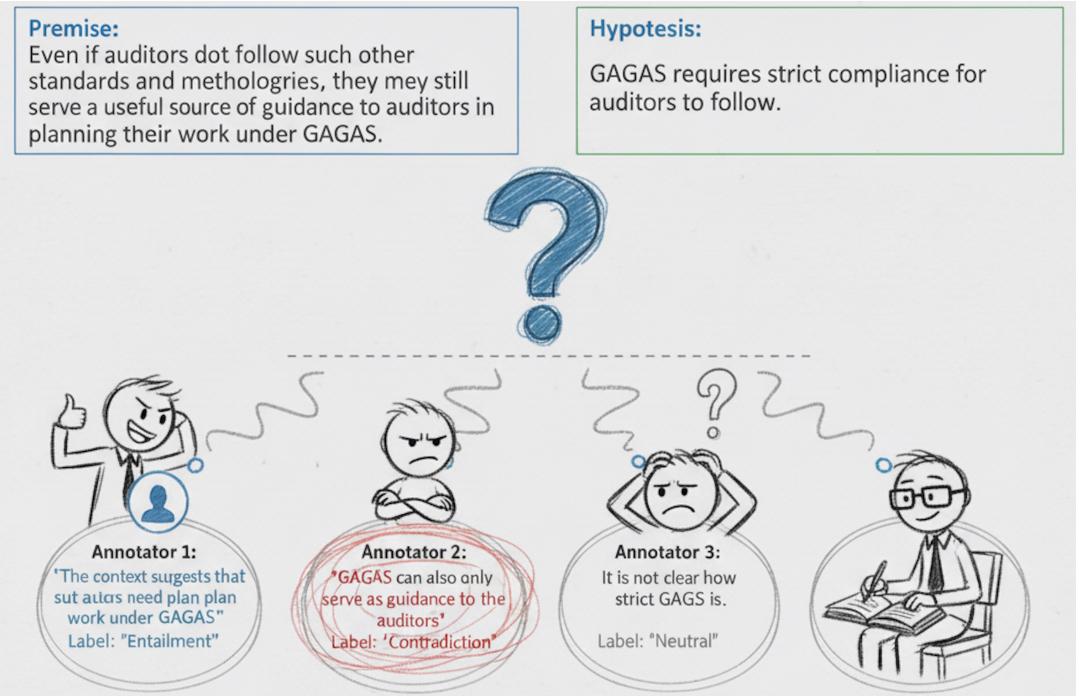
Threading the Needle: Reweaving Chain-of-Thought Reasoning to Explain Human Label Variation

Beiduo Chen, Yang Janet Liu, Anna Korhonen, Barbara Plank

MaiNLP lab, LMU Munich
University of Pittsburgh
LTL, University of Cambridge

EMNLP 2025: The 2025 Conference on Empirical Methods in Natural Language Processing
Nov 7, 2025

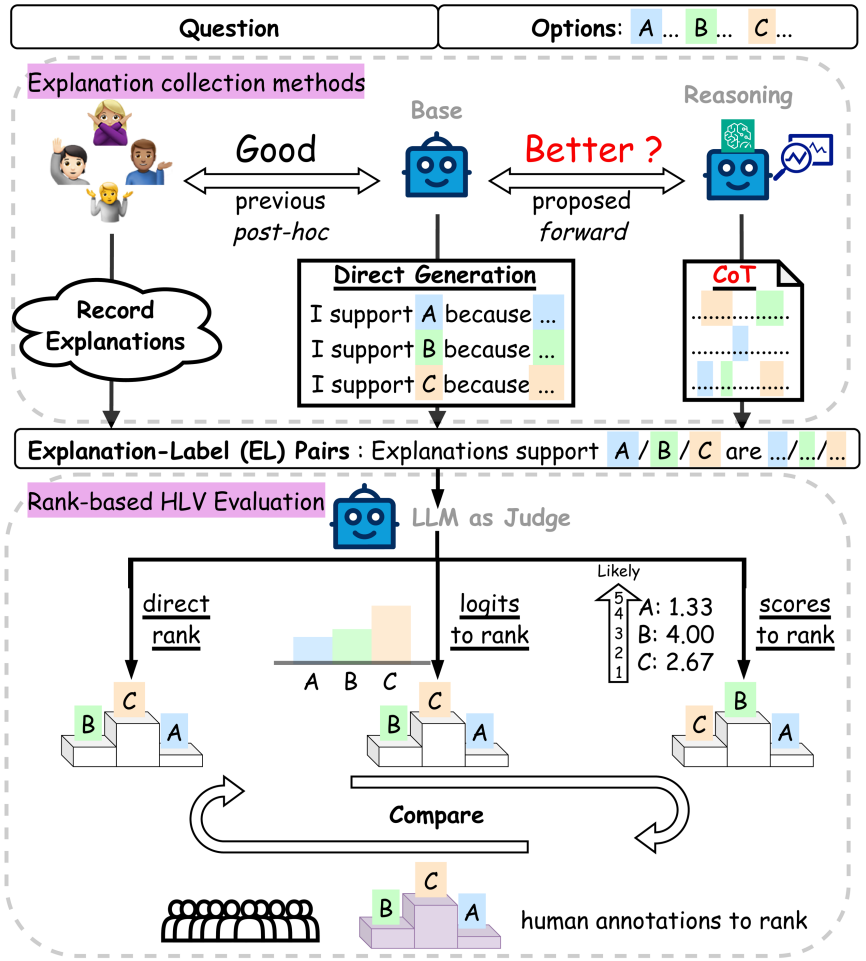
The Challenge with "Ground Truth"



Human Label Variation is a Signal, Not Noise

- In many NLP tasks, different annotators provide valid but divergent labels for the same instance.
- This phenomenon, **Human Label Variation** (HLV), reflects the diversity of human perspectives.
- **Our Goal**: To treat this variation as a valuable signal, enabling models to understand it rather than ignore it.
- Explanation-based method via annotators: effective but costy

Explanation-based Method via LLMs (Post-hoc)



Prior work often uses a **post-hoc** approach: an explanation is generated for a pre-determined label.

Key Problems:

- **Unnatural Process**: It reverses the human cognitive process of reasoning before deciding.
- **Risk of Hallucination**: Models might fabricate justifications for implausible options just to fulfill the prompt.
- **Lacks Comparison**: Explanations are generated in isolation, missing the comparative reasoning between options.

Explanations are Already in the "Thought" Process (Forward)

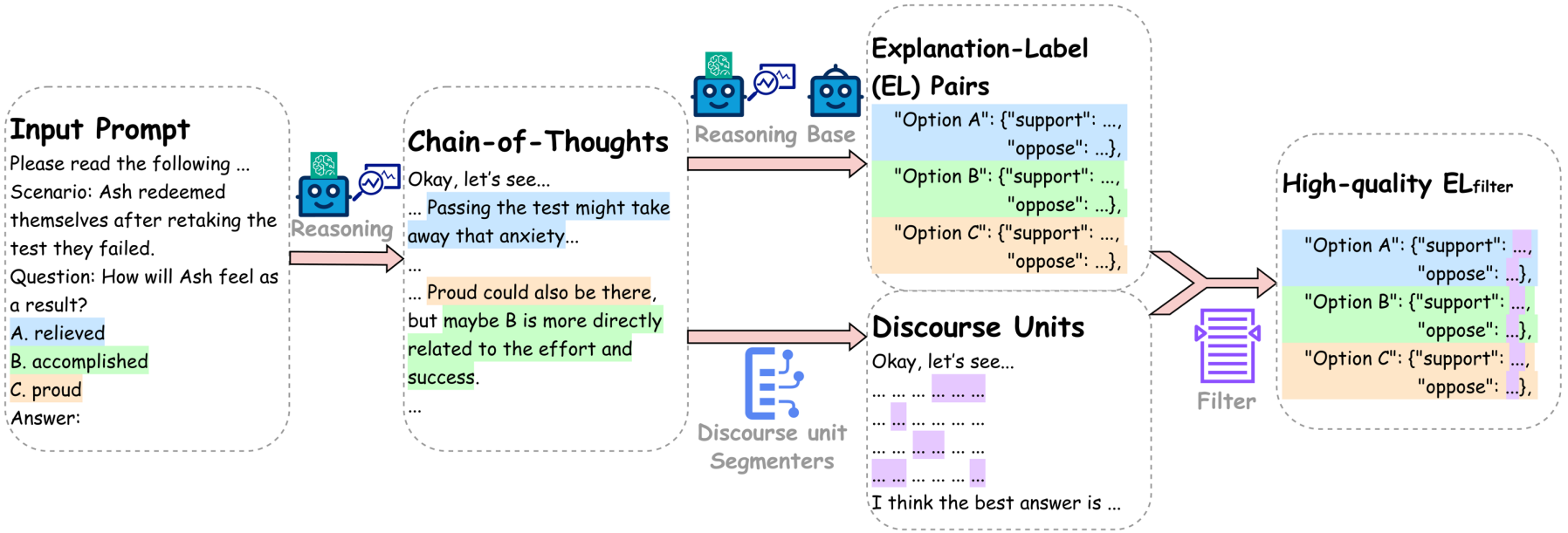
Input: Please read the following social scenario and the accompanying question, choose the most appropriate answer from the options provided and start your answer with a single letter. Scenario: Ash redeemed themselves after retaking the test they failed. Question: How will Ash feel as a result?
A. relieved B. accomplished C. proud Answer:

CoTs: Okay, let's see. The scenario is that Ash redeemed themselves by retaking a test they failed. The question is asking how Ash will feel as a result. The options are A. relieved, B. accomplished, and C. proud.
First, I need to understand what each of these emotions means in this context. Redeeming oneself after failing a test suggests that Ash probably worked hard to retake it and succeeded this time. So, the feelings here would be related to overcoming a challenge.
"Relieved"(A) would mean that Ash is feeling a release from the stress or worry about having failed before. Passing the test might take away that anxiety.
"Accomplished"(B) implies a sense of achievement, having completed something successfully, especially after effort. Since Ash had to retake the test, putting in the work to pass it now would make them feel like they've accomplished something.
"Proud"(C) is more about self-esteem or satisfaction in one's own achievements. It's a bit similar to accomplished but more focused on the personal pride aspect.
Now, considering the scenario: Ash didn't just pass a test; they redeemed themselves. That suggests that failing initially might have caused some shame or disappointment, so passing now would alleviate that. But redemption often involves not just relief but also a sense of achievement. So both A and B are possible.
But "proud" might be more about the internal sense of pride, whereas "accomplished" is about the feeling of having achieved something through effort. Which one fits better here? If Ash worked hard to retake and pass, then feeling accomplished makes sense. Relieved is possible, but maybe the primary feeling is accomplishment. Proud could also be there, but maybe B is more directly related to the effort and success.
I think the best answer is B. Accomplished.

Chain-of-Thought (CoT): A Rich Source of Latent Explanations

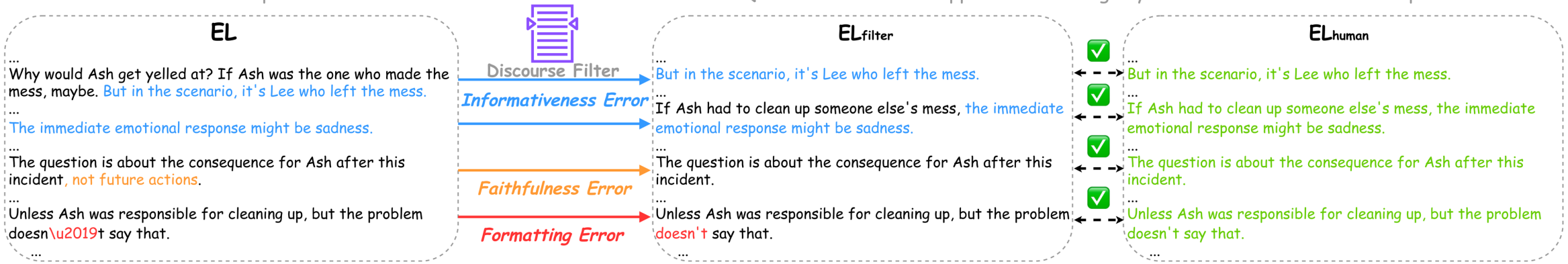
- A CoT is the reasoning process a model generates before providing a final answer.
- This "forward reasoning path" naturally contains analysis, comparison, and trade-offs among various options.
- **Our Core Insight**: Instead of generating new explanations from scratch, we can **extract and repurpose** these more authentic reasoning fragments that already exist within the CoT.

The CoT2EL Pipeline: From Chaos to Clarity



- Generate CoT: We first prompt a reasoning-tuned LLM for its detailed thought process.
- Initial Extraction: An LLM parser performs a first pass, structuring the CoT into supporting/opposing statements for each label.
- **Discourse Segmentation** : This initial output can be noisy. We introduce linguistically-motivated Discourse Segmenters to break the CoT into coherent semantic units.
- Align & Filter: We align the extracted explanations with these high-quality discourse units, filtering out noise and producing clean, faithful Explanation-Label pairs.

Discourse-guided Refinement



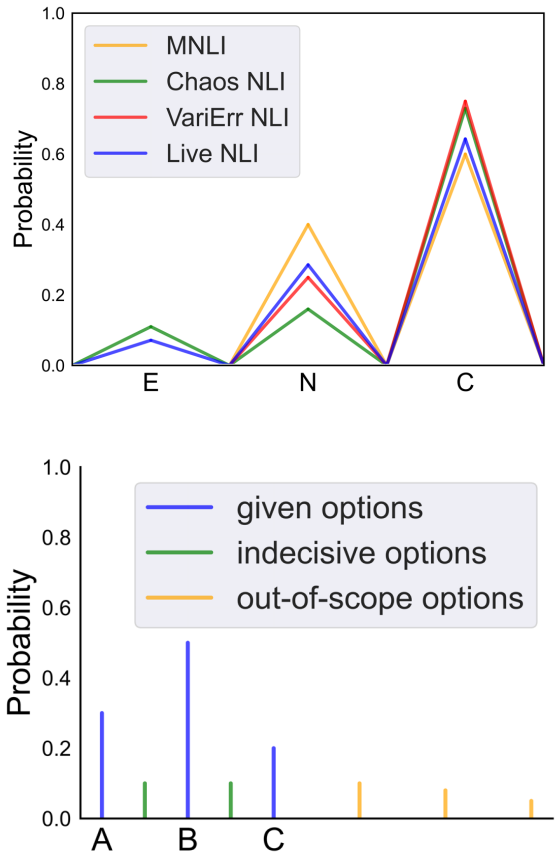
Three key issues:

- **Informativeness** Errors: Explanations are fragmented or contain irrelevant information.
- **Faithfulness** Errors: Statements are paraphrased or hallucinated, not matching the original CoT content.
- **Formatting** Errors: The structured JSON output is inconsistent or broken.

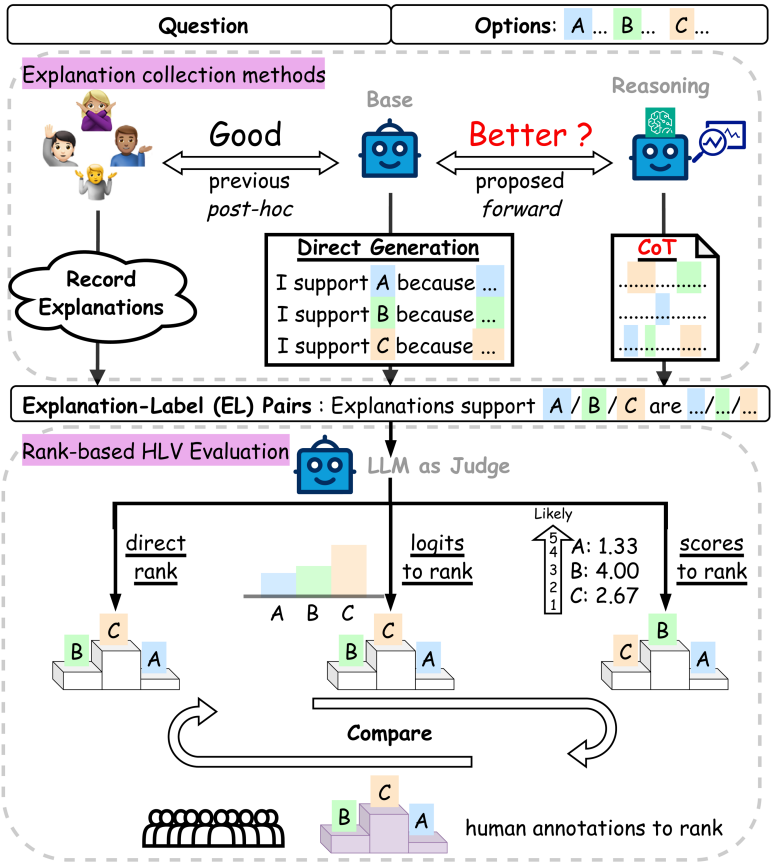
A coherent text, like a CoT, is composed of fundamental "thought units" called **discourse units**.

- We use two complementary discourse segmenters (based on RST and PDTB theories) to break the original CoT into a high-quality set of valid, coherent discourse units.
- We then take the initially parsed explanations and align each one with its closest match from our high-quality set of discourse units.

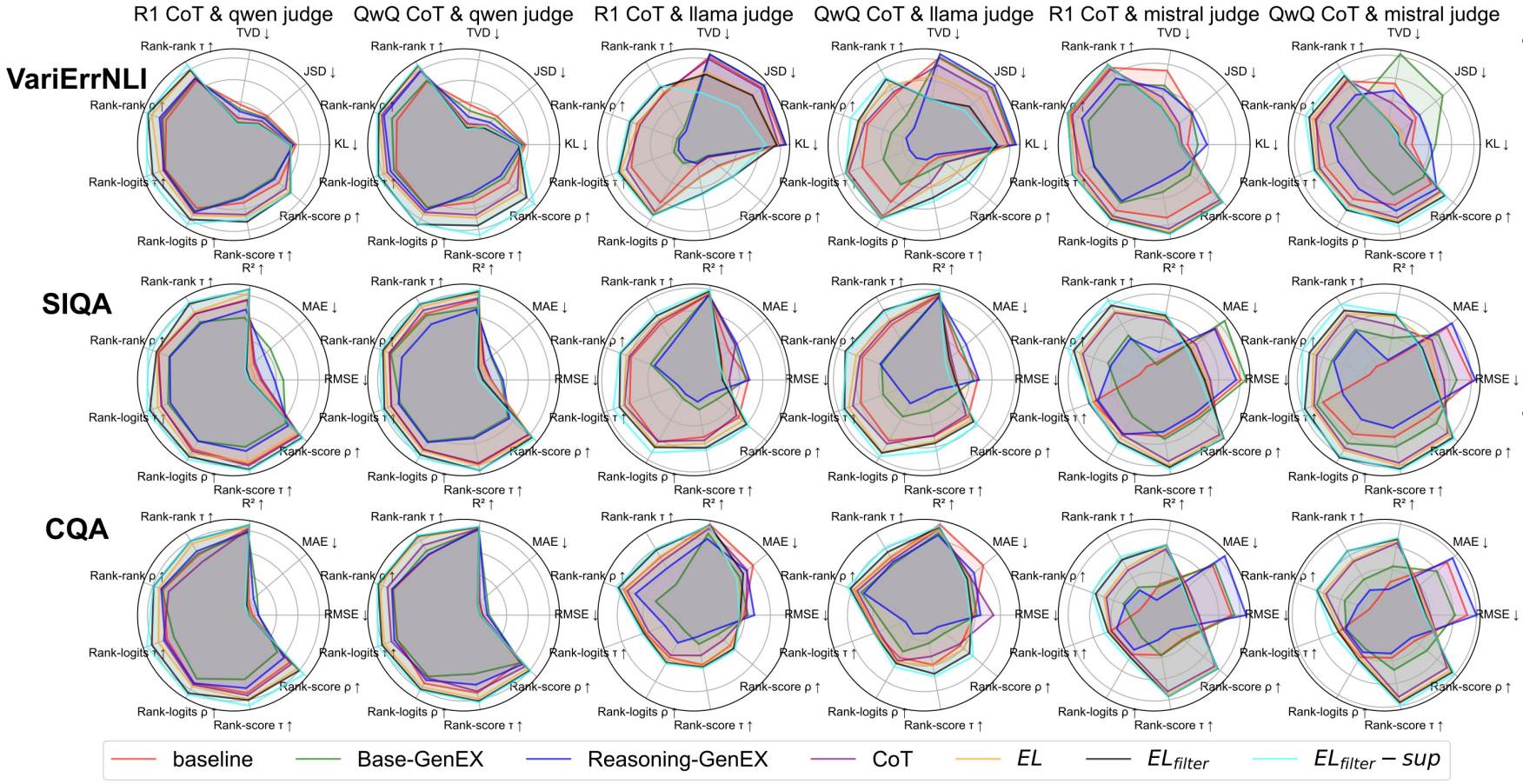
A More Robust HLV Evaluation: Ranks over Values



- Traditional HLV evaluation often focuses on matching exact probability distributions.
- We found that while these distributions vary significantly across different annotator groups, the **preference ranking** of the options is often remarkably consistent.
- **Our Evaluation Framework**: We focus on the correlation between the model's predicted ranking and the human-derived ranking, which provides a more robust measure of alignment.

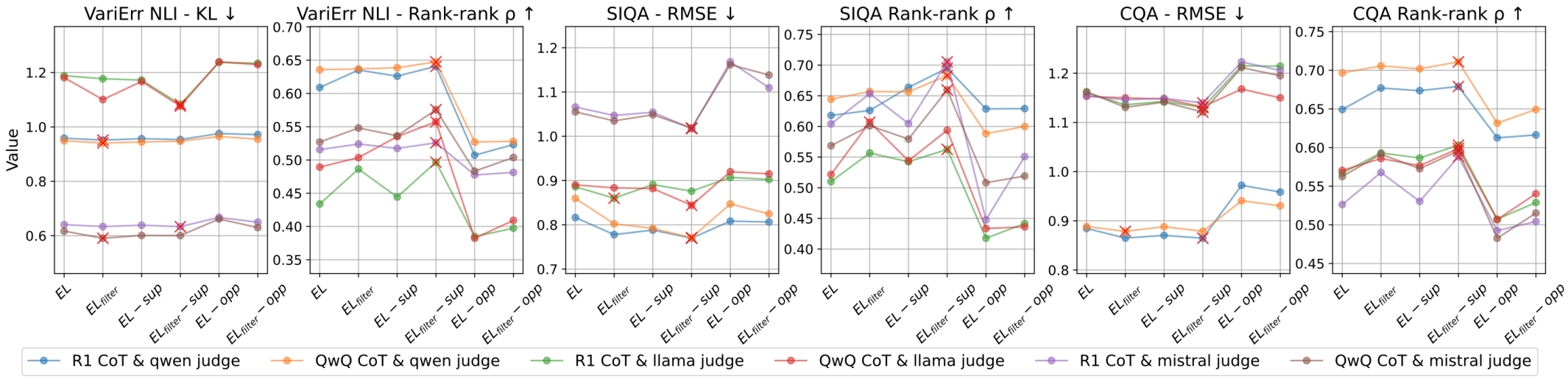


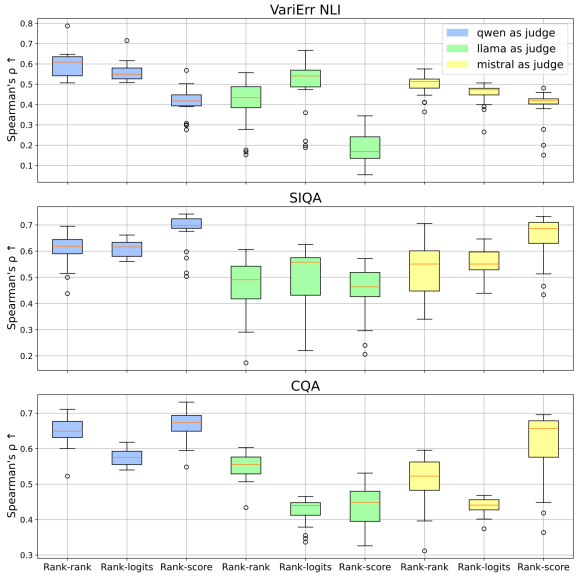
CoT Explanations Better Capture Human Preferences



- Across three datasets, our CoT2EL method consistently and significantly outperforms all baselines in aligning with human rankings and distributions.
- Crucially, it is far more effective than the traditional "Direct Generation" approach, demonstrating the superiority of the forward-reasoning paradigm.

Analyses





- Providing only Supportive explanations consistently outperformed providing both supportive and opposing explanations.
- The best ranking generation method for the LLM judge depended on the dataset's original annotation format.
- LLM judges performed significantly better when explanations were presented in a well-structured format compared to when they were given the raw, unstructured CoT output.

Conclusion & Contributions

- **A New Paradigm**: We shift from "post-hoc" justification to "forward" reasoning extraction from CoT for modeling HLV.
- **A Novel Pipeline** (CoT2EL): We creatively integrate discourse segmentation from linguistics to enhance the quality and faithfulness of extracted explanations from CoTs.
- **A More Robust Evaluation**: We advocate for and validate a rank-based evaluation framework that is more stable and interpretable for HLV.

Thank You !

Beiduo Chen

Contact: beiduochen@cis.lmu.de

Code: https://github.com/mainlp/CoT2EL