting Recall	0	1

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024: The Official Nuggets

Nugget	Response 1	Response 2
confused feelings are common	✗	✓
awe feelings are common	✗	✓
profound psychological responses	✗	✓
cause increased agitation	✗	✓
cause unusual dreams	✗	✓
interfere with decision-making	✗	✓
cause relationship difficulties	✗	✓
cause bursts of creativity	✗	✓
Resulting Recall	0	1

Potential Nuggets for Annotator with different Opinion

Nugget	Response 1	Response 2
astrological beliefs exist	✓	✓
no scientific backing	✓	✗
Resulting Recall	1	0.5

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024: The Official Nuggets

Nugget	Response 1	Response 2
confused feelings are common	✗	✓
awe feelings are common	✗	✓
profound psychological responses	✗	✓
cause increased agitation	✗	✓
cause unusual dreams	✗	✓
interfere with decision-making	✗	✓
cause relationship difficulties	✗	✓
cause bursts of creativity	✗	✓
Resulting Recall	0	1

Potential Nuggets for Annotator with different Opinion

Nugget	Response 1	Response 2
astrological beliefs exist	✓	✓
no scientific backing	✓	✗
Resulting Recall	1	0.5

How big are such effects?

Lets ask some reliability tests (before expensive experiments)

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024: Reliability Test Results

α	β	Test	τ of V_{strict}	Most Similar IR Measure
16	1	Corpus Bootstr.	0.88	AP@20
		LOGO	0.96	nDCG@20
		Pseudo Qrels	0.08	AP@20
		Unenthusiastic	0.06	P@1
	1	Optimistic	0.92	nDCG@100
		Pessimistic	0.91	nDCG
		Disgruntled	0.97	SetP
		Lazy	0.68	nDCG@3
		Fatigued	0.92	nDCG@5
		Markov	0.73	P@1
	16	Optimistic	0.95	nDCG@100
		Pessimistic	0.74	R@100
		Disgruntled	0.65	R@100
		Lazy	0.81	nDCG@3
		Fatigued	0.21	P@1
		Markov	0.91	nDCG@3

Reliability Tests for Information Retrieval Evaluations: RAG

Example from TREC RAG 2024: Reliability Test Results

α	β	Test	τ of V_{strict}	Most Similar IR Measure
16	1	Corpus Bootstr.	0.88	AP@20
		LOGO	0.96	nDCG@20
		Pseudo Qrels	0.08	AP@20
		Unenthusiastic	0.06	P@1
	1	Optimistic	0.92	nDCG@100
		Pessimistic	0.91	nDCG
		Disgruntled	0.97	SetP
		Lazy	0.68	nDCG@3
		Fatigued	0.92	nDCG@5
		Markov	0.73	P@1
	16	Optimistic	0.95	nDCG@100
		Pessimistic	0.74	R@100
		Disgruntled	0.65	R@100
		Lazy	0.81	nDCG@3
		Fatigued	0.21	P@1
		Markov	0.91	nDCG@3

Interpretation

- Often similar reliability to classical measures with top-k ≤ 20 documents

Conclusions

Summary

- ❑ Different evaluation paradigms have different reliability behaviour
- ❑ Important
 - Do not use reliability as a criteria to select your measure
 - Still: Select the measure based on the user model
 - But: Adjust your experiment design based on reliability behaviour

Conclusions

Summary

- ❑ Different evaluation paradigms have different reliability behaviour
- ❑ Important
 - Do not use reliability as a criteria to select your measure
 - Still: Select the measure based on the user model
 - But: Adjust your experiment design based on reliability behaviour

Future work

- ❑ Make measures that consider few documents (e.g., nDCG@10) more reliable towards noise?
- ❑ Make measures that consider many documents (e.g., Recall@1000) more sensitive to judgments?
- ❑ Identify new reliability tests specifically targeted at RAG

Conclusions

Summary

- ❑ Different evaluation paradigms have different reliability behaviour
- ❑ Important
 - Do not use reliability as a criteria to select your measure
 - Still: Select the measure based on the user model
 - But: Adjust your experiment design based on reliability behaviour

Future work

- ❑ Make measures that consider few documents (e.g., nDCG@10) more reliable towards noise?
- ❑ Make measures that consider many documents (e.g., Recall@1000) more sensitive to judgments?
- ❑ Identify new reliability tests specifically targeted at RAG

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>

[Balog'25]

Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. SIGIR 2025.

[Soboroff'01]

Ian Soboroff et al.: Ranking Retrieval Systems without Relevance Judgments. SIGIR 2001

[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.

[Clarke'24]

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