D3.2 - Report on Final Cascading IR/NLP Systems

Project: NEREO - Neural Information Retrieval and NLP Systems
Grant Agreement: PRIN 2022
Deliverable ID: D3.2
Work Package: WP3 - Cascading IR/NLP Systems
Due Date: M24
Lead Beneficiary: UNIROMA1

---

1. Executive Summary

This deliverable marks the completion of **Work Package 3 (Cascading IR/NLP Systems)**. In the second year, we capitalized on the foundational insights of Year 1 ("Power of Noise") to build advanced, self-correcting cascading architectures. We introduced **QPP-RA**, a method for using LLMs to assess and aggregate rankings, effectively realizing the "System-as-a-User" vision. Furthermore, we deepened our fundamental understanding of the Transformer architecture through a mechanistic analysis of **wavelet-like properties**, enabling more predictable control over the generation process.

2. Detailed Research Activities

2.1 QPP-RA: LLMs as Ranking Aggregators (Task 3.2)

Related Publications: ICTIR 2025

The NEREO proposal envisioned a system where the downstream component acts as a "user" of the upstream system, judging quality and requesting corrections.

- **Concept:** In *"QPP-RA: Aggregating Large Language Model Rankings"*, we implemented this by using LLMs to perform **Query Performance Prediction (QPP)**. We treat multiple LLMs as independent "judges" that re-rank the retrieval list based on their internal knowledge.
- **Aggregation:** We developed a robust **aggregation logic** (RA) that combines these LLM-generated rankings. This improves the final ranking quality significantly over any single model, effectively filtering out the "negatively relevant" documents that might fool a single reasoner but not an ensemble.
- **System-as-a-User:** This directly fulfills the WP3 goal of modeling the "System-as-a-User." The LLM aggregator effectively "critiques" the IR output, providing a feedback loop that stabilizes the cascading pipeline.

2.2 Mechanistic Interpretability of Transformers (Task 3.1)

Related Publications: ACL 2025 ("Beyond Position")

To control the "catastrophic interaction" between IR and NLP, we must understand how the NLP model (Transformer) processes its input (the retrieved documents).

- **Discovery:** In *"Beyond Position: the emergence of wavelet-like properties in Transformers"*, we analyzed the internal attention heads of Transformer models. We discovered that they spontaneously learn to encode positional information using patterns that resemble **Wavelet transforms**.
- **Impact:** This finding implies that Transformers have a multi-scale understanding of context (local vs. global). This explains why "random noise" (which disrupts local patterns) might be handled differently than "semantic distractors" (which mimic global patterns). This theoretical breakthrough

allows us to design "context injection" strategies that align with the model's internal wavelets, optimizing the information flow from IR to NLP.

2.3 Consistent Counterfactuals (Task 3.3)

Related Publications: IEEE Trans. Artif. Intell. 2025

We refined our explainability module to ensure high fidelity to the underlying data distribution.

- **Consistency:** In *"Consistent Counterfactual Explanations via Anomaly Control..."*, we addressed a major flaw in counterfactual generation: suggesting "impossible" changes (e.g., "increase age by 10 years and decrease seniority"). By incorporating **Anomaly Control**, our new framework ensures that the explanations generated by the cascading system are causal, tailored, and physically possible, enhancing user trust.

3. Impact on NEREO Objectives

1. **Objective O3 (End-to-End Optimization):** QPP-RA closes the loop in the cascading system. It transforms the pipeline from a unidirectional flow (IR -> NLP) to a bidirectional one where the NLP component actively refines the IR output.
2. **Objective O1 (Evaluation):** The mechanistic understanding of Wavelets provides a new lens for evaluation. We can now inspect *internal* attention maps to diagnose failures, offering a deeper evaluation than surface-level metrics.
3. **Objective O4 (Trustworthy AI):** The Consistent Counterfactuals work ensures that the system fits the *human-centric* criteria of Horizon Europe, providing valid and actionable recourse.

4. Conclusion

WP3 has evolved from identifying problems ("The Power of Noise") to engineering solutions ("QPP-RA"). The final Cascading IR/NLP system is not just a pipeline but an **intelligent agent** that can retrieve, evaluate, aggregate, and explain information. This represents the successful achievement of the NEREO scientific vision.

---

5. Scientific References

2025

- **Beyond Position: the emergence of wavelet-like properties in Transformers.**
  Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri.
  ACL 2025.
  URL

- **QPP-RA: Aggregating Large Language Model Rankings.**
  Filippo Betello, Matteo Russo, Paul Dütting, Stefano Leonardi, and Fabrizio Silvestri.
  ICTIR 2025.
  DOI: 10.1145/3731120.3744575

- **Consistent Counterfactual Explanations via Anomaly Control and Data Coherence.**
  Maria Movin, Federico Siciliano, Rui Ferreira, Fabrizio Silvestri, and Gabriele Tolomei.
  IEEE Trans. Artif. Intell., 6(4), 794-804.
  DOI: 10.1109/TAI.2024.3496616