Faculty of Science
Department of Mathematics

MA4198 PROJECT PROPOSAL (PROJECT CUM SEMINAR GROUP)

SUPERVISOR'S INFO

Name:	Nguyen Hung Minh Tan
Email:	tanmn@nus.edu.sg
Tel number:	+65 8308 2743
Office location:	S17-08-20

PROJECT ID: PS2520-02

TITLE

Behavioral Control in Large Language Models via Activation Steering

BRIEF DESCRIPTION OF PROJECT

Large Language Models (LLMs) are now widely used in domains such as education, healthcare, legal support, and creative applications. Despite their impressive capabilities, controlling specific behaviors, e.g., reducing harmful or toxic outputs, shaping refusal tendencies, or adjusting emotional tone, remains a significant challenge.

Most existing steering techniques manipulate activations using linear vector operations (e.g., adding or subtracting steering directions). However, these approaches depend heavily on hyperparameters and operate along a single, fixed direction, making their performance inconsistent and often difficult to tune. As a result, it is not always clear which method should be preferred in practice.

In this project, students will:

1. **Study and implement** key activation steering methods, including vector addition–based steering and directional ablation.

2. **Assess behavioral steering effectiveness** on tasks such as refusal/compliance adjustment, toxicity mitigation, and emotion modulation.

3. **Compare and benchmark** the performance of these baseline steering methods across multiple LLM architectures.

EXPECTATION/S

1. Read papers on activation steering in LLMs that Prof. Tan Nguyen's collected
2. Implement 3-4 activation steering methods
3. Benchmark those methods on at least 3 different LLMs
4. Submit a paper to a machine learning workshop or conference (e.g., ICML workshop or COLM/NeurIPS conference) after the project.

PREREQUISITE/S (at level 3000 or below, with at most one course at level 3000)

MA3270 or MA3252 or MA3264 or MA3238 or MA3238S or MA3236 or MA3227 or MA3220; MA1522 or MA1513 or MA1508E or MA1311 or MA2001 or MA2101 or MA2101S

READING REFERENCE/S

Angular Steering: https://arxiv.org/pdf/2510.26243
Steering Language Models With Activation Engineering: https://arxiv.org/pdf/2308.10248
Refusal in Language Models Is Mediated by a Single Direction: https://arxiv.org/pdf/2406.11717
Programming refusal with conditional activation steering: https://arxiv.org/pdf/2409.05907
Inference-Time Intervention: Eliciting Truthful Answers from a Language Model: https://arxiv.org/pdf/2306.03341