

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



Beiduo Chen

Yang Janet Liu

Anna Korhonen

Barbara Plank



MaiNLP, CIS, LMU Munich, Germany



Munich Center for Machine Learning, Germany

Department of Linguistics, University of Pittsburgh, USA



LTL, University of Cambridge, UK



Munich Center for Machine Learning

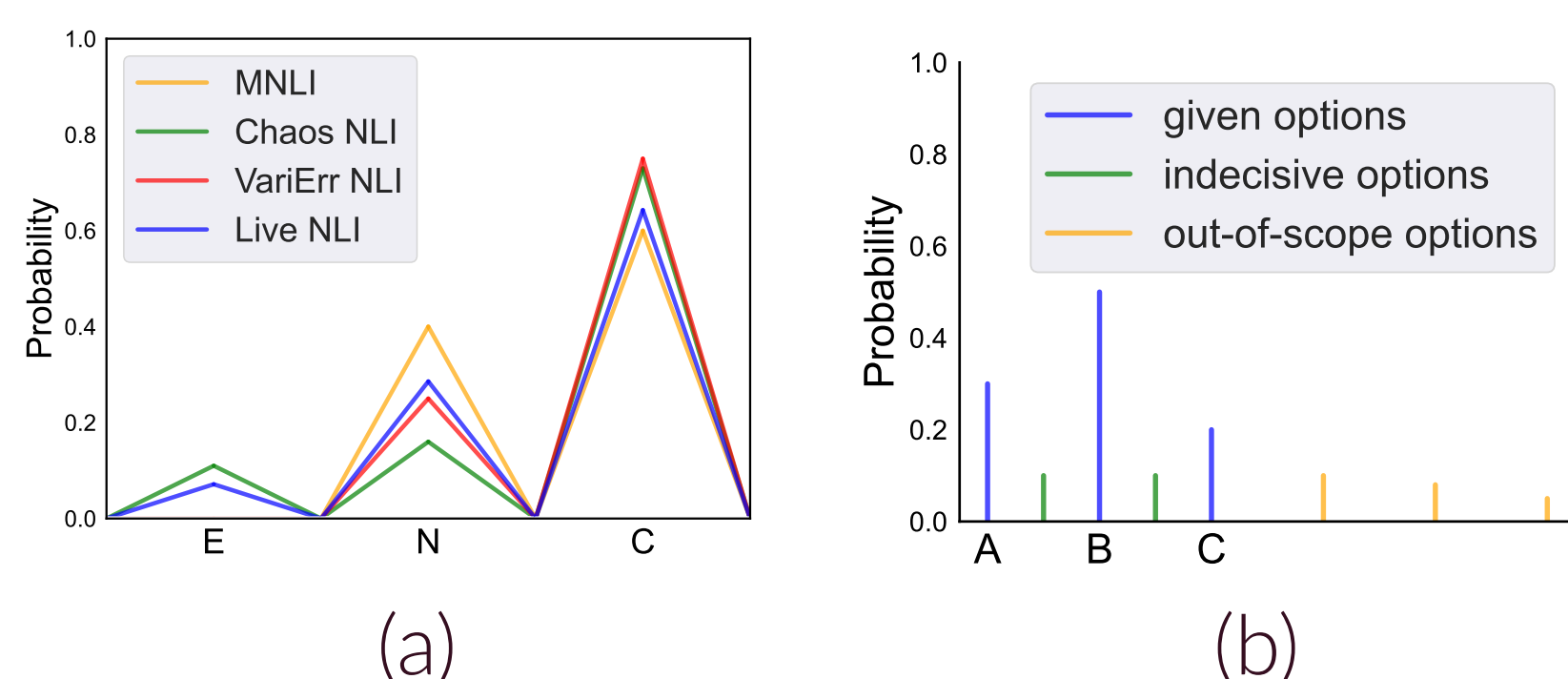
Introduction and Background

- Human Label Variation:** Different annotators may provide divergent yet valid labels, reflecting the diversity of human judgments.
- Reasoning-tuned Large Language Models:** LLMs trained to generate step-by-step reasoning (Chain-of-Thoughts) before answers, improving interpretability and decision-making.

CoT-based Explanation Extraction

- Modeling HLV via LLMs with human explanations:** Human-provided explanation-label pairs help LLMs capture annotation disagreements but are costly to collect.
- Explanation-Generation via LLMs as Post-hoc:** Prior methods ask LLMs to generate explanations after selecting a label—risking shallow or hallucinated reasoning.
- Chain-of-Thoughts from LLMs as Forward:** CoTs naturally embed rationales before answers in a forward process, enabling richer, label-specific explanations.

Rank-based HLV Evaluation



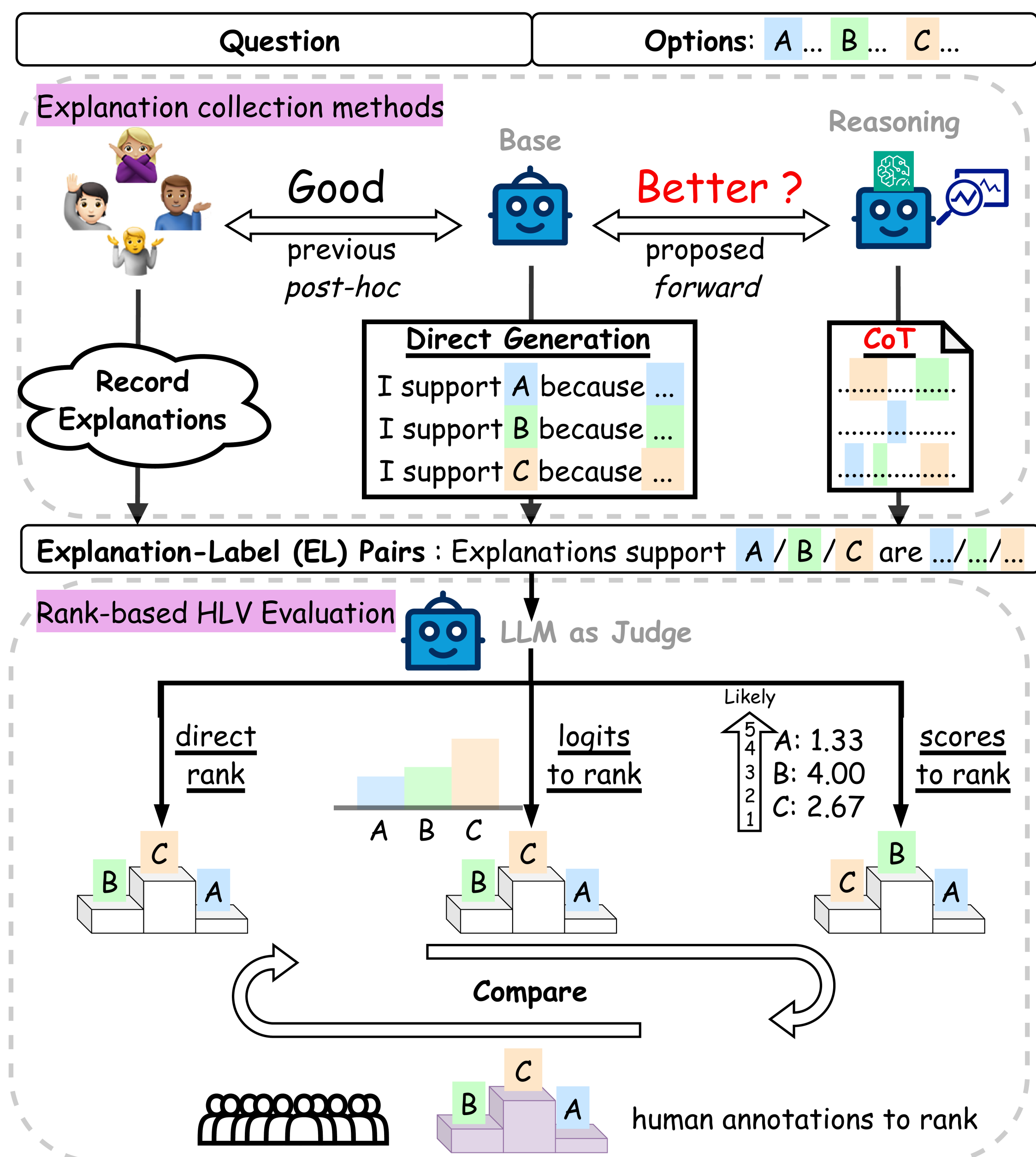
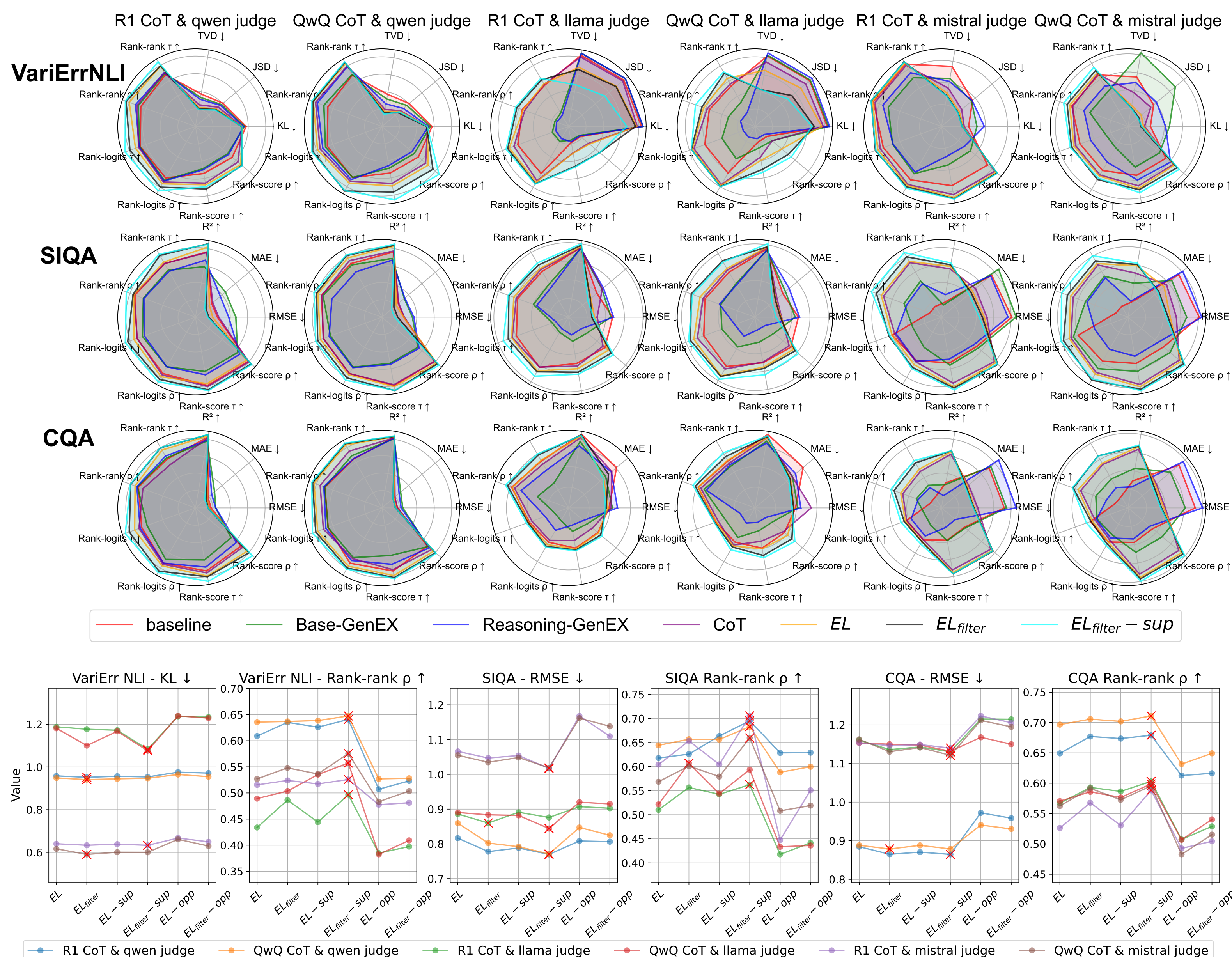
CoT2EL Pipeline

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.

The examples, drawn from Social IQa, are processed using R1, with CoT segments related to options A, B, and C highlighted in blue, green, and orange, respectively. For questions where all options appear plausible, the CoT reflects a step-by-step analysis involving comparison and evidence evaluation. These segments offer rich material for deriving explanations relevant to our study of the HLV phenomenon, enabling a deeper understanding of the task and the roots of HLV beyond just the final answers.

Main Results



Input Prompt

Please read the following ...
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:

Chain-of-Thoughts

Okay, let's see...
... Passing the test might take away that anxiety...
... Proud could also be there, but maybe B is more directly related to the effort and success.

Explanation-Label (EL) Pairs

"Option A": ("support": ..., "oppose": ...),
"Option B": ("support": ..., "oppose": ...),
"Option C": ("support": ..., "oppose": ...).

Discourse Units

Okay, let's see...
...
I think the best answer is ...

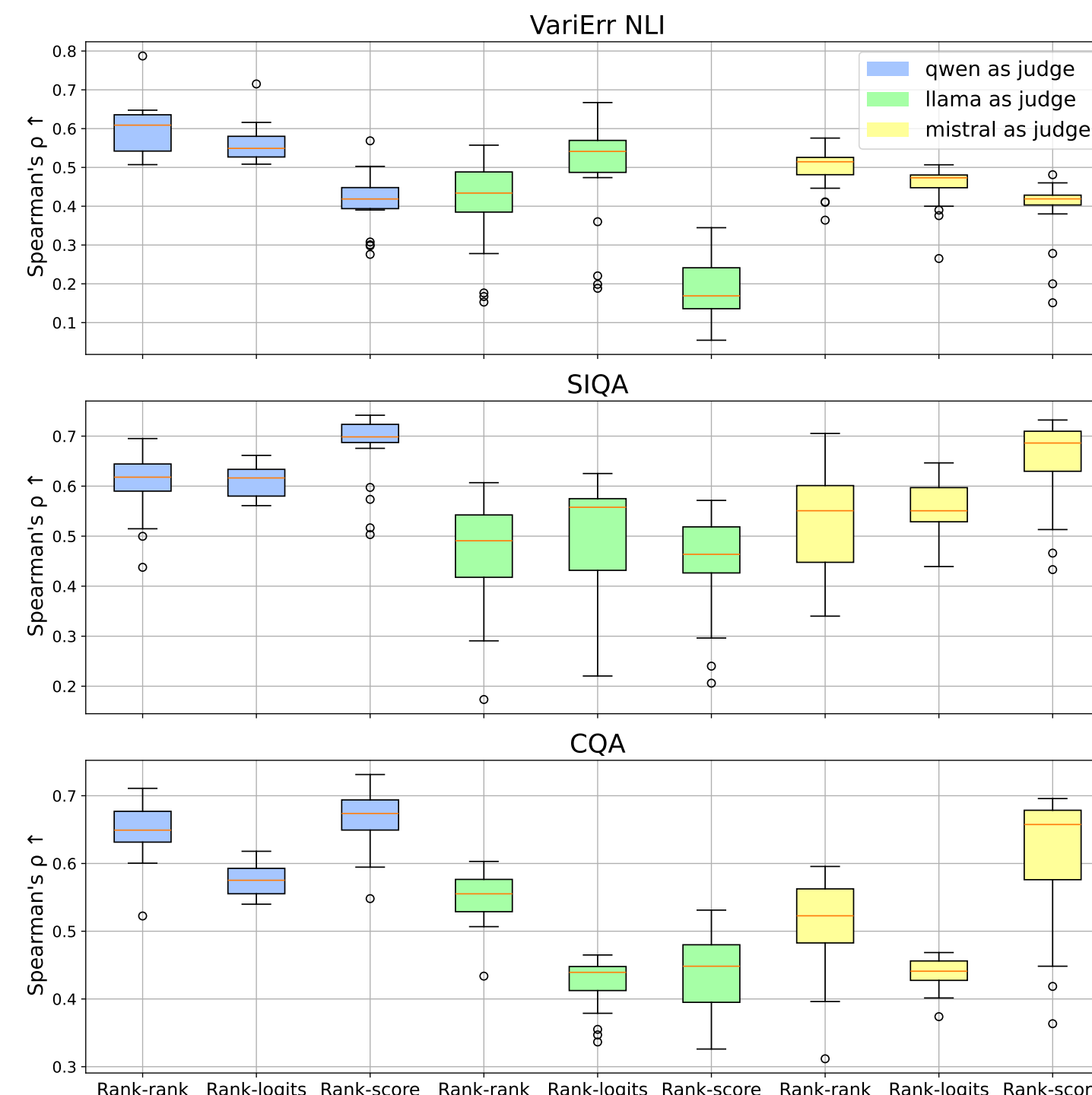
High-quality EL-filter

"Option A": ("support": ..., "oppose": ...),
"Option B": ("support": ..., "oppose": ...),
"Option C": ("support": ..., "oppose": ...).

EL	EL-filter	EL-human
Why would Ash get yelled at? If Ash was the one who made the mess, maybe. But in the scenario, it's Lee who left the mess. The immediate emotional response might be sadness. The question is about the consequence for Ash after this incident, not future actions. Unless Ash was responsible for cleaning up, but the problem doesn't say that.	But in the scenario, it's Lee who left the mess. If Ash had to clean up someone else's mess, the immediate emotional response might be sadness. The question is about the consequence for Ash after this incident. Unless Ash was responsible for cleaning up, but the problem doesn't say that.	But in the scenario, it's Lee who left the mess. If Ash had to clean up someone else's mess, the immediate emotional response might be sadness. The question is about the consequence for Ash after this incident. Unless Ash was responsible for cleaning up, but the problem doesn't say that.

Analyses and Takeaways

Datasets	VariErr NLI					SIQA					CQA				
	Distribution			Rank-rank		Score		Rank-rank			Score		Rank-rank		
Settings/Metrics	KL ↓	JSD ↓	TVD ↓	τ ↑	ρ ↑	RMSE ↓	MAE ↓	R2 ↑	τ ↑	ρ ↑	RMSE ↓	MAE ↓	R2 ↑	τ ↑	ρ ↑
baseline	1.0006	0.2644	0.2776	0.4971	0.5119	0.8630	0.7461	0.1300	0.5451	0.6069	0.9101	0.7417	0.4255	0.5395	0.6283
HumanEX	0.9408	0.2455	0.2448	0.7411	0.7872	0.8912	0.7730	0.0912	0.4047	0.4377	0.9209	0.7536	0.4205	0.4507	0.5225
R1 - CoT parser	0.9610	0.2576	0.2637	0.5597	0.5966	0.8222	0.7113	0.2429	0.5450	0.6169	0.8849	0.7298	0.4428	0.5716	0.6419
R1 - EL	0.9583	0.2566	0.2625	0.5693	0.6089	0.8164	0.7184	0.2479	0.5611	0.6179	0.8845	0.7298	0.4554	0.5957	0.6492
R1 - EL-filter-sup	0.9534	0.2552	0.2604	0.6050	0.6408	0.7698	0.6660	0.3176	0.6500	0.6951	0.8646	0.6956	0.4937	0.6114	0.6790
QwQ - CoT parser	0.9504	0.2534	0.2589	0.5698	0.6201	0.8607	0.7248	0.2536	0.6002	0.6346	0.9006	0.7326	0.4329	0.6253	0.6734
QwQ - EL	0.9488	0.2535	0.2583	0.5962	0.6357	0.8597	0.7220	0.2670	0.6089	0.6443	0.8882	0.7317	0.4357	0.6270	0.6966
QwQ - EL-filter-sup	0.9471	0.2528	0.2552	0.6104	0.6475	0.7709	0.6672	0.3212	0.6394	0.6830	0.8787	0.7197	0.4541	0.6378	0.7109



- supporting statements within CoTs are more predictive of human choices than the opposing ones;
- ranking strategies aligned with human annotation practices lead to better approximation of human label variation;
- well-structured input prompts significantly enhance the informativeness.



Paper



Code



The 2025 Conference on Empirical Methods in Natural Language Processing

ellis



European Research Council

Established by the European Commission

