D3.1 - Report on Initial Cascading IR/NLP Systems

Project: NEREO - Neural Information Retrieval and NLP Systems
Grant Agreement: PRIN 2022
Deliverable ID: D3.1
Work Package: WP3 - Cascading IR/NLP Systems
Due Date: M12
Lead Beneficiary: UNIROMA1

---

1. Executive Summary

This report presents the research activities and findings of **Work Package 3 (Cascading IR/NLP Systems)** for the first year of the NEREO project. WP3 focuses on the "downstream" component of the pipeline, where Information Retrieval outputs are consumed by Natural Language Processing reasoners (specifically, Large Language Models). Year 1 produced a paradigm-shifting result regarding the role of "noise" in Retrieval-Augmented Generation (RAG), and successfully delivered **DanteLLM**, a foundational model to support the Italian NLP ecosystem. These results directly address the project's core hypothesis about the "catastrophic interaction" between IR and NLP components.

2. Detailed Research Activities

2.1 The "Power of Noise" in RAG Systems (Task 3.1)

Related Publications: SIGIR 2024 ("The Power of Noise"), IIR 2024 ("Rethinking Relevance")

The central premise of RAG is that retrieving *relevant* documents helps the LLM generate better answers. In *"The Power of Noise: Redefining Retrieval for RAG Systems"*, we challenged this assumption through a large-scale systematic analysis.

- **Methodology:** We experimented with various "context compositions," feeding LLMs (Llama2, GPT-4) with mixtures of:
  - *Gold* documents (containing the answer).
  - *Related* documents (topically relevant but incorrect).
  - *Random* documents (completely unrelated noise).
- **Key Findings:**
  1. **The Noise Paradox:** Surprisingly, including **random documents** in the prompt context did not degrade performance. On the contrary, it improved accuracy by up to **35%** in certain settings compared to using only relevant docs.
  2. **Harmful Relevance:** Documents that were "related but wrong" (e.g., discussing the same entity but a different attribute) were observed to be **negatively relevant**, actively confusing the model and causing hallucinations.
- **Implications for NEREO:** This confirms our proposal's Limitation **L1** and **L3**. Current IR metrics (Precision@k, NDCG) reward "related" documents, but these are exactly the ones that harm the cascading system. This work defines the specification for a "Noise-Aware" Retriever (Task 3.2), which must distinguish between "benign noise" and "harmful distractors."

2.2 DanteLLM: Establishing an Italian Baseline (Task 3.2)

Related Publications: LREC/COLING 2024

Effectively evaluating cascading systems requires robust downstream reasoners. However, most open-source LLMs are English-centric.

- **Model Development:** In *"DanteLLM: Let's Push Italian LLM Research Forward!"*, we curated a massive Italian corpus and fine-tuned LLaMA-based models to create a dedicated Italian LLM.
- **Benchmarking:** We evaluated DanteLLM on a suite of tasks (summarization, categorization, reasoning), demonstrating that it significantly outperforms multilingual models of similar size.
- **Strategic Value:** DanteLLM serves as the "Downstream Reasoner" for all our future Italian-language experiments (e.g., Fact-Checking in Italian news), fulfilling **Objective O4** of providing tools/applications for the local research community.

2.3 Explainability in Cascading Systems (Task 3.3)

Related Publications: xAI 2024 ("Human-in-the-Loop...")

Trust in AI systems requires not just accuracy but explainability.

- **Counterfactual Recourse:** We developed a framework for **Personalized Counterfactual Recourse**. Instead of just outputting a decision, the system generates "recourse": telling the user *what* input features to change to flip the decision (e.g., "If you had 2 more years of experience, the loan would be approved").
- **Human-in-the-Loop:** We integrated human feedback to ensure these explanations are feasible, addressing the "Human-Centric AI" priority of Horizon Europe.

3. Impact on NEREO Objectives

- **Objective O1 (Evaluation):** The "Power of Noise" study has provided the empirical basis for a new evaluation framework. We now know that we must evaluate IR systems not just by *relevance* (NDCG) but by *downstream utility* (LLM Accuracy), and specifically penalize "related distractors."
- **Objective O3 (End-to-End Optimization):** By identifying that "noise" varies in impact, we have opened a path to optimize the retrieval stage to be "distractor-free" rather than "noise-free," a nuance that was previously misunderstood.
- **Objective O4 (Technology Transfer):** DanteLLM is a public asset (Open Science) that transfers state-of-the-art LLM capabilities to Italian academia and industry.

4. Conclusion and Next Steps

Year 1 of WP3 has been transformative. We have moved from a "naive" RAG approach to a sophisticated understanding of how IR and NLP interact. The insights from "The Power of Noise" will define the research agenda for Year 2, where we will build **Noise-Aware Retrievers** (WP2) and **Robust Aggregators** (WP3) to exploit these findings.

---

5. Scientific References

2024

- **The Power of Noise: Redefining Retrieval for RAG Systems.**
  Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri.
  SIGIR 2024.
  [DOI: 10.1145/3626772.3657834](#)

- **Rethinking Relevance: How Noise and Distractors Impact Retrieval-Augmented Generation.**
  Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri.
  IIR 2024.
  [URL](URL)

- **DanteLLM: Let's Push Italian LLM Research Forward!**
  Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri.
  LREC/COLING 2024.
  [URL](URL)

- **Human-in-the-Loop Personalized Counterfactual Recourse.**
  Carlo Abrate, Federico Siciliano, Francesco Bonchi, and Fabrizio Silvestri.
  xAI 2024.
  [DOI: 10.1007/978-3-031-63800-8_2](DOI: 10.1007/978-3-031-63800-8_2)