SOVEREIGN: How does back-translation paraphrasing affect the robustness of LLM question answering performance across diff

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

NLP practitioners often want to take existing trained models and apply them to data from new domains. While fine-tuning or few-shot learning can be used to adapt a base model, there is no single recipe for making these techniques work; moreover, one may not have access to the original model weights if it is deployed as a black box. We study how to improve a black box model's performance on a new domain by leveraging explanations of the model's behavior. Our approach first extracts a set of features combining human intuition about the task with model attributions generated by black box interpre

1 Introduction

Analysis of: Can Explanations Be Useful for Calibrating Black Box Models?. Research goal: How does back-translation paraphrasing affect the robustness of LLM question answering performance across different domains when evaluated on the MRQA and MultiQA datasets?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

8 papers retrieved. 9 claims extracted, 9 verified. Tribunal: 7.7/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
NLP practitioners often want to take existing trained models and apply them to data from new domains	✓	0.31
Fine-tuning or few-shot learning can be used to adapt a base model	✓	0.27
One may not have access to the original model weights if it is deployed as a black box	✓	0.31
The approach first extracts a set of features combining human intuition about the task with model attributions generated	✓	0.40
The approach uses a simple calibrator, in the form of a classifier, to predict whether the base model was correct or not	✓	0.28
The experimental results across all the domain pairs show that explanations are useful for calibrating these models	✓	0.35
The method is tested on extractive question answering and natural language inference tasks	✓	0.17
The method covers adaptation from several pairs of domains with limited target-domain data	✓	0.23
Calibration model transfers to some extent between tasks	✓	0.21

References

- https://doi.org/10.18653/v1/d19-5801
- https://doi.org/10.18653/v1/2022.acl-long.429
- https://doi.org/10.18653/v1/d19-5829