NEAT-IR: Neural Explainable Analysis Tool for Information Retrieval

Lev Sukherman
WPI
lsukherman@wpi.edu

Artem Frenk
WPI
afrenk@wpi.edu

Nina Klimenkova
WPI
nklimenkova@wpi.edu

Connor Jason
WPI
cjason@wpi.edu

Abstract

Neural IR models achieve strong performance but remain difficult to interpret. We present NEAT-IR, a black-box analysis framework that explains ColBERT's ranking behavior using 26 classical IR features (BM25, TF-IDF, IDF measures, positional signals). We analyze ColBERT through two complementary lenses: regression (predicting exact scores) and learning-to-rank (predicting relative order), evaluated on MS MARCO (48,250 query–passage pairs). Our key finding is a **score–rank gap**: classical features preserve ColBERT's rankings nearly perfectly ($NDCG@5 \approx 0.99$) yet explain only $R^2 \approx 0.28$ of score variance. Feature attribution reveals that regression and ranking models rely on distinct feature subsets: query-level IDF signals dominate score prediction, while document-matching features (BM25, cosine TF-IDF) drive ranking preservation. These findings suggest that ColBERT's ordinal behavior on MS MARCO is largely recoverable from classical signals, while neural contributions primarily affect score magnitude. NEAT-IR enables practitioners to diagnose when neural rankers deviate from classical patterns, supporting interpretable model auditing and informed hybrid pipeline design.

1 Introduction

Neural information retrieval models like ColBERT (Khattab and Zaharia, 2020) achieve state-of-the-art performance but remain difficult to interpret. These models combine lexical matching with semantic understanding learned from large corpora through attention mechanisms and contextualized embeddings (Devlin et al., 2019), yet it is unclear how much each contributes to final rankings. Recent explainability work (Anand et al., 2022; Fernando et al., 2019; Lundberg and Lee, 2017) focuses on model-internal signals such as attention weights and SHAP values, showing *how* models compute but not *whether* they follow interpretable IR principles. Practitioners need rigorous evaluation of explanation quality (Doshi-Velez and Kim, 2017) and methods grounded in domain knowledge.

We present NEAT-IR, defining explainability as *behavioral fidelity*: the degree to which a model's input–output mapping can be reproduced by interpretable, domain-native features (BM25, TF-IDF, positional signals), without claims about internal mechanisms. Our hypothesis: if ColBERT captures semantic matching beyond lexical overlap, classical features should poorly approximate its behavior; high approximation would suggest neural models remain grounded in classical IR principles (Robertson and Zaragoza, 2009). Unlike prior work that analyzes model internals (Formal et al., 2021), we treat ColBERT as a black box and quantify, for the first time, the score–rank gap: the divergence between score prediction and ranking preservation.

We analyze ColBERT through two complementary lenses: **regression** (predicting exact scores) and **ranking** (predicting relative order), designed to reveal different aspects of its behavior in sequence. Classical features preserve rankings nearly perfectly yet predict less than a third of score variance, showing that they capture ordinal structure but not score magnitudes. Feature attribution via intrinsic importance (global) and SHAP (local) explains why the two tasks diverge: the score–rank gap only becomes visible when cardinal and ordinal evaluations are placed side by side.

2 Related Work

Explainable Neural Ranking. Explainability methods for neural IR range from token-level saliency to model-agnostic surrogates. Prediction decomposition via DeepSHAP has been applied to neural rankers (Fernando et al., 2019), and local approximation via LIME offers model-agnostic explanations (Ribeiro et al., 2016). However, both methods operate on raw input features rather than familiar IR concepts such as BM25 or query coverage (Anand et al., 2023).

Classical vs. Neural IR. The transition from classical probabilistic models (Robertson and Zaragoza, 2009) to neural architectures (Vaswani et al., 2017) raised questions about what neural models learn beyond term matching. Dense retrievers achieve strong performance through learned semantic representations (Karpukhin et al., 2020), yet neural rerankers often rely on simple lexical patterns (Lin et al., 2021). ColBERT in particular has been shown to largely rediscover lexical term importance on in-domain data while struggling to generalize out-of-domain (Formal et al., 2022). Our work quantifies this relationship by predicting ColBERT's scores from classical features alone.

Score Semantics and Explainability Metrics. Neural rankers have been shown to optimize relative document orderings rather than calibrated absolute scores (Yu et al., 2025), a concern directly related to our score–rank gap. The neural contribution has also been framed as a residual correction atop BM25 (Gao et al., 2021), suggesting that classical signals form a strong baseline that neural models refine rather than replace. Complementary evaluation frameworks propose intrinsic and extrinsic explainability metrics for neural rankers, demonstrating that the most relevant models are not necessarily the most explainable (Pandian et al., 2024).

Feature-Based Analysis. Neural models have been shown to replicate classical patterns in vision (Zhou et al., 2015) and NLP (Rogers et al., 2020; Tenney et al., 2019), but no prior work examines this for IR-specific features such as BM25 or IDF aggregates.

3 Method

3.1 Feature Engineering

We extract 26 features in four categories: lexical matching (term/n-gram overlaps), statistical scoring (BM25, TF-IDF, IDF-based weighting), positional signals (where query terms appear in document), and document properties (length, vocabulary diversity). Table 1 summarizes these categories.

Category	**Count**	**Examples**
Statistical Scoring	11	BM25, TF-IDF, IDF
Lexical Matching	9	N-grams, coverage
Positional Signals	5	Position, span
Document Properties	4	Length, diversity

Table 1: Feature categories.

We select features using the elbow method (Figure 1) with a 1% marginal improvement threshold. At 10 features, marginal improvement falls below this threshold ($\Delta R^2 = 0.0013$). This configuration achieves $R^2 = 0.273$, comparable to the full 26-feature model while reducing dimensionality by 62%.

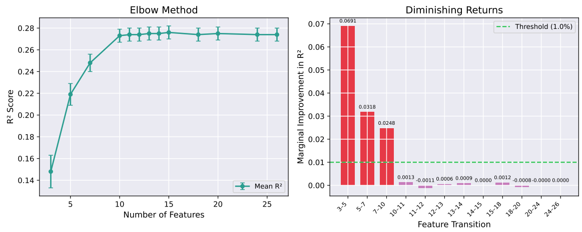

Figure 1: Elbow method for feature selection: R^2 performance (left) and marginal improvement (right).

3.2 Models and Evaluation

We train two types of models to analyze ColBERT from different perspectives:

Regression models predict exact ColBERT scores, measuring how well classical features approximate neural scoring. We train both Random Forest and XGBoost regressors; both achieve similar R^2 (Table 2).

Model	R^2	**MSE**
Random Forest	0.278 ± 0.006	6.223
XGBoost Reg	0.262 ± 0.007	6.363

Table 2: Classical features explain 28% of ColBERT's scoring variance.

Ranking model predicts relative document order within each query, measuring whether classical features preserve ColBERT's ranking decisions. We use XGBoost Ranker optimizing NDCG@5.

Figure 2 shows high correlation between regression models (Pearson: 0.962) and moderate correlation with the Ranker (Pearson: 0.644). We use Random Forest for feature importance analysis due to its interpretability.

All models use 5-fold GroupKFold cross-validation grouped by query_id. NDCG@5 remains stable across hyperparameter configurations (range: 0.001) while R^2 shows moderate sensitivity (range: 0.06). We therefore treat ranking preserva-

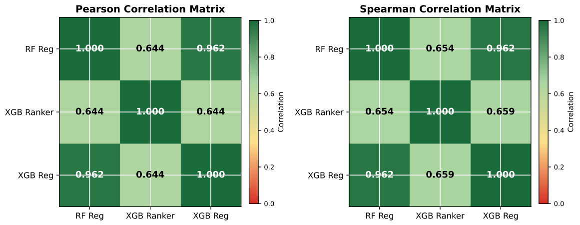

Figure 2: Inter-model correlation matrices (Pearson and Spearman).

tion as the primary robust finding and score prediction as an approximate diagnostic.

3.3 Evaluation Metrics

Score Approximation: R^2 measures how well classical features predict ColBERT scores. We note that our target is ColBERT's *behavior* (predicted scores), not human relevance judgments; NEAT-IR explains the model, not relevance itself.

Ranking Preservation: We use three metrics: (1) Kendall's τ for rank correlation; (2) Spearman's ρ for monotonic relationships; (3) NDCG@5 for weighted agreement at top-5 positions. High NDCG@5 with moderate R^2 would indicate that classical features capture ordinal but not cardinal properties of ColBERT scores.

Feature Importance: We analyze at two levels: **global** importance via Random Forest impurity decrease, and **local** explanations via per-prediction SHAP values.

4 Experiments

4.1 Setup

Dataset: MS MARCO v1.1 (Nguyen et al., 2016), 9,650 queries $\times$ 5 passages = 48,250 pairs.

Neural Model: ColBERT v2.0 (Khattab and Zaharia, 2020) scores serve as prediction targets. On the MS MARCO dev set, ColBERTv2 achieves MRR@10 = 0.397 and Recall@1000 = 0.984 (Santhanam et al., 2022), establishing it as a strong neural baseline whose ranking behavior is worth explaining.

Implementation: NLTK for tokenization, scikit-learn for TF-IDF and Random Forest, XGBoost for ranking.

4.2 Results

Score Prediction. Both models achieve $R^2 \approx 0.28$ and MSE $\approx$ 6.2 (Table 2), with similar correlation to the Ranker as shown in Figure 2.

Ranking Preservation. Table 3 shows ranking agreement between model predictions and ColBERT scores.

Model	τ	ρ	NDCG@5
XGB Ranker	0.38	0.44	0.9914

Table 3: Per-query ranking agreement with ColBERT.

Per-query correlations are moderate (τ = 0.38, ρ = 0.44) with high variance (std $\approx$ 0.4), yet XGBoost Ranker achieves NDCG@5 ≈ 0.99. This is because τ and ρ penalize disagreement at all ranks equally, while NDCG@5 focuses on top positions where users look first.

We call this the **score–rank gap**: moderate score prediction ($R^2 \approx 0.28$) coexists with near-perfect ranking preservation ($NDCG@5 \approx 0.99$). Classical features predict *which* documents rank higher, but not *how much higher* they should score.

Feature Importance. Figure 3 compares intrinsic importance between regression and ranker models.

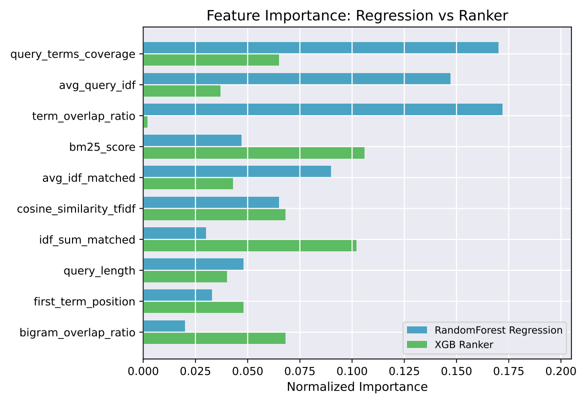

Figure 3: Intrinsic feature importance: regression (blue) vs. ranker (green).

Intrinsic importance: Regression prioritizes lexical coverage features (query_terms_coverage: 0.17, term_overlap_ratio: 0.17) and query-level IDF (avg_query_idf: 0.14). The ranker instead emphasizes statistical scoring (bm25_score: 0.10, idf_sum_matched: 0.08).

SHAP (local): We compute SHAP values for both models to explain per-prediction behavior. As shown in Figure 4, the regression model is dominated by query-level features (`avg_query_idf`: 0.45, `query_length`: 0.45), reflecting that score magnitude is largely determined by query characteristics rather than document content. In contrast, Figure 5 shows that the ranker relies primarily on document-matching features (`cosine_similarity_tfidf`: 0.09, `idf_sum_matched`: 0.07), which capture how well a document aligns with the query terms.

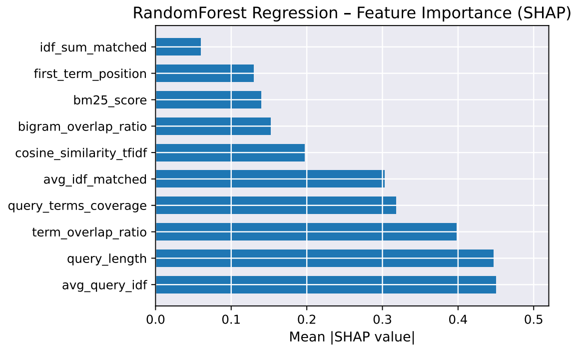

Figure 4: SHAP feature importance for the regression model (RandomForest).

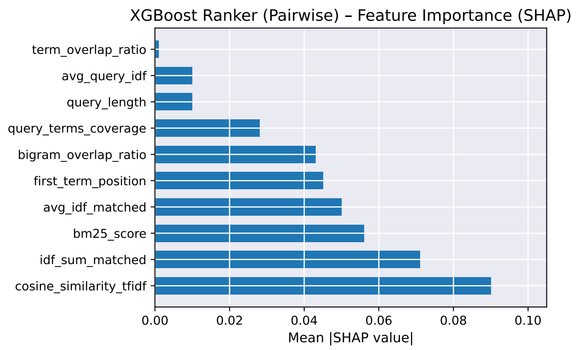

Figure 5: SHAP feature importance for the ranking model (XGBoost Ranker).

5 Discussion

5.1 Key Findings

Classical Feature Dominance. As shown in Section 4.2, classical IR features dominate both global and local importance measures across regression and ranking models. Despite using dense embeddings without explicit term weighting, ColBERT's output behavior on MS MARCO closely mirrors classical scoring patterns (Robertson and Zaragoza, 2009).

Query-Level IDF Weighting. The feature avg_query_idf shows the highest regression SHAP (0.45), tied with query_length. Query-level IDF aggregation is not a standard scoring signal; BM25 applies IDF per-term, not as a query summary. This suggests ColBERT adjusts scores based on query specificity: queries with rare terms receive higher scores.

Task-Dependent Features. The regression–ranking feature divergence (Section 4.2) provides a mechanistic account of the score–rank gap: query-level features set the score scale per query, while document-matching features determine relative order within it. Together, our three evaluation axes (fidelity R^2, ranking preservation NDCG@5, and feature attribution SHAP) follow best practices for interpretability evaluation (Doshi-Velez and Kim, 2017) and jointly characterize this gap. For context, a random ranker achieves $NDCG@5 \approx 0.59$ on this 5-candidate setup (assuming one relevant passage per query), confirming that the near-perfect ranking preservation reflects genuine classical–neural alignment rather than an artifact of the constrained pool.

5.2 Practical Implications

On the evaluated dataset, classical features achieve near-perfect top-5 ranking preservation (NDCG@5 $\approx$ 0.99) at lower computational cost. The divergence between NDCG@5 and per-query rank correlations ($\tau = 0.38$) is itself an instance of the score–rank gap: classical features reliably recover the relative order at positions that matter most to users, while exact pairwise orderings deeper in the candidate list are less consistent. If this pattern generalizes, it would support hybrid pipelines where classical features handle candidate selection and neural models add value through final reranking.

6 Conclusion

We introduced NEAT-IR, a framework that quantifies the classical-neural boundary in modern IR. By analyzing ColBERT through both regression (score prediction) and ranking (order preservation), we reveal the score–rank gap: classical features achieve 28% score explanation but 99% ranking preservation.

On MS MARCO, our analysis suggests that: (1) ColBERT's output behavior closely mirrors classical scoring patterns (BM25, TF-IDF), though this may partly reflect the dataset's lexical bias; (2) query difficulty adjustment represents a candidate neural contribution beyond classical features; (3) regression and ranking models rely on different feature subsets, explaining the score–rank gap.

NEAT-IR enables interpretable debugging using domain vocabulary. Practitioners can identify when neural models deviate from classical patterns, signaling semantic contribution versus lexical matching. The 99% ranking preservation supports hybrid architectures: classical features for candidate selection, neural models for reranking.

7 Future Work

Four directions follow naturally from our findings. First, the score–rank gap should be tested on semantically demanding benchmarks (BEIR, TREC-DL) and larger candidate sets (top-100) to determine whether recoverability by classical features reflects dataset bias or a more general property of ColBERT. Second, per-query failure analysis at deeper cutoffs (top-20, top-100) and white-box inspection of internal representations would provide a more complete picture of where and why ColBERT goes beyond classical ranking patterns. Third, query-level recoverability signals (IDF statistics, query length, lexical ambiguity) could help predict which queries benefit from neural models and which are handled well by classical features alone. Finally, the high ranking preservation we observe suggests exploring whether ColBERT can be distilled into a lightweight classical-feature ranker for factoid-heavy deployment, and whether such a surrogate degrades on the semantically demanding settings where neural models are most advantageous.

Limitations

Dataset: MS MARCO is dominated by factoid, short-answer questions where lexical matching is inherently effective (Thakur et al., 2021). Our near-perfect ranking preservation likely reflects this lexical bias: on semantically demanding benchmarks such as BEIR (Thakur et al., 2021) or TREC-DL (Craswell et al., 2020), where queries require multi-hop reasoning, paraphrase understanding, or entity resolution, we would expect classical recoverability to decrease substantially.

Candidate Set Size: With only 5 passages per query, NDCG@5 evaluates the entire candidate list, which may inflate ranking preservation relative to operational top-100 reranking settings where finer-grained ranking distinctions become necessary.

Architecture: We analyze only ColBERT; dense retrievers (DPR) or cross-encoders may exhibit different classical–neural boundaries.

Feature Space: Our 26 features may not capture all signals available to ColBERT. Low R^2 could partly reflect feature-set insufficiency rather than genuine semantic reasoning.

Explanation vs. Justification: High feature–output correlation does not imply that ColBERT internally computes classical signals. NEAT-IR is best used as a *diagnostic* tool, as its explanations may reflect model outputs rather than model reasoning.

Acknowledgments

This work originated as a group project for the Information Retrieval course taught by Prof. Kyumin Lee at Worcester Polytechnic Institute. We thank Prof. Lee for his instructions and feedback on this project, and the WPI Computer Science Department for computational resources.

References

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*.

Avishek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. 2023. Explainable information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3448–3451.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Proceedings of the 28th Text REtrieval Conference (TREC 2019)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the interpretability of neural retrieval models using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1005–1008.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A white box analysis of ColBERT. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR)*, pages 257–263.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2022. Match your words! a study of lexical matching in neural information retrieval. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*, pages 120–127.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR)*, pages 146–160.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Saran Pandian, Debasis Ganguly, and Sean MacAvaney. 2024. Evaluating the explainability of neural rankers. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR)*, pages 369–383.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the NeurIPS 2021 Datasets and Benchmarks Track*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Puxuan Yu, Daniel Cohen, Hemank Lamba, Joel R. Tetreault, and Alejandro Jaimes. 2025. Explain then rank: Scale calibration of neural rankers using natural language explanations from LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22716–22730.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2015. Object detectors emerge in deep scene CNNs. In *International Conference on Learning Representations*.