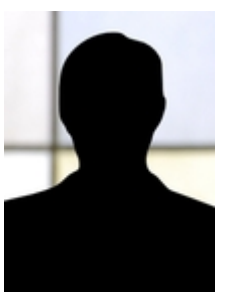
Are Large Language Models Reliable Argument Quality Annotators?

RATIO 2024

Nailia Mirzakhmedova

Marcel Gohsen



Chia Hao Chang

Benno Stein

Bauhaus-Universität Weimar

Webis

Argument Quality Assessment

Background

- **Argument** – a claim on a controversial topic supported by premises.

claim

" If you wanna hear my view, I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated. "

premises

Argument Quality Assessment

Background

- **Argument** – a claim on a controversial topic supported by premises.

claim

" If you wanna hear my view, I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated. "

premises

- **Argument Quality Assessment**
  - Rating an argument or ranking different arguments
  - Critical for any system leveraging argument mining

> *"In some sense, the question about the quality of an argument is the 'ultimate' one for argumentation mining."* (Stede and Schneider, 2018).

Argument Quality Assessment

Background

- **Argument** – a claim on a controversial topic supported by premises.

claim

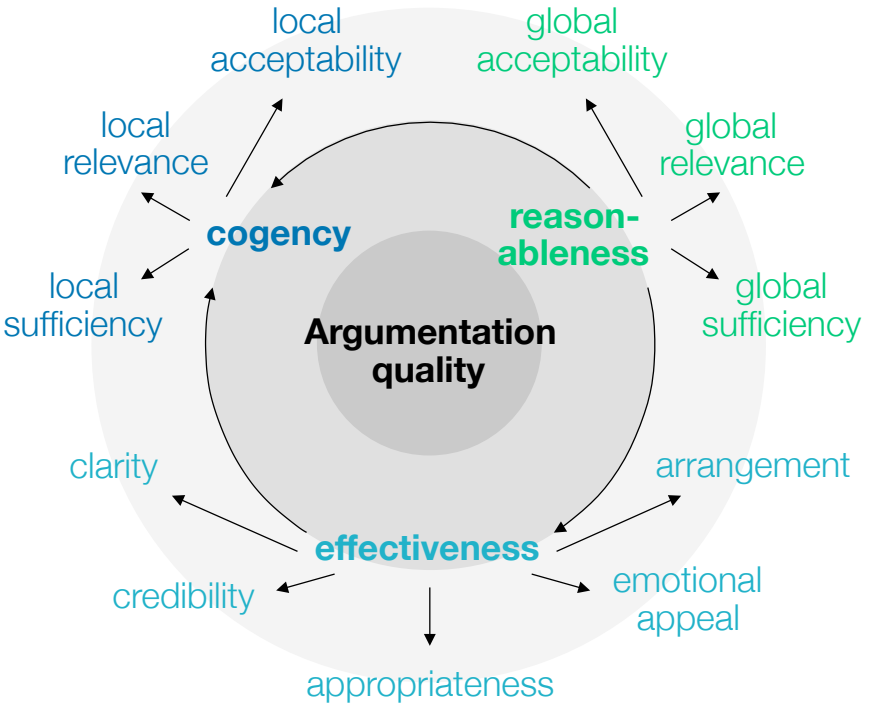
" If you wanna hear my view, I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated. " **premises**

- **Argument Quality Assessment**
  - Rating an argument or ranking different arguments **How?**
  - Critical for any system leveraging argument mining

"In some sense, the question about the quality of an argument is the 'ultimate' one for argumentation mining." (Stede and Schneider, 2018).

Argument Quality Assessment

Background



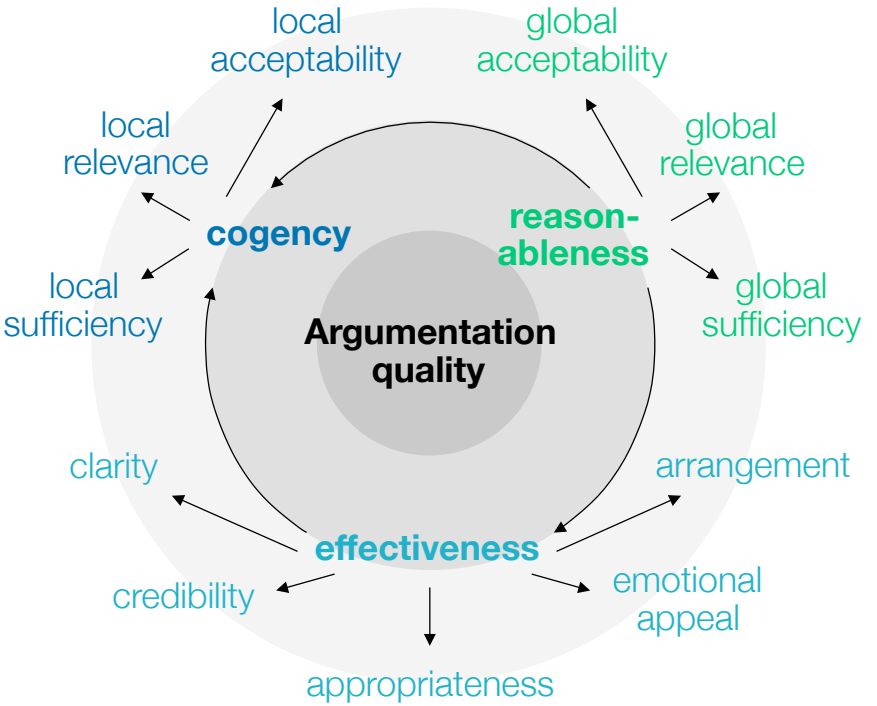
Taxonomy of argument quality dimensions
(Wachsmuth et al. 2017)

Argument Quality Assessment

Background

Challenges:

- ❑ Domain-specific knowledge
  - ⇨ Time-consuming
- ❑ Some dimensions are (highly) subjective
  - ⇨ Inconsistency in annotations
- ❑ Multiple annotators
  - ⇨ Expensive



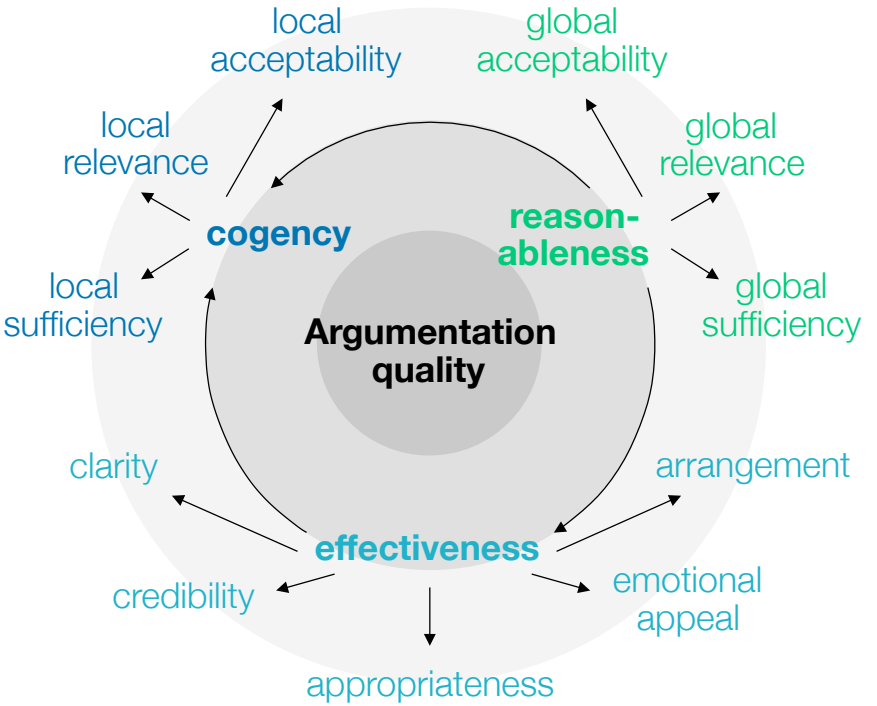
Taxonomy of argument quality dimensions
(Wachsmuth et al. 2017)

Argument Quality Assessment

Background

Challenges:

- ❑ Domain-specific knowledge
  - ⇨ Time-consuming
- ❑ Some dimensions are (highly) subjective
  - ⇨ Inconsistency in annotations
- ❑ Multiple annotators
  - ⇨ Expensive



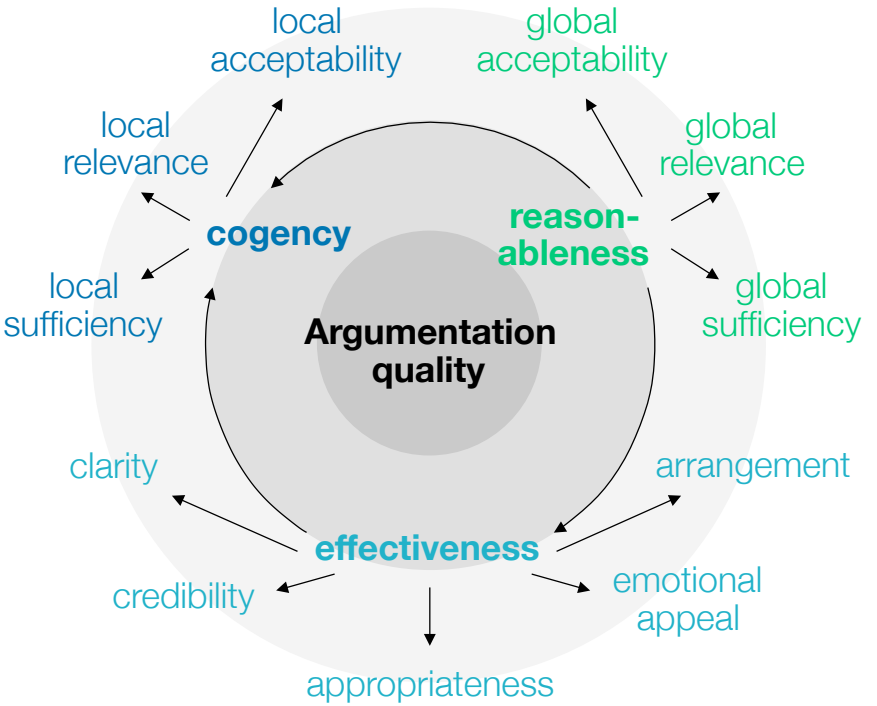
Taxonomy of argument quality dimensions
(Wachsmuth et al. 2017)

Argument Quality Assessment

Background

Challenges:

- ❑ Domain-specific knowledge
  - ⇨ Time-consuming
- ❑ Some dimensions are (highly) subjective
  - ⇨ Inconsistency in annotations
- ❑ Multiple annotators
  - ⇨ Expensive



Taxonomy of argument quality dimensions
(Wachsmuth et al. 2017)

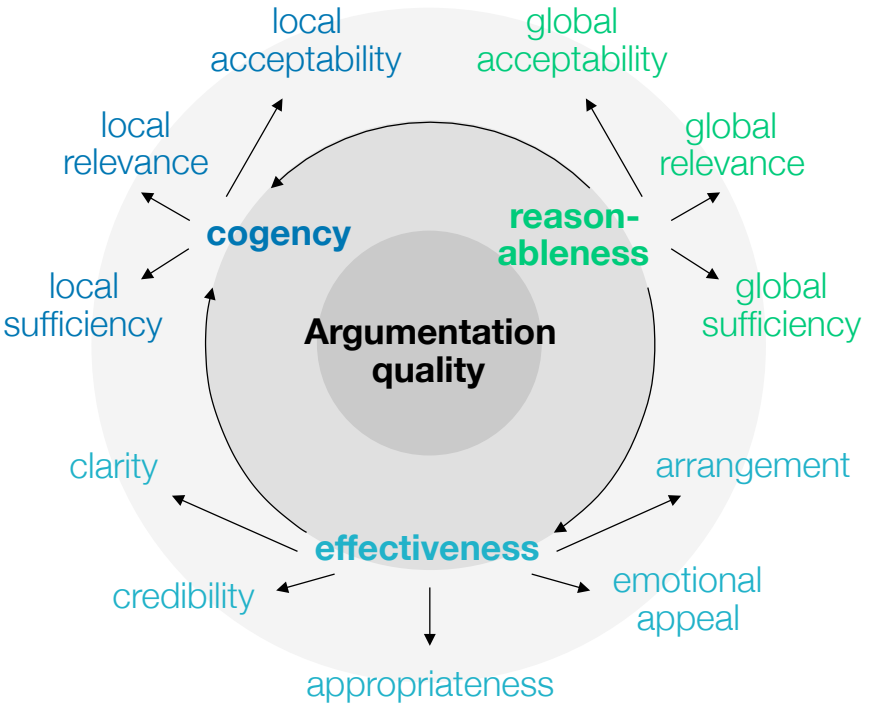
Argument Quality Assessment

Background

Challenges:

- ❑ Domain-specific knowledge
  - ⇨ Time-consuming
- ❑ Some dimensions are (highly) subjective
  - ⇨ Inconsistency in annotations
- ❑ Multiple annotators
  - ⇨ Expensive

Can we use LLMs to address these challenges?



Taxonomy of argument quality dimensions (Wachsmuth et al. 2017)

Are LLMs Reliable Argument Quality Annotators?

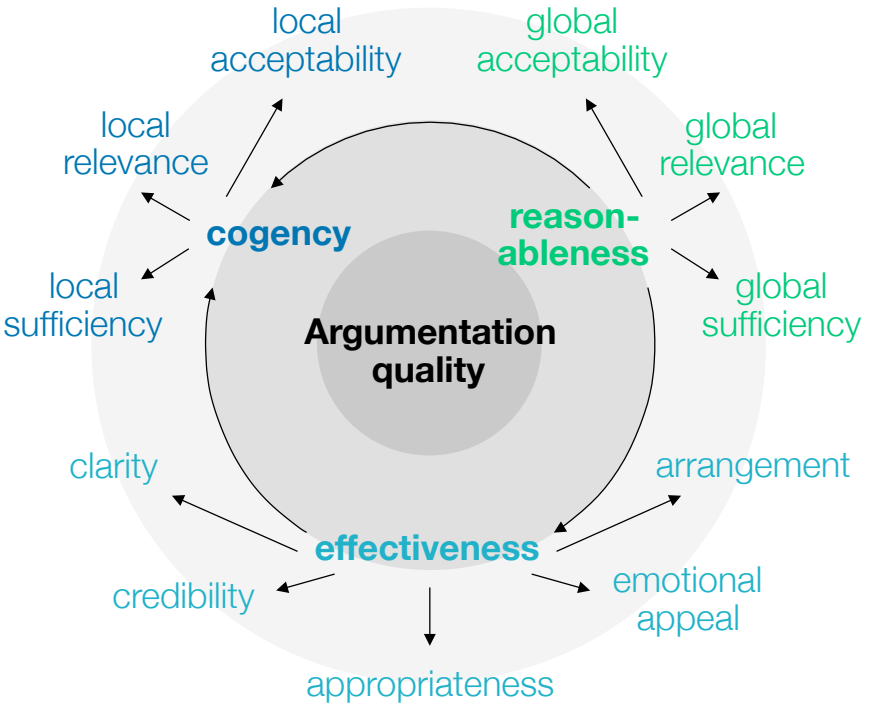
Background

Research Questions:

RQ1: Do LLMs provide consistent argument quality annotations?

RQ2: Do LLM annotations align with human annotations?

RQ3: Can we use LLMs as additional annotators to improve the resulting agreement?



Taxonomy of argument quality dimensions (Wachsmuth et al. 2017)

Are LLMs Reliable Argument Quality Annotators?

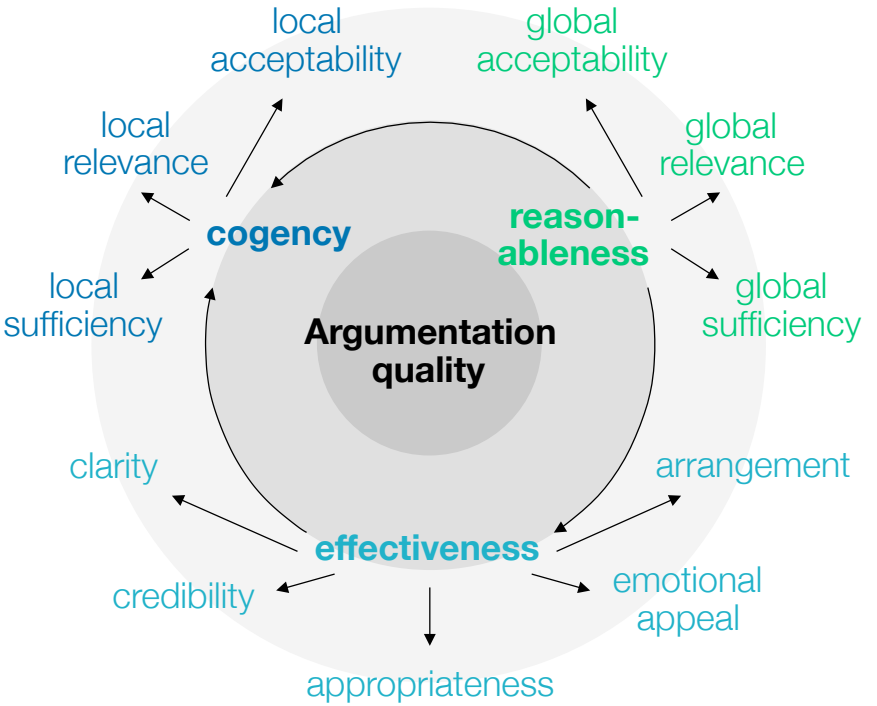
Background

Research Questions:

RQ1: Do LLMs provide consistent argument quality annotations?

RQ2: Do LLM annotations align with human annotations?

RQ3: Can we use LLMs as additional annotators to improve the resulting agreement?



Taxonomy of argument quality dimensions (Wachsmuth et al. 2017)

Are LLMs Reliable Argument Quality Annotators?

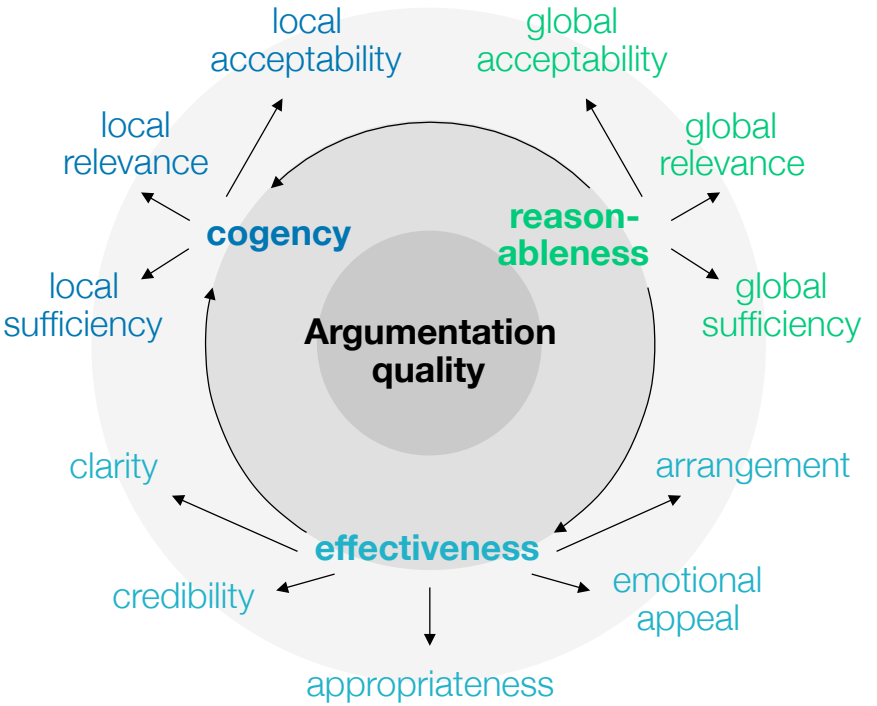
Background

Research Questions:

RQ1: Do LLMs provide consistent argument quality annotations?

RQ2: Do LLM annotations align with human annotations?

RQ3: Can we use LLMs as additional annotators to improve the resulting agreement?



Taxonomy of argument quality dimensions (Wachsmuth et al. 2017)

Are LLMs Reliable Argument Quality Annotators?

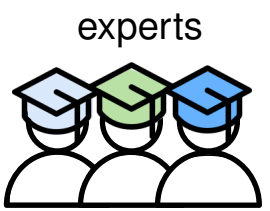
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Are LLMs Reliable Argument Quality Annotators?

Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale



Expert Annotators:
2 Postdocs & 1 PhD student

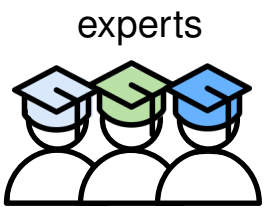
Are LLMs Reliable Argument Quality Annotators?

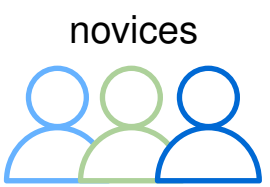
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students





Are LLMs Reliable Argument Quality Annotators?

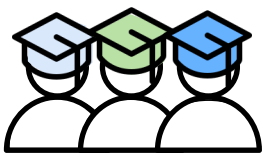
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

experts



novices

Simplified Guidelines Example
Local Acceptability:

Expert definition:
A premise of an argument should be seen as acceptable if it is worthy of being believed, i.e., if you rationally think it is true or if you see no reason for not believing that it may be true.

Novice definition:
The reasons are individually believable: they could be true.

"premises" renamed to "reasons"
"stance" and "issue" became "conclusion"

Are LLMs Reliable Argument Quality Annotators?

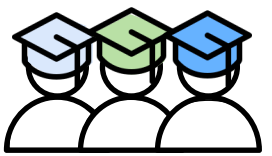
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

experts



novices

Simplified Guidelines Example
Local Acceptability:

Expert definition:
A premise of an argument should be seen as acceptable if it is worthy of being believed, i.e., if you rationally think it is true or if you see no reason for not believing that it may be true.

Novice definition:
The reasons are individually believable: they could be true.

"premises" renamed to "reasons"
"stance" and "issue" became "conclusion"

Are LLMs Reliable Argument Quality Annotators?

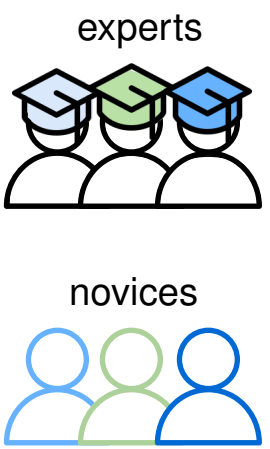
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

LLM Annotators:
PaLM 2 (`text-bison@001`)
GPT-3 (`gpt-3.5-turbo-0613`)



Are LLMs Reliable Argument Quality Annotators?

Experimental Setup

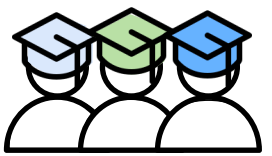
Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

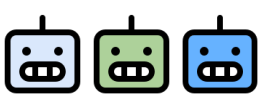
LLM Annotators:
PaLM 2 (`text-bison@001`)
GPT-3 (`gpt-3.5-turbo-0613`)

experts



novices

LLMs



Knowledge

- Issue
- Stance
- Argument
- Dimension Definition

Are LLMs Reliable Argument Quality Annotators?

Experimental Setup

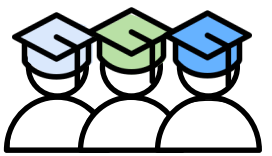
Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

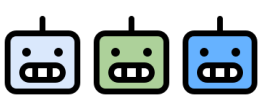
LLM Annotators:
PaLM 2 (`text-bison@001`)
GPT-3 (`gpt-3.5-turbo-0613`)

experts



novices

LLMs



Knowledge

- Issue
- Stance
- Argument
- Dimension Definition

Rating

- ❑ 3 – High
- ❑ 2 – Medium
- ❑ 1 – Low
- ❑ ? – Cannot Judge

Are LLMs Reliable Argument Quality Annotators?

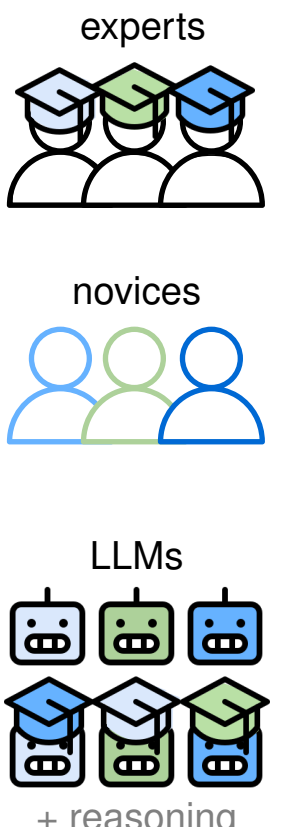
Experimental Setup

Dataset:
Dagstuhl-15512-ArgQuality
304 arguments
15 quality dimensions
3-point Likert scale

Expert Annotators:
2 Postdocs & 1 PhD student

Novice Annotators:
108 undergrad students

LLM Annotators:
PaLM 2 (`text-bison@001`)
GPT-3 (`gpt-3.5-turbo-0613`)



Knowledge

- Issue
- Stance
- Argument
- Dimension Definition

Rating

- ❑ 3 – High
- ❑ 2 – Medium
- ❑ 1 – Low
- ❑ ? – Cannot Judge

Are LLMs Reliable Argument Quality Annotators?

Prompt Design

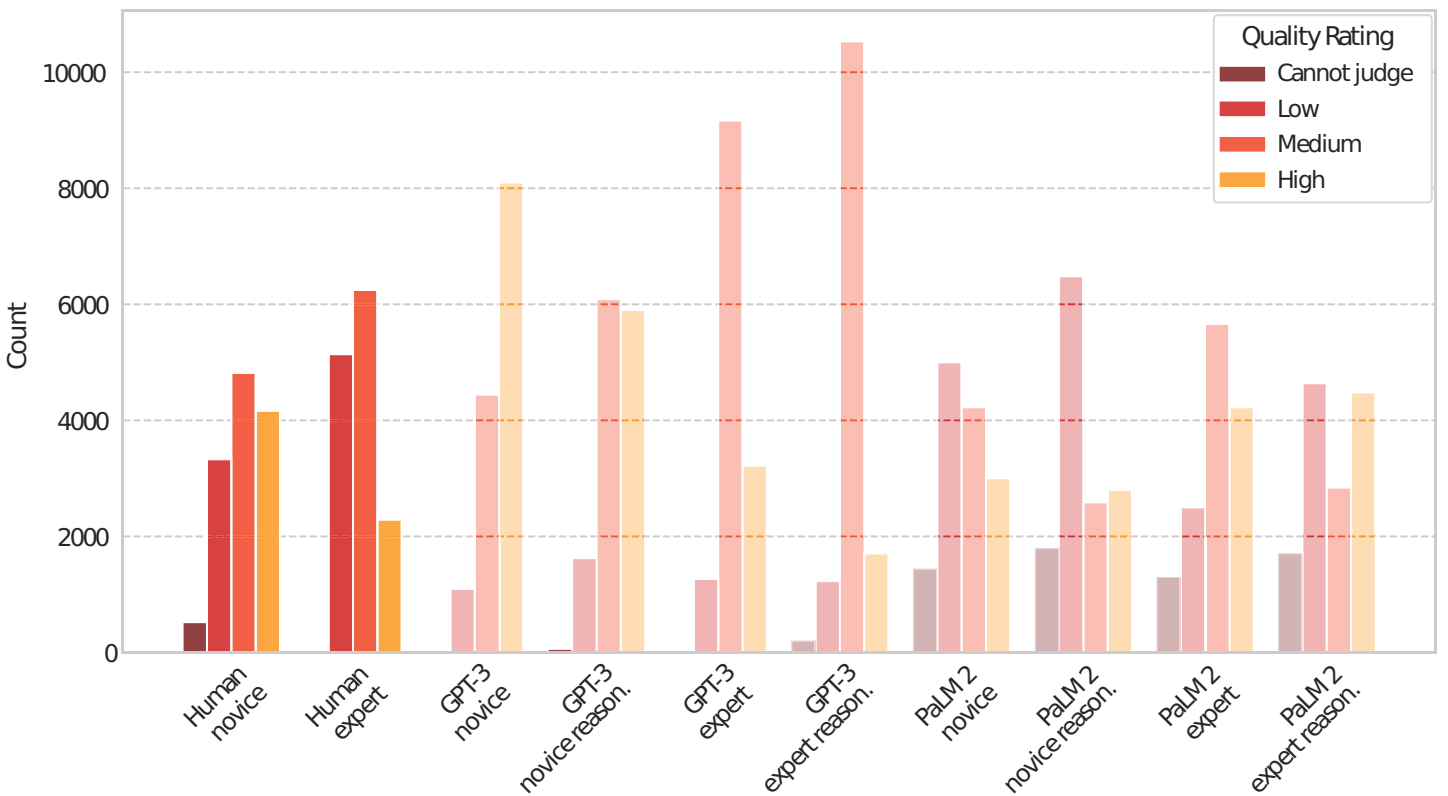
```
### Instruction:
Please answer the following question for the
given comment from an online debate forum on
a given issue.

### Issue: Is TV better than books?

### Stance: No, it isn't.

### Argument: <argument>

### Quality dimension definition:
Local Acceptability:  <expert definition>

### Question:
How would you rate the acceptability of the
premises of the author's argument?  Choose
one of the options below [and explain your
reasoning]:
3 - High
2 - Medium
1 - Low
?  - Cannot judge
```

```
### Instruction:
Please rate the quality dimension of the
given argument from an online debate forum.

### Conclusion: TV is not better than books.

### Reason(s): <argument>

### Quality dimension definition:
Local Acceptability:  <simplified definition>

### Task:
Choose one of the options below [and explain
your reasoning].

The acceptability of the reasons is ...:
3 - High
2 - Medium
1 - Low
?  - Cannot judge
```

- Each prompt variant was repeated (at least) 3 times
- Fixed hyperparameters:
  - τ = 0.3
  - top-k = 40
  - top-p = 1.0
  - max new tokens = 256

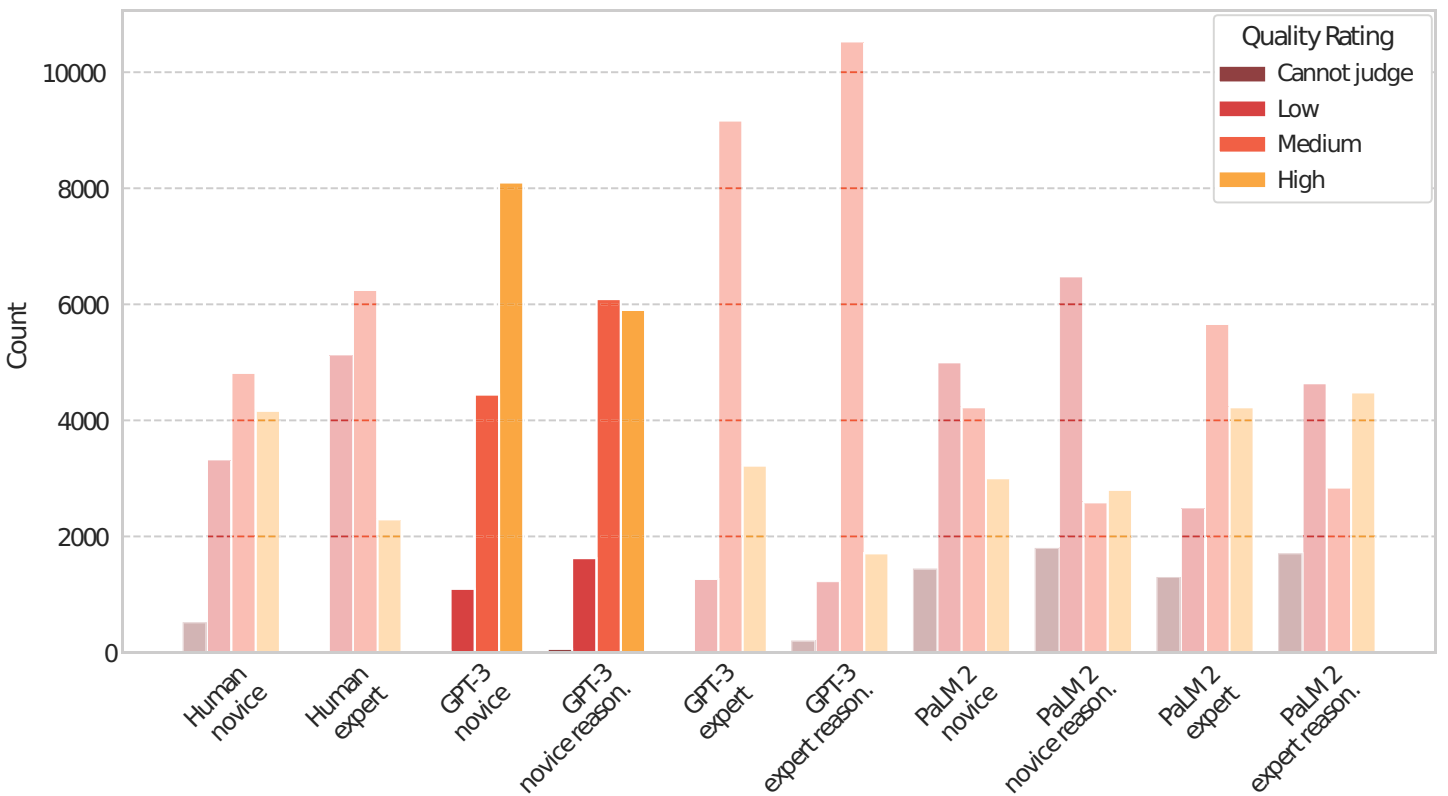
Are LLMs Reliable Argument Quality Annotators?

Distribution of Assigned Ratings



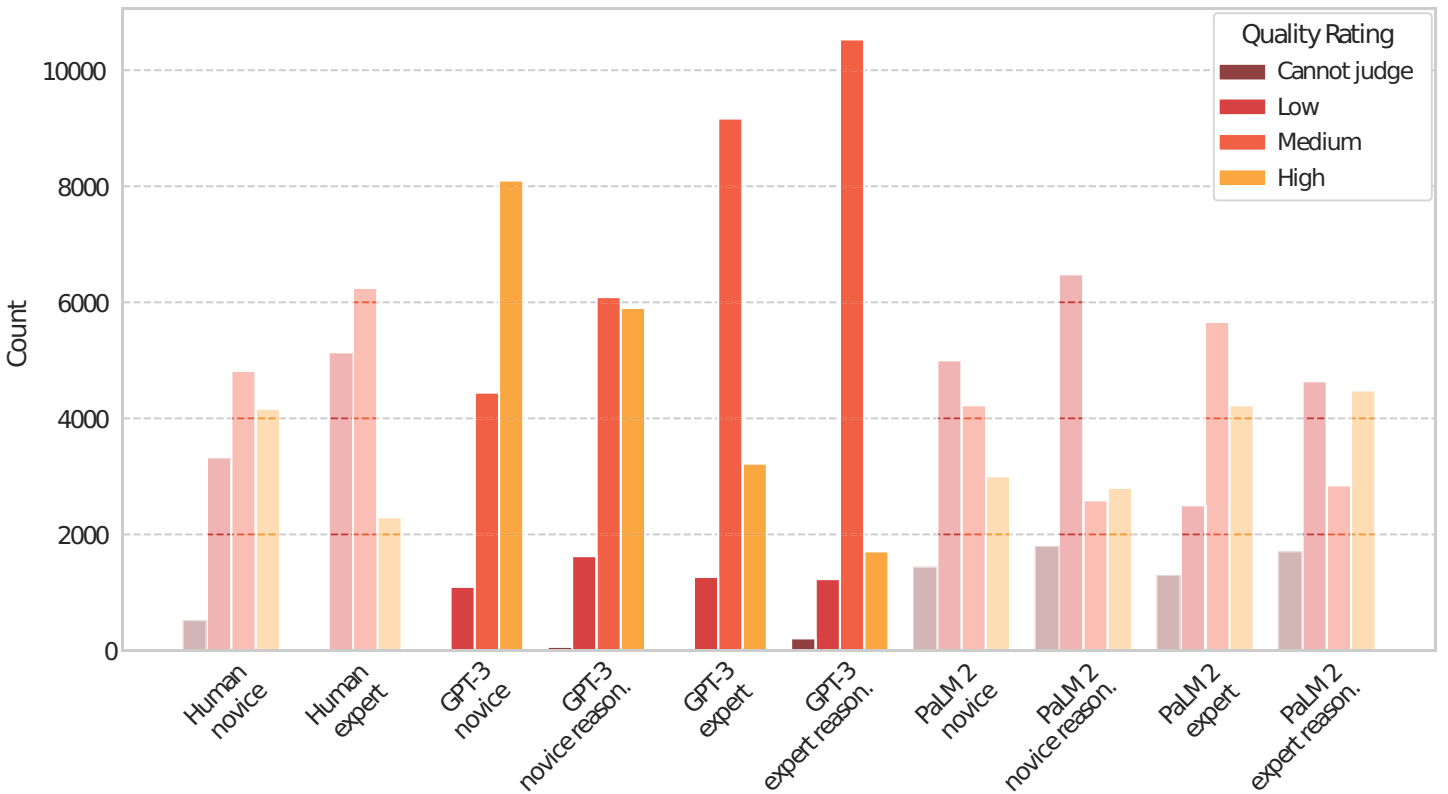
Are LLMs Reliable Argument Quality Annotators?

Distribution of Assigned Ratings



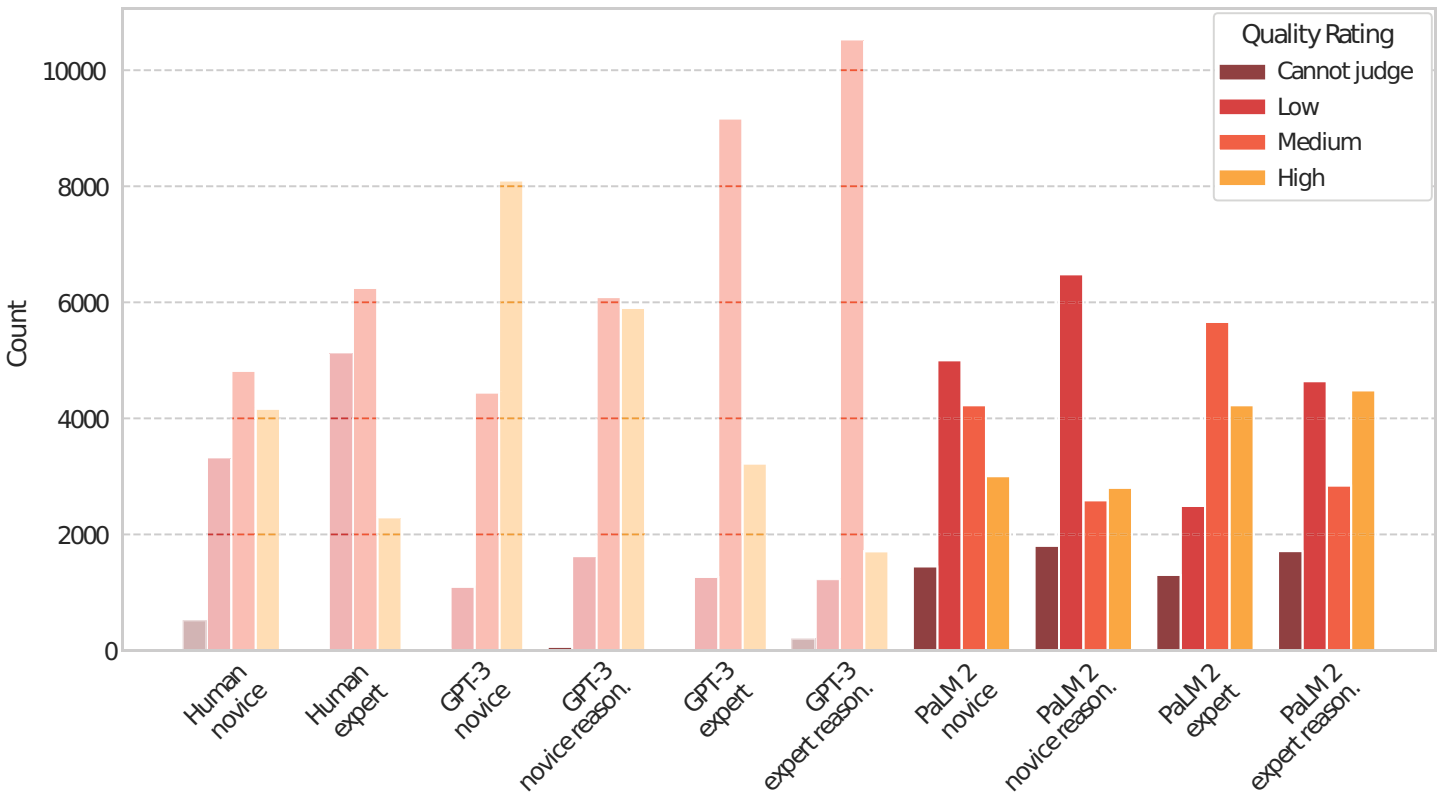
Are LLMs Reliable Argument Quality Annotators?

Distribution of Assigned Ratings



Are LLMs Reliable Argument Quality Annotators?

Distribution of Assigned Ratings



Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	**Human**		**GPT-3**				**PaLM 2**			
					Reasoning				**Reasoning**	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

RQ1. Do LLMs provide consistent argument quality annotations?

Quality Dimension	Human		GPT-3				PaLM 2			
					Reasoning				Reasoning	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Cogency	0.38	0.38	0.72	0.73	0.77	0.72	**0.99**	0.98	0.73	0.74
Local Acceptability	**0.43**	0.33	0.64	0.69	0.70	0.75	0.98	0.97	0.60	0.71
Local Relevance	0.36	0.41	0.70	0.59	0.76	0.61	0.98	0.98	0.78	0.68
Local Sufficiency	0.35	0.27	0.74	0.69	0.79	0.72	0.98	0.97	0.63	0.63
Effectiveness	0.41	0.33	0.72	0.70	0.77	0.74	0.98	**0.99**	0.78	0.79
Credibility	0.36	0.23	**0.79**	**0.79**	0.81	0.78	**0.99**	0.97	0.72	0.67
Emotional Appeal	0.35	0.21	0.73	0.56	0.74	0.70	0.98	0.97	0.64	0.72
Clarity	0.27	0.25	0.72	0.69	0.71	0.69	**0.99**	**0.99**	**0.82**	0.80
Appropriateness	0.39	0.17	0.66	0.50	0.68	0.58	**0.99**	**0.99**	0.75	0.81
Arrangement	0.39	0.26	0.68	0.66	0.71	0.69	**0.99**	**0.99**	0.69	0.65
Reasonableness	0.35	**0.45**	0.73	0.78	0.81	0.74	0.97	0.97	0.70	0.76
Global Acceptability	0.37	0.39	0.72	0.77	0.77	0.74	0.98	0.97	0.66	0.70
Global Relevance	0.38	0.26	0.69	0.71	0.81	0.70	**0.99**	0.98	0.74	**0.85**
Global Sufficiency	0.27	0.17	0.72	0.69	0.72	0.75	0.98	0.96	0.62	0.47
Overall Quality	0.41	0.44	0.77	0.77	**0.82**	**0.81**	0.98	0.97	0.77	0.78
Overall α	0.37	0.40	0.76	0.73	0.78	0.74	0.99	0.98	0.76	0.78

Within-group inter-annotator agreement (Krippendorf's α)

Are LLMs Reliable Argument Quality Annotators?

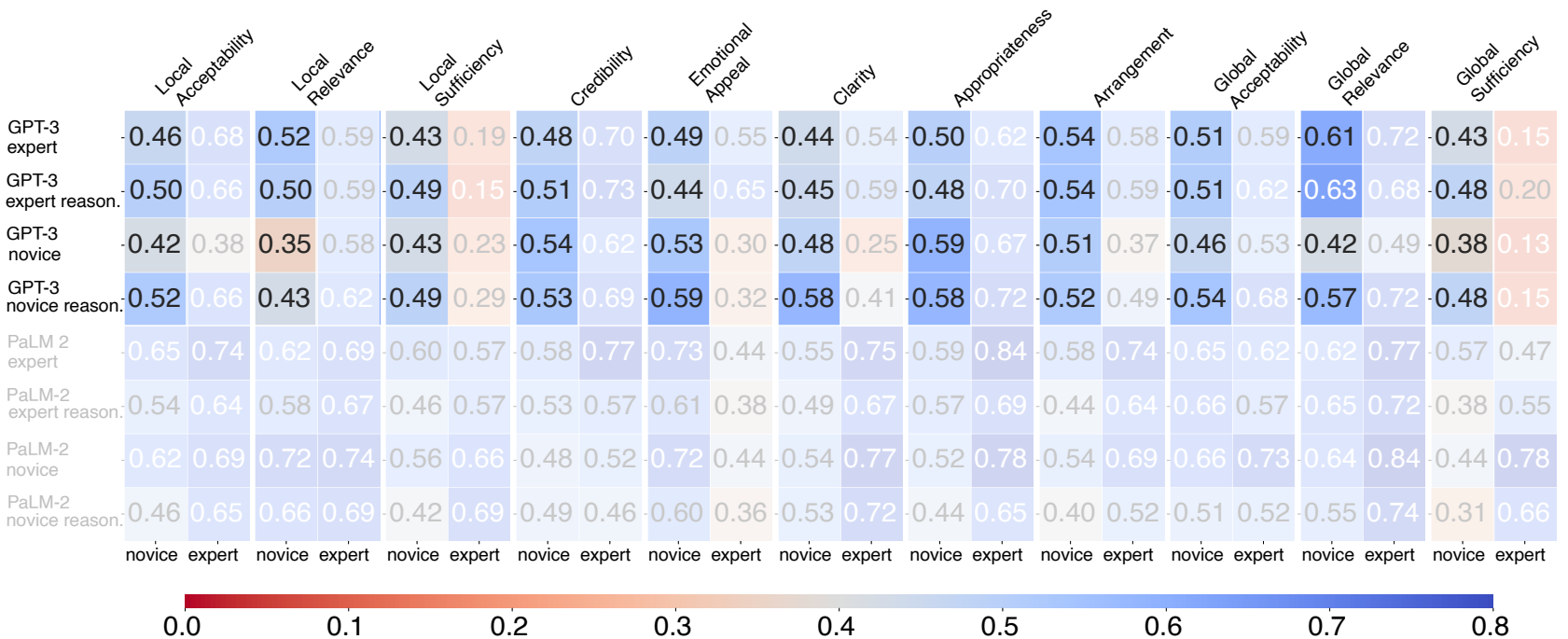
RQ2. Do LLM annotations align with human annotations?

Quality Dimension	Expert	Novice
Cogency	40%	35%
Local Acceptability	27%	38%
Local Relevance	33%	33%
Local Sufficiency	37%	29%
Effectiveness	42%	**39%**
Credibility	38%	28%
Emotional Appeal	43%	30%
Clarity	29%	30%
Appropriateness	17%	34%
Arrangement	27%	34%
Reasonableness	41%	**39%**
Global Acceptability	32%	34%
Global Relevance	22%	32%
Global Sufficiency	**45%**	27%
Overall Quality	44%	43%

Percentage of arguments with perfect agreement within each group of human annotators

Are LLMs Reliable Argument Quality Annotators?

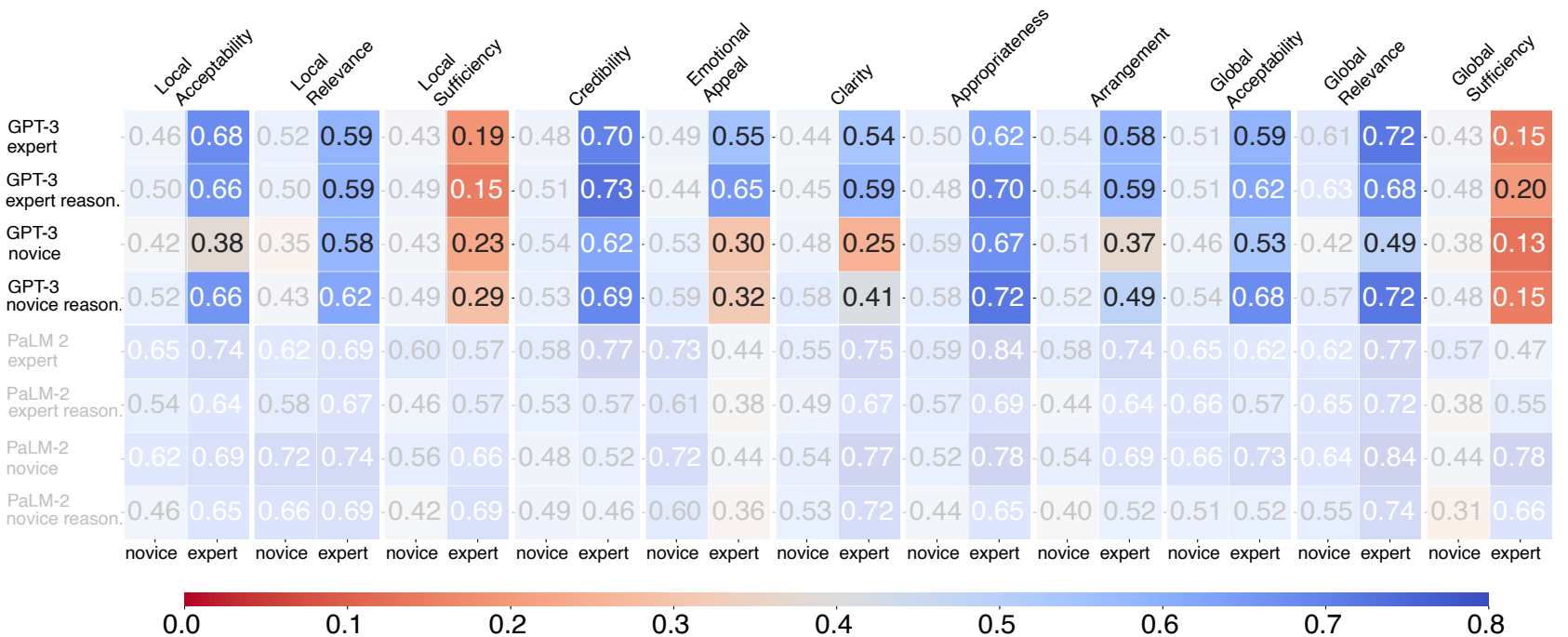
RQ2. Do LLM annotations align with human annotations?



Krippendorff's α between GPT-3 and novices

Are LLMs Reliable Argument Quality Annotators?

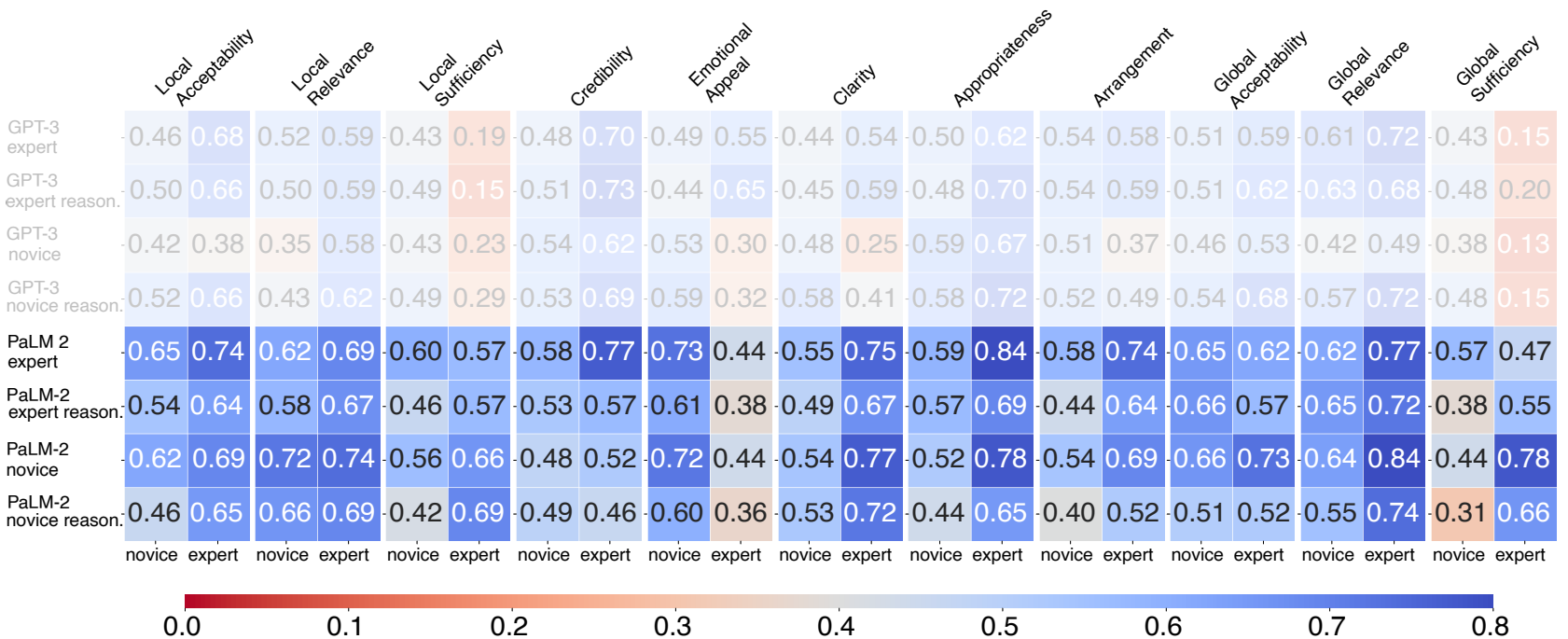
RQ2. Do LLM annotations align with human annotations?



Krippendorff's α between GPT-3 and experts

Are LLMs Reliable Argument Quality Annotators?

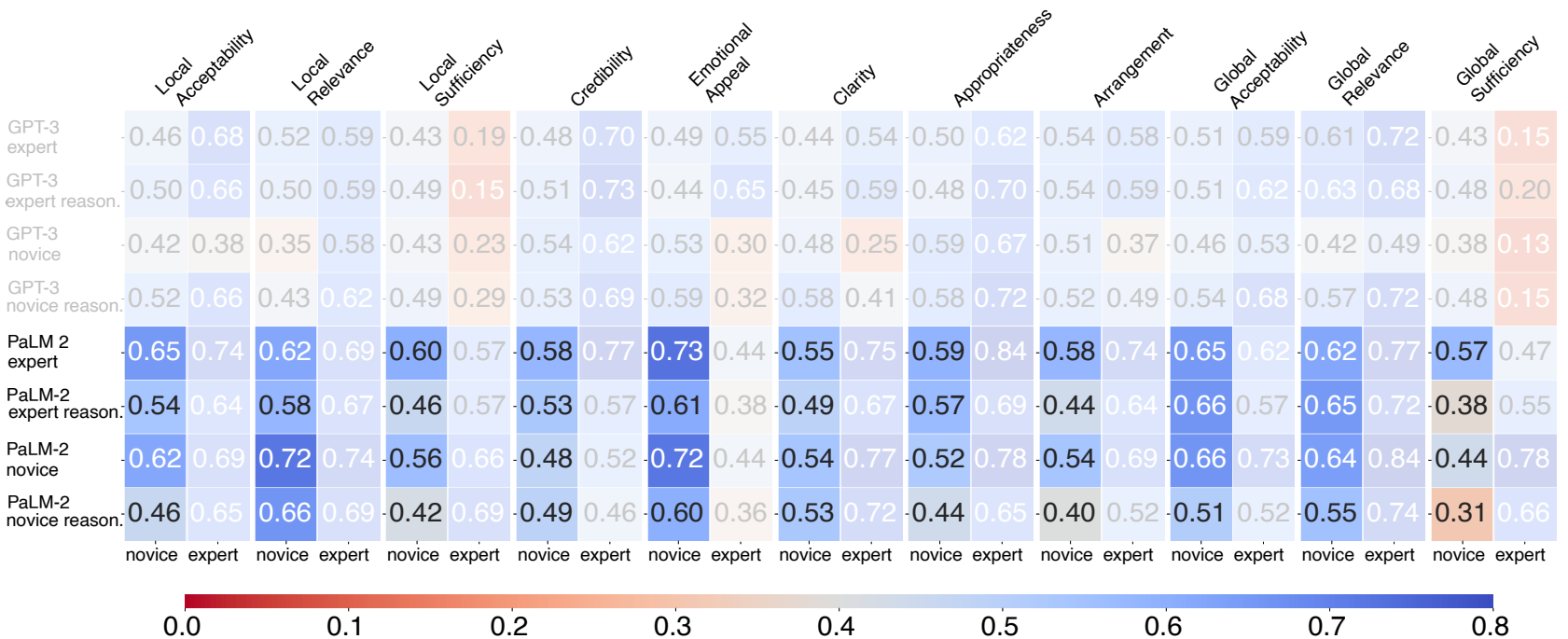
RQ2. Do LLM annotations align with human annotations?



Krippendorff's α between PaLM 2 and human annotators

Are LLMs Reliable Argument Quality Annotators?

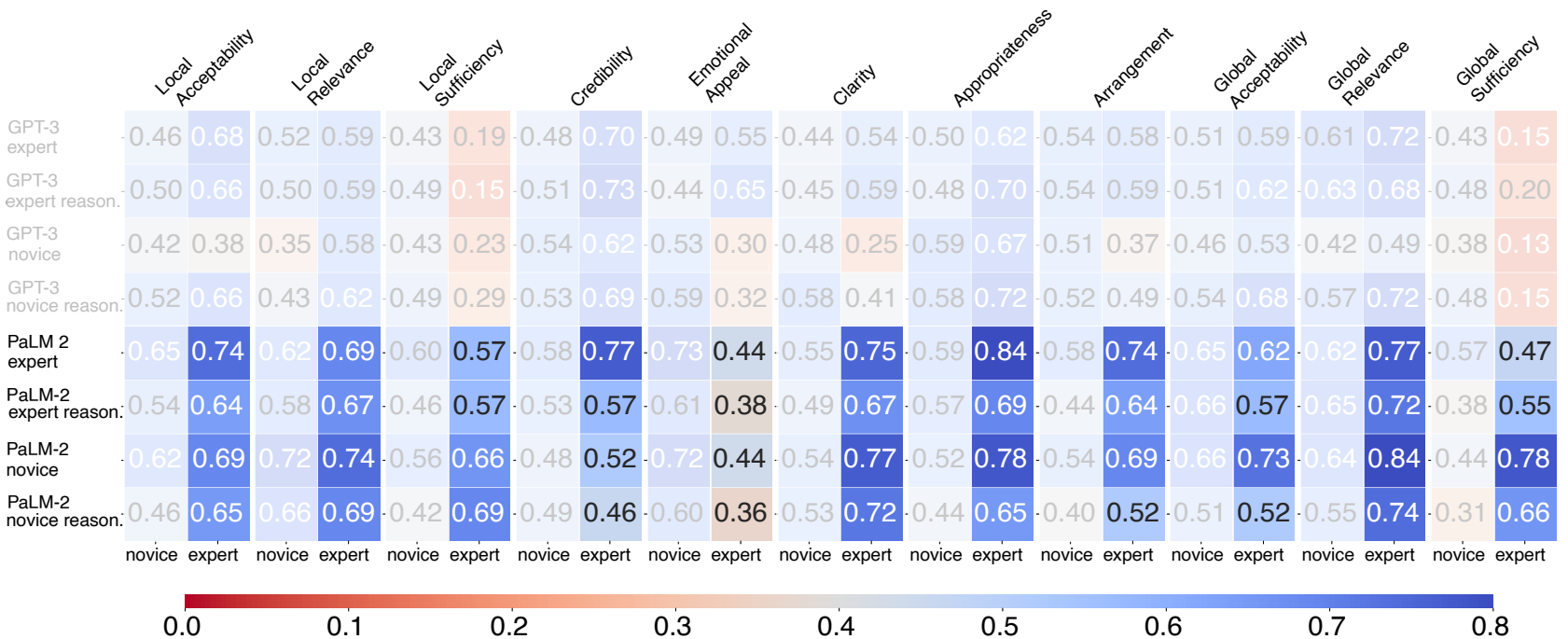
RQ2. Do LLM annotations align with human annotations?



Krippendorff's α between PaLM 2 and novices

Are LLMs Reliable Argument Quality Annotators?

RQ2. Do LLM annotations align with human annotations?



Krippendorff's α between PaLM 2 and experts

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

RQ3. Can we use LLMs as additional annotators to improve the agreement?

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human experts	0.40	0.40	0.40	0.40
+1 annotation	0.32*	0.37*	0.30*	0.37*
+2 annotations	0.31*	0.40	0.32*	0.40
+3 annotations	0.33	0.44*	0.35	0.44*
+4 annotations	0.35	0.47*	0.39	0.47*
+5 annotations	0.37	0.50*	0.42*	0.50*

Annotations	Expert		Novice	
	GPT-3	PaLM 2	GPT-3	PaLM 2
Human novices	0.37	0.37	0.37	0.37
+1 annotation	0.27*	0.29*	0.27*	0.27*
+2 annotations	0.26*	0.33	0.29*	0.30
+3 annotations	0.28*	0.37*	0.33	0.35*
+4 annotations	0.30*	0.41*	0.37	0.39*
+5 annotations	0.32*	0.45*	0.40*	0.43*

Overall Krippendorf's α change after adding annotations made by LLMs to the human annotations.
Significant changes ($p < 0.05$) are marked with *.

Are LLMs Reliable Argument Quality Annotators?

Conclusion

- LLMs provide consistent argument quality annotations
- LLM annotations show various degrees of agreement with humans
- Using LLMs as additional annotators can significantly improve the agreement between human annotators

Are LLMs Reliable Argument Quality Annotators?

Conclusion

- LLMs provide consistent argument quality annotations
- LLM annotations show various degrees of agreement with humans
- Using LLMs as additional annotators can significantly improve the agreement between human annotators

Code and Data

https://github.com/webis-de/RATIO-24

Are LLMs Reliable Argument Quality Annotators?

Conclusion

- LLMs provide consistent argument quality annotations
- LLM annotations show various degrees of agreement with humans
- Using LLMs as additional annotators can significantly improve the agreement between human annotators

Code and Data

https://github.com/webis-de/RATIO-24

Thank you!