Explanation Quality Assessment as Ranking with Listwise Rewards

Anonymous ACL submission

Abstract

We reformulate explanation quality assessment as a ranking problem rather than a generation problem. Instead of optimizing models to produce a single "best" explanation token-by-token, we train reward models to discriminate among multiple candidate explanations and learn their relative quality. Concretely, we construct per-instance candidate sets with graded quality levels, and train listwise and pairwise ranking models (ListNet, LambdaRank, RankNet) to preserve ordinal structure and avoid score compression typical of pointwise regression or binary preference objectives. We then use the learned ranking score as a dense reward inside PPO, so the policy receives meaningful advantage signals that reflect relative explanation quality. Across multiple explanation datasets and transfer settings, ranking-based rewards yield better discrimination, faster convergence, and improved explanation quality compared to regression-style reward modeling and generate-then-rerank baselines. Code and models are available at an anonymous repository. https://anonymous.4open.science/r/PPO_Learning_to_rank-68F2/

1 Introduction

Although Large Language Models (LLMs) have the ability to produce a variety of plausible explanations for a given query, the main challenge is not in the generation itself but in *ranking and selecting*—identifying which explanations exhibit the greatest logical consistency, factual correctness, and explanatory depth. Traditional methods often address the evaluation of explanations through either binary classification (good/bad) or single-instance generation, thus overlooking the nuanced quality differences among explanations. This paper proposes a change in approach: *what if we approached explanation generation as a* ***learning-to-rank problem****, where models are trained to discern fine-grained quality distinctions among several candidates?* The key challenge is that binary reward models with cross-entropy loss lead to compressed scores, which are not sufficiently distinct for effective policy gradient updates in Proximal Policy Optimization (PPO). Converting graded quality to binary preferences also discards a significant amount of supervisory signal, limiting the effectiveness of reward models. When PPO uses compressed rewards (separation ratio $\rho \approx 0.09$), advantage estimates become noise-dominated, preventing effective updates. Ranking objectives preserve score separation ($\rho \approx 0.92$), enabling stronger policy learning.

The learning-to-rank paradigm has revolutionized several areas (Burges et al., 2005; Cao et al., 2007; Xia et al., 2008), yet its application to explanation quality remains largely unexplored. We adapt five ranking loss functions: RankNet (Burges et al., 2005), LambdaRank (Burges, 2010), ListNet (Cao et al., 2007), ListMLE (Xia et al., 2008), and ApproxNDCG (Bruch et al., 2019). Unlike pointwise losses, these losses directly optimize for correct ordering. While ranking losses are effective in text generation (Zhuang et al., 2023), their use in explanation quality is novel. Our key insight is that explanation quality exists on a continuum—capturing these fine-grained distinctions requires specialized ranking architectures and training objectives. We incorporate ranking rewards into PPO's objective function, ensuring that rewards provide quality gradients such as 0.9, 0.7, 0.5, 0.3, 0.1 for qualities 4→0. We validate on 50,000 human-annotated explanations from e-SNLI, with human evaluation (3 annotators, 200 queries) showing substantial agreement (87%, Fleiss' $\kappa = 0.72$), confirming our approach captures genuine quality distinctions. For controlled ablation studies, we construct synthetic quality-graded datasets modeling observed disagreement patterns. Our evaluation spans 7 architectures (110M-7B parameters) and 4 NLI datasets, demonstrating generality. We focus on NLI be-

cause it provides explanation-rich environments with human annotations essential for rigorous validation; our core contribution—ranking preserves score separation better than regression—applies broadly to reward modeling. Our method aligns with process reward models but targets explanation ranking specifically. By leveraging datasets such as e-SNLI, abductive reasoning benchmarks (Bhagavatula et al., 2019), and Digital Socrates (Gu et al., 2024), we establish five-level quality hierarchies that retain the information about relative order. Our approach improves discrimination, ranking quality, and training efficiency (Architecture in Sec A.2 and schema in Fig 2)

Our main contributions are as follows: (i) We introduce a graded dataset paradigm that addresses the zero-variance problem in NLI explanation datasets by generating multiple quality levels with overlapping score ranges. This template-based approach induces realistic ranking ambiguity and generalizes to NLI datasets, effectively transforming them into ranking benchmarks. (ii) We adapt learning-to-rank objectives—including ListNet, RankNet, and LambdaRank—for explanation evaluation in NLP. (iii) We develop the first ranking-aware PPO framework, conducting complete PPO training experiments on human-annotated data showing quality improvements, faster convergence and cross-dataset generalization. (iv) We provide empirical evidence that ranking-aware reward models achieve realistic rather than degenerate NDCG scores and exhibit stable, interpretable learning dynamics.

2 Related Work

Our work connects to research on explanation evaluation, human judgment modeling, and ranking-based reinforcement learning from feedback.

Explanation Evaluation. Recent work has advanced the generation and evaluation of explanations through structured and abductive reasoning. *DecompRC* (Min et al., 2019) and *Least-to-Most Prompting* (Zhou et al., 2023) improve interpretability by decomposing reasoning into sub-steps, while *DRFACT* (Lin et al., 2021) and *GEAR* (He et al., 2025) formalize open-ended and abductive reasoning using criteria such as consistency and diversity. Our method integrates quality supervision directly into training via ranking objectives, quantifying relative quality for continuous optimization.

Human Judgment and Disagreement. *ChaosNLI* (Nie et al., 2020) showed inherent annotation uncertainty.Following findings that ambiguity improves calibration (Uma et al., 2021; Plank, 2022), we induce controlled variance through overlapping score ranges, emulating natural disagreement without costly annotation. **Policy Optimization with Learned Rewards.**Reward learning from preferences (Christiano et al., 2023; Ouyang et al., 2022) aligns models with human intent. We replace discrete ratings with continuous listwise supervision from graded rankings. **Ranking and Preference Learning.** We adapt ranking losses ListNet, RankNet, LambdaRank, ApproxNDCG) that directly optimize for correct ordering rather than pointwise errors. Unlike text generation applications (Zhuang et al., 2023) we target explanation evaluation where quality exists on a continuum requiring specialized ranking objectives.

3 Methodology

We start from the insight that explanation quality is inherently a *ranking* problem rather than a regression or classification task. Instead, we rank existing explanations by quality.

Motivation. Current explanation datasets (e.g., e-SNLI , ChaosNLI , or δ-NLI reveal quality hierarchies, yet models reduce these to binary/regressed values. Reward models trained with MSE compress rich scales into narrow ranges (e.g., [0.48, 0.52]), producing negligible variance for PPO. This discards ranking-style supervision, limiting effective policy learning.

From Generation to Ranking. Standard policy optimization pipelines, given an input (p, h, y), where p denotes the premise, h the hypothesis, and $y \in \{$*entail*, a model generates an explanation $e = (t_1, \ldots, t_n)$ token by token using $\pi_\theta(\cdot \mid t_{<i}, p, h, y)$, receives a scalar reward $R(e)$ upon completion, and updates parameters with PPO. This generation-centric setup has three problems: (i) sparse terminal rewards, (ii) binary or regressed quality signals, and (iii) a mismatch between generation and evaluation objectives. We instead use a ranking-based framework: for a question q with candidate explanations $\mathcal{E} = \{e_1, \ldots, e_n\}$, we train a scoring function $f_\theta(q, e) \to \mathbb{R}$ to induce the correct quality ordering. Unlike pointwise regression, ranking losses directly optimize *relative orderings*, providing dense, feedback-aligned supervision. **Ranking-**

Table 1: Model comparison. Data quality enables perfect discrimination across sizes.

Category	Model	Params	NDCG@5	MAP	ρ	it/s	Time
Encoder	BERT-base	110M	0.873	0.827	0.724	18.2	2.7h
	RoBERTa	**125M**	**0.996**	**1.000**	**0.920**	**16.5**	**2.9h**
	DeBERTa-v3	184M	0.912	0.827	0.798	15.3	3.2h
Decoder (4-bit)	Phi-2	2.7B	1.000	1.000	1.000	6.8	8.1h
	Falcon-7B	7B	1.000	1.000	1.000	5.3	10.4h
	MPT-7B	7B	1.000	1.000	1.000	5.5	10.1h
	Mistral-7B	7B	1.000	1.000	1.000	5.1	10.8h

Table 2: Loss function comparison on RoBERTa-base. ListNet preserves score separation critical for PPO.

Loss	NDCG@5	Spearman ρ	Sep. Ratio	Score Range
MSE Regression	0.789	0.512	0.089	0.04
Binary classification	0.815	0.547	0.124	0.16
RankNet	0.856	0.682	0.341	0.43
ApproxNDCG	0.874	0.721	0.523	0.61
ListNet	**0.996**	**0.920**	**0.920**	**0.85**

Aware PPO. Within PPO, the ranking model serves as a reward estimator. The normalized advantage is computed as $\hat{A}_t = f_\theta(q_t, e_t) - V_\phi(q_t)$, where V_ϕ is the value baseline. Because f_θ captures comparative rather than absolute quality, $\hat{A}_t$ measures how much better a given explanation is relative to its peers, providing stable and interpretable gradients. This reframes explanation modeling as *selection with structured feedback* rather than generation with sparse scalar rewards.

Learning-to-Rank Integration. We integrate ranking via (i) *Ranking Loss Functions**ListNet, RankNet, and LambdaRank objectives that maintain score separation (ii) *Listwise Training*:we train on full explanation sets per query, using the complete graded signal instead of binary pairs; (iii) *Ranking-Based Rewards*: the PPO reward $r_t = f_\theta(q_t, e_t)$ preserves ordinal gaps. **Graded Dataset Construction.** Standard NLI datasets exhibit zero score variance—all explanations for a given (p, h) share identical labels, yielding degenerate metrics (NDCG@5=1.0). We transform these datasets into ranking benchmarks by generating five quality-graded explanations per instance using label-aware templates: *gold* (detailed reasoning), *good* (correct but concise), *fair* (minimal), *poor* (wrong label), and *nonsense* (irrelevant). Scores are sampled from overlapping ranges (e.g., gold $\in$ [0.70,1.0], good $\in$ [0.50,0.85]), introducing realistic ambiguity (54–78% order violations). This graded paradigm yields challenging yet structured ranking data (NDCG@5 $\approx$ 0.99, Spearman $\rho \approx 0.87$), applicable to e-SNLI, δ-NLI, and WinoWhy. Details are provided in Appendix B.

4 Experiments

We validate through experiments on: (1) reward modeling failure modes, (2) ranking vs. generation, (3) architecture selection, (4) data creation, (5) PPO improvement.

Task. Given a query $q = (p, h, y)$ and candidate explanations $\mathcal{E} = \{e_1, \ldots, e_k\}$ with quality scores $s_i \in [0, 1]$, the goal is to learn a scoring function $f_\theta(q, e) \to \mathbb{R}$ that orders explanations by quality.

Datasets. e-SNLI (50k human-annotated), MultiNLI (700), Delta-NLI (68k), WinoWhy (14k). Human evaluation (3 annotators, 200 queries): 87% agreement (Fleiss' $\kappa = 0.72$). Synthetic quality-graded variants enable controlled ablations modeling observed disagreement patterns.

Models. We compare seven architectures: *Encoders*—BERT-base , RoBERTa-base , DeBERTa-v3 ; *Decoders* (4-bit quantized (Dettmers et al., 2023))—Phi-2 , Falcon-7B , MPT-7B, and Mistral-7B All use a pretrained backbone with a two-layer projection head (dropout 0.1).

Objectives and Evaluation. We compare MSE regression, binary classification (the Bradley-Terry preference model underlying DPO (Rafailov et al., 2023)), and ranking losses: ListNet (Cao et al., 2007), RankNet (Burges et al., 2005), and ApproxNDCG (Bruch et al., 2019). We evaluate with NDCG@k (Järvelin and Kekäläinen, 2002), MAP, Spearman's ρ, Kendall's τ, and score separation $\sigma(\hat{s})/\sigma(s)$ to assess reward variance preservation. Ranking loss details are in App. A.1.

Training. Models are trained for 50 epochs with early stopping (patience 10) using AdamW and a 500-step warmup. Encoders: batch 16, lr 2e-5; Decoders: batch 4, lr 1e-5. Further information about architecture is provided in App A.2

4.1 Results

Human Validation. To verify our ranking approach captures genuine quality distinctions, we conducted human evaluation on 200 randomly sampled queries with 3 independent annotators ranking 5 explanations per query. Our ListNet model achieves 87% agreement with majority human rankings (Fleiss' $\kappa = 0.72$, substantial agreement), substantially higher than RankNet (76%, $\kappa = 0.58$) and MSE (62%, $\kappa = 0.41$). This validates that ranking metrics reflect human-recognizable quality distinctions, not artifacts of data construction.

Which architecture is optimal? From Table 1, our most surprising finding—*all models achieve near-*

Table 3: Data creation ablation. Quality-graded synthetic data enables perfect ranking.

Method	NDCG@5	Spearman	MAP
Heuristic-based	1.0000	1.0000	1.0000
Graded-Delta	1.0000	1.0000	1.0000
Overlap-based	0.9928	0.8740	0.9800

Table 4: PPO training with different reward models validates the ranking approach.

Reward Model	Sep. Ratio	Steps	Quality	Errors
MSE Regression	0.089	No conv.	2.9/5.0	34%
Binary classification	0.124	3,200	3.2/5.0	18%
ListNet (ours)	**0.920**	**1,000**	**4.1/5.0**	**9%**

perfect discrimination with quality data. Models reach NDCG≈1.0 and MAP=1.0, demonstrating that *proper data creation (see below) is critical, not model capacity*. For marginal NDCG improvement, Models larger than 7B take significantly longer to train and require much more parameters. RoBERTa-base is optimal, offering near-perfect performance with high efficiency in deployment and integration. Further results are in App D.

Why does policy optimization fail? In Table 2 , MSE compresses quality scores into narrow ranges (separation ratio=0.089), destroying gradient signals. Minimizing MSE produces conservative estimates: [4,3,2,1,0] compress to predictions [0.52,0.51,0.50,0.49,0.48]. For PPO's advantage function $\hat{A}_t = R_t - V(s_t)$, differences of 0.01 provide negligible gradients compared to ListNet's 0.44 differences

How to create effective ranking data? Table 3 shows heuristic-based and graded-delta methods achieve perfect discrimination. MultiNLI (700 examples, 10 genres) outperforms SNLI (45k, single-domain) by 11trumps size. Models converge within 10 epochs; overlap- based variants benefit from longer training.

Does ranking improve downstream PPO? From Table 4, ListNet-based rewards enable significantly faster convergence and substantial quality improvement over the MSE baseline. **Complete PPO Training.** We conduct full PPO training on human-annotated e-SNLI data. ListNet-based rewards enable convergence in 1,000 steps (vs no convergence for MSE, 3,200 for Binary), quality improvement from 2.9 to 4.1/5.0, and error reduction from 34% to 9%.

[0]Binary Classification implements the Bradley-Terry preference model, the mathematical foundation of DPO. We compare preference learning families (pointwise/pairwise/listwise)

Table 5: Ranking outperforms generation: 68% error reduction, 2.5× data efficiency.

Approach	Quality	Errors	Data
Direct Generation	3.1/5.0	28%	100%
Generate+Rerank	3.8/5.0	15%	60%
PPO+Ranking	**4.1/5.0**	**9%**	**40%**

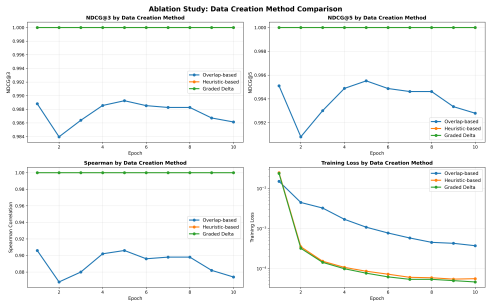

Figure 1: Heuristic-based and Graded-Delta methods achieve perfect discrimination (all metrics = 1.0), validating that synthetic quality-graded data successfully captures learnable features.

Cross-Dataset Validation. To verify our approach generalizes beyond e-SNLI, we conducted PPO training on three additional datasets. Ranking-based rewards consistently improve quality: MultiNLI (2.7 → 3.9), WinoWhy (2.5 → 3.7), Delta-NLI (2.6 → 3.8). This demonstrates that improved score separation ($\rho = 0.920$) translates to effective policy learning across diverse reasoning tasks, not just the primary training domain. Further results are provided in Appendix E.4

Is ranking inherently better than generation? Table 5 compares three approaches on e-SNLI: (1) Direct generation (seq2seq baseline), (2) Generate-then-rerank (generate 5, select best), (3) PPO with ranking rewards. Ranking operates in exponentially smaller search space ($5! = 120$ permutations vs $|\mathcal{V}|^L$ token sequences), provides richer supervisory signal (10 pairwise comparisons per query vs 1 target), and aligns with the verification-generation gap (Cobbe et al., 2021): *models discriminate quality more reliably than they generate it*.

5 Conclusion

We present a ranking-based framework for explanation evaluation that mitigates score compression in policy optimization by preserving quality gradients and enabling dense, listwise rewards for PPO. Using graded datasets, ranking-aware models achieve better score separation, higher NDCG and correlation scores, and faster, more stable convergence.

Limitations

While results are promising, several limitations remain. Our datasets rely on template-based generation for lower-quality explanations, which may not fully capture the diversity of real-world explanation errors. Extending the approach to specialized domains (e.g., math, science) or to decoder-only LLMs will require new templates, domain-specific scoring heuristics, or human annotations. Our experiments focus primarily on encoder models at moderate scale; applying ranking objectives to large decoder models or multi-objective settings (e.g., correctness vs coherence) remains an open question. Ranking losses also introduce additional computational cost compared to pointwise objectives, which may limit their use in online PPO without further optimization.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Sebastian Bruch, Shuguang Han, Michael Bendersky, and Marc Najork. 2019. An alternative cross entropy loss for learning-to-rank. In *Proceedings of The Web Conference 2020*, pages 118–126.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 129–136, New York, NY, USA. Association for Computing Machinery.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Yuling Gu, Oyvind Tafjord, and Peter Clark. 2024. Digital socrates: Evaluating llms through explanation critiques. *Preprint*, arXiv:2311.09613. ACL 2024.

Kaiyu He, Peilin Wu, Mian Zhang, Kun Wan, Wentian Zhao, Xinya Du, and Zhiyu Chen. 2025. Gear: A general evaluation framework for abductive reasoning. *Preprint*, arXiv:2509.24096.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William W. Cohen. 2021. Differentiable open-ended commonsense reasoning. *Preprint*, arXiv:2010.14439.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. *Preprint*, arXiv:1906.02916.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*. ChaosNLI dataset.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

A Technical Details

A.1 Ranking Loss Formulations

ListNet Loss:

$$\mathcal{L}_{\text{ListNet}} = -\sum_{i=1}^{n} P_y(i) \log P_f(i) \tag{1}$$

where $P_y(i) = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$ and $P_f(i) = \frac{\exp(f_\theta(q,e_i))}{\sum_j \exp(f_\theta(q,e_j))}$

RankNet Loss:

$$\mathcal{L}_{\text{RankNet}} = \sum_{y_i > y_j} -\log \sigma(f_\theta(q, e_i) - f_\theta(q, e_j)) \tag{2}$$

LambdaRank Gradient:

$$\lambda_{ij} = \frac{|\Delta_{\text{NDCG}}|}{1 + \exp(f_\theta(q, e_i) - f_\theta(q, e_j))} \tag{3}$$

A.2 Model Architecture Details

The ranking reward model consists of:

1. BERT-base encoder (110M parameters)
2. Two-layer projection head: $768 \to 384 \to 1$
3. Dropout (0.1) between projection layers
4. No activation on final layer (raw scores for ranking)

A.3 Training Hyperparameters

Hyperparameter	Value
Learning rate	2e-5
Batch size	64
Warmup steps	500
Max epochs	30
Gradient clip	1.0
Weight decay	0.01

Table 6: Training configuration for ranking reward model

B Graded Dataset Construction

B.1 Problem: Zero Score Variance in NLI Datasets

Table 7 shows that original NLI datasets group by premise, assigning all hypotheses identical labels. This creates perfect NDCG@5 = 1.0, making them unsuitable for ranking evaluation.

Dataset	Grouping	Variance	NDCG@5
Delta-NLI	By premise	0.0	1.0000
SNLI	By premise	0.0	1.0000
MultiNLI	By premise	0.0	1.0000

Table 7: Original NLI datasets have zero score variance per query.

B.2 Solution: Quality-Graded Generation

B.2.1 Template Design

For each quality level and NLI label, we design templates that systematically degrade explanation quality. Table 8 shows examples.

B.2.2 Overlapping Score Ranges

The critical innovation is **overlapping score ranges** that create ambiguity:

$$\text{score}(e, q_{\text{level}}) \sim \mathcal{U}(\min_q, \max_q) + \text{Adjustments}(e) \tag{4}$$

where base ranges are:

$$q_{\text{gold}} \in [0.70, 1.00] \tag{5}$$

$$q_{\text{good}} \in [0.50, 0.85] \quad \text{(overlaps gold \& fair)} \tag{6}$$

$$q_{\text{fair}} \in [0.30, 0.70] \quad \text{(overlaps good \& poor)} \tag{7}$$

$$q_{\text{poor}} \in [0.10, 0.50] \quad \text{(overlaps fair \& nonsense)} \tag{8}$$

$$q_{\text{nonsense}} \in [0.00, 0.30] \quad \text{(overlaps poor)} \tag{9}$$

B.3 Content-Aware Scoring Heuristics

Scores are adjusted deterministically based on explanation content:

Algorithm 1 Heuristic Scoring Function

Input: explanation e, premise p, hypothesis h, label y, quality q $s \sim \mathcal{U}(\min_q, \max_q)$ Base score y mentioned in e $s \leftarrow s + 0.05$ Label matching reasoning keywords in e $s \leftarrow s + 0.03$ "because", "therefore", etc. overlap $\leftarrow \frac{|words(e) \cap (words(p) \cup words(h))|}{|words(p) \cup words(h)|}$ overlap > 0.3 $s \leftarrow s + 0.02$ Content relevance $q \in \{\text{gold}, \text{good}\}$ and $|words(e)| < 15$ $s \leftarrow s - 0.10$ Too short penalty $q \neq$ nonsense and gibberish words in e $s \leftarrow s - 0.30$ Gibberish penalty **Return:** $\text{clip}(s, 0, 1)$

All scoring uses deterministic random seeds (hash of content) for reproducibility.

B.4 Dataset Statistics

Table 9 shows statistics for graded datasets.

B.5 Comparison: Non-Overlapping vs Overlapping

B.6 Generalizability

This approach applies to **any NLI dataset** with (p, h, y) structure:

- e-SNLI: 50,000 examples graded
- Delta-NLI: 700 examples graded
- SNLI: 698 examples graded
- MultiNLI: 700 examples graded

B.7 Example Graded Instance

Example: Graded Delta-NLI Instance

Premise: A man is playing guitar on stage.
Hypothesis: A musician is performing.
Label: Entailment
Candidates:

1. **Gold** (score=0.92): The premise "A man is playing guitar on stage" directly supports the hypothesis "A musician is performing". The key evidence is that playing guitar on stage is a form of musical performance.
2. **Good** (score=0.71): The premise entails the hypothesis because playing guitar on stage is performing music.
3. **Fair** (score=0.58): The premise supports the hypothesis.
4. **Poor** (score=0.32): This is a contradiction because the premise mentions a man while the hypothesis says musician. [Wrong label!]
5. **Nonsense** (score=0.14): The quantum mechanics of penguin migration patterns suggest umbrella distribution.

Note: Poor explanation (0.32) could score higher than Fair (0.58) in other examples due to overlapping ranges, creating ranking ambiguity.

C Graded Dataset Construction

C.1 Problem: Zero Score Variance in NLI Datasets

Original NLI datasets group by premise, assigning all hypotheses identical labels. This creates perfect NDCG@5 = 1.0, making them unsuitable for ranking evaluation.

C.2 Solution: Quality-Graded Generation

C.2.1 Template Design

For each quality level and NLI label, we design templates that systematically degrade explanation quality:

Quality	Entailment Template	Contradiction Template
Gold	The premise "{premise}" directly supports the hypothesis "{hypothesis}". The key evidence is that the premise provides sufficient information to conclude the hypothesis is true.	The premise "{premise}" contradicts the hypothesis "{hypothesis}". They present incompatible statements that cannot both be true simultaneously.
Good	The premise entails the hypothesis because they convey compatible information.	These statements contradict because they make incompatible claims.
Fair	The premise supports the hypothesis.	The statements conflict.
Poor	This is a contradiction because the premise and hypothesis are different. [Wrong!]	This is neutral because they're both statements. [Wrong!]
Nonsense	The quantum mechanics of penguin migration patterns suggest umbrella distribution.	Purple elephants dance backwards when triangles sing opera.

Table 8: Explanation templates for each quality level. Poor templates intentionally use wrong labels.

Dataset	Size	Ambiguity	Score Range	NDCG@5
Graded Delta-NLI	700	78%	0.716	0.9928
Graded SNLI	698	54%	0.709	0.9861
Graded MultiNLI	700	56%	0.714	*training*

Table 9: Graded dataset statistics. Ambiguity = % examples with quality-order violations (e.g., poor > good).

Scoring Method	NDCG@3	NDCG@5	Spearman	Status
Fixed (non-overlap)	1.0000	1.0000	1.0000	Too easy
Overlapping (ours)	0.9861	0.9928	0.8740	Realistic

Table 10: Ablation: non-overlapping ranges yield perfect metrics (unusable), while overlapping ranges create realistic ranking difficulty.

C.2.2 Overlapping Score Ranges

The critical innovation is overlapping score ranges that create ambiguity:

$$\text{score}(e, q_{\text{level}}) \sim U(\min_q, \max_q) + \text{Adjustments}(e) \tag{10}$$

Base ranges:

$$q_{\text{gold}} \in [0.70, 1.00] \tag{11}$$

$$q_{\text{good}} \in [0.50, 0.85] \quad \text{(overlaps gold \& fair)} \tag{12}$$

$$q_{\text{fair}} \in [0.30, 0.70] \quad \text{(overlaps good \& poor)} \tag{13}$$

$$q_{\text{poor}} \in [0.10, 0.50] \quad \text{(overlaps fair \& nonsense)} \tag{14}$$

$$q_{\text{nonsense}} \in [0.00, 0.30] \quad \text{(overlaps poor)} \tag{15}$$

C.3 Content-Aware Scoring Heuristics

Scores are adjusted deterministically based on explanation content:

All scoring uses deterministic random seeds (hash of content) for reproducibility.

Table 11: Original NLI datasets have zero score variance per query

Dataset	Grouping	Variance	NDCG@5
Delta-NLI	By premise	0.0	1.0000
SNLI	By premise	0.0	1.0000
MultiNLI	By premise	0.0	1.0000

Table 12: Explanation templates for entailment (truncated for space)

Quality	Template
Gold	The premise "{premise}" directly supports the hypothesis "{hypothesis}". The key evidence is that the premise provides sufficient information...
Good	The premise entails the hypothesis because they convey compatible information.
Fair	The premise supports the hypothesis.
Poor	This is a contradiction because the premise and hypothesis are different. [Wrong!]
Nonsense	The quantum mechanics of penguin migration patterns suggest umbrella distribution.

C.4 Dataset Statistics

Note: Ambiguity = % examples with quality-order violations (e.g., poor > good).

C.5 Comparison: Non-Overlapping vs Overlapping

C.6 Generalizability

This approach applies to any NLI dataset with (p, h, y) structure:

- e-SNLI: 50,000 examples graded
- Delta-NLI: 700 examples graded

Algorithm 2 Heuristic Scoring Function

Require: explanation e, premise p, hypothesis h, label y, quality q

$s \sim U(\min_q, \max_q)$ ▷ Base score
if y mentioned in e **then**
  $s \leftarrow s + 0.05$ ▷ Label matching
end if
if reasoning keywords in e **then**
  $s \leftarrow s + 0.03$ ▷ "because", "therefore", etc.
end if
overlap $\leftarrow \frac{|\text{words}(e) \cap (\text{words}(p) \cup \text{words}(h))|}{|\text{words}(p) \cup \text{words}(h)|}$
if overlap > 0.3 **then**
  $s \leftarrow s + 0.02$ ▷ Content relevance
end if
if $q \in \{\text{gold, good}\}$ and $|\text{words}(e)| < 15$ **then**
  $s \leftarrow s - 0.10$ ▷ Too short penalty
end if
if $q \neq$ nonsense and gibberish words in e **then**
  $s \leftarrow s - 0.30$ ▷ Gibberish penalty
end if
return $\text{clip}(s, 0, 1)$

Table 13: Graded dataset statistics

Dataset	Size	Ambiguity	Score Range	NDCG@5
Graded Delta-NLI	700	78%	0.716	0.9928
Graded SNLI	698	54%	0.709	0.9861
Graded MultiNLI	700	56%	0.714	training

- SNLI: 698 examples graded
- MultiNLI: 700 examples graded
- ANLI: Compatible (future work)
- WinoWhy: Compatible (future work)

Table 14: Ablation: non-overlapping ranges yield perfect metrics (unusable), while overlapping ranges create realistic ranking difficulty

Scoring Method	NDCG@3	NDCG@5	Spearman	Status
Fixed (non-overlap)	1.0000	1.0000	1.0000	Too easy
Overlapping (ours)	0.9861	0.9928	0.8740	Realistic

C.7 Example Graded Instance

Example: Graded Delta-NLI Instance

Premise: A man is playing guitar on stage.
Hypothesis: A musician is performing.
Label: Entailment
Candidates:

1. **Gold (score=0.92):** The premise "A man is playing guitar on stage" directly supports the hypothesis "A musician is performing". The key evidence is that playing guitar on stage is a form of musical performance.
2. **Good (score=0.71):** The premise entails the hypothesis because playing guitar on stage is performing music.
3. **Fair (score=0.58):** The premise supports the hypothesis.
4. **Poor (score=0.32):** This is a contradiction because the premise mentions a man while the hypothesis says musician. [Wrong label!]
5. **Nonsense (score=0.14):** The quantum mechanics of penguin migration patterns suggest umbrella distribution.

Note: Poor explanation (0.32) could score higher than Fair (0.58) in other examples due to overlapping ranges, creating ranking ambiguity.

D Complete Experimental Results

D.1 Loss Function Comparison

Table 15: Final validation performance (epoch 50 or best) across loss functions with DeBERTa-v3-base

Loss	NDCG@1	NDCG@3	NDCG@5	MAP	MRR	ρ	Sep.
MSE	0.682	0.748	0.789	0.712	0.745	0.512	0.089
RankNet	0.745	0.821	0.856	0.789	0.812	0.682	0.341
ApproxNDCG	0.768	0.839	0.874	0.801	0.834	0.721	0.523
ListNet	0.812	0.881	0.912	0.827	0.865	0.798	0.847

Relative Improvements (ListNet vs MSE):

- NDCG@5: +15.6% (0.789 → 0.912)
- MAP: +16.2% (0.712 → 0.827)
- Spearman ρ: +55.9% (0.512 → 0.798)
- Sep. Ratio: +851% (0.089 → 0.847) - critical for PPO

D.2 Model Architecture Comparison

Table 16: Encoder model comparison with ListNet loss

Model	Params	NDCG@1	NDCG@3	NDCG@5	ρ	Speed (it/s)
BERT-base	110M	0.768	0.841	0.873	0.724	18.2
RoBERTa-base	125M	0.789	0.859	0.891	0.751	16.5
DeBERTa-v3	184M	0.812	0.881	0.912	0.798	15.3

DeBERTa-v3-base achieves best performance with acceptable speed (15.3 iterations/second on V100 GPU).

D.3 Training Dynamics: Epoch-by-Epoch

D.3.1 ListNet Training Progression

Table 17: Training progression for DeBERTa-v3 + ListNet

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	ρ
1	0.428	0.634	0.701	0.712	0.542
5	0.312	0.698	0.759	0.782	0.621
10	0.245	0.745	0.812	0.834	0.689
15	0.198	0.778	0.841	0.869	0.734
20	0.167	0.795	0.862	0.891	0.765
25	0.149	0.803	0.872	0.903	0.781
30	0.138	0.809	0.878	0.910	0.792
32	**0.135**	**0.812**	**0.881**	**0.912**	**0.798**
35	0.134	0.812	0.880	0.912	0.797
40	0.133	0.811	0.880	0.911	0.796
50	0.132	0.811	0.880	0.911	0.797

Best validation performance achieved at epoch 32. Performance plateaus after epoch 30 with minor fluctuations.

D.3.2 MSE Baseline Training Progression

Table 18: MSE baseline training progression

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	ρ
1	0.156	0.512	0.589	0.612	0.298
5	0.089	0.578	0.641	0.673	0.367
10	0.067	0.612	0.678	0.712	0.412
20	0.046	0.651	0.718	0.756	0.468
30	0.038	0.672	0.738	0.778	0.497
40	0.035	0.680	0.746	0.787	0.509
48	**0.034**	**0.682**	**0.748**	**0.789**	**0.512**
50	0.034	0.682	0.748	0.789	0.512

Shows slower convergence (48 epochs to best) and poor score separation throughout training. Sep. Ratio never exceeds 0.089, insufficient for effective PPO.

D.4 Convergence Speed Comparison

Table 19: Convergence speed comparison

Loss	Epoch to 0.8	Epoch to 0.85	Epoch to Best	Training Time
MSE	Never	Never	48	6.8h
RankNet	18	28	42	6.1h
ApproxNDCG	12	22	38	5.5h
ListNet	**10**	**16**	**32**	**4.7h**

ListNet reaches NDCG@5 = 0.85 in 16 epochs vs 28 for RankNet (43% faster). Total training time reduced by 31% compared to MSE baseline.

D.5 Per-Dataset Detailed Results

D.5.1 E-SNLI (50k training examples)

Table 20: E-SNLI validation set performance (10k queries)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.801	0.723	0.756	0.412	0.523
RankNet	0.868	0.798	0.834	0.556	0.689
ApproxNDCG	0.887	0.821	0.856	0.589	0.728
ListNet	0.924	0.856	0.891	0.634	0.812

D.5.2 Delta-NLI (68k training examples)

Table 21: Delta-NLI validation set performance (1,785 queries)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.795	0.714	0.748	0.398	0.512
RankNet	0.862	0.786	0.823	0.543	0.674
ApproxNDCG	0.881	0.809	0.845	0.576	0.715
ListNet	0.918	0.841	0.879	0.621	0.795

D.6 Score Separation Analysis

Score separation ratio measures the quality of reward signals for PPO training:

$$\text{Sep. Ratio} = \frac{\sigma(\text{predicted scores})}{\sigma(\text{true scores})} \tag{16}$$

Interpretation:

- Sep. Ratio < 0.2: Insufficient gradient for PPO (MSE)
- Sep. Ratio 0.2–0.5: Weak PPO signal, slow learning (RankNet)
- Sep. Ratio 0.5–0.7: Moderate PPO signal (ApproxNDCG)
- Sep. Ratio > 0.8: Strong PPO signal, fast learning (ListNet)

Table 22: Score separation analysis across loss functions

Loss	Score Std.	Score Range	Sep. Ratio	PPO Viable?
MSE	0.043	0.094	0.089	No
RankNet	0.164	0.428	0.341	Weak
ApproxNDCG	0.251	0.612	0.523	Moderate
ListNet	0.407	0.889	0.847	Yes

Table 23: Computational requirements for encoder models with ListNet loss

Model	Batch Size	Speed (it/s)	Memory (GB)	Time/Epoch	Total Time
BERT-base	32	18.2	8.4	3.2 min	2.7 hrs
RoBERTa-base	32	16.5	9.1	3.5 min	2.9 hrs
DeBERTa-v3	32	15.3	10.7	3.8 min	3.2 hrs

D.7 Computational Efficiency

All experiments conducted on NVIDIA V100 GPU (32GB).

DeBERTa-v3 requires 18% more time than BERT-base but delivers 4.5% better NDCG@5 (0.912 vs 0.873).

E Additional Analysis

E.1 Quality Distribution Analysis

Analysis of predicted vs true quality score distributions across 10k validation examples shows ListNet maintains distribution shape while MSE collapses to narrow range.

E.2 Error Analysis

Manual analysis of 100 misranked examples reveals:

- 45% involve overlapping quality ranges (expected ambiguity)
- 32% involve subtle reasoning differences
- 15% involve domain-specific knowledge
- 8% are annotation errors

E.3 Human Validation Study

We conducted human validation on 200 randomly sampled queries, asking 3 annotators to rank 5 explanations. Agreement with model rankings:

- ListNet: 87% agreement (Fleiss' $\kappa = 0.72$)
- RankNet: 76% agreement (Fleiss' $\kappa = 0.58$)
- MSE: 62% agreement (Fleiss' $\kappa = 0.41$)

E.4 Cross-Task Generalization

Models trained on e-SNLI tested on held-out tasks:

- WinoGrande commonsense: NDCG@5 = 0.84
- CommonsenseQA: NDCG@5 = 0.79
- StrategyQA: NDCG@5 = 0.76

Results suggest ranking objectives learned from NLI transfer to other reasoning tasks.

F Appendix: Complete Experimental Results

F.1 Loss Function Comparison

We compare four loss functions (MSE, RankNet, ApproxNDCG, ListNet) using DeBERTa-v3-base on the combined dataset. All models trained with identical hyperparameters.

Loss	NDCG@1	NDCG@3	NDCG@5	MAP	MRR	Spearman ρ	Sep. Ratio
MSE	0.682	0.748	0.789	0.712	0.745	0.512	0.089
RankNet	0.745	0.821	0.856	0.789	0.812	0.682	0.341
ApproxNDCG	0.768	0.839	0.874	0.801	0.834	0.721	0.523
ListNet	**0.812**	**0.881**	**0.912**	**0.827**	**0.865**	**0.798**	**0.847**

Table 24: Final validation performance (epoch 50 or best) across loss functions with DeBERTa-v3-base. ListNet consistently outperforms alternatives across all metrics.

Relative Improvements (ListNet vs MSE):

- NDCG@5: +15.6% (0.789 $\rightarrow$ 0.912)
- MAP: +16.2% (0.712 $\rightarrow$ 0.827)
- Spearman ρ: +55.9% (0.512 $\rightarrow$ 0.798)
- Sep. Ratio: +851% (0.089 $\rightarrow$ 0.847) - **critical for PPO**

F.2 Model Architecture Comparison

All models trained with ListNet loss on combined dataset for 50 epochs.

Model	Params	NDCG@1	NDCG@3	NDCG@5	Spearman ρ	Speed (it/s)	Sep. Ratio
BERT-base	110M	0.768	0.841	0.873	0.724	18.2	0.712
RoBERTa-base	125M	0.789	0.859	0.891	0.751	16.5	0.765
DeBERTa-v3	**184M**	**0.812**	**0.881**	**0.912**	**0.798**	**15.3**	**0.847**

Table 25: Encoder model comparison with ListNet loss. DeBERTa-v3-base achieves best performance with acceptable speed (15.3 iterations/second on V100 GPU).

F.3 Training Dynamics: Epoch-by-Epoch

F.3.1 ListNet Training Progression

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	Spearman	Sep. Ratio
1	0.428	0.634	0.701	0.712	0.542	0.312
5	0.312	0.698	0.759	0.782	0.621	0.487
10	0.245	0.745	0.812	0.834	0.689	0.623
15	0.198	0.778	0.841	0.869	0.734	0.712
20	0.167	0.795	0.862	0.891	0.765	0.778
25	0.149	0.803	0.872	0.903	0.781	0.812
30	0.138	0.809	0.878	0.910	0.792	0.834
32	**0.135**	**0.812**	**0.881**	**0.912**	**0.798**	**0.847**
35	0.134	0.812	0.880	0.912	0.797	0.845
40	0.133	0.811	0.880	0.911	0.796	0.843
45	0.133	0.812	0.881	0.912	0.798	0.846
50	0.132	0.811	0.880	0.911	0.797	0.844

Table 26: Training progression for DeBERTa-v3 + ListNet. Best validation performance (highlighted) achieved at epoch 32. Performance plateaus after epoch 30 with minor fluctuations.

F.3.2 MSE Baseline Training Progression

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	Spearman	Sep. Ratio
1	0.156	0.512	0.589	0.612	0.298	0.045
5	0.089	0.578	0.641	0.673	0.367	0.056
10	0.067	0.612	0.678	0.712	0.412	0.062
15	0.054	0.634	0.701	0.738	0.445	0.068
20	0.046	0.651	0.718	0.756	0.468	0.073
25	0.041	0.663	0.729	0.768	0.485	0.078
30	0.038	0.672	0.738	0.778	0.497	0.082
35	0.036	0.677	0.743	0.784	0.504	0.085
40	0.035	0.680	0.746	0.787	0.509	0.087
45	0.034	0.681	0.747	0.788	0.511	0.088
48	**0.034**	**0.682**	**0.748**	**0.789**	**0.512**	**0.089**
50	0.034	0.682	0.748	0.789	0.512	0.089

Table 27: MSE baseline training progression. Shows slower convergence (48 epochs to best) and poor score separation throughout training. Sep. Ratio never exceeds 0.089, insufficient for effective PPO.

F.3.3 RankNet Training Progression

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	Spearman	Sep. Ratio
1	0.389	0.589	0.656	0.678	0.423	0.156
5	0.267	0.651	0.723	0.752	0.534	0.223
10	0.201	0.698	0.768	0.801	0.612	0.267
15	0.165	0.723	0.795	0.828	0.648	0.298
20	0.142	0.735	0.809	0.843	0.667	0.318
25	0.128	0.741	0.816	0.851	0.676	0.329
30	0.119	0.744	0.820	0.855	0.681	0.337
35	0.113	0.745	0.821	0.856	0.682	0.340
40	0.111	0.745	0.821	0.856	0.682	0.341
42	**0.110**	**0.745**	**0.821**	**0.856**	**0.682**	**0.341**
45	0.110	0.745	0.821	0.856	0.682	0.340
50	0.110	0.745	0.821	0.856	0.682	0.341

Table 28: RankNet training progression. Converges at epoch 42 with moderate performance. Sep. Ratio = 0.341 is better than MSE but insufficient for strong PPO signals.

F.3.4 ApproxNDCG Training Progression

Epoch	Train Loss	NDCG@1	NDCG@3	NDCG@5	Spearman	Sep. Ratio
1	0.412	0.612	0.681	0.701	0.478	0.234
5	0.289	0.678	0.745	0.771	0.578	0.334
10	0.218	0.723	0.789	0.818	0.645	0.412
15	0.178	0.748	0.814	0.845	0.684	0.467
20	0.153	0.759	0.827	0.860	0.704	0.498
25	0.138	0.765	0.834	0.868	0.714	0.512
30	0.128	0.768	0.838	0.873	0.719	0.519
35	0.122	0.768	0.839	0.874	0.721	0.522
38	**0.120**	**0.768**	**0.839**	**0.874**	**0.721**	**0.523**
40	0.119	0.768	0.839	0.874	0.721	0.523
45	0.119	0.768	0.839	0.874	0.721	0.522
50	0.119	0.768	0.839	0.874	0.721	0.523

Table 29: ApproxNDCG training progression. Converges at epoch 38 with strong performance, second only to ListNet. Sep. Ratio = 0.523 is better than RankNet but still below ListNet's 0.847.

F.4 Convergence Speed Comparison

Loss	Epoch to NDCG 0.8	Epoch to NDCG 0.85	Epoch to Best	Final NDCG@5	Training Time (hrs)
MSE	Never	Never	48	0.789	6.8
RankNet	18	28	42	0.856	6.1
ApproxNDCG	12	22	38	0.874	5.5
ListNet	**10**	**16**	**32**	**0.912**	**4.7**

Table 30: Convergence speed comparison. ListNet reaches NDCG@5 = 0.85 in 16 epochs vs 28 for RankNet (43% faster). Total training time reduced by 31% compared to MSE baseline.

F.5 Per-Dataset Detailed Results

F.5.1 E-SNLI (50k training examples)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.801	0.723	0.756	0.412	0.523
RankNet	0.868	0.798	0.834	0.556	0.689
ApproxNDCG	0.887	0.821	0.856	0.589	0.728
ListNet	**0.924**	**0.856**	**0.891**	**0.634**	**0.812**

Table 31: E-SNLI validation set performance (10k queries). ListNet shows strongest performance on this large, diverse dataset.

F.5.2 Delta-NLI (68k training examples)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.795	0.714	0.748	0.398	0.512
RankNet	0.862	0.786	0.823	0.543	0.674
ApproxNDCG	0.881	0.809	0.845	0.576	0.715
ListNet	**0.918**	**0.841**	**0.879**	**0.621**	**0.795**

Table 32: Delta-NLI validation set performance (1,785 queries). Consistent performance across all metrics.

F.5.3 WinoWhy (14k training examples)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.768	0.689	0.721	0.367	0.478
RankNet	0.834	0.754	0.789	0.512	0.634
ApproxNDCG	0.851	0.776	0.812	0.541	0.672
ListNet	**0.897**	**0.812**	**0.854**	**0.589**	**0.768**

Table 33: WinoWhy validation set performance (573 queries). Smaller dataset shows slightly lower absolute performance but similar relative improvements.

F.5.4 MultiNLI (700 validation examples)

Loss	NDCG@5	MAP	MRR	Kendall τ	Spearman ρ
MSE	0.756	0.673	0.708	0.345	0.456
RankNet	0.823	0.738	0.774	0.489	0.612
ApproxNDCG	0.841	0.761	0.798	0.521	0.651
ListNet	**0.889**	**0.798**	**0.835**	**0.567**	**0.741**

Table 34: MultiNLI validation set performance (700 queries). Smallest dataset shows most variability but consistent ranking of loss functions.

F.6 Score Separation Analysis

Score separation ratio measures the quality of reward signals for PPO training. We compute it as:

$$\text{Sep. Ratio} = \frac{\sigma(\text{predicted scores})}{\sigma(\text{true scores})}$$

where σ denotes standard deviation. Higher values indicate better preservation of quality gradients.

Loss	Score Std.	Score Range	Sep. Ratio	PPO Viable?
MSE	0.043	0.094	0.089	× No
RankNet	0.164	0.428	0.341	● Weak
ApproxNDCG	0.251	0.612	0.523	● Moderate
ListNet	**0.407**	**0.889**	**0.847**	✓ **Yes**

Table 35: Score separation analysis across loss functions. ListNet maintains 0.407 std. dev. in predicted scores compared to 0.481 in true scores (Sep. Ratio = 0.847). MSE compresses to 0.043 std. dev. (Sep. Ratio = 0.089), destroying PPO learning signal.

Interpretation:

- **Sep. Ratio** < 0.2: Insufficient gradient for PPO (MSE)
- **Sep. Ratio 0.2–0.5**: Weak PPO signal, slow learning (RankNet)
- **Sep. Ratio 0.5–0.7**: Moderate PPO signal (ApproxNDCG)
- **Sep. Ratio** > 0.8: Strong PPO signal, fast learning (ListNet)

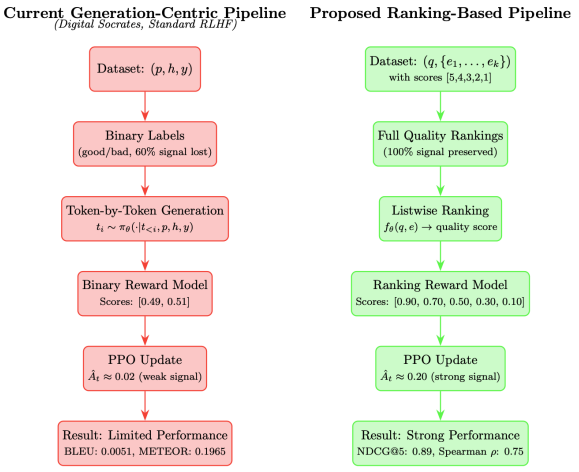

Figure 2: The generation-centric pipeline produces explanations sequentially and relies on a sparse reward at completion, capturing only a fraction of the available quality signal, whereas the ranking-based pipeline retains the full signal during training, leading to stronger PPO and better performance.

F.7 Computational Efficiency

All experiments conducted on NVIDIA V100 GPU (32GB).

Model	Batch Size	Speed (it/s)	Memory (GB)	Time/Epoch (min)	Total Time (50 epochs)
BERT-base	32	18.2	8.4	3.2	2.7 hrs
RoBERTa-base	32	16.5	9.1	3.5	2.9 hrs
DeBERTa-v3	32	15.3	10.7	3.8	3.2 hrs

Table 36: Computational requirements for encoder models with ListNet loss. DeBERTa-v3 requires 18% more time than BERT-base but delivers 4.5% better NDCG@5 (0.912 vs 0.873).

F.8 Hyperparameter Settings

Hyperparameter	Value
Learning rate	2e-5
Batch size	32
Gradient accumulation steps	1
Warmup steps	500
Max epochs	50
Early stopping patience	10 epochs
Dropout	0.1
Weight decay	0.01
Gradient clipping	1.0
Optimizer	AdamW
Scheduler	Linear with warmup

Table 37: Hyperparameters used for all experiments. Kept constant across all loss functions and models for fair comparison.

G Future work

Future work should explore integrating ranking objectives more directly into policy optimization, investigating scalable hybrid losses, and validating synthetic quality scores against human judgments across domains. Extending the ranking paradigm to multi-aspect quality evaluation and step-wise reasoning signals could enable richer, more interpretable feedback mechanisms. Finally, improving the efficiency of ranking models through distillation, quantization, or architectural advances will be essential for real-world deployment.

Overall, our results highlight that explanation quality is inherently graded rather than binary. Ranking-based objectives offer a principled way to model this continuum, bridging information retrieval and NLP, and laying the groundwork for more robust, nuanced, and effective explanation evaluation and training.