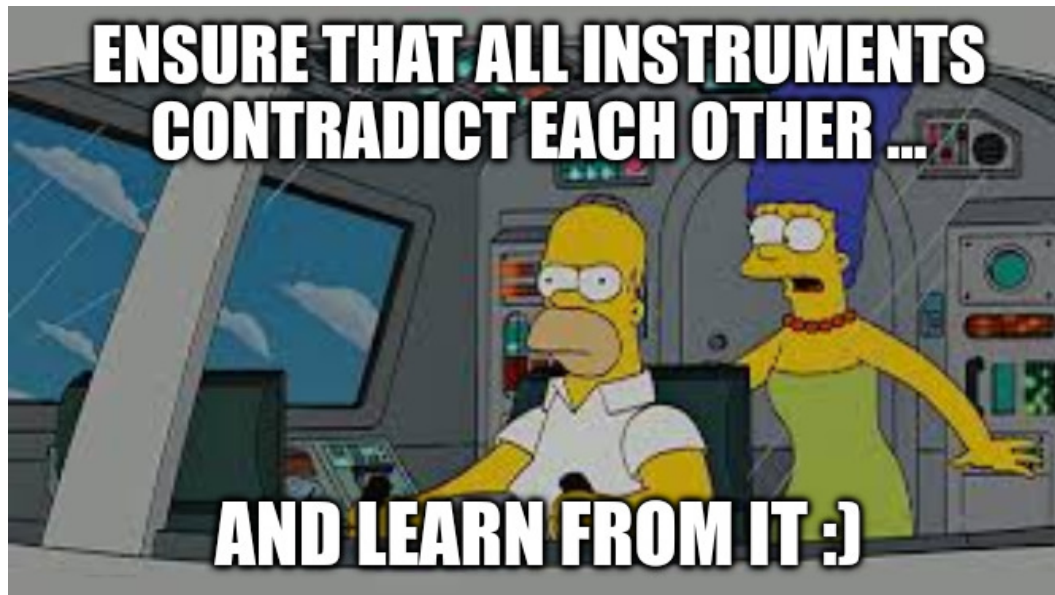


Evaluation of Retrieval Augmented Generation via Argumentative Nuggets



Maik Fröbe

Friedrich-Schiller-Universität Jena

@webis_de

www.webis.de

RAG Evaluation via Argumentative Nuggets



Does the solar eclipse affect mental health?



RAG Evaluation via Argumentative Nuggets



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...

RAG Evaluation via Argumentative Nuggets



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...](#)

RAG Evaluation via Argumentative Nuggets



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?

RAG Evaluation via Argumentative Nuggets



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?

The TREC RAG evaluation prefers **Response 2**.
Lets learn why...

RAG Evaluation via Argumentative Nuggets

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?

RAG Evaluation via Argumentative Nuggets

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)

RAG Evaluation via Argumentative Nuggets

The Official TREC RAG 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

RAG Evaluation via Argumentative Nuggets

The Official TREC RAG 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

RAG Evaluation via Argumentative Nuggets

The Official TREC RAG 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 is the Impact of such Disagreement?

We make a connection to existing reliability tests

RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

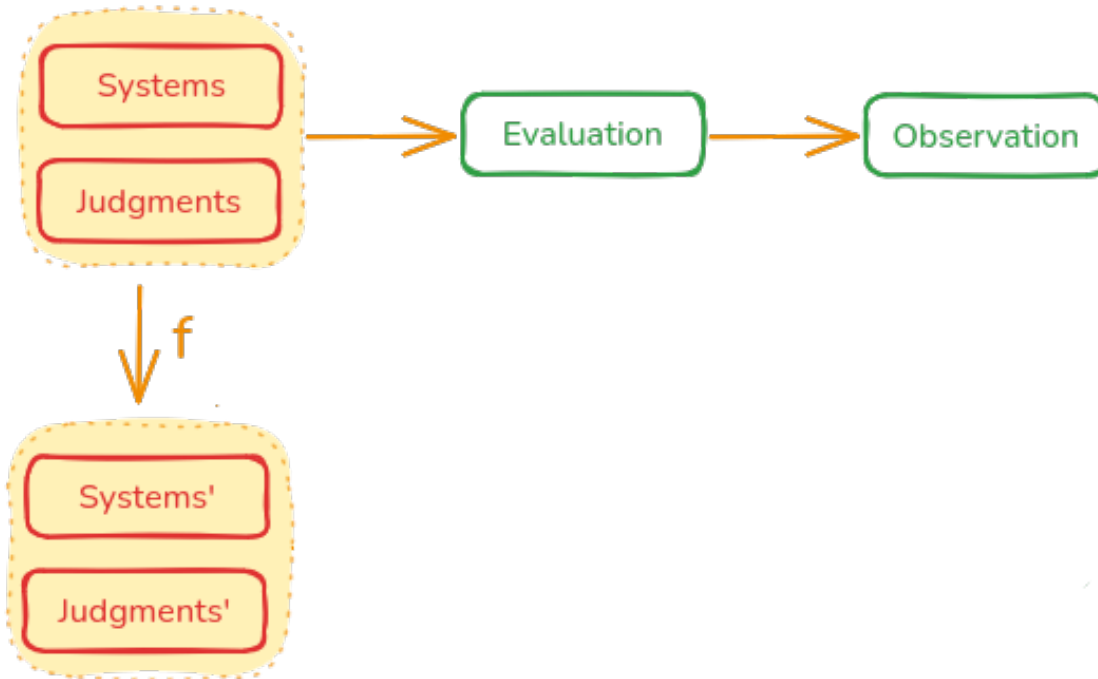
Experiment



RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

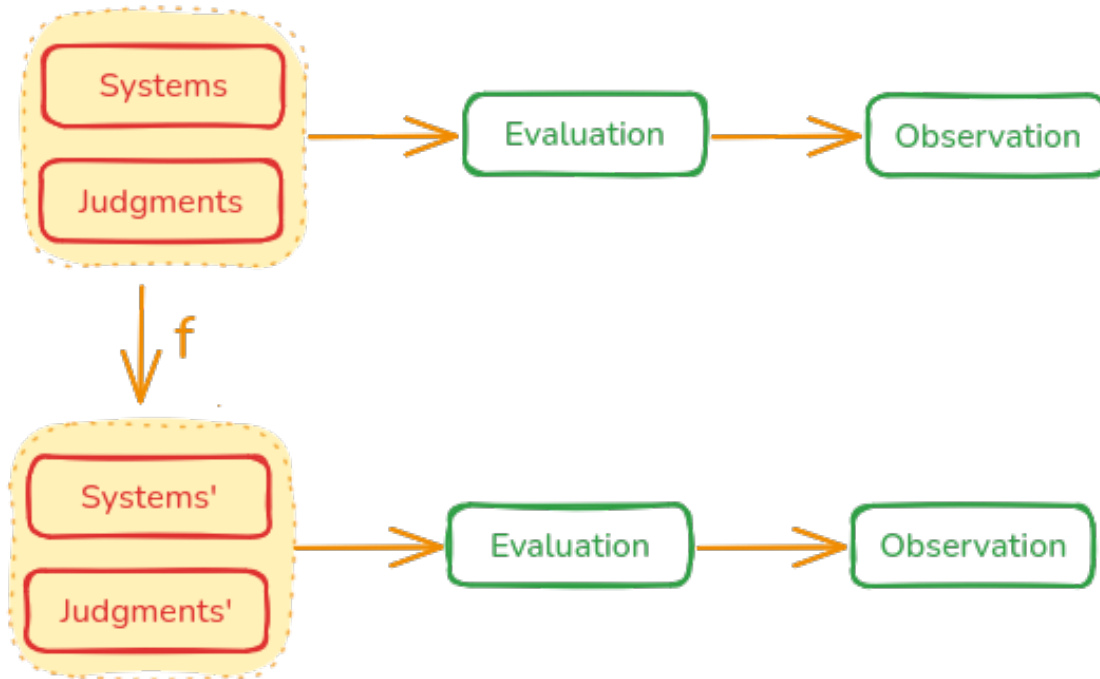
Experiment



RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

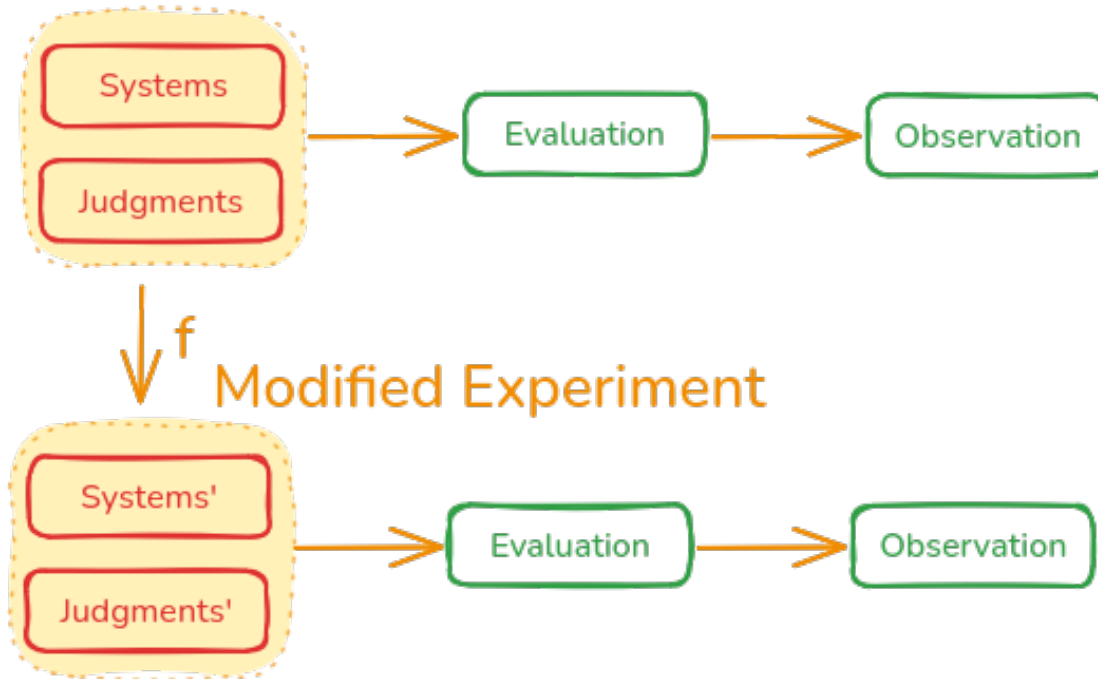
Experiment



RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

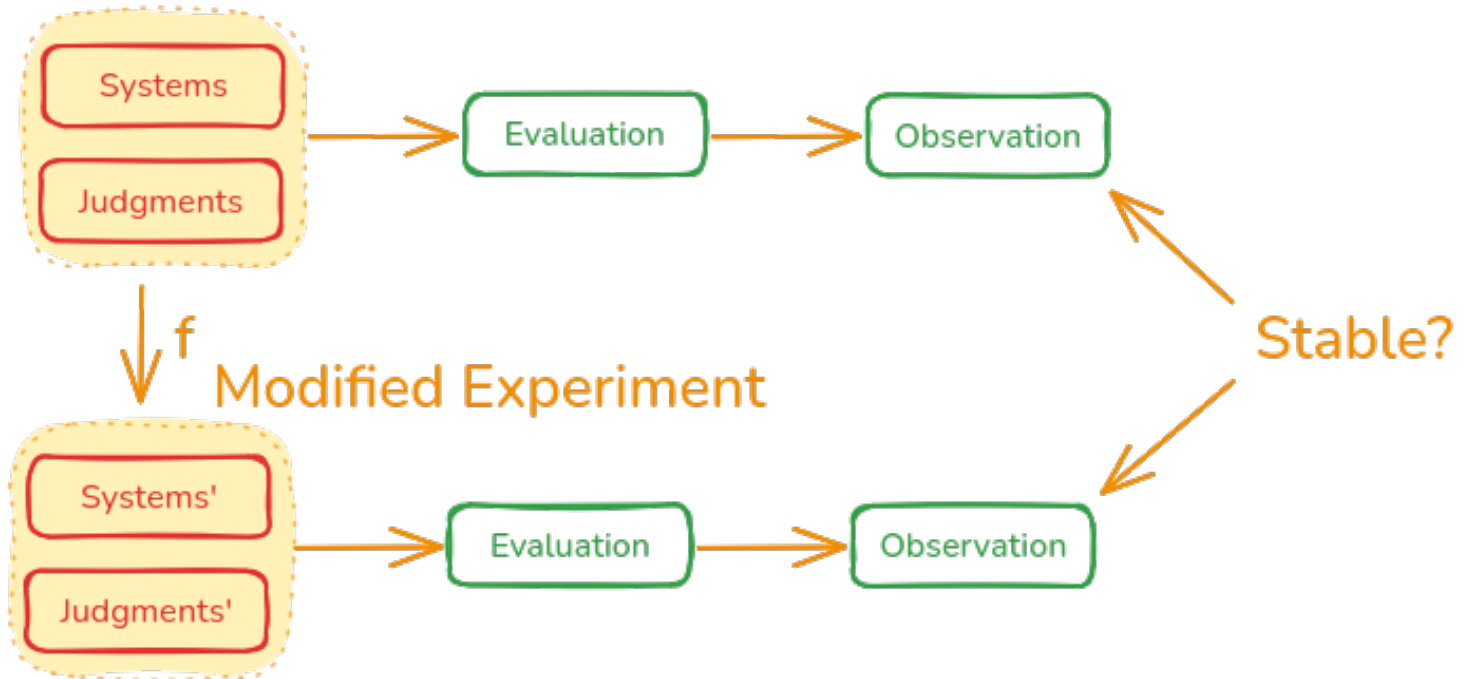
Experiment



RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

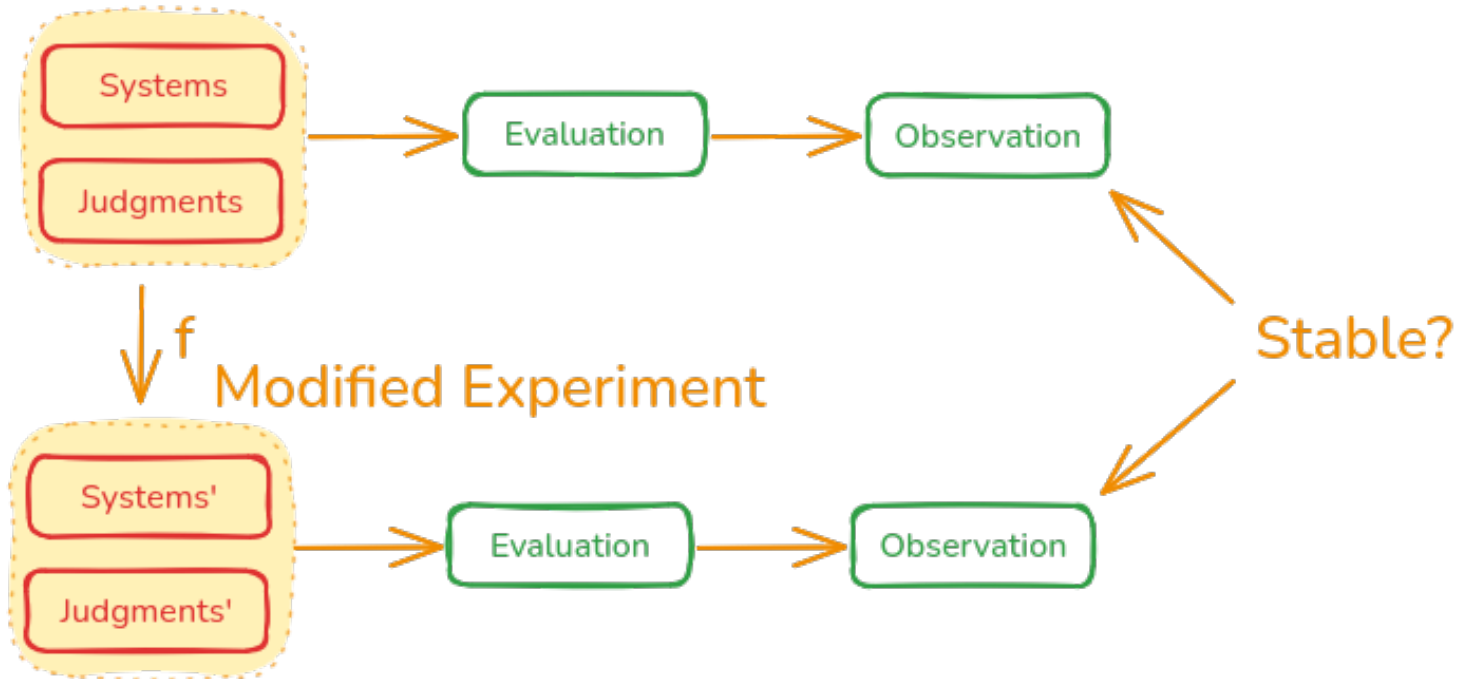
Experiment



RAG Evaluation via Argumentative Nuggets

Background on Reliability Tests

Experiment



The Function f is a Reliability Test

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

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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:

$$\tau = 1$$

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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:

$$\tau = 0.8$$

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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

RAG Evaluation via Argumentative Nuggets

Measure if observations are stable with 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:

$$\tau = -0.1$$

RAG Evaluation via Argumentative Nuggets

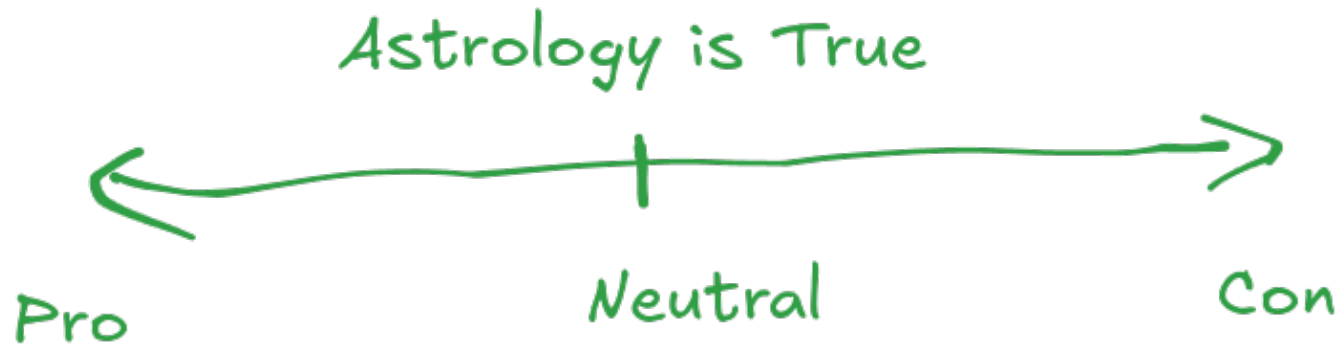
Argumentative Nuggets for Argumentative Queries



Does the solar eclipse affect mental health?



Before annotation: Make argumentative stance explicit



RAG Evaluation via Argumentative Nuggets

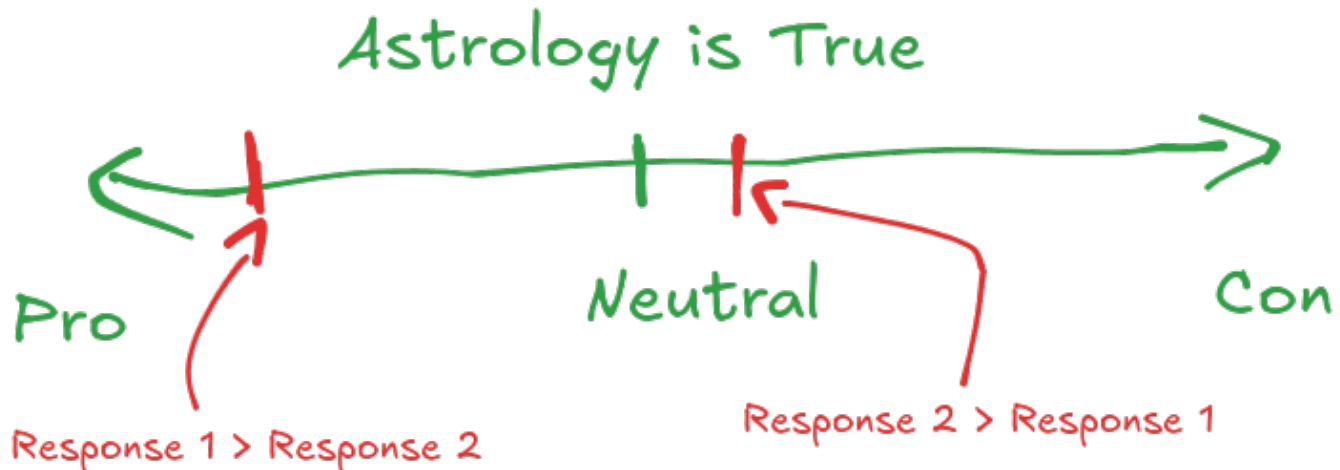
Argumentative Nuggets for Argumentative Queries



Does the solar eclipse affect mental health?



Before annotation: Make argumentative stance explicit



RAG Evaluation via Argumentative Nuggets

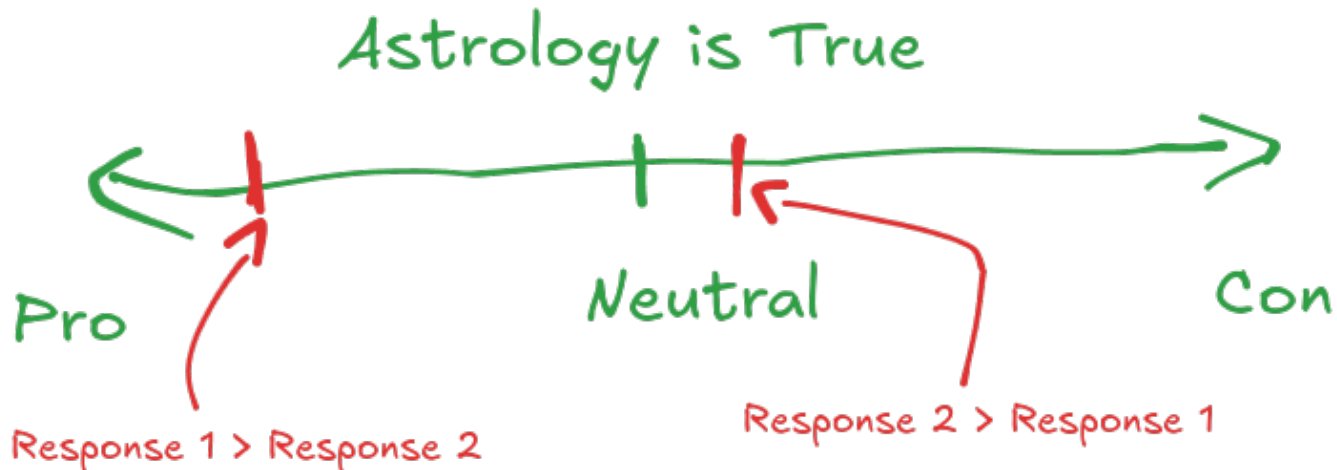
Argumentative Nuggets for Argumentative Queries



Does the solar eclipse affect mental health?



Before annotation: Make argumentative stance explicit



Goal (Ongoing Bsc. Thesis)

Measure reliability explicitly model system and user stances

RAG Evaluation via Argumentative Nuggets

What reliability should we expect?

- ❑ There is not much experience with RAG evaluations

RAG Evaluation via Argumentative Nuggets

What reliability should we expect?

- ❑ There is not much experience with RAG evaluations

But: ≤ 30 years of experience with the classical Cranfield-style IR evaluations.

- ❑ We decided to use the reliability of classical IR evaluations as context

RAG Evaluation via Argumentative Nuggets

What reliability should we expect?

- ❑ There is not much experience with RAG evaluations

But: ≤ 30 years of experience with the classical Cranfield-style IR evaluations.

- ❑ We decided to use the reliability of classical IR evaluations as context

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)
- ❑ 42 000 CPU hours (more than 4 years; but highly parallelized)

RAG Evaluation via Argumentative Nuggets

What reliability should we expect?

- ❑ There is not much experience with RAG evaluations

But: ≤ 30 years of experience with the classical Cranfield-style IR evaluations.

- ❑ We decided to use the reliability of classical IR evaluations as context

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)
- ❑ 42 000 CPU hours (more than 4 years; but highly parallelized)

Idea: Run all tests on your experiment and compare against public repository

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

- ❑ Gives us context: We can interpret if the stance has a big or small impact

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

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

Results

Task	Measure	Reliability
IR	nDCG	0.486
IR	MRR	0.024

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

Results

Task	Measure	Reliability
IR	nDCG	0.486
IR	MRR	0.024
RAG	vital_strict	0.056

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

Results

Task	Measure	Reliability
IR	nDCG	0.486
IR	MRR	0.024
RAG	vital_strict	0.056

How accross all different reliability tests?

- ❑ We measured the similarity accross all tests
- ❑ RAG evaluation mostly similar to IR evaluation for top- $k \leq 20$ documents
 - Noise and subjectivity expected to have a non-negligible impact

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

- ❑ Simulate how different stances impact evaluation

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

- ❑ Simulate how different stances impact evaluation

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.