LLM Judges for Retrieval Augmented Argumentation

SCAN ME

EMORY UNIVERSITY

Kaustubh Dhole Eugene Agichtein

Department of Computer Science, Emory University, USA

IR Intelligent Information Access Lab

Motivation

- Most RAG evaluations rely on single-score metrics, which often lack interpretability.
- These evaluations are typically conducted on short contexts and answers, limiting their applicability to longer documents.
- Tasks like argumentation involve long, noisy documents, posing challenges for traditional RAG metrics.

Query: Should the Federal Minimum Wage Be Increased **Stance:** Pro
Argument: The Federal Minimum Wage Should Be Increased...Raising the federal minimum wage to $15 would significantly benefit low-wage workers and their families..... Current federal minimum wage levels are insufficient to support a basic standard.... As noted by the Economic Policy Institute (EPI) ... federal minimum wage of $7.25 an hour cannot bring their family above the federal poverty line with full-time work which is currently pegged at $15 080 annually—significantly below the poverty threshold for a family with a child **[1]**....
Evidence Documents:
[1] Ben Zipperer, *Gradually Raising the Minimum Wage to $15 Would Be Good for Workers, Good for Businesses, and Good for the Economy*, epi.org, Feb. 7, 2019] **[2]** The Economist, "*Pay Dirt,*" economist.com, May 20, 2015 **[3]** ..... **[4]** ...

Contributions

- We demonstrate **feasibility** of model based evaluation for argumentation
- We extend our analysis to the **retrieval augmented setting** for long arguments
- We explore the **sensitivity of the proposed models to retrieval noise**

Related Work

- Wachsmuth et al. (2017) created a 15 level schema for a high quality argument
- Mirzakhmedova et al. (2024) show that LLMs can be prompted for each of the 15 dimensions to achieve moderate agreement with expert annotators

Computational argumentation quality assessment in natural language, Wachsmuth et al. (2017a)

Argumentation Quality Assessment: Theory vs. Practice, Wachsmuth et al. (2017)

Are Large Language Models Reliable Argument Quality Annotators? Mirzakhmedova et al. (2024)

LLM Judge for *Short* Arguments (validation experiments)

Dataset: *UKPConvArg1* corpus

Query: Should Physical Education Be Mandatory in Schools? **Stance:** Pro
Argument: PE should be compulsory because it keeps us constantly fit and healthy. If you really dislike sports, then you can quit it when you're an adult. But when you're a kid, the best thing for you to do is study, play and exercise. If you prefer to be lazy and lie on the couch all day then you are most likely to get sick and unfit. Besides, PE helps kids be better at teamwork.

LLM-Judges:

- We probe GPT4o with 15 theoretical argumentation dimensions provided to annotators to generate argument quality ratings in a listwise manner
  - ***(Argument, Stance) + Documentation → LLM → 15 quality ratings***
- Listwise approach is ***more efficient*** as it generates multiple annotations at once requiring lesser tokens and a single inference call.
- We generate annotator ratings from different LLMs by sampling with different temperatures.

Argumentation Quality\Annotators	LLM Annotator 1/Expert	LLM Annotator 2/Expert	LLM Annotator 3/Expert	LLM Annotator 4/Expert	Crowd of 4 Listwise LLMs / Expert
Overall Quality	0.34	0.37	0.37	0.35	**0.42**
Average of all Metrics (excluding OQ)	0.31	0.31	0.32	0.31	**0.37**

Argumentation Quality\Annotators	Human Crowd / Expert Wachsmuth et al. (2017)	LLMs Pointwise / Expert Mirzakhmedova et al. (2024) GPT3.5	LLMs Pointwise / Expert Mirzakhmedova et al. (2024) Palm2	Crowd of 4 Listwise LLMs / Expert (Ours)
Cogency	.27	-.15	.13	**.38**
Local Acceptability	.49	.43	**.59**	.41
Local Relevance	.42	.41	**.47**	.29
Local Sufficiency	.18	-.4	**.47**	.38
Effectiveness	.13	-.22	.2	**.37**
Credibility	.41	**.62**	.42	.46
Emotional Appeal	.45	**.48**	.17	.42
Clarity	.42	.39	**.47**	.24
Appropriateness	.54	**.57**	.53	.5
Arrangement	**.53**	.43	.51	.25
Reasonableness	.33	.21	.43	**.44**
Global Acceptability	.54	.42	.43	**.47**
Global Relevance	.44	**.64**	.58	.23
Global Sufficiency	-.17	-.27	**.46**	.37
Overall Quality	**.43**	.02	.29*	**.42**
Average of rubrics	.36	.25	**.42**	.37

Inter-annotator agreement (Krippendorff's α) between (a) between experts and individual LLM annotators. argument quality dimension (b) human experts and LLM annotations for each fine-grained *Mirzakhmedova et al. (2024) also show that providing novice documentation improves the overall quality to upto .55

LLM Judge for *Long* Arguments (new dataset)

- **Dataset:** We create a **new** RAG corpus based on the popular debate website ProCon.org.
  - Controversial Topic: q, Stance: s, Evidence Documents: D*, Argument: A*
- **RAG system:**
  - **Retriever:**
    - BM25 + Stance Conditioned Instruction LLM reranker (listwise)
    - Index: Webpages Scraped from Evidence URLs or Evidence Bing Searches
  - **Generator:** gpt4o-mini: q, s, D (retrieved context) → A
- **Evaluation:**
  - **Evaluation Metrics:** RAG Triad, Argument Quality
- **LLM Judges**
  - Pairwise Preference: Compare RAG- vs. Expert- written Argument (Table on Right)
  - Pointwise: Evaluate Argument Quality Based on 15 dimensions

LLM Judges	Description	Fine-Grained
TruLens (TruEra, 2024)	A single score metric	✖
Rubric-RAG-Boolean (Ours)	Evaluates by summing the presence (True) of attributes	✔
G-Eval-RAG (Adaptation of Liu et al, 2023 for RAG)	Generates evaluation criteria before evaluation	✖
Rubric-RAG-Rating (Ours)	Evaluates by summing the ratings (1 to 5) of the attributes	✔
I-Eval (Ours)	Evaluates by making the model first generate the three metrics (RAG Triad) in a Rubric-RAG style and then the final preference	✔
I-Eva-Direct (Ours)	Evaluates by making the model first generate the three metrics (RAG Triad) in a Direct Manner and then the final preference	✔

Metrics used to compute – context relevance, answer relevance, answer groundedness, and argument pairwise preference.

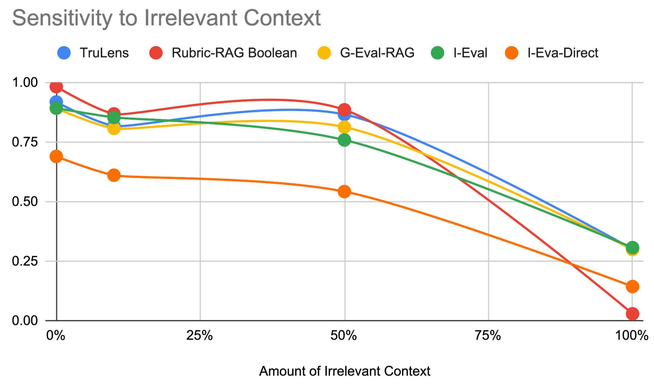
Preliminary results

Description	Arg15_Boolean 14 Metrics (Boolean) + Argument Quality (1 to 5)	Arg15_Likert All 15 Metrics (1 to 5)	Arg15_Likert_with_RAG_Triad RAG Triad + 14 Metrics (1 to 5) + Argument Quality (1 to 5)
Agreement of different ***pointwise*** LLMJudges with Expert Ratings			
Krippendorf's α	0.428	0.157	**0.461**

Agreement with Human Pairwise Preferences	TruLens	Rubric-RAG-Boolean	G-Eval-RAG	Rubric-RAG-Rating	I-Eval	I-Eva-Direct
3-way	0.012	0.023	0.059	0.132	-0.014	**0.204**
2-way*	0.346	0.318	0.220	0.288	0.006	**0.424**

Agreement of different ***pairwise preference*** LLMJudges against Human Arguments

*Annotators and model both are asked to generate a preference among Argument1, Argument2 and Both (3-way). To resolve ties for a 2-way agreement, a preference of 'both' is converted to two datapoints for each argument respectively.



Sensitivity to Retrieval Noise:

- With more *retrieval noise*, an LLM-judge assessing ***context_relevance*** should decrease
- Result: for increasing levels of retrieval noise, fine-grained metrics like I-eval and I-eval-direct monotonically decrease, but single-score metrics do not

Summary and Future Work

- Crowd of LLM-Judges can be effective to evaluate arguments
- Popular metrics used for NLG can be adapted for the RAG setting ensuring interpretabilty, sensitivity to retrieval noise
- Our dataset and analysis are useful for other tasks such as claim verification, educational assessment, literature review, legal analysis

EMORY UNIVERSITY https://ir.cs.emory.edu/

Emory Intelligent Information Access Lab

Workshop on Task Focused IR in the Era of Generative AI, 2024, Microsoft Research

Partially supported by Accelerating Foundation Models Research Microsoft