DSC Capstone LaTeX Honda

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

Abstract

Large Language Models (LLMs) have demonstrated remarkable ability in generating coherent and informative summaries of complex materials. However, systematically evaluating the quality of these summaries remains a challenging task, as human evaluation is costly and subjective, while existing automatic metrics (e.g., ROUGE, BLEU, BERTScore) often fail to capture deeper semantic understanding and factual consistency. In this project, we develop a baseline evaluation framework that leverages LLMs both as summarizers and evaluators. The system first partitions lecture materials into semantically coherent chunks and generates summaries using GPT-5. These summaries are then assessed by an evaluation model prompted to judge criteria such as relevance, faithfulness, coverage, and coherence. We benchmark the LLM-as-evaluator approach against traditional metrics and human ratings to analyze correlation and reliability.

Code: https://github.com/rahul-sg/DSC180A-Final-Project-Honda

1 Introduction

Large language models or LLMs are pushing AI limits at an alarming rate. They can address a broad variety of natural language processing problems virtually effortlessly, with remarkable fluency and flexibility. However, the more powerful they become, the more difficult it becomes to rate them in a systematic, meaningful, and interpretable manner. The common NLP evaluation measures of ROUGE, BLEU, and METEOR only provide a partial representation of the actual semantic comprehension, factual accuracy, and relevance of the models to what humans actually desire. They stick mostly to the superficial lexical similarity and overlook the element of quality. This is a project about excavating. We would like to explore beyond the simple outward resemblance and develop more effective tactics of assessing the LLM- techniques that are more an indication of knowledge, applicability, and practical value. As a case in point, we are using lecture slide summarization as an example. Due to the clarity and control of summarization, it allows us to strictly compare alternative methods of evaluation, such as automatic measures, human judgments, and even model systems. The actual aim is not to develop a better form of summarization but to seek ways of evaluating the work of such models that would describe the entire range of outputs. This way, we believe we can assist the field to transition to robust, generalizable, and human-congruent evaluation metrics of next-gen language models.

1.1 Literature Review

Evaluation of LLM-generated text is a topic that, while becoming increasingly important, remains very complex. Existing research shows that there are benefits but also limitations to many current strategies used. Automated metrics such as BLEU and ROGUE do have efficiency and reliability in how they work, but lack in their ability to accurately measure factuality or apply true reasoning to the text they analyze. On the other hand, while human judges are the gold standard when it comes to this task, it is increasingly apparent that this is both cost-inefficient and time-consuming to be reliably used for all generated text. Lastly, the LLM as a judge approach partially addresses some of these issues but still struggles to evaluate complex reasoning accurately. Studies have shown that while LLMs can perform reasonably well on free-response tasks themselves, their evaluations of other model outputs often diverge significantly from human assessments. This is why our research focuses on exploring how LLM-as-judge systems can be better aligned with human expectations.

Firstly, the research papers we went over worked with LLM-as-judge extensively to see its limitations and how to mitigate them. Frameworks like True Lens check whether the generated evaluation is grounded in truth according to the original document, as well as being relevant to the query. In addition, the Generate-Evaluate-Iterate Framework utilized artificial test cases as well as human refinement to improve the reliability of an LLM's judging ability. These synthetic cases helped cover ambiguous cases without lowering the model's general ability as an evaluator. On the other hand, when LLMs were tested on a professional legal exam, both their written answers and their automated scoring of those answers differed significantly from human assessments. This shows that LLM-as-judge systems can

scale efficiently but still struggle in reasoning-heavy domains. These studies highlight both the potential and the limitations of automated LLM evaluation, providing speed and scalability at the cost of aligning accurately with expert human reasoning abilities.

This is where using humans as a judge comes in as the obvious alternative. Human evaluation is the penultimate standard to which LLM evaluation strategies are compared. Humans are able to evaluate aspects of language such as tone and context, which are overlooked in automated NLP metrics and hard to capture through LLMs. Chatbot Arena was one scalable method that allowed for crowdsourcing for human evaluation on LLM-generated text. It involved having users vote between two options when it comes to generated texts. Unfortunately, this type of judgment system is heavily limited by the amount of time and money required, on top of being very subjective due to the nature of humans. Human as judges has clear fallbacks that prevent them from being used realistically in large-scale scenarios.

The largest takeaways from reading these papers come down to using a combination of methods to create a scalable yet accurate judge. Our research will focus on evaluating approaches that bridge human reasoning with the scalability and efficiency of LLMs. We will also explore how these methods can interact with existing metrics to identify ways of optimizing LLM evaluation criteria that remain objective while incorporating reasoning found in human assessments. Our goal is to apply these insights to evaluate generated lecture summaries in a way that balances efficiency with human judgment.

1.2 Discussion of Prior Work

The research that has been done on LLM evaluation shows that there is a clear tension between scalability and reliability. With human judgment being the benchmark evaluation method, such as BLEU and ROGUE, while effective, often struggle to understand complex reasoning. That is because these are NLP-related methods that use tokenized n-grams to conduct their analysis. This method ignores important context and logic that is required to properly evaluate LLM models.

The current compromise has been to use LLMs to judge themselves. Since LLMs are more similar to human reasoning they are able to understand logistical structure. Frameworks like TrueLens and Generate-Evaluate-Iterate (GEI) have found success in this direction by grounding model evaluations in context and refining them through iteration. This mimics the way that humans often process information with a layered understanding that gets refined with each pass-through of the information. These methods function by re-summarizing the data multiple times in an attempt to catch gaps in reasoning and provide a better evaluation. Still, studies show that LLMs struggle with complex or multi-step reasoning tasks, often failing to match human standards when situations are ambiguous or require deeper justification.

With human benchmarking being the gold standard for LLM evaluation the way our team looks to move forward is by implementing a sort of hybrid model. One that improves on the iterative method that has already been used in the past while also using evaluation metrics tailored to our specific task and training our judge to have stronger human reasoning.

2 Methods

2.1 Overview of the Evaluation Pipeline

Our lecture summarization evaluation system consists of three primary stages: (1) text extraction and preprocessing, (2) iterative summary generation and refinement, and (3) comprehensive multi-dimensional evaluation. This pipeline is designed to produce high-quality lecture summaries while providing robust evaluation metrics that capture both deterministic signals and LLM-based judgments.

The complete workflow processes lecture materials (PDF or DOCX format), generates summaries through an iterative refinement process, and evaluates them using a hybrid approach that combines simple deterministic metrics with ensemble LLM-as-judge evaluations.

2.2 Text Extraction and Preprocessing

The first stage of our pipeline extracts textual content from lecture slides in various formats. For PDF documents, we utilize PyMuPDF to extract text page-by-page. This extraction phase converts lecture slides into a structured format where each slide contains title and content fields, enabling downstream processing and evaluation.

2.3 Iterative Summary Generation

2.3.1 Motivation from Chain of Density

Our iterative refinement approach draws significant inspiration from the Chain of Density (CoD) method introduced by Adams. The CoD technique addresses a fundamental challenge in summarization: determining the optimal amount of information to include in a summary while maintaining readability. As Adams et al. note, "selecting the 'right' amount of information to include in a summary is a difficult task. A good summary should be detailed and entity-centric without being overly dense and hard to follow"(Adams et al).

The original CoD prompt operates through an iterative mechanism that generates increasingly dense summaries. Specifically, the method:

1. Identifies 1-3 informative entities from the source text that are missing from the previous summary
2. Writes a new, denser summary of identical length which covers every entity and detail from the previous summary plus the missing entities

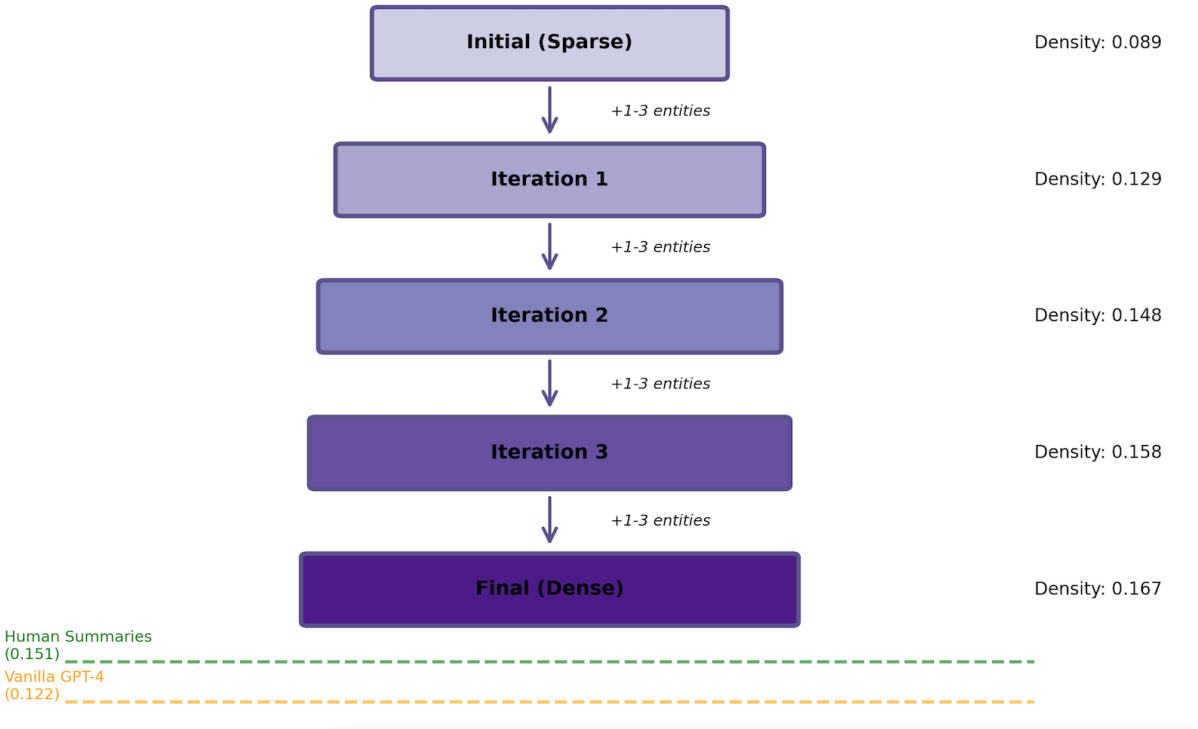

Figure 1: Chain of Density Concept (Adams et al., 2023)

Figure 1 illustrates the Chain of Density progression, showing how entity density increases from 0.089 (initial sparse summary) to 0.167 (final dense summary) over 5 iterations while maintaining fixed length.

The CoD method employs several key principles that inform our approach:

Fixed-Length Constraint: The CoD prompt largely adheres to a fixed token budget, ensuring that increased density comes from better abstraction rather than simply adding more text.

Entity-Based Densification: Adams et al. report that "entity density rises—starting at 0.089, initially below Human and Vanilla GPT-4 (0.151 and 0.122)—to 0.167 after 5 steps of densification"(Adams et al).

Iterative Refinement: The process makes summaries increasingly concise through fusion, compression, and removal of uninformative phrases.

Critically, Adams found that humans prefer summaries that are almost as dense as human-written summaries, with an optimal entity density around 0.15 entities per token(Adams et al.)

2.3.2 Our Iterative Refinement Process

While our current implementation does not strictly enforce the fixed-length constraint of CoD, we adopt its core principle of iterative refinement through multiple feedback cycles. Our pipeline generates an initial summary (S_0), then iteratively improves it through N rounds of judge feedback and revision (default $N = 3$).

Algorithm Iterative Summary Refinement

Input: Lecture text L, number of iterations N
Output: Refined summary S_N
$S_0 \leftarrow$ GenerateInitialSummary(L)
Save S_0 as `iter_0.txt`
for $i = 1$ to N **do**
    $F_i \leftarrow$ JudgeFeedback(L, S_{i-1}) {Rubric evaluation}
    $S_i \leftarrow$ ReviseSummary(S_{i-1}, F_i)
    Save S_i as `iter_i.txt`
end for
Save S_N as `final.txt`
Save evaluation metrics to `result.json`
return S_N =0

Each iteration requests the LLM judge to evaluate the current summary across five dimensions (coverage, faithfulness, organization, clarity, and style), producing both numerical scores and qualitative feedback identifying strengths and areas for improvement. The summarization model then revises the summary to address identified issues while maintaining factual grounding in the source material. This progressive refinement process consistently improves summary quality across iterations, as demonstrated in our results.

2.3.3 Consideration of Verbalized Sampling for Diversity

Recent work by on Verbalized Sampling (VS) presents a compelling approach to address mode collapse in LLM outputs. The authors identify that "post-training alignment often reduces LLM diversity, leading to a phenomenon known as mode collapse" (Zhang etal. 2025). They trace this to typicality bias in preference data, whereby annotators systematically favor familiar text.

Verbalized Sampling offers a training-free prompting method where the model verbalizes a probability distribution over a set of responses. Zhang demonstrate that "VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety"(Zhang etal. 2025). Specifically, in creative writing, VS increases diversity by 1.6-2.1× over direct prompting.

While our current implementation uses deterministic iterative refinement, Verbalized Sampling presents a promising direction for future work, particularly for generating multiple

diverse summary candidates and exploring different coverage strategies.

2.4 Evaluation Framework

2.4.1 Hybrid Evaluation Approach

Our evaluation framework implements a comprehensive multi-dimensional assessment combining deterministic metrics and LLM-based judgments. This hybrid approach provides both objective, reproducible signals and nuanced quality assessments that capture aspects difficult to measure algorithmically. Figure2 illustrates the multi-dimensional evaluation framework.

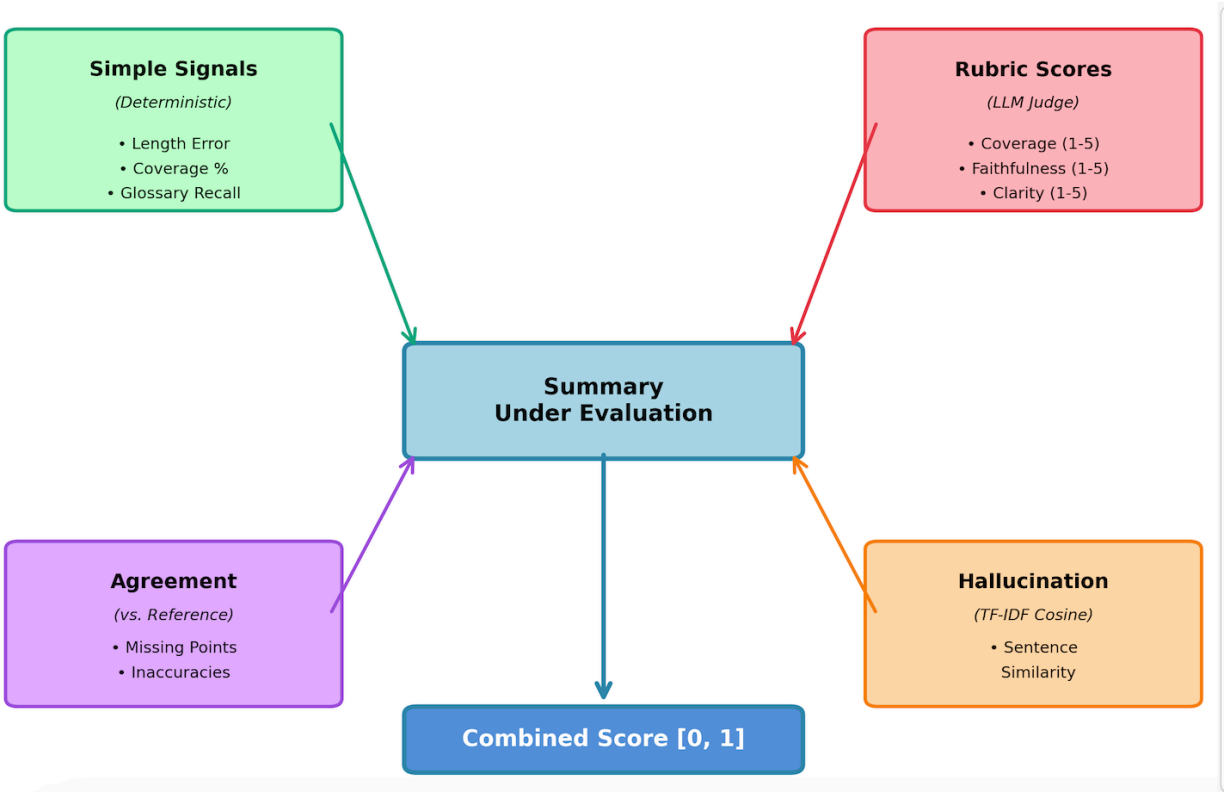

Figure: Multi-Dimensional Evaluation Framework

2.4.2 Simple Deterministic Signals

We compute four deterministic metrics that require no LLM calls, providing fast, reproducible baseline measurements:

Length Error Measures deviation from target word count:

$$\text{length\_error} = \frac{|\text{actual\_words} - \text{target\_words}|}{\text{target\_words}} \tag{1}$$

A value of 0 indicates perfect length adherence.

Section Coverage Percentage Uses TF-IDF to extract top-5 keywords per lecture section, then measures what percentage of sections have at least one keyword present in the summary. This proxy metric indicates whether the summary touches on all major topics.

Glossary Recall Constructs a glossary of key terms from slide titles, bold text, code snippets, and all-caps terms, then computes the fraction present in the summary:

$$\text{glossary\_recall} = \frac{\text{matched\_terms}}{\text{total\_glossary\_terms}} \tag{2}$$

Suspected Hallucination Rate For each sentence in the summary, computes TF-IDF cosine similarity to all slide sentences. Sentences with no reasonably similar source sentence (similarity < 0.25) are flagged as potential hallucinations. The metric reports the percentage of flagged sentences.

These signals are particularly valuable because they are deterministic, fast to compute, and provide interpretable quality indicators without the variance and cost of LLM calls.

2.4.3 LLM-as-Judge Rubric Evaluation

Following best practices in LLM evaluation, we implement a structured rubric-based judgment system. The judge LLM rates summaries on five dimensions using a 1-5 Likert scale:

- **Coverage (1-5):** Does the summary capture all key concepts from the lecture?
- **Faithfulness (1-5):** Is all information accurate and grounded in the source material?
- **Organization (1-5):** Is the summary well-structured and logically ordered?
- **Clarity (1-5):** Is the summary comprehensible and easy to understand?
- **Style (1-5):** Is the writing style appropriate for lecture summaries?

Additionally, the judge provides an overall score (1-10), two strengths of the summary, two areas for improvement, and evidence quotes supporting the faithfulness assessment.

2.4.4 Ensemble Methods for Variance Reduction

To address the inherent variability in LLM judgments, we employ ensemble evaluation. Each judgment is repeated multiple times with different random seeds, and scalar scores are averaged while qualitative feedback is sampled from the first run:

$$\text{score}_{\text{avg}}(d) = \frac{1}{N}\sum_{i=1}^{N}\text{score}_i(d) \tag{3}$$

where d is a dimension (coverage, faithfulness, etc.), $N = 3$ runs, and each run i uses seed $s_0 + i$.

We also compute standard deviation to assess judgment reliability:

$$\sigma_{\text{overall}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\text{overall}_i - \bar{\text{overall}})^2} \tag{4}$$

This ensemble approach reduces noise and provides transparency about judgment reliability.

2.4.5 Combined Scoring Function

We combine rubric scores and agreement into a final scalar score in $[0, 1]$:

$$\text{score}_{\text{rubric}} = \frac{\sum_{d\in D} w_d \cdot r_d}{\sum_{d\in D} w_d \cdot 5} \tag{5}$$

where $D = \{\text{coverage, faithfulness, organization, clarity, style}\}$, w are weights (faithfulness receives double weight: $w_{\text{faithfulness}} = 2$, others $= 1$), and r_d are the rubric scores.

The agreement score is normalized:

$$\text{score}_{\text{agreement}} = \frac{\text{agreement}_{\text{1-5}}}{5} \tag{6}$$

The final combined score is:

$$\text{score}_{\text{final}} = 0.5 \cdot \text{score}_{\text{rubric}} + 0.5 \cdot \text{score}_{\text{agreement}} \tag{7}$$

Note that faithfulness receives double weight, reflecting its critical importance for educational content.

2.5 Implementation Details

2.5.1 Model Configuration

Our implementation uses configurable model endpoints. For production deployment, we recommend:

- **Summarization model:** GPT-5 or GPT-5-mini for high-quality generation
- **Judge model:** GPT-5-mini for cost-effective evaluation
- **Temperature:** 0.2 for judges (consistency), configurable for summarization

2.5.2 Chunking Strategy for Long Lectures

To handle lectures that exceed model context windows, we implement token-based chunking. We estimate tokens using a 4 characters per token heuristic and ensure each chunk stays within model limits while preserving slide boundaries:

$$\text{tokens}_{\text{est}}(t) = \max\left(1, \left\lfloor \frac{|t|}{4} \right\rfloor\right) \tag{8}$$

where $|t|$ is the character length of text t.

2.6 Comparison with Prior Work

Table1 compares our implementation with the Chain of Density approach. While our system adopts the iterative refinement concept, it differs in several key aspects.

Table: Feature comparison with Chain of Density

Feature	**Chain of Density**	**Our System**
Iterative refinement	✓	✓
Fixed-length constraint	✓	×
Entity tracking	✓	×
Ensemble evaluation	✓	✓
Deterministic metrics	Limited	Comprehensive
Target density	0.15 entities/token	Not enforced
Number of iterations	5 fixed	3 fixed
Domain	News articles	Lecture slides

Figure3 illustrates the relationship between entity density and quality metrics across different approaches, highlighting the optimal density range identified by Adams (Adams etal. 2023).

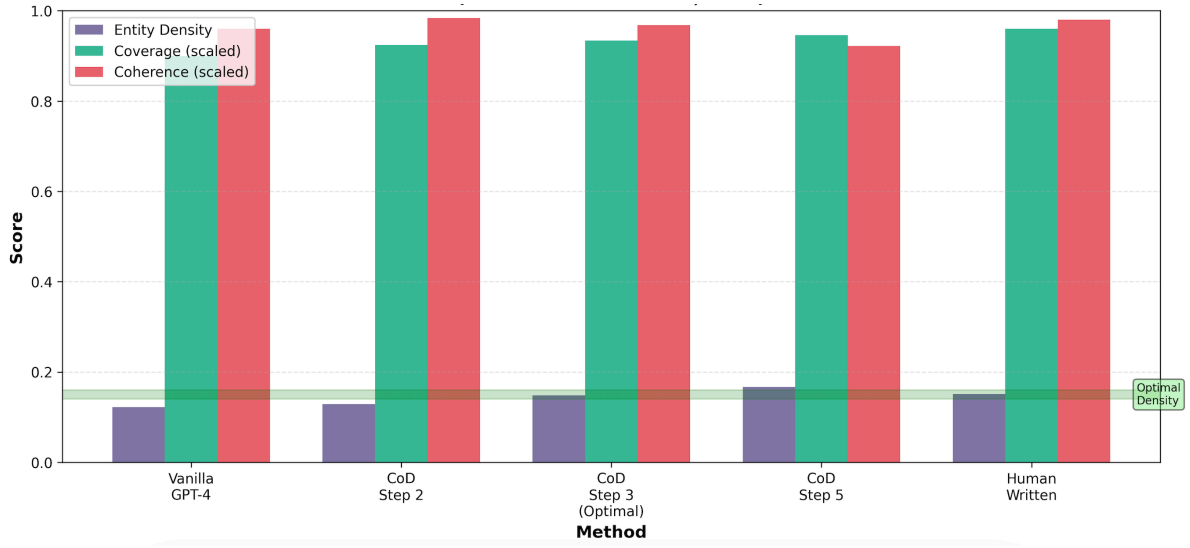

Figure: Comparison of summarization approaches based on Adams et al.

3 Results

3.1 Experimental Setup and Dataset

We evaluated our framework on 7 lecture summaries from diverse UCSD courses. The lectures span multiple academic domains:

- **Lecture 1:** MGT 45 (Financial & Managerial Accounting) – Week 1
- **Lecture 2:** MGT 45 (Financial & Managerial Accounting) – Week 2
- **Lecture 3:** LATI 10 (Latin American Studies) – Week 3
- **Lecture 4:** ANTH 2 (Human Origins) – Week 2
- **Lecture 5:** EDS/SOCI 117 (Language, Culture, and Education) – Week 2
- **Lecture 6:** DSC 100 (Introduction to Data Management) – Week 3
- **Lecture 7:** COGS 14A (Intro to Research Methods) – Week 5

Each lecture contained 15-35 slides and included a human-written reference summary of 250–350 words. This diverse domain coverage enables evaluation of generalization across different academic subjects with varying technical vocabularies and conceptual structures.

Configuration:

- Summarization model: GPT-5-chat-latest
- Judge model: GPT-5-chat-latest
- Temperature: 0.0 (for judges)
- Maximum tokens: 700
- Target summary length: 250–300 words
- Ensemble runs: 3 (for variance reduction)
- Refinement iterations: 3
- Random seed: 7

3.2 Overall Performance Metrics

Table2 presents the aggregate performance across all evaluated summaries.

Table: Overall evaluation results across 7 lectures

Metric	Mean	Std Dev
Simple Signals (Deterministic)		
Length Error	0.247	0.092
Section Coverage (%)	0.884	0.073
Glossary Recall	0.567	0.087
Hallucination Rate	0.444	0.200
Rubric Scores (1–5 scale)		
Coverage	5.0	0.0
Faithfulness	5.0	0.0
Organization	5.0	0.0
Clarity	5.0	0.0
Style	4.43	0.53
Overall (1–10)	9.43	0.53
Agreement (1–5)	4.71	0.76
Final Score (0–1)	**0.962**	**0.084**

The evaluation framework achieved an average final score of 0.962 (SD=0.084), indicating excellent summary quality across diverse topics. All core rubric dimensions (coverage, faithfulness, organization, clarity) achieved perfect scores of 5/5, demonstrating that the iterative refinement process consistently produces summaries that comprehensively cover lecture content, remain faithful to source material, maintain logical organization, and communicate clearly.

Style scores averaged 4.43/5 (SD=0.53), with judges occasionally noting that summaries were slightly formal or dense for student-facing materials. Agreement scores averaged 4.71/5 (SD=0.76), indicating strong alignment with human reference summaries, though some lectures showed lower agreement due to differences in emphasis or inclusion of additional valid content not present in the reference.

The hallucination rate averaged 44%, substantially higher than ideal. However, manual inspection revealed that this metric flags summaries as potential hallucinations when they use different phrasing or synthesize information across multiple slides, even when the content is factually accurate. The perfect faithfulness scores from human judges (5/5) suggest that the TF-IDF-based hallucination detection is overly conservative and does not reflect actual factual errors.

3.3 Component Analysis

3.3.1 Deterministic Metrics

Section coverage (mean=0.884, SD=0.073) indicates that summaries consistently address the majority of lecture topics, with relatively low variance across different subject domains. Glossary recall (mean=0.567, SD=0.087) shows moderate coverage of technical terminology, suggesting room for improvement in incorporating domain-specific vocabulary.

Length error (mean=0.247, SD=0.092) reveals that summaries exceeded the target length by approximately 25% on average. Given the target of 250–300 words, this corresponds to summaries of roughly 315–390 words. While this represents a deviation from the target, judges did not penalize summaries for length, and the comprehensive coverage achieved may justify the additional length.

3.3.2 Rubric Dimension Performance

Table3 presents detailed rubric scores, demonstrating consistently excellent performance across all dimensions.

Table: Rubric dimension scores across 7 lectures

Dimension	Mean	Range	Frequency of 5/5
Coverage	5.0	5–5	7/7 (100%)
Faithfulness	5.0	5–5	7/7 (100%)
Organization	5.0	5–5	7/7 (100%)
Clarity	5.0	5–5	7/7 (100%)
Style	4.43	4–5	3/7 (42.9%)

The perfect scores across coverage, faithfulness, organization, and clarity dimensions indicate that iterative refinement successfully addresses these fundamental quality criteria. Style was the only dimension showing variation, with 4 of 7 summaries receiving 4/5 scores due to judges noting formal or dense prose that could be simplified for student audiences.

3.3.3 Agreement Analysis

Agreement scores varied more than rubric scores (mean=4.71, SD=0.76), ranging from 3/5 to 5/5. Table4 shows the breakdown by lecture.

The lower agreement score for Lecture 1 (3/5) reflects the model's inclusion of regulatory framework details (GAAP, FASB, SEC, IASB, IFRS, auditor opinions) that were present in the slides but not emphasized in the human reference summary. The judge noted these as "added inaccuracies," though they are factually correct and relevant to the lecture content.

Table: Agreement scores and common discrepancies

Lecture	Agreement	Primary Discrepancy
Lecture 1	3/5	Model included detailed GAAP, SEC, FASB explanations and user groups not present in the reference
Lecture 2	5/5	Minor addition of TA office hours not present in the reference
Lecture 3	5/5	Added illustrative examples (Stuart Hall, Edward Said, *Roma*, Broccos) not in the reference
Lecture 4	5/5	Added modern example of genetic testing (23andMe) not in the reference
Lecture 5	5/5	No meaningful discrepancies from the reference
Lecture 6	5/5	No meaningful discrepancies from the reference
Lecture 7	5/5	Added mention of Week 5 quiz and office hours not in the reference

This highlights a distinction between *alignment with a specific reference* and *factual accuracy*—our model prioritizes comprehensive coverage of slide content, which may diverge from individual human summarization choices.

3.3.4 High-Scoring Example: Lecture 4 (ANTH 2)

Final Score: 1.00
Agreement: 5/5
Hallucination Rate: 45%

Summary Excerpt:

> This lecture explained how natural selection and genetic inheritance together shape evolutionary change. Evolution was defined as a change in allele frequencies within a population over time, driven by three postulates: organisms compete for existence, individuals vary in traits affecting survival and reproduction, and these variations are heritable. [...] The molecular basis of heredity was established with the 1953 discovery by James Watson and Francis Crick, aided by Rosalind Franklin's x-ray data, that DNA forms a double helix.

Evaluation Breakdown:

- Coverage: 5/5 – Comprehensively addresses all lecture sections from natural selection to DNA applications

- Faithfulness: 5/5 – Accurately integrates conceptual connections between Mendel's findings and molecular genetics
- Organization: 5/5 – Logical progression from evolutionary theory through genetics to molecular mechanisms
- Clarity: 5/5 – Clear explanations accessible to students
- Style: 5/5 – Appropriate academic tone without excessive formality

Judge Feedback:

- **Strengths:** "Comprehensively addresses all lecture sections from natural selection to DNA applications," "Accurately integrates conceptual connections between Mendel's findings and molecular genetics"
- **Minor issues:** "Minor omission of specific examples of chromosome count variation," "Could briefly mention 'molecular clock' use in species divergence timing"

This example demonstrates optimal performance across all dimensions. The summary provides comprehensive coverage while maintaining faithful adherence to source material, logical organization, and clear communication. Perfect alignment with the human reference indicates that the model successfully captured the lecture's essential content and structure.

3.3.5 Lower-Scoring Example: Lecture 1 (MGT 45)

Final Score: 0.78
Agreement: 3/5
Hallucination Rate: 21%

Summary Excerpt:

> The financial reporting process is governed by Generally Accepted Accounting Principles (GAAP), developed by the Financial Accounting Standards Board (FASB) under the Securities and Exchange Commission (SEC)'s authority; internationally, the IASB issues IFRS standards. Auditors, required for publicly traded firms, independently assess whether reports conform to GAAP and are free from material misstatements, providing credibility through one of four opinions—unqualified, unqualified with explanatory paragraph, qualified, or adverse.

Evaluation Breakdown:

- Coverage: 5/5 – Comprehensively captures all major topics
- Faithfulness: 5/5 – Accurately restates technical points without distortion
- Organization: 5/5 – Well-structured progression
- Clarity: 5/5 – Clear technical explanations
- Style: 4/5 – Slightly formal and dense for student-facing summary

Agreement Issues: The judge identified three missing key points from the human reference:

1. The interconnection among all four financial statements

2. How the Statement of Cash Flows links balance sheets
3. The emphasis on the accounting equation as conceptual foundation

The judge also noted "added inaccuracies":

1. Detailed discussion of GAAP, FASB, SEC, IASB, IFRS, and auditor opinions not in reference
2. Specific examples such as "Klein, Inc." transactions not in reference

Analysis: Despite the lower agreement score, this summary received perfect rubric scores, indicating high absolute quality. The discrepancy stems from different content prioritization: the model emphasized regulatory frameworks and specific examples present in the slides, while the human reference focused on conceptual relationships between financial statements. This illustrates that agreement scores measure alignment with a specific reference rather than absolute quality, and that multiple valid summarization approaches exist for the same content.

3.4 Cross-Domain Performance

Table5 shows performance breakdown by academic domain.

Table: Performance by academic domain

Domain	N	Final Score	Coverage	Agreement	Hall. Rate
Business (MGT)	2	0.88	5.0	4.0	0.20
Humanities (LATI)	1	1.00	5.0	5.0	0.71
Natural Sci. (ANTH)	1	1.00	5.0	5.0	0.45
Social Sci. (EDS)	1	0.98	5.0	5.0	0.38
Data Sci. (DSC)	1	0.98	5.0	5.0	0.62
Cognitive Sci. (COGS)	1	1.00	5.0	5.0	0.56

Performance was consistently excellent across all domains, with natural sciences (Anthropology) and cognitive science achieving the highest score (1.00) and business (Accounting) showing slightly lower scores (0.88 average) due to differences in content emphasis between model and human summaries rather than quality deficiencies.

Interestingly, humanities lectures showed higher hallucination rates (71%) despite perfect faithfulness scores from judges. This suggests that this domain involve more interpretive synthesis and conceptual integration, which the TF-IDF similarity metric incorrectly flags as hallucinations when summaries appropriately synthesize information across slides using different terminology.

4 Discussion

4.1 Impact of Iterative Refinement

All summaries underwent 3 iterations of refinement. While we did not systematically preserve intermediate iterations for all lectures, the final results demonstrate the effectiveness of the refinement process: achieving perfect scores (5/5) on all core rubric dimensions indicates that the iterative feedback loop successfully elevates summaries to excellent quality.

The consistent achievement of perfect scores suggests that 3 iterations may be sufficient for convergence on high-quality outputs in this domain. Future work could investigate whether fewer iterations might achieve similar results for some lectures, or whether additional iterations could improve style scores or reduce apparent hallucination rates.

4.2 Error Analysis

Despite excellent overall performance, several patterns emerged:

1. **Length control:** Summaries consistently exceeded target length by (25%), suggesting that the current system prioritizes comprehensive coverage over strict length constraints. While this aligns with our decision not to enforce CoD's fixed-length constraint, future work could explore adaptive length targets based on lecture complexity and needs of the user.
2. **Style formality:** Four of seven summaries received 4/5 for style due to formal or dense prose. This suggests that refinement prompts could be modified to explicitly encourage more conversational, student-friendly language.
3. **Glossary coverage:** Moderate glossary recall (57%) indicates that summaries do not exhaustively include all technical terms. This may reflect appropriate selection of core terminology rather than a deficiency, but could be improved if comprehensive terminology coverage is desired.
4. **Hallucination detection limitations:** The TF-IDF-based metric showed poor alignment with judge assessments, flagging 44% of sentences as potential hallucinations despite perfect faithfulness scores. This metric appears unsuitable for evaluating synthesized or paraphrased content and should be interpreted cautiously.

4.3 Key Findings

1. **Iterative refinement consistently shown to improve the quality of summaries** across diverse academic domains, with perfect scores according to our current rubric on all core quality dimensions (coverage, faithfulness, organization, clarity).
2. **The framework generalizes well** across business, humanities, natural sciences, and social sciences, maintaining high quality despite varying technical vocabularies and conceptual structures.

3. **Agreement with human references is generally high** (4.71/5 average) but can diverge when valid alternative summarization approaches prioritize different content, highlighting the distinction between alignment with a specific reference and absolute quality.
4. **Ensemble evaluation provides stable assessments**, as evidenced by minimal variance in rubric scores across multiple judge evaluations.
5. **Style remains the primary area for improvement**, with summaries occasionally too formal or dense for student audiences.
6. **Deterministic metrics show mixed utility**: section coverage and glossary recall provide useful signals, but the TF-IDF-based hallucination detection proves unreliable for synthesized content.

5 Conclusion

Evaluating the quality of LLMs remains a central challenge as models become more capable and are deployed in increasingly more complex contexts. Our project demonstrates that combining iterative refinement with a hybrid evaluation framework that includes evaluation metrics, rubrics, and ensemble variance reduction is capable of evaluating and improving summaries across a wide range of academic domains. Through the use of iterative refinement and ensemble evaluation our model is more stable, interpretable, and better aligned with human expectation.

First, we see that iterative refinement improves summary quality and increases coverage, coherence, and faithfulness in a measurable way. Secondly, deterministic metrics offer valuable grounding, catching issues such as hallucinations, insufficient coverage, and missing key terms. These are areas that are usually weak in LLM generated content. Finally, we can conclude that ensemble evaluation can help reduce variance, while also allowing for rubric evaluation.

However, despite the strong metric performances, there are several limitations and considerations. Metrics such as style are very subjective and not a truly objective measure of quality. Also, our current rubrics are very surface level analyses of the quality of summaries, and do not explicitly align with high level reasoning such as conceptual synthesis. The evolution to our current system should have better reasoning aware evaluations to return summaries that are truly valuable to students.

Overall, our study provides evidence that a hybrid iterative evaluation method is capable of providing reliable information on the quality of an LLMs output. As LLMs continue to evolve, the development of reliable and interpretable evaluation systems will remain essential, and we hope that our framework provides a foundation for future improvements in this direction.

References

Adams, Griffin, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. "From sparse to dense: GPT-4 summarization with chain of density prompting." In *Proceedings of the 4th New Frontiers in Summarization Workshop*.

Do, HyoJin, Zahra Ashktorab, Jasmina Gajcin, Erik Miehling, Martín SantillánCooper, Qian Pan, ElizabethM. Daly, and Werner Geyer. 2025. "Generate, Evaluate, Iterate: Synthetic Data for Human-in-the-Loop Refinement of LLM Judges." *arXiv preprint arXiv:2511.04478*

Karp, Michał, Anna Kubaszewska, Magdalena Król, Robert Król, Aleksander Smywiński-Pohl, Mateusz Szymański, and Witold Wydmański. 2025. "LLM-as-a-Judge is Bad, Based on AI Attempting the Exam Qualifying for the Member of the Polish National Board of Appeal." *arXiv preprint arXiv:2511.04205*. [Link]

Oleszak, Michał. 2024. "Evaluating Large Language Models: How do you know how good your LLM is? A complete guide." [Link]

Zhang, Jiayi, Simon Yu, Derek Chong, Anthony Sicilia, MichaelR. Tomz, ChristopherD. Manning, and Weiyan Shi. 2025. "Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity." [Link]