Week-3: Choosing the Right Research Method for My AI-Based NLI Study

Choosing the Right Research Method for My AI-Based NLI Study

Introduction

Research method selection constitutes a pivotal decision in academic inquiry: it structures how evidence is gathered, the standards by which results are evaluated, and the extent to which conclusions can be generalized and replicated. In the context of my MSc. project, I investigate *Native Language Identification (NLI)* within user-generated English text by developing a bias-aware, generalizable framework that integrates Large Language Model (LLM) embeddings, topic debiasing, and open-set recognition. This post articulates the justification for adopting a **quantitative experimental–comparative design** and explains how this approach enables systematic *assessment and evaluation* of project outcomes including model accuracy, fairness, and robustness.

Project Summary

This study designs and evaluates a **hybrid LLM-based NLI system** on **Reddit-L2**, targeting three evaluation axes: (i) *performance* on known languages (accuracy; macro-F1), (ii) *fairness* through mitigation of topic leakage (e.g., Named-Entity Masking and linear concept erasure), and (iii) *generalization* to unseen native languages via **open-set recognition**. The contribution lies in a controlled, comparative framework that produces *quantitative, reproducible* evidence across in-domain and cross-topic regimes, thereby aligning method choice with the project's evaluation objectives.

Why the Study Is Quantitative (and Experimental–Comparative)

The core outputs I must evaluate are numeric- `accuracy`, `macro-F1`, `bias-leakage score`, and `false-positive rate on unseen L1s`. My proposal defines explicit hypotheses and compares multiple model variants under controlled conditions (same data splits; fixed preprocessing). This is a textbook match to a **quantitative, experimental–comparative design**.

> **Design logic:** manipulate the *independent variable* (model family / debiasing / open-set mechanism) and observe the effect on *dependent variables* (metrics above) across *in-domain* and *cross-topic* settings.

Why Quantitative Is the Most Suitable Method

To meet the requirement for the research project to follow an quantitative method on *assessing and evaluating outcomes*, the method must yield objective, reproducible evidence and allow controlled comparisons. Below is the extended logic that underpins my choice.

1) Alignment with a recognized framework

Urban & van Eeden-Moorefield (2018) characterise quantitative studies by *objective measurement*, *hypothesis testing*, *low researcher–participant interaction*, and *generalization*. My project maps to these traits one-to-one.

Criterion	Quantitative expectation	My study
Epistemology	Positivist; one discoverable truth	Seeks measurable improvements in fairness/accuracy
Core logic	Deductive; hypothesis-driven	Predefined hypotheses on debiasing and open-set gains
Data form	Numeric variables & scales	Accuracy, macro-F1, bias-leakage, FPR
Design control	Standardized procedures	Fixed splits, seeds, preprocessing, evaluation scripts

Criterion	Quantitative expectation	My study
Bias handling	Method controls for bias	NER masking; concept erasure; identical pipelines per condition
Generalization	External validity targeted	Cross-topic regime; TOEFL11 transfer (if licensed)
Analysis	Statistical comparison	Mean±SD over seeds; significance tests; ablations

2) Direct support for evaluation

- **Assessing outcomes:** KPIs are explicit and comparable across models.
- **Fairness evidence:** Bias-leakage is a quantitative signal, enabling objective auditing.
- **Reproducibility:** Version-controlled code/config ensures reviewers can rerun analyses.

3) Why not qualitative / mixed approach

- **Qualitative:** Suited to human experiences/interpretations; my study is computational with no participants.
- **Mixed-methods:** Valuable later (e.g., user interviews on fairness perceptions), but adds scope/complexity without improving metric validity for my research work’s evaluation plan.

Reflection: Choosing quantitative strengthens internal validity (control), external validity (cross-topic tests), and reliability (multi-seed runs)

Conditions/Comparators I Will Evaluate

To evaluate the research hypotheses, the following models will be compared under identical settings:

1. **Traditional Baseline:** character/word/POS n-grams with Logistic Regression or SVM.
2. **Zero-shot LLM:** Prompt-based inference using GPT-like architectures.
3. **Hybrid (LLM Embeddings + Debias):** BERT/RoBERTa embeddings combined with Named-Entity Masking and Linear Concept Erasure.
4. **Hybrid + Open-set:** The hybrid model extended with thresholding and embedding-distance novelty detection.

Each configuration will employ consistent data partitions, preprocessing, and evaluation metrics to ensure fair comparison.

Data I Will Gather

The study primarily utilizes the **Reddit-L2** corpus, a large-scale dataset of English texts written by non-native speakers. The corpus provides an opportunity to investigate topic bias and language transfer effects at scale. A secondary dataset, *TOEFL11*, may be used for external validation. All text data undergo anonymization,

normalization, and entity masking to maintain ethical and methodological consistency.

Tools & Techniques

The project leverages modern NLP and ML toolkits — **PyTorch**, **HuggingFace Transformers**, and **scikit-learn —** for training and evaluation. Debiasing employs Named-Entity Masking and Linear Concept Erasure, while open-set recognition uses probabilistic thresholding and Mahalanobis distance in embedding space. Computation will be GPU-accelerated, and all runs parameterized for reproducibility.

Analysis & Evaluation Plan

Evaluation follows a two-regime protocol- **in-domain** (same-topic) and **cross-topic** (out-of-domain). Performance metrics include Accuracy and Macro-F1, fairness is assessed via Bias-Leakage Score, and open-set capability through False Positive Rate. Reliability is supported by repeated trials under different random seeds, reporting mean, standard deviation and performing statistical tests to confirm significance. Ablation studies quantify the contribution of each debiasing component.

Validity, Reliability, and Replicability

- **Internal validity:** Controlled splits and identical preprocessing across conditions.
- **External validity:** Cross-topic evaluation and (if licensed) TOEFL11 transfer tests.

- **Reliability:** Seed control; repeated runs; consistent scoring pipelines.
- **Replicability:** Version-controlled code, configs, and evaluation scripts.

Key Takeaway

The quantitative experimental–comparative method provides the structure and statistical integrity necessary to evaluate AI model outcomes objectively. It supports systematic measurement, hypothesis testing, and replicability – ensuring that performance, fairness, and generalization results are interpretable and academically defensible.

References

1. Ahmad, F. (2025) *Final Research Project Proposal Report (PROM04 – Assignment-2)*, University of Sunderland.
2. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A. and Chodorow, M. (2013) 'TOEFL11: A corpus of non-native English', *ETS Research Report Series*, 2013(2), pp. 1–15.
3. Koppel, M., Schler, J. and Zigdon, K. (2005) 'Determining an author's native language by mining a text for errors', *Proceedings of KDD*, pp. 624–628.
4. Kumar, S., Wintner, S., Smith, N.A. and Tsvetkov, Y. (2019) 'Topics to avoid: Demoting latent confounds in text classification', *EMNLP-IJCNLP*, pp. 3931–3941.
5. Rabinovich, E., Ordan, N. and Wintner, S. (2018) 'Native

language cognate effects in L2 English lexical choice', *TACL*, 6, pp. 329–342.

6. Urban, J.B. and van Eeden-Moorefield, B.M. (2018) *Designing and Proposing Your Research Project*. Washington, DC: APA.
7. Yaghoobzadeh, Y., Hertel, J. and Tsvetkov, Y. (2024) 'The medium is not the message: Deconfounding text embeddings via linear concept erasure', *arXiv*:2403.05025.
8. Zhan, X., Liu, L., Wang, J. and Pan, S.J. (2021) 'A joint learning framework for open-set domain adaptation', *TPAMI*, 44(9), pp. 4626–4640.
9. Zhang, W. and Salle, A. (2023) 'Native Language Identification with Large Language Models', *arXiv*:2312.07819.