Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (SD TIB PAN)

SUBJECT:
Intent-Aware Large Language Models for Efficient Search

SUPERVISOR:
Prof. Arjen P. de Vries (main supervisor)
Radboud University, Radboud, Netherlands
arjen.devries@ru.nl

Wojciech Kusa, PhD (auxiliary supervisor)
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB),
ul. Kolska 12, Warszawa
wojciech.kusa@nask.pl

DESCRIPTION:
Search systems built on Large Language Models (LLMs), including retrieval-augmented generation, conversational search and agentic research assistants, usually run the same costly process for every query: retrieve documents, re-rank them, fill the context window and generate a long answer. This is wasteful as many queries are navigational or simple factual look-ups that need a single retrieval step or none at all, while only a minority require multi-step search and extended reasoning. User intent is a long-established and cheap signal in information retrieval, and it is a natural way to decide how much computation a query deserves. Existing intent taxonomies and classifiers, however, were designed for ranked lists of documents, not for generative systems. This project asks how intent can be predicted reliably at query time, and how it can then be used to make LLM-based search both more accurate and considerably cheaper.

The first objective is to improve intent detection itself. The starting point is ORCAS-I, which splits informational queries into factual, instrumental and abstain sub-classes and labels ten million queries using weak supervision. Later work shows that LLMs find more instances of each intent class than weak supervision but make more false positives, and that fine-tuning on weak labels works better than in-context learning. The project will study how to combine the two: using LLM predictions as additional labelling functions, distilling large models into classifiers small enough to run on every query, calibrating confidence so the system can abstain, and extending the taxonomy to conversational, multi-intent and non-English queries.

The second objective is to use the predicted intent to control the search pipeline. Intent can decide whether to retrieve at all, how many passages to pass to the model, which index or tool to call, whether to rewrite the query or ask a clarifying question, and whether to start a multi-step agent loop. These decisions will be compared with methods that estimate effort from question complexity or model uncertainty, such as Adaptive-RAG, Self-RAG and FLARE. Results will report answer quality together with latency, generated tokens and cost per query, rather than quality alone.

The third objective is to move intent prediction inside the model. The project will test intent prediction as an auxiliary task during supervised fine-tuning, intent expressed as special tokens that condition the answer, and reinforcement learning with a reward that combines answer quality with computational cost. Probing of internal representations will show how much intent information the model already encodes and when explicit supervision helps. The expected outcomes are new annotated resources for intent in generative search, open models and routing recipes, and evidence on how much computation current LLM search systems spend without benefit to the user.

REQUIREMENTS:

- MSc degree in computer science, computational linguistics, mathematics, artificial intelligence or a related field
- Strong programming skills in Python, with experience in PyTorch and the Hugging Face ecosystem
- Familiarity with information retrieval concepts, evaluation methodology and standard toolkits
- Knowledge of current research on LLMs, retrieval-augmented generation and model fine-tuning
- Ability to work with large query logs and to run experiments on GPU clusters
- Advanced level of English (spoken and written)

BIBLIOGRAPHY:

- Alexander D., Kusa W., de Vries A. P. (2022) ORCAS-I: Queries Annotated with Intent using Weak Supervision
- Alexander D., de Vries A. P. (2025) In a Few Words: Comparing Weak Supervision and LLMs for Short Query Intent Classification
- Asai A. et al. (2024) Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflection
- Broder A. (2002) A Taxonomy of Web Search
- Craswell N. et al. (2020) ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search
- Jeong S. et al. (2024) Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity
- Jiang Z. et al. (2023) Active Retrieval Augmented Generation
- Jin B. et al. (2025) Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning
- Ma X. et al. (2023) Query Rewriting for Retrieval-Augmented Large Language Models