
Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Industry 4.0 (38.th cycle)

# Efficient Adaptation of Large Language Models in Natural Language Processing

**Marco Braga**

* * * * *

**Supervisors**
Prof. Gabriella Pasi, Supervisor
Prof. Alessandro Raganato, Co-supervisor

Politecnico di Torino
April 30, 2026

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

.....

Marco Braga

Turin, April 30, 2026

# Summary

The rapid growth of Large Language Models (LLMs) has significantly improved performance across a wide range of Natural Language Processing (NLP) tasks, including Information Retrieval (IR). Despite their strong generalisation capabilities, LLMs still require domain- and task-specific fine-tuning to achieve competitive performance in highly specialised scenarios. LLMs exhibit some generalisation abilities, though these are not uniformly reliable across tasks and domains. In particular, adapting these models to downstream domains and tasks has become challenging as full fine-tuning is computationally expensive, memory-intensive and difficult to scale across multiple domains and tasks. Over the past few years, substantial research has been made into efficient fine-tuning of LLMs, resulting in various proposed approaches. Nevertheless, it remains an ongoing research area with several unresolved issues and challenges.

Motivated by this, in this dissertation we introduce novel efficient strategies for domain and task adaptation of LLMs, focusing on parameter efficient methods that reduce both training time and costs while preserving effectiveness. First, we provide a systematic review of Parameter-Efficient Fine-Tuning (PEFT) approaches, analysing how efficiency is defined and evaluated in the literature and identifying key open challenges. Building on these insights, we propose AdaKron, a novel adapter architecture based on Kronecker-product parametrisation, designed to increase expressiveness while updating less than 1% of model parameters. We further extend this approach with MAdaKron, which integrates a Mixture-of-Experts mechanism into AdaKron to improve adaptation quality through expert routing without increasing computational overhead. To address multi-domain information retrieval, we introduce DRAMA, a parameter-efficient retrieval framework based on domain-specific adapters and an adaptive gating mechanism that dynamically allocates model capacity depending on the input query domain.

In addition, to go beyond parameter-efficient approaches, we explore training-free model composition techniques, studying Task Arithmetic and task vectors for zero-shot retrieval adaptation and showing how domain- and language-specific retrieval models can be obtained through model merging without additional fine-tuning. Furthermore, through an analysis of task vectors, this thesis investigates what neural ranking models learn when fine-tuned on query-document pairs. Our

analysis reveals that prompt-token embeddings encode a large portion of the relevance adaptation signal in T5-based rerankers, enabling a lightweight adaptation strategy that achieves effectiveness comparable to full fine-tuning. Finally, we contribute new large-scale resources for reproducible evaluation, including SE-PQA, a dataset for personalized community question answering, and its synthetic extension generated with LLMs.

Overall, this dissertation advances the development of scalable and sustainable domain adaptation methods for LLMs, combining structured parameter-efficient architectures, dynamic retrieval adaptation, training-free model merging, and benchmark resources to support future research in efficient NLP and IR.