



INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

**Petrichor: Semi Supervised Fine-Tuning of  
Textual Entailment Based Boolean Question-Answering  
for Pre-trained Language Models**

A Dissertation by

Mr. Tony Praveen Jesuthasan

Supervised by

Mr. Torin Wirasinghe

Submitted in partial fulfilment of the requirements for the BSc (Hons) in Computer  
Science degree at the University of Westminster.

**May 2022**

## ABSTRACT

Pre-trained Language Models (PLM) have taken the Natural Language Processing (NLP) domain by storm since its inception during the latter years of the previous decade. Consisting of two training stages: pre-training (unsupervised) and fine-tuning (supervised), these language models require quite a large amount of annotated data for the latter process. Procurement of such data is expensive and an immensely time-consuming process, thus hampering the use of these powerful models especially in domains where annotated corpora is scarce. Boolean Question-Answering, an NLP task, is known for being notoriously difficult in nature to solve as it relies on the textual entailment between a question and passage to infer an answer. Obtaining a labelled dataset of this sort is extremely difficult as it requires long hours of human expertise, combing through each question and its relevant passage before deducing the right answer. This hindrance also limits the use of PLMs to solve the aforementioned NLP task. These aspects present the need for a solution capable of solving the labelled data requirements of the fine-tuning process of a language model and the Boolean Question-Answering task. A promising approach that has demonstrated the ability to reduce the annotated data requirement across any medium or task is semi-supervised learning. This dissertation presents *Petrichor* – a hybrid architecture that pairs a PLM with a Generative Adversarial Network (GAN) for semi-supervised fine-tuning of textual entailment based Boolean Question Answering to solve this gap. Experimental results indicate that the proposed architecture is capable of producing similar performance levels (F1-scores between 70-80%) when utilizing either 100% or 10% labelled data samples of a dataset. Benchmarking results showcase *Petrichor*'s ability to outperform functionally similar models that suffer massive performance drops when radically reducing annotated data quantity.

**Keywords** – Natural Language Processing; Pre-trained Language Models; Generative Adversarial Networks; Boolean Question-Answering; Textual Entailment; Semi-Supervised Learning; Data Annotation.