Improving Attributed Long-form Question Answering with Intent Awareness

ICLR 2026

Xinran Zhao, Aakanksha Naik, Jay DeYoung, Joseph Chee Chang, Jena D. Hwang,
Tongshuang Wu, Varsha Kishore

Outline

- The role of intents in human writing
  - Motivation & Taxonomy
- Intent-awareness improves deep research performance
  - inference & training
- Intents guide the understanding of models
  - Analyze behavior & Navigate readers
- Discussion

Motivation

- LLMs are trained with text from academic papers
  - However, beyond the words “on paper”, human academic writing is a complex cognitive process that also includes planning *what to write*, i.e., intents (Flower and Hayes, 1981)
- Recent study on LaTeX keystroke (Wang et al., 2025) shows that
  - 10% of human keystrokes are spent on planning the initial thoughts / logic / organizations
- We aim to reveal the intents of writing, as part of this complex thought process in the academic writing pipeline

Taxonomy

We consider types of writing intents at different granularities

- Citation intent
  - Sentence-level: we follow the widely-studied citation intent classification literature, e.g., background, uses, comparison, etc in Jurgens et al., (2018)
- Paragraph intent
  - (Multi-)paragraph-level: why people write certain contents, e.g., definition, argumentation, compare, narration, etc in *discours mode* (Song et al., 2017)

Literature	Categories
ACL-ARC (Jurgens et al., 2018)	Background, Motivation, Uses, Extension, Comparison or Contrast, Future
Discourse mode (Song et al., 2017)	Exposition, Definition, Argumentation, Compare or Contrast, Cause-Effect, Prompt-Solution, Evaluation, Narration

How to leverage intents?

- Ai2 ScholarQA RAG pipeline
  - At inference, we just ask the model to add the in-line intent with {tag, type, rationale}:
    - … Previous content ‹bcit› [type] rationale ‹ecit› [original citation] later content …
    - … Previous paragraph ‹bpit› [type] rationale ‹epit›
  - At fine-tuning,
    - We add in-line tagged intents to the training data
- Experiment setup
  - Metrics: answer quality and citation querlity
  - Question: In robotics, what are the leading methods for learning terrain traversability costs automatically
  - Prompt: Generate a report … Include inline citation from given references … output a formatted report
  - Example of the answer (formatted): {sections: {title, text, citations: {id, snippets, title, metadata}}}

Just ask for intents

- We first test the long-form attributed report generation performance with inline intents together with the answers
- + intents denote that we also prompt the model to generate answers with intents
- There is consistent performance improvement with best performing large models.

	SQA-CS-V2	Deep Scholar Bench	Research QA
Gemini 2.5 Pro	88.1	54.8	71.9
- With intent	**89.7**	**57.8**	**74.0**
Claude Opus 4	85.4	58.1	74.3
- With intent	**89.0**	**59.9**	**75.7**

Fine-tuning with intents

- We fine-tune smaller-sized models with large model generated reports with in-line intents
- Variants include: (1) no training; (2) baseline SFT without intents; (3)intent-aware SFT
- Variants are trained with controlled steps and compared on SQA-CS-V2. Intent-aware SFT leads to large margin over the baseline training with reports only

	No Training	Baseline SFT	Intent-aware SFT
Qwen 3 - 8B	80.7	83.2	**88.6**
Llama 3.1 - 8B	66.4	84.4	**89.2**
Qwen 3 - 4B	80.9	83.4	**87.0**

Intents help analyze model behavior

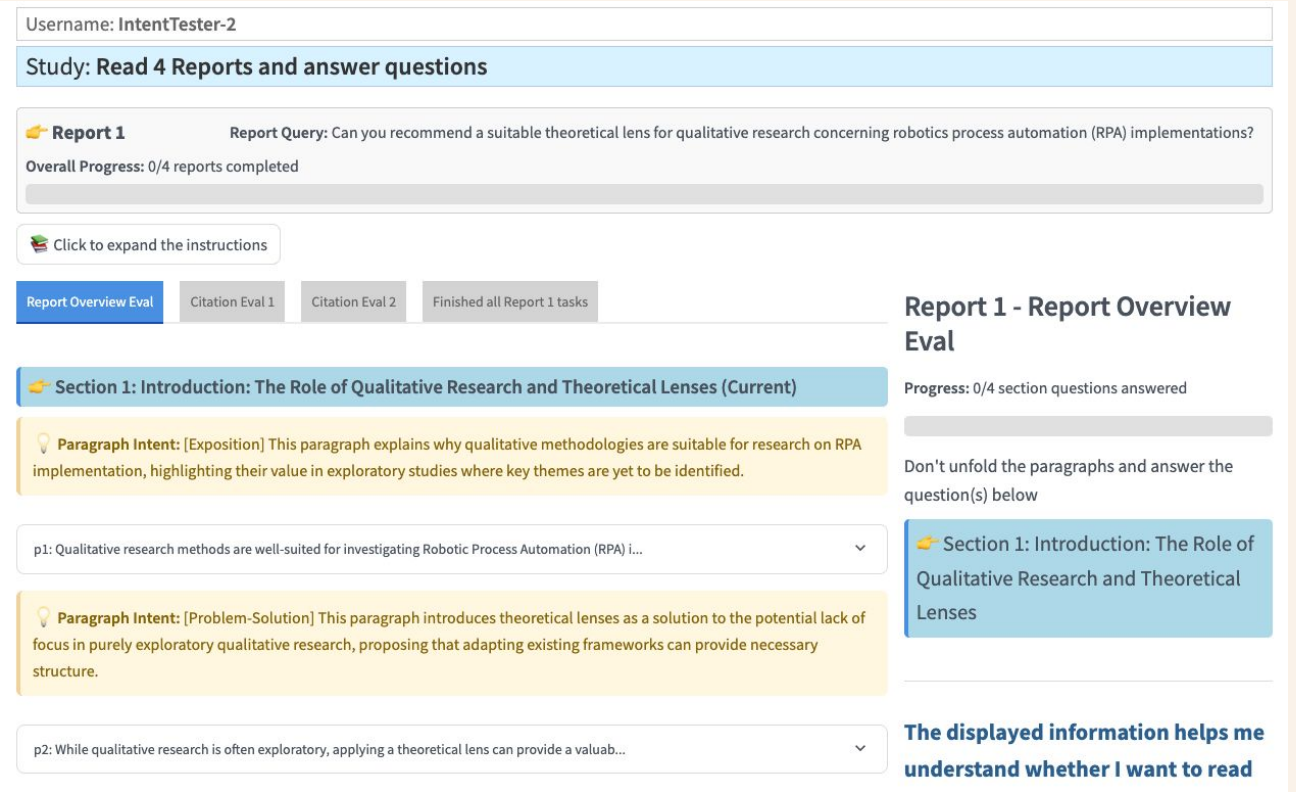
- Through the tag-based system, we can understand the model behavior differences
- Key findings:
  - Model citation aligns with humans - *Background* and *Uses* dominate
  - Models significantly underuse *Comparison or Contrast*
  - There are model specific differences - Gemini rarely produces *Extension* or *Future work*

Citation (%)	o3	gemini	opus-4.1	*Human ref*
Background	28.2	29.6	21.1	51.9
Motivation	10.6	7.1	6.8	5.0
Uses	40.4	55.9	47.4	18.5
Extension	6.9	0.7	12.8	3.7
Comparison	4.7	4.8	3.8	17.5
Future	4.2	0.9	2.8	3.5
(error)	5.0	0.9	5.3	0.0

Paragraph (%)	o3	gemini	opus-4.1
Expos.	41.5	51.5	39.9
Def.	7.0	7.1	7.3
Argu.	11.6	8.6	5.1
Comp.-Contr.	6.4	6.1	9.7
Cause-Eff.	6.1	2.6	5.4
Prob.-Sol.	14.5	13.4	22.8
Narr.	2.7	5.2	1.3
Eval.	9.6	5.4	8.5
(error)	0.0	0.0	0.0

Intents help navigate users

- A glimpse of our reading system
- With 20 participants and 70 reports, they report +16% and +23.2% readability on paragraphs and citations



Discussion

- Potential future directions
  - Hierarchy of intents beyond sentences and paragraphs
  - Intent as a Diagnostic and Analysis Layer (data & reward)
  - Generalization to other domains
- A short conclusion
  - Our work introduces intent awareness to the inference and training pipeline of long-form attributed report generation
  - Beyond performance, intents help analyze model behavior and navigate readers in the wordy reports
- Thank you for listening! More details are in our paper!

Thank you!

- ----------