Iterative Refinement of "LLM-as-Judge" Framework

Akshay Medidi · Rahul Sengupta · Zachary Thomason · Zeyu (Edward) Qi · Mentors: Rajeev Chhajer · Ryan Lingo

amedidi@ucsd.edu · rasengupta@ucsd.edu · zthomason@ucsd.edu · zeqi@ucsd.edu · rajeev_chhajer@honda-ri.com · ryan_lingo@honda-ri.com

Motivation

- **Task:** Evaluate LLM-generated lecture slide summaries for student learning.
- **Why it matters:** Summaries must be faithful, useful, and reliable in real settings.
- **Goal:** Build a scalable evaluation process that also drives summary improvement.

Challenge & Our Approach

- **Gap:** ROUGE/BLEU capture overlap, while human review is costly and slow.
- **Issue:** LLM-as-Judge is scalable but can vary across prompts and runs.
- **Approach:** A hybrid framework combining deterministic signals, rubric-based judging, and iterative refinement.
- **Outcome:** Produces interpretable scores and improves summary quality over iterations.

Targets common failure modes:

Omission	Hallucination	Low Usefulness
misses key concepts from slides	adds unsupported or incorrect claims	vague, verbose, or hard to study from

Why Hybrid Evaluation?

What is a Hybrid Framework?

- **Hybrid evaluation** combines deterministic signals with rubric-based LLM judging.
- Signals provide grounded checks; the judge adds semantic scoring and actionable feedback for refinement.

Formula for Hybrid Framework:

Deterministic Signals + **Rubric LLM Judge** + **Iterative Refinement** =

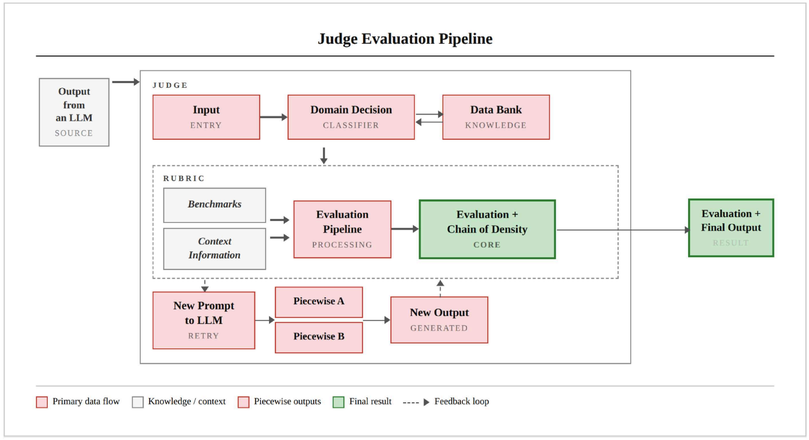
Hybrid Framework

Traditional vs. Hybrid Framework:

Traditional	Hybrid
Slide Text + Base Summary	Slide Text + Base Summary
↓	↓
ROUGE/BLEU or Human Review	Signals + Rubric LLM Judge
↓	↓
Coverage/Faithfulness gaps or high cost	Feedback-guided Refinement

Result: scalable, interpretable evaluation with better summaries

Method: End-to-End Refinement Pipeline



- **Input:** lecture text/structure and current summary S_t
- **Iterate:** judge feedback updates prompts until quality plateaus
- **Output:** final summary with score trace and diagnostics

Final Scoring System

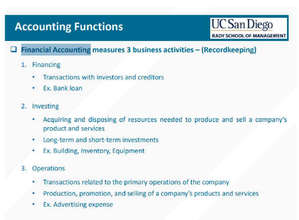
$$
\begin{aligned}
C &= domain_{rubric} && \text{rubric score}\\
M &= (0.8 \cdot base + 0.2 \cdot coverage) - \beta \cdot 2^{h} && \text{metric blend}\\
Q_{raw} &= 0.7C + 0.3M, \qquad \alpha_{eff} = \alpha \cdot m_{domain} && \text{domain-aware risk scale}\\
d_{raw} &= 1 - \alpha_{eff} \cdot h && \text{raw risk discount}\\
d &= \mathrm{clip}(d_{raw}, d_{min}, d_{max}), \quad d_{min} = 0.75,\ d_{max} = 1.0 && \text{bounded trust factor}\\
Q_{risk} &= Q_{raw} \cdot d, \qquad final\_score\_0to1 = Q_{risk} && \text{final ranking score}
\end{aligned}
$$

C	M	h	α
rubric quality	metric blend	hallucination level	base risk sensitivity
β	m_{domain}	d	Q_{risk}
hallucination penalty	domain strictness	trust factor $[0.75, 1.0]$	final adjusted score

Typical setting: $(\alpha = 0.05, \beta = 0.0)$; stricter m_{domain} for technical domains.

Summary Refinement Case Study

(A) Lecture Opening Slide (Input)



(B) Initial Summary Sentence

Financial accounting measures three main business activities: financing (transactions with investors and creditors), investing (acquiring productive resources), and operating (activities related to producing and selling goods and services).

Initial summary: accurate coverage, but dense and less scannable for quick review.
Observed limitation: limited emphasis cues for prioritizing high-yield concepts.
Why refine? high coverage, but weak scanability and limited instructional framing.

(C) Refined Summary Sentence

Financial accounting measures three core business activities:

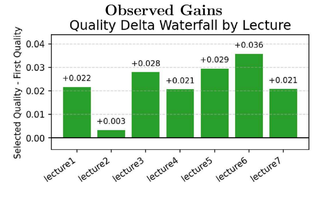
- **Financing:** transactions with investors/lenders (stock, borrowing).
- **Investing:** acquisition or disposal of operating assets.
- **Operating:** day-to-day production, marketing, and sales activities.

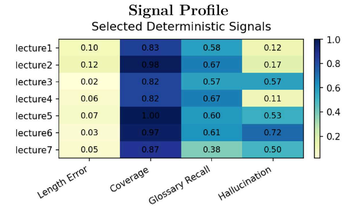
(D) Judge Feedback + Signals

Metric	Value
Overall Score	9/10
Coverage	4/5
Organization	5/5
Length Error	0.1029
Risk-Adjusted Score	0.8270

Strength/Issue: clear activity structure; add course-specific reporting context.

Key Findings & Results





Shared takeaway: refinement produces consistent quality gains, while signal-level variation (recall and hallucination) indicates where additional prompt tuning helps most.

- **Outcome:** quality delta is positive across lectures, with strongest lift at Lecture 6.
- **Diagnostics:** coverage remains strong, while recall/hallucination vary by lecture.

Conclusion

Our hybrid, risk-aware *LLM-as-Judge* pipeline serves as both an evaluator and an improvement engine for lecture summaries.

- **Impact:** on a **7-lecture, 5-domain** benchmark, refinement improves selected quality in **100% of lectures** (7/7), with a mean improvement of **+2.29%** and a peak improvement of **+3.6%**.
- **Interpretability:** signal diagnostics show consistently strong coverage, while recall and hallucination vary by lecture to expose actionable tuning targets.
- **Practical value:** the framework is transparent, data-driven, and suitable for real educational deployment.

Future Direction:

Calibrate → Robustify → Adapt

- **Human calibration:** align rubric and pairwise judgments with student/instructor ratings.
- **Robustness checks:** evaluate cross-model and cross-seed stability with confidence intervals.
- **Adaptive policy learning:** tune $(\alpha, \beta, m_{domain})$ by domain and objective.

References

— Adams et al. "From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting." arXiv:2309.04269 (2023).

— De Chezelles et al. "The BrowserGym Ecosystem for Web Agent Research." arXiv:2412.05467 (2024).

— Yadav et al. "Automatic Text Summarization Methods: A Comprehensive Review." arXiv:2204.01849 (2022).

— Karp et al. "LLM-as-a-Judge is Bad, Based on AI Attempting the Exam Qualifying for the Member of the Polish National Board of Appeal." arXiv:2511.04205 (2025).

Resources

Project Website

GitHub Repo