Beyond Note-Taking: Empowering Examiners with a Diarization-Aware and Transparent AI Assistant for Oral Exams

Maximilian Dauner, Ricky Liu, Gudrun Socher

Hochschule München University of Applied Sciences, 80335 Munich, Germany
Munich Center for Digital Sciences and AI (MUC.DAI)
maximilian.dauner0@hm.edu, ricky.liu@hm.edu, gudrun.socher@hm.edu

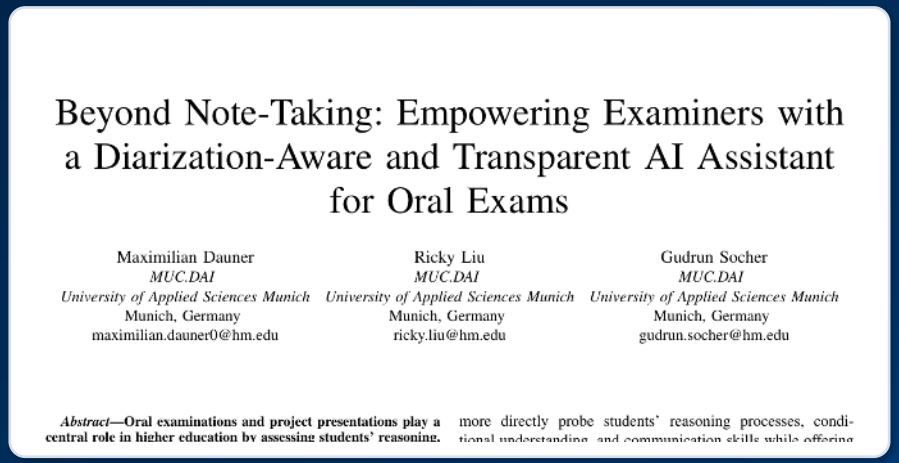
IEEE EDUCON 2026

HM MUC.DAI Munich Center for Digital Sciences and AI

Overview: Problem, Contribution and Take-away

A privacy-preserving, human-in-the-loop pipeline for oral assessment that supports documentation, rubric-linked feedback and transparent review while keeping final grading under examiner control.

Maximilian Dauner · Ricky Liu · Gudrun Socher
MUC.DAI, University of Applied Sciences Munich



Beyond Note-Taking: Empowering Examiners with a Diarization-Aware and Transparent AI Assistant for Oral Exams

Maximilian Dauner
MUC.DAI
University of Applied Sciences Munich
Munich, Germany
maximilian.dauner0@hm.edu

Ricky Liu
MUC.DAI
University of Applied Sciences Munich
Munich, Germany
ricky.liu@hm.edu

Gudrun Socher
MUC.DAI
University of Applied Sciences Munich
Munich, Germany
gudrun.socher@hm.edu

***Abstract*—Oral examinations and project presentations play a central role in higher education by assessing students' reasoning,** more directly probe students' reasoning processes, conditional understanding, and communication skills while offering

Privacy-preserving **Diarization-aware** **Rubric-grounded**

Why it matters

Oral exams become more important as written assignments are easier to automate with generative AI.

Core contribution

On-premise transcription, retrieval, and rubric-based feedback generation in one transparent workflow.

Main take-away

Strongest immediate value for formative support and documentation, not autonomous summative grading.

Motivation: Why Oral Assessment Needs Structured Support

- LLMs make written assignments easier to automate, increasing the relevance of oral exams and project presentations.
- Oral formats probe live reasoning, synthesis, and communication more directly than text-only submissions.
- In practice, oral assessment is difficult to scale: feedback is brief, documentation is weak, and grading can be subjective.
- These tensions are amplified in multilingual and diverse classrooms, where consistency and perceived fairness matter.

Assessment tension

Challenge

Time-constrained feedback, limited documentation, and subjective grading in oral formats.

Opportunity

Use speech and language technology to capture evidence, structure evaluation, and improve transparency.

Design requirement

Keep processing on institutional infrastructure and preserve explicit human oversight for final grades.

Research Objective and Main Contributions

Research objective: Evaluate whether a privacy-preserving, diarization-aware AI pipeline can support oral assessment in a transparent, useful, and reviewable way.

Contribution 1

A fully on-premise pipeline tailored to oral exams, thesis defenses, and project presentations.

Contribution 2

An empirical comparison of instructor rubric scores and pipeline rubric-constrained scores across milestones.

Contribution 3

A joint technical and pedagogical analysis of documentation, transparency, feedback quality, calibration, and human oversight.

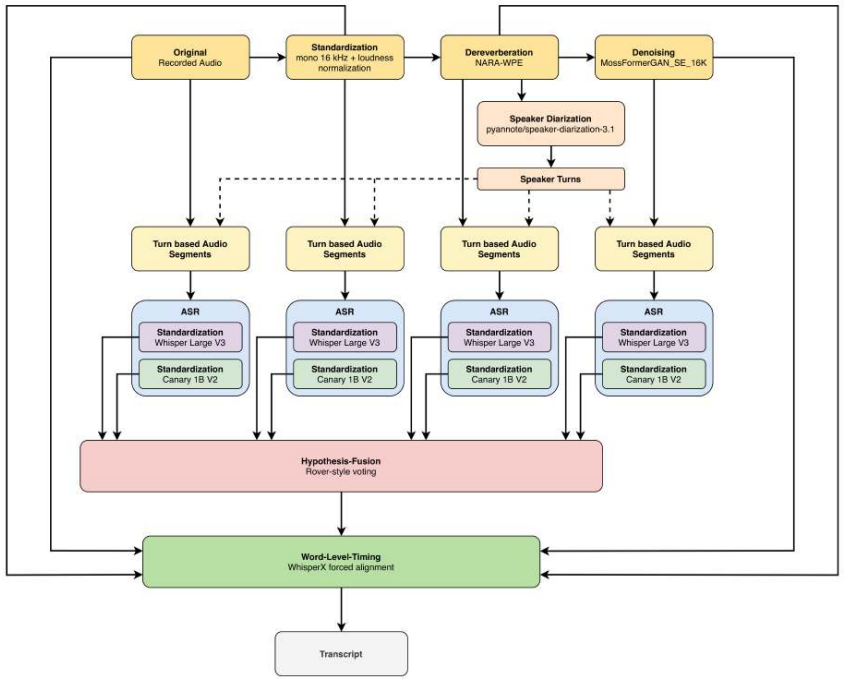
Scientific framing for the talk

Problem → System → Study design → Results → Interpretation → Future work

Audio Processing and Diarization-Aware Transcription

- Audio standardization: mono, 16 kHz, loudness normalization to reduce recording variability
- Optional dereverberation and single-channel speech enhancement to improve intelligibility
- Speaker diarization identifies who spoke when and creates speaker-homogeneous turns
- Two complementary ASR systems transcribe each turn and are fused with a ROVER-style voting scheme
- Forced alignment yields a speaker-attributed, time-aligned transcript for downstream evaluation

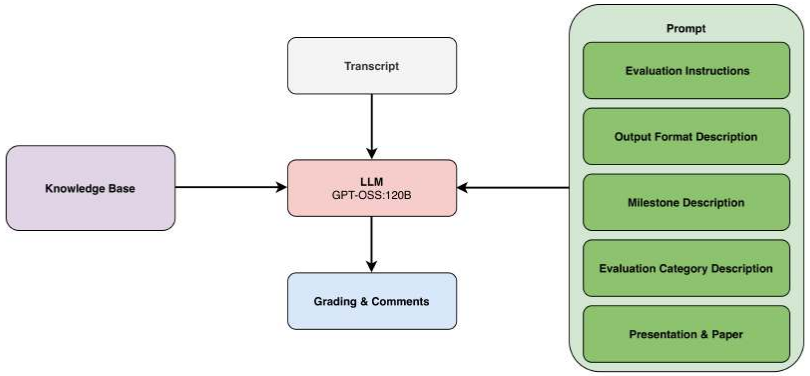
Key models: pyannote.audio · Whisper large-v3 · Canary-1B-v2 · WhisperX



ig. 1. Overview of the audio processing pipeline designed to maximize transcript fidelity for downstream assessment. Recordings are standardized (mono

Figure 1 from the paper: the pipeline maximizes transcript fidelity before grading.

Retrieval-Augmented, Rubric-Constrained Evaluation



ig. 2. Overview of the LLM document pipeline, comprising a retrieval knowledge base built from course materials, an extensive system prompt, and th iarization-aware transcript. The LLM produces rubric-aligned grades and detailed, actionable feedback.

or common educational file formats. For PDFs, native, nachine-readable text is extracted whenever possible, with an ptical character recognition (OCR) fallback only when pages ontain insufficient native text (as in scanned handouts or em- edded figures). This strategy improves coverage while avoid- ng unnecessary OCR-induced errors [43]. For slide decks,

reduce irrelevant context in the prompt, choices that have bee systematically studied in recent best-practice evaluations [36 [37]. Recent state-of-the-art variants further incorporate adap tive retrieval and self-critique mechanisms to decide whe retrieval is needed and to improve citation faithfulness, thereb reducing unsupported generations in knowledge-intensive set

Figure 2 from the paper: transcript + knowledge base + prompt → rubric-aligned grades and comments.

1. Document context extraction

Text, slide content, tables, and rubrics are extracted and cleaned for reliable evidence packaging.

2. Retrieval-augmented grounding

Course materials are chunked into a knowledge base so evaluation can be grounded in relevant instructional evidence.

3. Embedding-based key-point coverage

Expected concepts are compared against student responses to quantify whether core content was addressed.

4. Structured LLM output

The local evaluator returns rubric-level points, criterion-specific feedback, an overall summary, and warnings or uncertainty flags.

Milestones and Assessment Focus

M1 - Research Proposal

Focus

Framing the research problem

What students did

Students were assessed on early project framing, including research question, context, PRISMA, and presentation quality

Tested categories

- Research question / hypothesis
- Context & background
- PRISMA systematic review
- Presentation quality / slides

M3 - Methods Analysis

Focus

Developing and justifying the method

What students did

Students were assessed on methods, analysis logic, paper development, and in-class presentation

Tested categories

- Methods & analysis
- Paper development and written structure
- Presentation in class

M4 - Final Presentation

Focus

Presenting and defending the completed work

What students did

Students were assessed on the completed work, including paper quality, rigor, visuals, slides, and oral delivery

Tested categories

- Paper completeness & structure
- Content quality & rigor
- Writing quality & professionalism
- Figures, tables & visual communication
- PRISMA documentation & integration
- Presentation slides quality
- Presentation / oral delivery

Evaluation Setup and Study Design

Cohort and design

Graduate seminar · descriptive evaluation only · subset of N = 11 complete student cases · three assessed milestones: M1, M3, and M4.

Evaluator setup

Institutional GPU server with NVIDIA A100 accelerators · strict university examiner prompt · low temperature · structured schema for points, comments, summary, and warnings.

Milestone score structure

Milestone	Max pts	Criteria
M1	20	4
M3	15	3
M4	35	7

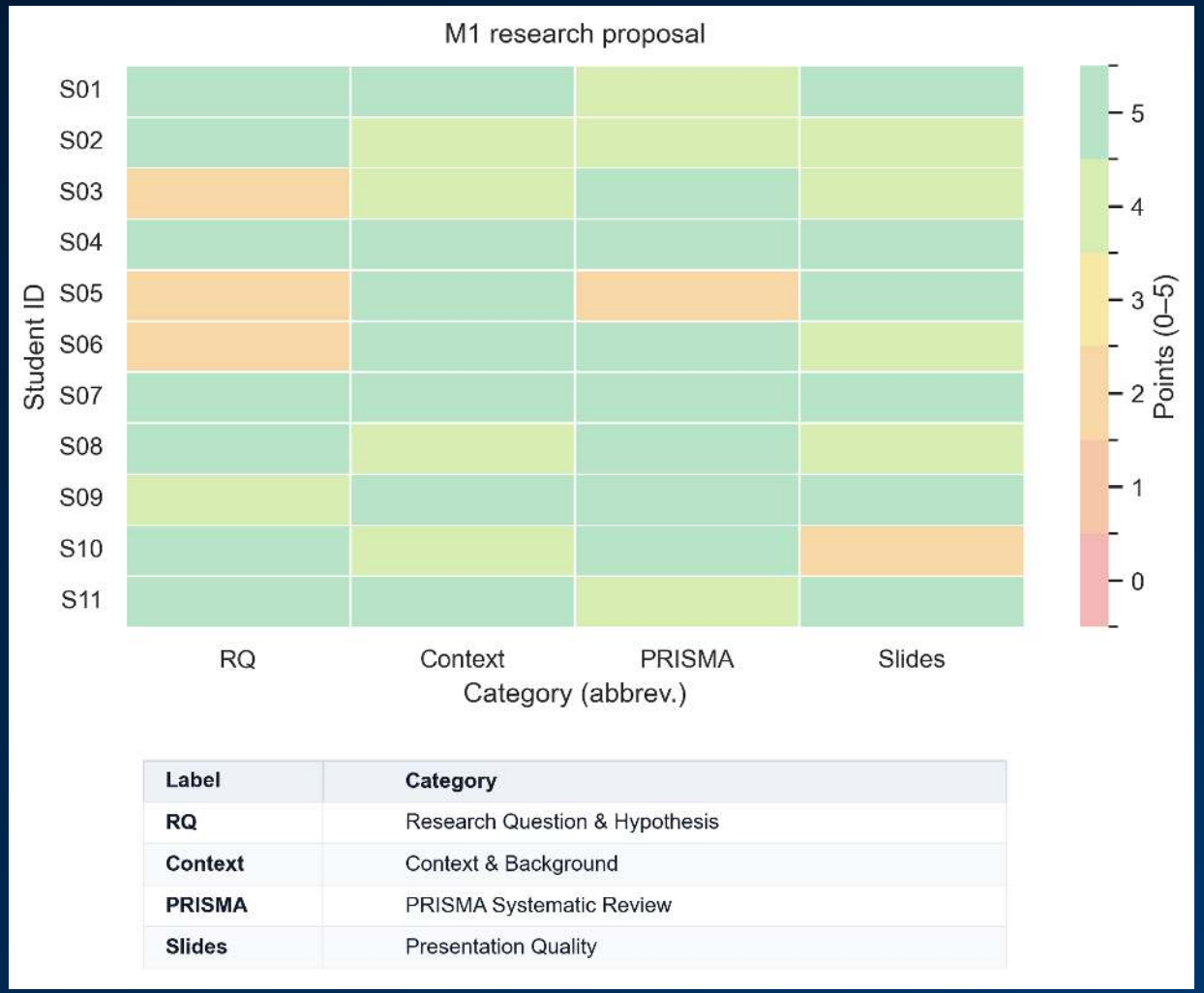
Human reference vs. AI evaluator

Instructor-assigned rubric scores were compared with pipeline-generated rubric-constrained scores from a locally hosted LLM evaluator.

Feedback meta-evaluation

An external LLM judge (GPT-5.2) scored feedback on a 0-5 scale for transparency, actionability, specificity, learning value, rubric alignment, tone, and overall helpfulness.

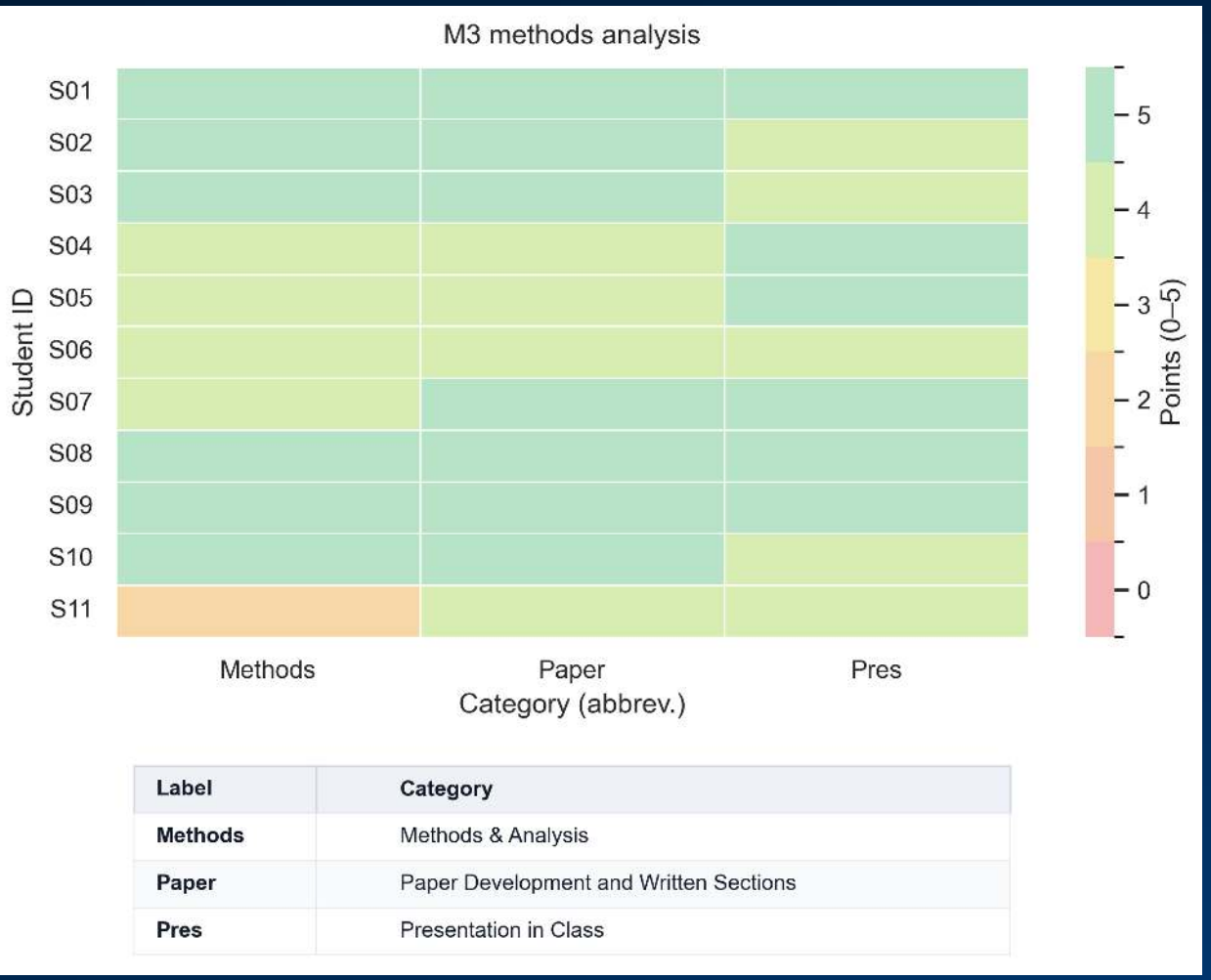
Instructor vs. Pipeline Rubric Scores (M1)



Label	Category
RQ	Research Question & Hypothesis
Context	Context & Background
PRISMA	PRISMA Systematic Review
Slides	Presentation Quality

Label	Category
RQ	Research Question & Hypothesis
Context	Context & Background
PRISMA	PRISMA Systematic Review
Slides	Presentation Quality

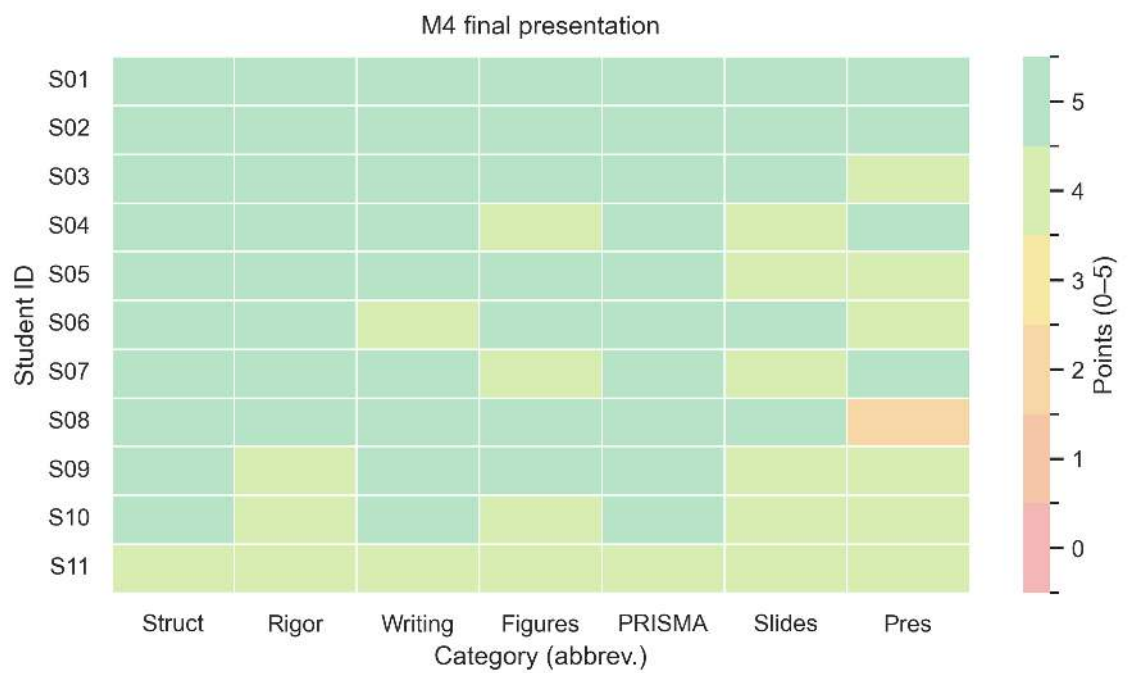
Instructor vs. Pipeline Rubric Scores (M3)



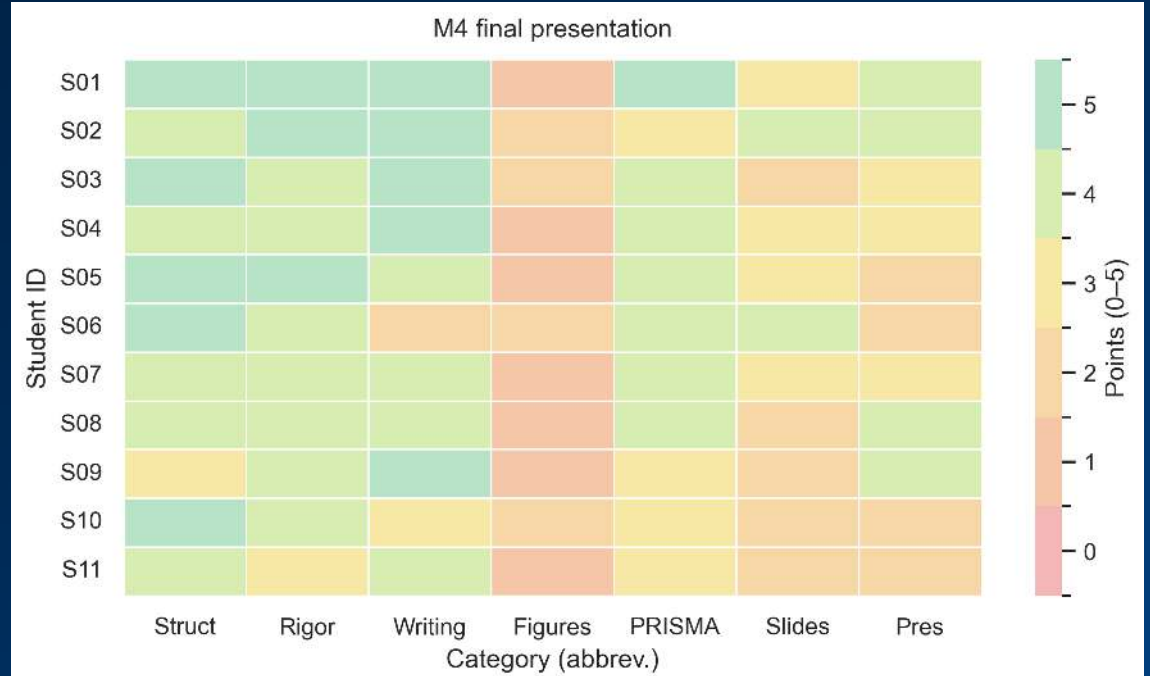
Label	Category
Methods	Methods & Analysis
Paper	Paper Development and Written Sections
Pres	Presentation in Class

Label	Category
Methods	Methods & Analysis
Paper	Paper Development and Written Sections
Pres	Presentation in Class

Instructor vs. Pipeline Rubric Scores (M4)



Label	Category
Struct	Paper Completeness and Structure
Rigor	Content Quality and Rigor
Writing	Writing Quality and Professionalism
Figures	Figures, Tables, and Visual Communication
PRISMA	PRISMA Documentation and Integration
Slides	Presentation Slides Quality
Pres	Presentation



Label	Category
Struct	Paper Completeness and Structure
Rigor	Content Quality and Rigor
Writing	Writing Quality and Professionalism
Figures	Figures, Tables, and Visual Communication
PRISMA	PRISMA Documentation and Integration
Slides	Presentation Slides Quality
Pres	Presentation

Average Rubric Scores Across Milestones

Average total scores

Milestone	Instructor	Pipeline	Δ
M1	17.5	13.6	**-3.9**
M3	13.5	10.8	**-2.7**
M4	32.6	23.5	**-9.1**

Interpretation

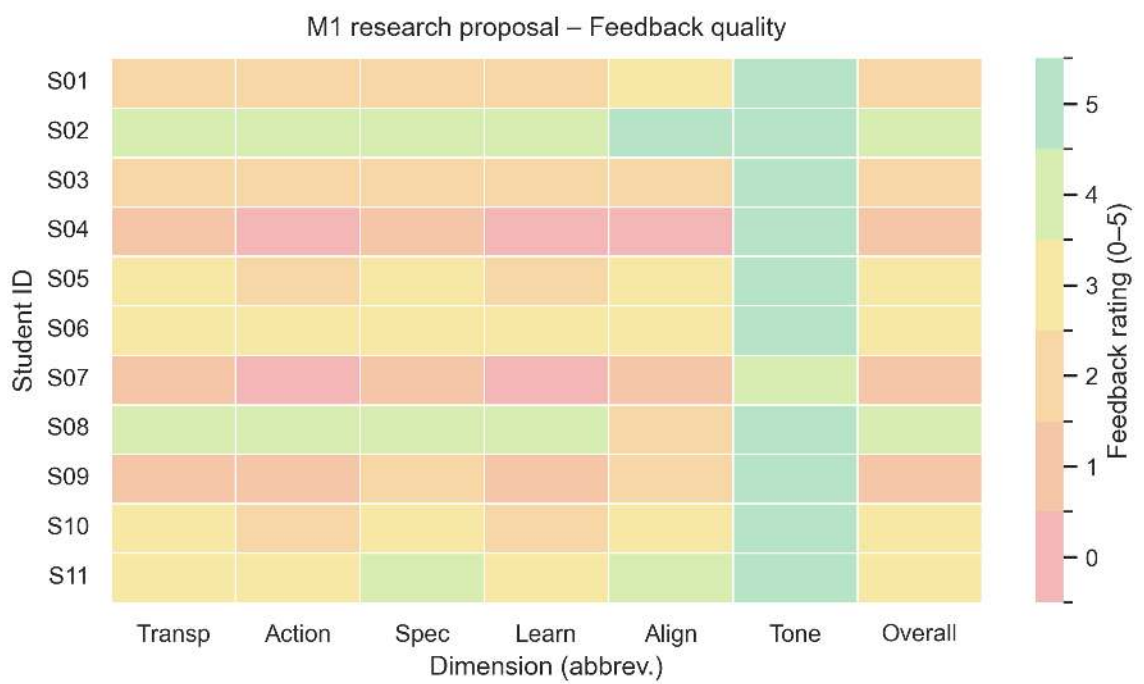
The gap is present in all three milestones and becomes largest in M4, indicating a systematic conservative scoring tendency of the pipeline.

(a) M1: Research proposal. Instructor (left) vs. pipeline (right) rubric scores; larger gaps mainly in delivery/slide-related criteria.

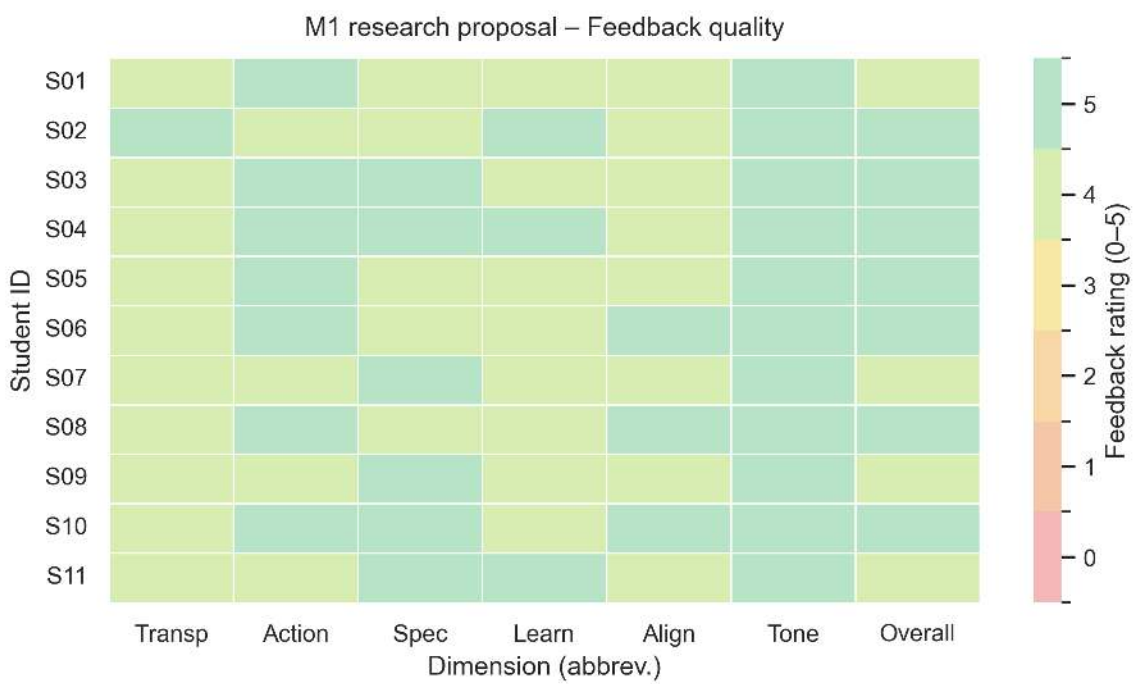
(b) M3: Methods analysis. Instructor (left) vs. pipeline (right) rubric scores; pipeline tends to score more conservatively overall.

Figure 3: instructor (left) vs. pipeline (right) heatmaps across M1, M3, and M4.

Feedback Quality Ratings: Instructor vs. Pipeline (M1)



Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)



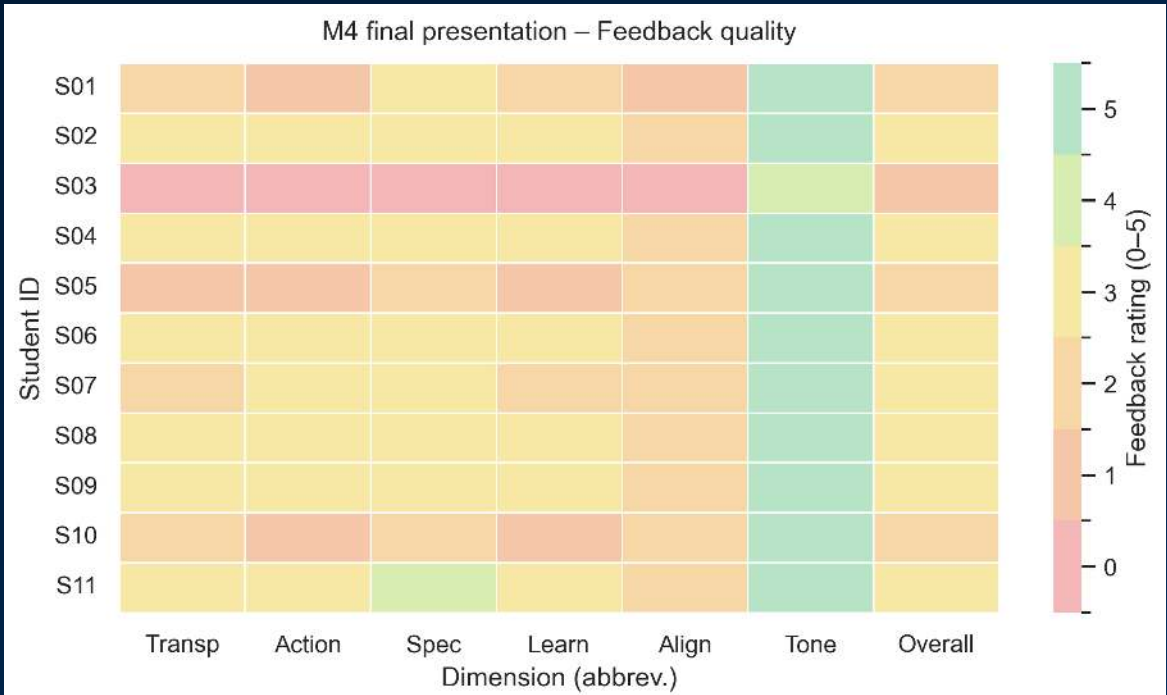
Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)

Feedback Quality Ratings: Instructor vs. Pipeline (M3)

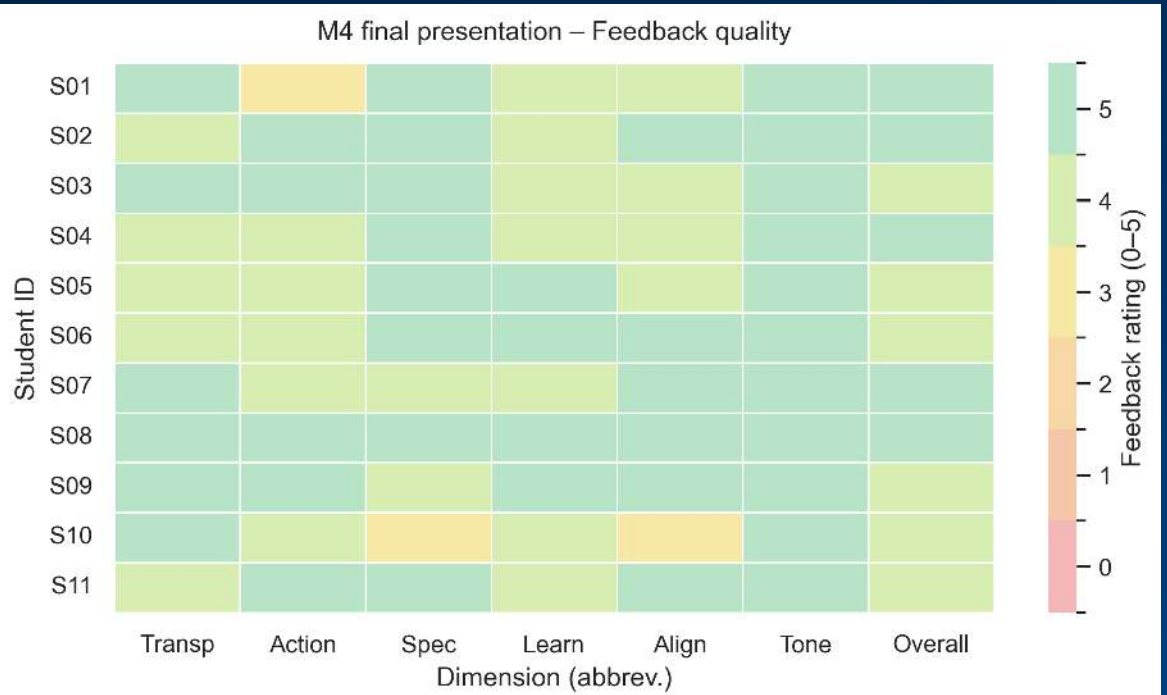
Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)

Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)

Feedback Quality Ratings: Instructor vs. Pipeline (M4)

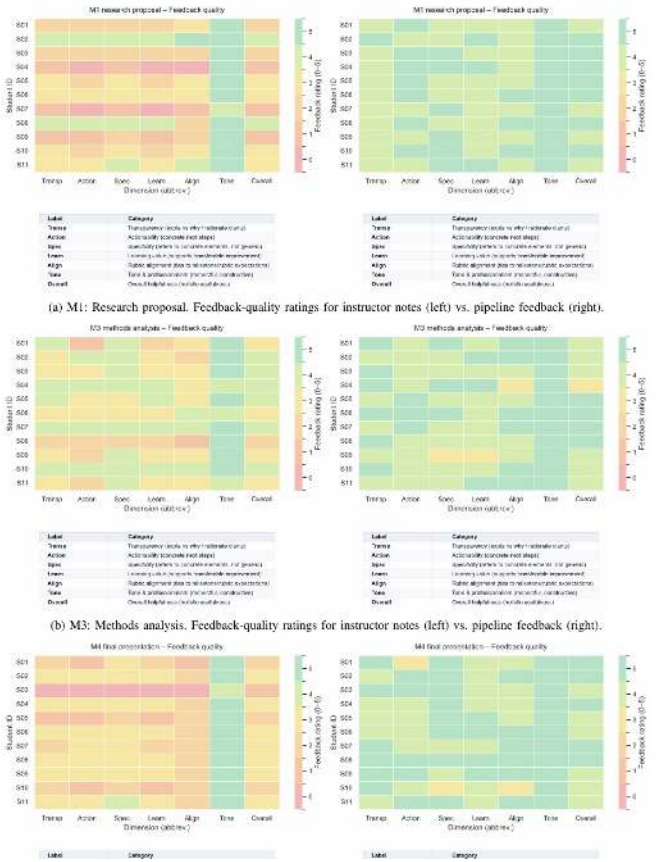


Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)



Label	Category
Transp	Transparency (explains why / rationale clarity)
Action	Actionability (concrete next steps)
Spec	Specificity (refers to concrete elements, not generic)
Learn	Learning value (supports transferable improvement)
Align	Rubric alignment (ties to milestone/rubric expectations)
Tone	Tone & professionalism (respectful, constructive)
Overall	Overall helpfulness (holistic usefulness)

Feedback Results and Interpretation



(a) M1: Research proposal. Feedback-quality ratings for instructor notes (left) vs. pipeline feedback (right).

(b) M3: Methods analysis. Feedback-quality ratings for instructor notes (left) vs. pipeline feedback (right).

Figure 4: feedback-quality heatmaps for instructor notes (left) and pipeline feedback (right).

Highest-rated dimension

Tone and professionalism remain consistently high for both human and AI-generated feedback.

Pipeline strength

Pipeline feedback is typically judged as more specific and more actionable than brief instructor notes; tone remains high for both.

Important nuance

Judge scores are descriptive tendencies, not definitive evidence of superiority or grading validity.

Conclusion: What the Study Shows, Limits and Future Work

What this work shows

A fully on-premise, human-in-the-loop decision-support pipeline can enhance oral assessment by improving transcript fidelity, documentation, and rubric-linked feedback.

What it does not show

The current evidence does not justify autonomous high-stakes grading: the pipeline is stricter overall and weakest on visually grounded and delivery-oriented criteria.

Future work

Compare several local LLMs, including the Qwen 3.5 series, with respect to scoring behavior, feedback quality, and calibration. Extend the study to multiple lectures to examine robustness and generalizability across cohorts and course settings.

Thank you for your attention!

Maximilian Dauner, Ricky Liu, Gudrun Socher

Hochschule München University of Applied Sciences
80335 Munich, Germany

Maximilian Dauner
PhD-Student
maximilian.dauner0@hm.edu

IEEE EDUCON 2026

HM

MUC.DAI
Munich Center for
Digital Sciences and AI