FutureData4EU

Training Future Big Data Experts for Europe

UNIMORE

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Co-funded by the European Union

"FutureData4Eu (Grant. Agreement n. 101126733) Co-Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them."

Beyond Information Exchange

Fine-Tuning LLMs for Metadiscourse Control in Academic Writing

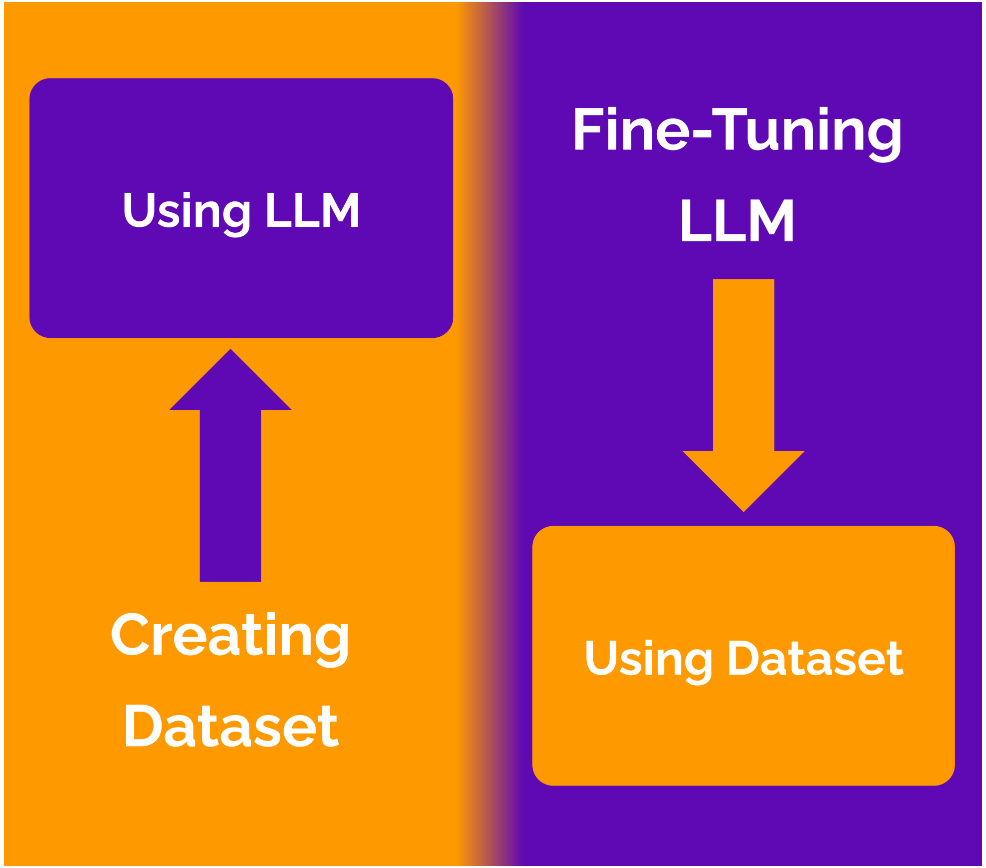
Javad Haditaghi

Tutor: Prof. Cavalieri

Co-tutor : Prof. Bondi

A General Overview

- Fine-tune LLM(?) for controlled metadiscourse in based on context in academic writing
- Create comprehensive dataset (20,000 sentences) using Hyland's framework (2018)
- Implement three-level annotation system (high/medium/low)
- Ensure cross-disciplinary balance and annotation quality through IAA
- Apply supervised fine-tuning with optimized parameters



FutureData4EU
Training Future Big Data Experts for Europe

A General Schema

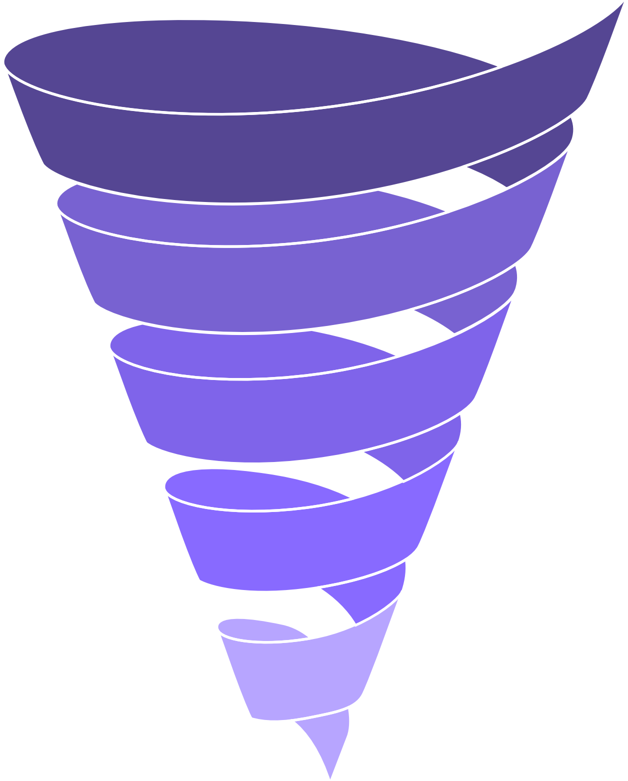
Create Annotated Dataset
Developing a dataset with metadiscourse annotations

Implement IAA
Ensuring consistency in annotations through inter-annotator agreement

Supervised Fine-Tuning
Training the LLM using the annotated dataset

Evaluate Model
Assessing the LLM's performance in metadiscourse control

Refine Model
Improving the LLM based on evaluation results



Why Create a New **Metadiscourse Dataset?**

01

Existing open source datasets lack metadiscourse depth and disciplinary diversity

02

Hyland's (2018) model stresses interactive & interactional features (Broad rather than Narrow)

03

Current resources: low inter-annotator agreement, manual bias, ML-unfriendly; Lack of clear data statement (profiling)

04

Goal: benchmark-quality dataset ready for LLMs, ML, DL, & academic writing tools

Corpus Compilation & Annotation Dimensions

+200
Dissertations

29
Disciplines

+20,000
Annotated Instances

Annotation dimensions:

- Sentence
- Metadiscourse Category
- Metadiscourse Feature
- Section (IMRaD)
- Moves & Steps (Swales, 2004; Coto et al.,2020; Yang & Allison, 2003)
- Target (Hyland, 2018)
- Rhetorical Strength
- Sentence Position/ Paragraph Location
- Writer Background (Native, Non-native)

Annotation
Protocol & Tools

GPT 4

Hallucination

Stochastic Behavior

Rationale-driven Collaborative Few-shot Prompting with Iterative Validation Loop (Wu et al., 2025)

Data Profiling & Analysis

Reliability, Validity, & Robustness

01

Inter Annotator Agreement

Manual pilot phase;
Krippendorff's Alpha
Cohen's or Fleiss' Kappa
Artstein (2017)

02

Datasheets for Datasets

technical and structural dimensions of datasets
(Gebru et al., 2018)

03

Data Statements for NLP

linguistic and ethical profiling
(Bender & Friedman, 2018)

04

Stat features

Showing meta-level features
(Uddin & Lu, 2024)

What Happens Without Data Profiling?

The DiseaseAlert Failure Story (Bender & Friedman, 2018)

A hospital in the U.S. developed an early-warning system for infectious diseases based on Twitter data — it worked well locally and was released as open-source.

01

Problem began when a hospital in Abuja, Nigeria adopted the system. Despite using local tweets, the model failed to detect outbreaks, causing false alerts and loss of trust.

02

Root Cause? Not a bug. Not bad code.

A dataset.

The language ID component used a model trained on:

Only highly edited US/UK English

And came with no data statement

03

What Went Missing in the Dataset?

1. No mention of dialectal or regional language coverage
2. No info on genre, domain, or data source
3. No way for users to evaluate suitability or bias

Even a high-performing model can fail catastrophically if it's trained on a narrow, undocumented dataset.

Data Profiling isn't a luxury
It is the safety feature.

Anticipated
Model Capabilities

Style Adaptation

- Adjust hedging vs. boosting language
- Control engagement markers density
- Modulate attitude markers strength
- Adapt to native/non-native preferences

Context Sensitivity

- Section-appropriate metadiscourse use
- Move/step-specific language patterns
- Discipline-aware rhetorical conventions
- Position-sensitive feature placement

Training Pipeline & Process

Data Preprocessing

Feature engineering
Contextual Embedding
Density metrics
rhetorical progression patterns

Model Initialization

RoBERTa-large?
T5?
GPT?

Fine-Tuning Training

Option 1 = Standard fine-tuning approach
Option 2 = Custom architecture building

Validation & Testing

precision/recall/F1
native vs. non-native speaker survey
Error Analysis

Model Deployment

A simple web-based editor
React JS/ Next JS
Django

Expected Results

Scientific Results – Non-Native Friendly – Low Hedging

Control Settings:

Hedging/Boosting: 0.8 (confident)
Engagement: 0.3 (low)
Attitude: 0.6 (moderate)
Complexity: 0.4 (non-native friendly)
Section: Results
Discipline: STEM

Annotated Output:

The experiment clearly (Booster – Interactional) demonstrates an increase in efficiency under the new algorithm. As shown in Table 2 (Endophoric Marker – Interactive), the optimized model reduces computation time by an average of 18%. This improvement likely (Hedge – Interactional) results from better resource allocation. It is important (Attitude Marker – Interactional) to note that no significant errors were observed during testing. The system consistently (Booster – Interactional) outperformed the baseline across all test cases. These results support (Booster – Interactional) the hypothesis presented in Section 2 (Endophoric Marker – Interactive).

Impacts & Innovations

01

Supports AI-assisted writing education

02

Broad Disciplinary Coverage

03

Bridges NLP with genre/rhetoric studies

04

Addresses lack of metadiscourse depth, poor profiling, domain narrowness

Acknowledgement

This is a secret :D

FutureData4EU

Training Future Big Data Experts for Europe

References

Artstein, Ron. "Inter-annotator agreement." *Handbook of linguistic annotation* (2017): 297-313.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, *46*, 90-106.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing* (2nd ed.). Bloomsbury Academic.

Ruiying, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for specific purposes*, *22*(4), 365-385.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Uddin, S., & Lu, H. (2024). Dataset meta-level and statistical features affect machine learning performance. *Scientific Reports*, *14*, Article number 1670. https://doi.org/10.1038/s41598-024-51825-x

Wu, J., Wang, X., & Jia, W. (2025, April). Enhancing text annotation through rationale-driven collaborative few-shot prompting. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

FutureData4EU

Training Future Big Data Experts for Europe

Funding:

COESIONE ITALIA 21-27 EMILIA-ROMAGNA

Co-funded by the European Union

REPVBBLICA ITALIANA

RegioneEmilia-Romagna

Partner:

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

UNIVERSITÀ CATTOLICA del Sacro Cuore

Università degli Studi di Ferrara

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DI PARMA

POLITECNICO MILANO 1863