

D2.2 - Report on Final Upstream IR Systems

Project: NEREO - Neural Information Retrieval and NLP Systems

Grant Agreement: PRIN 2022

Deliverable ID: D2.2

Work Package: WP2 - Upstream IR Systems

Due Date: M24

Lead Beneficiary: UNIPi

1. Executive Summary

This deliverable constitutes the final report for **Work Package 2 (Upstream IR Systems)**. Building on the robust foundations laid in Year 1, the research activities in Year 2 shifted focus towards **efficiency** and **interpretability**. As cascading systems become more complex, the "Upstream" component must be able to retrieve information rapidly (low latency) and transparently. We successfully delivered **E2Rank**, a layer-wise efficient reranker, and **Eclipse**, a novel method for interpreting dense embeddings. These technologies ensure that the upstream system is scalable and its decisions are understandable, fulfilling the final objectives of the NEREO project.

2. Detailed Research Activities

2.1 E2Rank: Efficient Layer-Wise Reranking (Task 2.1)

Related Publications: ECIR 2025

cascading systems often suffer from high latency because the retrieved documents must be processed by heavy downstream models. Therefore, the upstream re-ranking stage must be as fast as possible.

- **Mechanism:** In "*E2Rank: Efficient and Effective Layer-Wise Reranking*", we proposed a novel **Early Exiting** strategy for Transformer-based rerankers (like BERT). E2Rank dynamically decides when to stop processing a document. If a document is clearly relevant (or clearly irrelevant) after just 3 layers, the model "exits" and outputs a score, saving the computation of the remaining 9 layers.
- **Performance:** Our experiments show that E2Rank can reduce inference time by **50-60%** with negligible loss in ranking accuracy (NDCG). This efficiency is critical for **Objective O3** (End-to-End Optimization), enabling real-time neural databases.

2.2 Eclipse: Interpretable Dense Retrieval (Task 2.2)

Related Publications: ICTIR 2025 ("Eclipse")

Dense retrieval (embedding-based search) is powerful but opaque. It is often unclear *which* query terms matched *which* document features.

- **Dimension Importance:** In "*Eclipse: Contrastive Dimension Importance Estimation with Pseudo-Irrelevance Feedback*", we developed a method to analyze the semantic embeddings produced by models like DPR or ANCE. Eclipse identifies which specific dimensions of the 768-dimensional vector are responsible for the relevance score.
- **Pseudo-Irrelevance Feedback:** By contrasting relevant documents with "pseudo-irrelevant" ones (hard negatives), Eclipse highlights the dimensions that encode discriminatory features. This helps in filtering out "negatively relevant" documents (matching on the wrong dimensions), directly addressing **Limitation L2** and **L3**.

2.3 Advanced Model Merging (Task 2.1)

Related Publications: CVPR 2025 ("Task Singular Vectors")

To create a versatile upstream system that handles multiple domains (e.g., News, Medical, Scientific), we explored **Model Merging** techniques.

- **Task Singular Vectors (TSV):** We introduced TSV to merge independently trained models without "task interference." By orthogonalizing the parameter updates associated with different tasks, we created a unified Upstream Encoder that performs well across diverse distributions without catastrophic forgetting. This supports the **Generalizability** of the NEREO platform.

3. Impact on NEREO Objectives

1. **Objective O3 (Efficiency & Optimization):** E2Rank directly enables the practical deployment of cascading systems. Without efficient upstream retrieval, the cost of the downstream LLM would make the overall system prohibitively slow.
2. **Objective O1 (Evaluation & Understanding):** Eclipse provides a new tool for evaluating *why* retrieval works (or fails). It moves evaluation from a black-box score to a dimension-wise analysis, allowing for finer-grained optimization.
3. **Objective O2 (Negative Relevance):** By identifying the dimensions that cause false positives (via Eclipse), we can explicitly dampen their influence, reducing the retrieval of "damaging" documents.

4. Conclusion

WP2 has delivered a comprehensive suite of Upstream IR technologies. We have moved from **Robustness** (Y1) to **Efficiency** (Y2) and **Interpretability** (Y2). The final "NEREO Upstream System" is a high-performance, explainable neural retriever that provides the optimal starting point for any cascading AI application.

5. Scientific References

2025

- **E2Rank: Efficient and Effective Layer-Wise Reranking.**
Cesare Campagnano, Antonio Mallia, Jack Pertschuk, and Fabrizio Silvestri.
ECIR 2025.
[DOI: 10.1007/978-3-031-88714-7_41](https://doi.org/10.1007/978-3-031-88714-7_41)
- **Eclipse: Contrastive Dimension Importance Estimation with Pseudo-Irrelevance Feedback for Dense Retrieval.**
Giulio D'Erasmus, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto.
ICTIR 2025.
[DOI: 10.1145/3731120.3744579](https://doi.org/10.1145/3731120.3744579)
- **Task Singular Vectors: Reducing Task Interference in Model Merging.**
Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà.
CVPR 2025.
[DOI: 10.1109/CVPR52734.2025.01742](https://doi.org/10.1109/CVPR52734.2025.01742)
- **Are Convolutional Sequential Recommender Systems Still Competitive? Introducing New Models and Insights.**

Federico Siciliano, Antonio Purificato, Filippo Betello, Nicola Tonellotto, and Fabrizio Silvestri.
IJCNN 2025.

[DOI: 10.1109/IJCNN64981.2025.11229036](https://doi.org/10.1109/IJCNN64981.2025.11229036)

- **A Theoretical Analysis of Recommendation Loss Functions under Negative Sampling.**

Giulia Di Teodoro, Federico Siciliano, Nicola Tonellotto, and Fabrizio Silvestri.

IJCNN 2025.

[DOI: 10.1109/IJCNN64981.2025.11228603](https://doi.org/10.1109/IJCNN64981.2025.11228603)

- **Projection-Displacement-Based Query Performance Prediction for Embedded Space of Dense Retrievers.**

Suchana Datta, Guglielmo Faggioli, Nicola Ferro, Debasis Ganguly, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto.

ACM Trans. Inf. Syst., 44(1), Article 7.

[DOI: 10.1145/3746617](https://doi.org/10.1145/3746617)

- **Neural Prioritisation for Web Crawling.**

Ophir Frieder, Nicola Ferro, Joel Mackenzie, Marc Najork, Edie Rasmussen, and Nicola Tonellotto.

ICTIR 2025.

[DOI: 10.1145/3731120.3744597](https://doi.org/10.1145/3731120.3744597)

- **Efficient Recommendation with Millions of Items by Dynamic Pruning of Sub-Item Embeddings.**

Nicola Tonellotto, Aleksandr Petrov, and Craig Macdonald.

SIGIR 2025.

[DOI: 10.1145/3726302.3729963](https://doi.org/10.1145/3726302.3729963)

- **A Reproducibility Study of PLAID.**

Sean MacAvaney and Nicola Tonellotto.

arXiv preprint.

[URL: arXiv:2404.14989](https://arxiv.org/abs/2404.14989)

2024

- **Dimension Importance Estimation for Dense Information Retrieval.**

Giulio D'Erasmus and Nicola Tonellotto.

SIGIR 2024.

[DOI: 10.1145/3626772.3657691](https://doi.org/10.1145/3626772.3657691)

- **Faster Learned Sparse Retrieval with Block-Max Pruning.**

Antonio Mallia, Torsten Suel, and Nicola Tonellotto.

SIGIR 2024.

[DOI: 10.1145/3626772.3657906](https://doi.org/10.1145/3626772.3657906)