TEXAS A&M UNIVERSITY
CORPUS CHRISTI

MS Thesis Defense:	Vishakan Umapathy
Title:	A Comparative Evaluation Framework for a Question Answering System Using Large Language Models
Date and time:	Wednesday, July 8, 2026 2:00 pm
Place:	Zoom only meeting

Abstract:

This thesis develops and evaluates methods for assessing large language models (LLMs) on closed-context question answering using the Stanford Question Answering Dataset 2.0 (SQuAD 2.0). It addresses two problems that distort how modern generative models are scored on this benchmark: (i) surface-form metrics such as Exact Match (EM) and token-level F1 penalize answers that are semantically correct but lexically different from the reference, and (ii) adversarial "unanswerable" annotations have aged into label noise that modern reasoning models routinely see through, so that correct, context-grounded answers are marked wrong. To address these, the thesis proposes a two-part framework. The first part is a dual-metric evaluation that augments standard lexical scoring with an LLM-as-a-judge, producing a Semantic Equivalence Score (SES) that credits meaning-preserving paraphrases and a Re-Validation Score (RVS) that checks whether a disagreeing answer is nonetheless entailed by the passage. The second part is the Agentic Consensus Protocol (ACP), in which heterogeneous LLM agents resolve disagreement not by voting but by structured debate, proposing independent positions and then challenging, defending, peer-scoring, and converging until a formal σ-ρ rule certifies durable agreement. The full study reports prompt-optimization experiments, a zero-shot baseline of two frontier models on a fixed 10,000-example evaluation set, parameter-efficient fine-tuning of three open-weight models, and a failure-recovery evaluation in which ACP, applied to 1,364 questions the strongest single model had already failed, recovers correct answers on a substantial fraction (23.17% EM and roughly 50% under the semantic measures). Across every setting, the consistent ordering of EM below F1 below the semantic scores quantifies how much genuine comprehension lexical metrics fail to credit. The result is a transparent, reproducible evaluation methodology for closed-context QA that rewards semantic accuracy, treats abstention as a first-class outcome, and analyzes multi-agent consensus behavior.