Q2 Project Proposal

Rahul Sengupta
rasengupta@ucsd.edu

Zachary Thomason
zthomason@ucsd.edu

Zeyu (Edward) Qi
zeqi@ucsd.edu

Akshay Medidi
amedidi@ucsd.edu

Ryan Lingo
ryanlingo@gmail.com

1 Problem Statement

As LLMs continue to improve and develop, they generate increasingly large volumes of text that must be evaluated. This has become a significant issue because evaluation methods are not keeping pace with the rapid growth of LLM capabilities. No matter the task an LLM is attempting to accomplish, its value is only as strong as our ability to evaluate its output effectively.

2 Evidence of the Issue

Individual evaluation approaches are not effective enough on their own. LLM-as-a-judge is the most promising solution due to its ability to process large amounts of information. However, as Shankar and Husain (2025) note in *Application-Centric AI Evals for Engineers and Technical Product Managers*, engineers often struggle to get LLM-as-a-judge systems to work reliably in practice. This is due to the highly sensitive nature of LLMs, which require precise guidance and specificity to produce correct evaluations. Even with such guidance, issues persist, including hallucinations, topic drift, and flawed reasoning.

Other common approaches also suffer from major limitations. Human evaluation is too slow and costly; having humans review the growing number of LLM outputs is infeasible because the time required to read and assess each instance scales poorly. Task-specific evaluation metrics such as BLEU and ROUGE-1 also introduce problems, as described by Oleszak (2024) in *Evaluating Large Language Models*. ROUGE-1 measures surface-level overlap rather than meaning, enabling summaries that achieve high scores despite being unhelpful or incorrect. BLEU score, used for translation, similarly evaluates n-gram matching rather than true semantic fidelity. Individually, each of these approaches has significant shortcomings that must be addressed.

3 Current Model

Our evaluation and refinement framework is designed to mitigate the inconsistency and bias that arise when using a single LLM as an automatic judge. Instead of relying on one evaluation method, we combine three components. First, we use an ensemble LLM-as-judge rubric that scores summaries along coverage, faithfulness, organization, clarity, and style. Second, we incorporate an agreement metric that compares generated summaries against a human-written reference. Third, we compute deterministic signals that capture word count, redundancy, and slide coverage.

These signals are unified into a scalar performance score and fed into an iterative refinement loop in which the model repeatedly critiques and improves its own summaries. Additionally, we use pairwise tournament-style comparisons and selective refinement to stabilize judgments further. Together, these components yield more consistent scoring and higher-quality summaries than any single evaluation strategy.

4 Final Model

For the final model, we aim to improve upon our existing model by introducing a fully realized Chain of Density procedure, an AI persona judge that mimics human evaluation, and an adaptive rubric that gets updates during each refinement iteration. The judge will be implemented as a grading persona and will evaluate summaries using a structured rubric that establishes clear criteria, contains human examples, and has penalties for hallucination and omission. The rubric will continuously improve during the iterative process through meta evaluation steps that evaluate and update the rubric to better evaluate the specific domain. The model will also only continue iterations until the evaluation score stops improving, eliminating the fixed iterations used in the previous model. We believe that with this combination of methods the model will eliminate the need for human benchmarking, it will be more scalable, more stable, and more resistant to hallucination and inconsistency.

5 Data

In order to collect the data needed to complete this project, we will start by generating a small collection of high-quality human-generated summaries of some subset of lecture slides. They will be used as the calibration examples to fit the AI persona judge to the same amount of evaluation as a human. Once this initial calibration is done, the evaluation system should be able to scale without the contribution of some other human references.

We shall now perform the task of deriving content units out of each lecture deck in order to facilitate our implementation of Chain-of-Density. Such content units enable us to quantify the coverage, density, and enhancement through refinements.

It will also produce some forms of synthetic data, such as adversarial summaries, low-quality summaries, and deliberately bad outputs. These stress-test the ability of the evaluation system to be able to punish the omissions, hallucinations, and redundancies correctly.

When trained, the AI persona will create summaries of new lecture decks that have the appearance of references, and this will allow training and assessment of the AI persona on a large scale without human supervision. This artificial reference will be used to confirm that the assessment system extrapolates the small human-labeled sample.

The combination of these datasets allows making sure that (1) the model is aligned with human judgment, (2) the Chain-of-Density process is provided with the required structural information, and (3) the evaluation pipeline can be tested both in regular and edge-case modes.