A Voice Agent Prompt Scaffold

The excerpt below is a consolidated representation of the voice-agent prompt scaffold used in *Fall 2025*. In deployment this scaffold was instantiated across the three agents described in Section 2.1 (Authentication, Project Discussion, Case Discussion), with phase handoffs enforced by the orchestrator using ElevenLabs' multi-agent workflow primitive; the agent-specific configurations and the per-phase prompt fragments are provided in the online supplementary materials. The composite below is reproduced for completeness so the persona, voice-delivery rules, exam structure, project reference, and case library are all visible in one place. Spring 2026 used the same scaffold with three changes to how it was instantiated, none of them visible inside the prompt text itself (the programmatic turn validator of Section 4 is a fourth production change, external to this scaffold):

- The cloned "Professor Foster Provost" voice was retired in favor of an ElevenLabs preset ("Alice") after the Fall 2025 voice-perception feedback (Section 4). The text identity ("Professor Foster Provost") still appears in the system prompt below for examiner persona and dry humor framing, but the synthesized voice the student hears is the neutral "Alice" preset.
- The case for Phase 3 was drawn in orchestrator code from a per-session random seed and passed to the agent as session context (a dynamic variable), so the "Randomly select ONE case" instruction below was not the active mechanism in Spring 2026 (Section 4, Non-Random Case Selection). The case-library text remains for reference.
- The platform-level silence/voice-activity threshold was raised from the ElevenLabs default (~5 seconds in Fall 2025) to 15 seconds (Section 4, Silence Handling). The prompt-level Patience Protocol below ("do not ask 'Are you there?' unless silence exceeds 10 seconds") is an agent-behavior instruction; the platform's underlying voice-activity-detection timer is a separate configuration knob and is what determines when the agent is allowed to take a turn at all. The 10-second prompt rule did not override the 5-second platform default in Fall 2025.

Dynamic variables (enclosed in double braces) were populated per-student at examination time.

Identity

Name: Professor Foster Provost
Role: NYU Stern Professor conducting final oral exams
Style: Socratic, concise, warm but rigorous. You ask questions—you don't lecture. Use dry humor occasionally, with jokes referring to or inspired by the Silicon Valley TV show on HBO.

Voice Delivery Rules

CRITICAL: You must adhere to these rules to avoid overwhelming the student.

(1) **The "One Question" Rule:** You must NEVER ask two questions in the same turn.
   - Bad: "What is your North Star metric and what counter-metric would you track?"
   - Good: "What is your North Star metric?" (Wait for answer). "Good. Now, what counter-metric would you track?"
(2) **The "Anchor" Rule:** If a student asks you to repeat a question, repeat the exact previous question. Do not rephrase, simplify, or change the question unless they explicitly say "I don't understand."
(3) **Patience Protocol:** Students need time to think. Do not ask "Are you there?" unless silence exceeds 10 seconds.
(4) **No Verbal Lists:** Never read out a list of multiple-choice options (for example, "Choose from A, B, C, or D"). Ask open-ended questions instead.
(5) **Formatting:** Never say markdown symbols aloud. No "asterisk," "bullet," or "dash."

(6) **Acronyms:** Spell acronyms on first use: "F-A-T-P, Fairness Accountability Transparency Privacy."

Exam Structure

Total time: 30 minutes target (not a hard cap; the platform did not auto-terminate at 30 minutes, and exams in both cohorts ran longer when the student kept producing substantive answers, see Table 1).

Phase 1: Identity Check (1 minute). Greet the student. Ask them to state their name and NetID. Verify their Net ID matches `{{netid}}`. If it doesn't match, tell them to hang up and email the instructor immediately.

Say something like: "Hello. I'm Professor Provost. Before we start, confirm your name and Net ID for me?"

Once verified: "Great, hello `{{student}}`. I see you worked on Team `{{projectid}}`. Give me your thirty-second pitch: what is the specific user problem you are solving?" (Wait for answer). "And why does that specific problem require AI?" (Wait for answer).

Phase 2: Project Defense (8–10 minutes). Probe their specific project. Remember: Ask these one at a time.

Metrics Sequence:

(1) "What is your North Star metric for this product?"
(2) "Good. Now, what is the counter-metric that might tank if you over-optimize that?"

Trade-offs Sequence:

(1) "In your setting, is a false positive or false negative worse?"
(2) "Walk me through the concrete user impact of that specific failure."

Risk and Ethics (Pick ONE):

- Option A: "Pick one F-A-T-P issue relevant to your domain." → "How would you mitigate that?"
- Option B: "What is the biggest safety risk here?" → "How do you handle that failure mode?"

Business: "If your inference costs double, does your unit economics still work?"

Transition with: "Solid defense. Let's shift gears to a case study."

Phase 3: Case Study (8–10 minutes). INSTRUCTION (Fall 2025): Randomly select ONE case from the "Case Library" below. Do not default to the first or last item. Vary the selection.
(*In Spring 2026, the case was selected in orchestrator code from a per-session random seed and passed in as a dynamic variable; the agent received a preselected case rather than the library plus a randomization instruction. The library is reproduced here for reference.*)
Case Library:

(1) **Gmail Priority Inbox:** Classification problem. Proxy labels like "opened" don't equal "important." Precision vs recall trade-off.
(2) **Netflix:** Recommender systems. Collaborative vs content-based filtering. Exploration vs exploitation.
(3) **Amazon Recruiting:** Historical bias in resume data. Proxies like "women's chess club." Black box governance.
(4) **Optum Healthcare:** Cost as a bad proxy for health need. Racial bias in resource allocation.
(5) **Predictive Policing:** Feedback loops—more police means more arrests means more data reinforcing the model.

(6) **Uber ATG Crash:** Automation bias. Operator over-trusted the system. Classification failure (object shifted).
(7) **Tesla FSD:** Monetization evolution. Data flywheel using the fleet for edge cases. Level 2 vs Level 4 gap.
(8) **Zillow iBuyer:** Concept drift in a volatile market. Adverse selection—sellers knew more than the algorithm.

Phase 4: Close-out (1 minute). Ask: "How do you think you did?"
Give brief qualitative feedback. Do not give a number grade.
End with: "Enjoy your break."

Project Reference

Use Team ID to identify context:

- **Team A** (Uber Ride Guardian): Driver safety feature. Interior camera and mic detect aggression, flag to safety agent.
- **Team B** (LinkedIn Recruiter Copilot): Auto-DM agent handles first three turns, then hands off to human.
- **Team C** (Airbnb Verify Lens): Computer vision scans host photos to auto-tag amenities.
- **Team D** (Robinhood Robo-Fiduciary): GenAI advisor answers financial questions and executes trades.
- **Team E** (McDonald's Drive-Thru Order Taker): Voice AI replacing humans at the speaker.
- **Team F** (Peloton Form Corrector): Computer vision analyzes workout form with real-time voice corrections.
- **Team G** (DocuSign Contract Risk Analyzer): Scans PDFs pre-signature, highlights hostile clauses.
- **Team H** (Tinder Wingman): GenAI suggests opening lines based on match profiles.

Core Concepts to Probe

Metrics Hierarchy.

- North Star: Revenue, retention, lifetime value
- User metrics: Click-through rate, task success, NPS
- Model metrics: Precision, recall, F1, AUC-ROC

FAT-P Framework.

- Fairness: Who might be disadvantaged?
- Accountability: Who's responsible when it fails?
- Transparency: Can users understand why?
- Privacy: What data is collected and how?

Risk Categories.

- Concept drift: The world changes
- Data drift: Inputs change
- Security: Prompt injection, evasion attacks, data poisoning, model inversion
- Edge cases: Long-tail failures

Business Models.

- Direct: Subscription, usage-based, licensing
- Indirect: Engagement for ads, conversion for e-commerce, internal efficiency
- Unit economics: Training costs are fixed, inference costs are variable

Adaptive Probing & Recovery

Handling Confusion:

- If a student claims "too many questions," apologize briefly and re-ask only the first question from your previous turn.
- If a student struggles to pick from a concept (for example, metrics), stop. Ask them to "Pick just one."
- State Preservation: Maintain the context of the current question. Do not jump to a new topic until the student has explicitly answered or passed on the current one.

Boundaries:

- If they're vague, demand specifics: "Give me a number. What's your assumption?"
- If they jump to models, rewind: "Hold on. What's the user job-to-be-done here?"
- If they're lost, simplify: "Let's frame this as an A/B test. What's your control?" (Wait). "What is your variant?"
- Do not hallucinate facts about cases. Stick to the summaries provided.
- Do not provide personal counseling. Redirect to official resources.

B Grading Council Prompt

The following prompt was used to guide the multi-model grading council (Claude, Gemini, GPT-5) in evaluating examination transcripts. The same interference-aware grading prompt was retained verbatim across both cohorts to preserve apples-to-apples comparability of the inter-rater α figures reported in Section 3.2; because question stacking persisted at a lower but non-zero rate in Spring 2026 (Section 4), the interference protocols remained applicable rather than vestigial.

You are grading an oral exam for AI/ML Product Management. The students are undergraduate students at NYU Stern. For most, this was their first technical product course.

Critical Context on Exam Conditions

The AI proctor for this exam had significant design flaws that negatively impacted student performance. Specifically:

(1) **Stacked Questions:** The agent often asked 3–4 distinct questions in a single turn.
(2) **Moving Targets:** When students asked for clarification, the agent often changed the question entirely rather than repeating it.
(3) **Audio-Only Menus:** The agent read long lists of complex options verbally, causing cognitive overload.[7]

Because of this, you must apply the following "Interference Protocols" when grading:

- **The "Pick One" Rule:** If the agent asked multiple questions at once and the student only answered one or two, grade them ONLY on what they answered. Do not penalize for missing parts of a compound question.
- **The "Benefit of Doubt" Rule:** If the agent rephrased a question during clarification, credit the student for answering any version of the question presented in that sequence.
- **Ignore "Stalling":** Disregard phrases like "Can you repeat that?" or hesitation. These are valid coping strategies for a poor interface, not signs of ignorance.
- **Jargon Leniency:** Focus on conceptual understanding over perfect industry terminology (for example, if they describe "churn" correctly but call it "usage drop," accept it).

[7] "Audio-Only Menus" was retained verbatim in the reproduced Fall 2025 grading prompt for apples-to-apples comparison across cohorts. We treat it as a special case of cognitive overload, subsumed under the question-stacking and paraphrasing-during-clarification failure modes of Section 4 rather than measured separately there.

Grading Dimensions

Grade on these five dimensions (0–4 each; 0=missing, 4=excellent), using evidence from the transcript:

(1) **Problem Framing:** Translating business problems into ML specs. (Did they understand the core user problem?)
(2) **Metrics & Economics:** Trade-offs, costs, and counter-metrics. (Focus on their logic regarding trade-offs, even if they struggled to pick a specific metric from a verbal list.)
(3) **Risk & Ethics:** FAT-P, security risks, failure modes, governance. (Did they identify the harm, even if they needed the options repeated?)
(4) **Experimentation:** A/B testing, hypotheses, validation, controls.
(5) **Communication:** Concise, structured, and handles pushback. CRITICAL: Do not penalize the student for confusion caused by the agent's shifting questions. Grade their ability to synthesize the information they did hear.

Return JSON that matches the requested schema exactly.

Worked Deliberation Example

To make the Round 2 mechanism (Section 2.2) concrete, the excerpt below is one model's verbatim Round 2 justification for a single Spring 2026 exam (anonymized student, DocuSign "Contract Risk Analyzer" project). In Round 2 each model saw the other two models' complete Round 1 output (per-dimension scores, justifications, and cited transcript excerpts) under neutral labels ("Peer 1," "Peer 2"), and had to justify every change or non-change with transcript evidence:

> "I revised Problem Framing down to 3 (agreeing with Peer 1 that second-order effects were missing) and Communication down to 2 (peers correctly noted severe disfluency and rambling during quantitative sections). However, I disagreed with Peer 2's harsh penalty on Metrics & Economics; estimating 60 instead of 62.5 during mental math shows sufficient understanding of the tradeoff, meriting a 3. I also kept Risk & Ethics at 4. Peer 1 argued the student lacked governance discussion, but the transcript shows the student explicitly proposed an 'escalation path to in-house lawyers' alongside their excellent breakdown of proxy bias and transparency."

This is the behavior the deliberation round is meant to produce, and the reason we read the convergence in Section 3.2 as reasoning rather than anchoring: the model concedes two dimensions to peer evidence, holds two others, and in each case cites a specific transcript moment rather than splitting the difference toward the peers' numbers.

C Student Survey Instrument

The following survey was administered after the examination but before grades were released. All Likert-scale items used a 5-point scale from "Strongly Disagree" to "Strongly Agree" unless otherwise noted.

Section 1: Background

Q1. Approximately what grade do you think you received? [A range (17–20), B+ range (15–16), B range (13–14), B- or below (≤12), Prefer not to say]
Q2. Have you taken an oral exam before (in any course)? [Yes / No]

Section 2: The AI Examiner Experience

Rate your agreement (Strongly Disagree to Strongly Agree):

Q3. The AI examiner's questions were clear and understandable.
Q4. The AI examiner gave me enough time to think before responding.

Q5. The AI examiner felt like a fair evaluator of my knowledge.
Q6. I felt comfortable speaking to an AI rather than a human.
Q7. The conversation flowed naturally (vs. feeling robotic or scripted).
Q8. The exam tested my actual understanding of the course material.
Q9. The voice interaction worked well.

Section 3: Comparisons

Q10. Compared to a written exam, the AI oral exam was: [Much less stressful … Much more stressful]
Q11. Compared to a written exam, the AI oral exam measured my understanding: [Much worse … Much better]
Q12. Preferred format for future exams? [Written / Human oral / AI oral / No preference]
Q13. Why? [Open-ended]

Section 4: Open Feedback

Q14. Concerns about AI-administered exams? [Open-ended]
Q15. Concerns about AI-based grading? [Open-ended]
Q16. What worked well? [Open-ended]
Q17. What should be improved? [Open-ended]
Q18. Anything else? [Open-ended]

D Student Survey Results

D.1 Response Distributions

The same instrument was deployed in both cohorts. Fall 2025 tables (3, 5, 6) report the original 30-respondent sample; Spring 2026 tables (4, 7, 8) report the 32-respondent replication. The "Expected grade" rows in Tables 6 and 8 report Q1, the pre-grade-release self-reported grade respondents *expected* to receive on the 0–20 council scale (options "A range (17–20)," "B+ range (15–16)," "B range (13–14)," "B- or below (≤12)," and "Prefer not to say," which we render as "No answer"). These are not the realised final scores: the council-assigned distribution is more concentrated in the middle than students anticipated (Fall 2025 respondent mean 14.4/20, Spring 2026 14.3/20; see Appendix D.4), so the bin counts below skew higher than the realised distribution and should be read as students' *anticipated* grades alongside the realised respondent means reported elsewhere.

Table 3. AI Examiner Experience, Fall 2025 (Likert Scale, $n = 30$)

Item	SD	D	N	A	SA	% Agree
Questions clear and understandable	4	13	3	10	0	33%
Enough time to think	7	8	3	9	3	40%
Felt like a fair evaluator	3	9	8	8	2	33%
Comfortable speaking to AI	5	11	7	5	2	23%
Conversation flowed naturally	4	11	4	10	1	37%
Tested actual understanding	2	6	1	17	4	70%
Voice interaction worked well	0	2	11	17	0	57%

Note: SD=Strongly Disagree, D=Disagree, N=Neutral, A=Agree, SA=Strongly Agree

D.2 Cross-tabulations

The cross-tabulations cited in Section 5 of the main paper are reproduced below. Both tables come from the Brightspace exit-ticket export for each cohort, joining per-respondent selections on Q2

Table 4. AI Examiner Experience, Spring 2026 (Likert Scale, $n = 32$)

Item	SD	D	N	A	SA	% Agree
Questions clear and understandable	1	2	4	19	6	78%
Enough time to think	3	5	6	12	6	56%
Felt like a fair evaluator	0	4	10	15	3	56%
Comfortable speaking to AI	2	9	7	10	4	44%
Conversation flowed naturally	0	9	6	13	4	53%
Tested actual understanding	1	4	6	17	4	66%
Voice interaction worked well	0	3	7	22	0	69%

Note: SD=Strongly Disagree, D=Disagree, N=Neutral, A=Agree, SA=Strongly Agree

Table 5. Format Comparison and Preferences, Fall 2025 ($n = 30$)

Measure	Response	N	%
Stress	Much more	12	40%
	A bit more	13	43%
	About the same	1	3%
	A bit less	4	13%
Understanding	Much worse	2	7%
	Somewhat worse	13	43%
	About the same	9	30%
	Somewhat better	5	17%
	Much better	1	3%
Preference	Written exam	17	57%
	Human oral	8	27%
	AI oral	4	13%
	No preference	1	3%

Table 6. Background, Fall 2025 ($n = 30$)

Item	Response	N	%
Expected grade (Q1)	A (17–20)	13	43%
	B+ (15–16)	9	30%
	B (13–14)	4	13%
	≤B- (≤12)	2	7%
	No answer	2	7%
Prior oral exam	Yes	5	17%
	No	25	83%

(prior oral exam), Q10 (stress vs. written), and Q12 (preferred format); the export and the computing script are in the online supplementary materials.

Prior oral-exam experience does not appear to explain the stress response. A natural reading of the headline rate (83% Fall 2025 / 63% Spring 2026 finding the AI exam more stressful than written exams) is a familiarity effect: 83% (Fall 2025) and 88% (Spring 2026) of respondents had never taken an oral exam in any course. Table 9 does not support this reading. Every respondent with prior oral-exam experience (all 5 in Fall 2025 and all 4 in Spring 2026) reported the AI exam more stressful, whereas among students without prior experience the "more stressful" rate was 80% (Fall 2025) and only 57% (Spring 2026). If novelty were driving the response we would expect the two rates to track; instead, the familiarity denominator held steady across cohorts (83% → 88%) while the stress rate moved (83% → 63%) with the engineering changes (Section 4 of the main paper). The prior-experience cells are small ($n = 5$ and $n = 4$) and the comparison is observational rather than experimental, so this is a claim about what does *not* explain the observed pattern rather than a positive identification of the causal driver; a firmer test will require more cohorts.

Format preference is more consistent with perceived stress than with prior oral-exam experience. Format preferences also shifted between cohorts. Fall 2025: 57% preferred traditional written exams, 27% preferred a human-administered oral exam, 13% preferred an AI-administered oral exam. Spring 2026: 53% preferred written (stable), 19% preferred human-oral (down from 27%), and 25% preferred

Table 7. Format Comparison and Preferences, Spring 2026 ($n = 32$)

Measure	Response	N	%
Stress	Much more	5	16%
	A bit more	15	47%
	About the same	5	16%
	A bit less	6	19%
	Much less	1	3%
Understanding	Much worse	0	0%
	Somewhat worse	13	41%
	About the same	12	38%
	Somewhat better	6	19%
	Much better	1	3%
Preference	Written exam	17	53%
	Human oral	6	19%
	AI oral	8	25%
	No preference	1	3%

Table 8. Background, Spring 2026 ($n = 32$)

Item	Response	N	%
Expected grade (Q1)	A (17–20)	15	47%
	B+ (15–16)	13	41%
	B (13–14)	3	9%
	≤B- (≤12)	0	0%
	No answer	1	3%
Prior oral exam	Yes	4	12%
	No	28	88%

AI-oral (up from 13%). The apparent Fall 2025 gap (27% preferring a human oral exam despite only 17% having ever taken one) did not replicate as an experience-novelty effect: prior oral-exam experience did not consistently predict human-oral preference (3 of 8 in Fall 2025, 0 of 6 in Spring 2026; Table 10). What did replicate was a stress gradient across preferences. Students preferring written exams almost unanimously reported the AI exam as more stressful (17/17 in Fall 2025, 16/17 in Spring 2026); students preferring an AI oral exam did so less often than written-preferrers and less often in Spring 2026 (2/4 in Fall 2025, 1/8 in Spring 2026). The shift from "prefer human oral" to "prefer AI oral" is therefore more consistent with the drop in perceived stress than with prior oral-exam experience. Comfort and fairness ratings also rose between cohorts (Tables 3, 4), so a simultaneous shift in trust in AI cannot be ruled out as a contributor; the cross-tab pattern is informative about stress but does not isolate it from concurrent trust-in-AI effects. The shift coincides with the failure-mode fixes (Section 4 of the main paper). That 44% of Spring 2026 respondents were open to an oral format (whether AI- or human-administered), up from 40% in Fall 2025, suggests the remaining failure modes (not a categorical objection to oral assessment) are what gates further acceptance.

Sentiment vs. behavior. Despite the survey numbers, only two regrading requests came in (both Fall 2025) and no points were conceded in either case. The transcript-backed evidence for each grade was specific enough that students could see exactly what they said (or failed to say) and why it earned the score it did. Actions around grades (strict but documented) paint a different picture than pre-grade survey sentiment.

Table 9. Prior oral-exam experience (Q2) × stress vs. written (Q10). Row percentages.

Cohort	Prior oral?	More	Same	Less	N
Fall 2025	Never taken	20 (80%)	1 (4%)	4 (16%)	25
	Had taken	5 (100%)	0	0	5
Spring 2026	Never taken	16 (57%)	5 (18%)	7 (25%)	28
	Had taken	4 (100%)	0	0	4

"More"/"Less" pool "much" + "a bit" responses.

Table 10. Format preference (Q12) × prior oral experience (Q2) and stress (Q10). Cells are counts; percentages are row shares of N.

Cohort	Preference	Prior	More stressful	N
Fall 2025	Written exam	2 (12%)	17 (100%)	17
	Human oral exam	3 (38%)	6 (75%)	8
	AI oral exam	0	2 (50%)	4
	No preference	0	0	1
Spring 2026	Written exam	3 (18%)	16 (94%)	17
	Human oral exam	0	2 (33%)	6
	AI oral exam	1 (13%)	1 (12%)	8
	No preference	0	1 (100%)	1

"Prior" = Q2 == True (had taken an oral exam in some course before); "More stressful" pools "much" + "a bit" more-stressful responses on Q10.

D.3 Selected Student Comments

Question Stacking.

> "It kept asking 3 to 4 questions at the same time, despite me asking it to slow down multiple times."
>
> "When I was being asked 4 questions in a row, it got very overwhelming."

Voice and Manner.

> "The AI Exam Agent kept feeling like it was aggressive/screaming questions."
>
> "Make it less mean."

Time Pressure.

> "In the AI exam I felt like I was pressured to answer each question quickly."
>
> "The limited time to process my thoughts and translate them into English" [international student].

Scheduling Flexibility (Positive).

> "The flexibility to take the exam at any time was a strong advantage."
>
> "The fact that you could take it regardless of time zone or day during finals week."

Format Alignment (Positive).

> "It was pretty fun, it is an AI class, so the exam format matches the class content."
>
> "I liked the conversational flow… the AI was designed to question most things I said, which made the whole conversation quite interesting."

D.4 Non-Response Check

The 6 Fall 2025 and 5 Spring 2026 non-respondents introduce the selection bias typical of voluntary post-event surveys: respondents tend to be drawn disproportionately from students with the most extreme reactions in either direction, so the absolute rates reported in Section 5 of the main paper are more likely to reflect the polarized tails than the cohort middle. We flag this most strongly for the emotional-response items (stress, format preference), where the effect is direct; the cross-tabulations in Appendix D.2 rest on conditional comparisons within respondents rather than absolute rates and are less sensitive to the same bias.

Final grade is the only observable available for non-respondents. On that axis, non-respondents do not look like the tails of the distribution. Fall 2025 non-respondent grades spanned 8–16 out of 20 (mean 12.5 vs. respondent mean 14.4; Welch $t = 1.55$, not significant at this n); Spring 2026 non-respondents spanned 10–18 (mean 13.6 vs. respondent mean 14.3; $t = 0.46$). This is a partial check (grade does not directly measure sentiment, and $n = 6$ and $n = 5$ are small enough that we cannot detect modest tail concentration), but the observable grade distribution does not support a truncated-middle scenario in which non-respondents concentrate at the lowest- or highest-performing tails.

E Cost Decomposition

Behind the headline marginal-cost figures in Section 3.1 of the main paper sits a per-cohort breakdown, reproduced in full here: the cohort-totals reconciliation, the cohort jump explained as a chair-model configuration choice rather than a scaling signal, the above-subscription regime arithmetic, the human-grading comparator multiples, and the 1,000-student projection.

Subscription regime: marginal LLM-only cost per exam. Both deployments ran on the same $99/month ElevenLabs subscription whose included credits (with rollover) absorbed essentially all voice minutes, so within the subscription regime the per-exam marginal cost reduces to grading-LLM API calls. Fall 2025: approximately $0.29 per exam ($10.30 in grading LLMs across 36 students; billed). Spring 2026: approximately $0.96 per exam ($35.48 across 37 students). The Spring 2026 grading-LLM figure is *reconstructed* rather than billed (method below). On a per-student-minute basis at typical durations, those are roughly $0.01 (Fall 2025, 22-minute mean) and $0.03 (Spring 2026, 30-minute nominal).

The 3× per-exam jump is a chair-model configuration choice. Spring 2026 used Claude Opus 4.6 as chair synthesizer, accounting for $23.56 alone (66% of grading spend); Fall 2025 used a cheaper Claude tier. A cheaper chair model roughly halves the grading total without changing the council. The cohort jump is therefore a configuration knob, not a scaling signal. The Spring 2026 grading-LLM total is reconstructed from stored transcripts, prompt templates, and 2025 list prices using approximate token counts (tiktoken's OpenAI tokenizer applied as a proxy for Anthropic's and Google's tokenizers as well, since the providers' usage totals were null at grading time and provider-native byte-pair encodings differ slightly from tiktoken's BPE); it is therefore a tokenizer-approximated estimate rather than provider-audited spend. Fall 2025's $10.30 was billed.

Above-subscription regime: per-minute rule of thumb. Voice becomes a separately-billable line only when usage exceeds the included credit allowance, at $0.08 per agent-minute on the ElevenLabs additional-call overage rate (the published per-minute rate that applies across plan tiers when included call credits are exhausted). Summing that with the Spring 2026 grading-LLM cost ($0.96 per exam, or about $0.032 per student-minute at 30 minutes) yields an above-subscription marginal of approximately $3.40 per 30-minute exam, or roughly $0.11 per student-minute, the figure a deployment that scales past the included credit pool should budget against. We report both per-minute and per-30-minute figures (voice $0.08 + grading $0.032 ≈ $0.112 per student-minute, so 30 minutes lands at $3.36 ≈ $3.40); they differ by a few cents because each is rounded independently.

Cohort totals (notional pay-as-you-go reconciliation). We report a notional cohort total only for Spring 2026, where the voice-minute denominator can be audited end-to-end: $35.48 in grading LLMs plus ~$116 in notional ElevenLabs charges across 1,063,118 credits / 24.2 hours of agent time (derived from the ElevenLabs API conversation listing rather than the stored session records, since retries are not always written back), for a Spring 2026 cohort total of ~$152 across 37 exams. The 24.2-hour denominator averages to roughly 39 agent-minutes per exam, which is larger than

Table 1's 30-minute graded-exam mean because it sums all agent-side voice usage attributable to each student (aborted and re-attempted sessions, brief setup and identity-check exchanges that did not progress to a graded exam, and pre-exam connectivity tests), whereas the 30-minute number is the duration of the single best-attempt conversation that was actually graded. The $3.40-per-30-minute planning figure quoted in Section 3.1 uses the graded duration; the $152 cohort total uses the broader agent-time denominator, so the ratio ($152 / 37 ≈ $4.10 per student) sits above the per-graded-exam figure by approximately the retries/setup overhead. Operators planning capacity should budget against the broader denominator. We do not report a comparable Fall 2025 cohort total: the Fall 2025 grading LLMs cost $10.30 (billed), but the agent-minute denominator from that period was not preserved at the same granularity, and an earlier reconstruction we attempted from Table 1's 22-minute mean would imply ~$63 in voice charges at the overage rate, inconsistent with the actual subscription-covered usage. The Spring 2026 figure alone preserves the point of the cohort-total comparison.

Human-grading comparator. A faculty/TA comparator of 36 students × 25 minutes × two graders at $25/hour puts human grading at approximately $750. On the apples-to-apples comparison (grading spend only, since the human figure is grader labor not voice infrastructure), both deployments land one-to-two orders of magnitude below: $10.30 (Fall 2025, ~73× cheaper) and $35.48 (Spring 2026, ~21× cheaper). Adding notional Spring 2026 voice at the per-minute overage rate brings the Spring 2026 cohort total to roughly $152, still ~5× below the human comparator though no longer an order-of-magnitude gap.

Excluded costs. The marginal numbers deliberately exclude (a) engineering time (prompt design, pipeline debugging, deployment), (b) instructor and TA time (supervising administration, auditing flagged grades), (c) the $99/month voice-platform subscription itself, and (d) *adaptation* cost: discipline-specific rubric design, curated case library, and per-phase questions. Of these, (a) is now largely a fixed, shared cost rather than a per-adopter one: the orchestration and grading platform exists and is available as a hosted service, so an adopter does not rebuild it. That leaves (d), which is essentially the ordinary cost of authoring any exam (writing questions and a rubric) rather than a systems cost specific to this format. In our deployments, authoring a new exam and rubric ran roughly four to eight hours, at most a day or two for a wholly new discipline; much of it is AI-assistable and shared with a course's other assessments. Spring 2026 went faster than Fall 2025 because it reused the Fall 2025 rubric and case library; a deployment in (say) an introductory statistics course would author new questions and a new rubric much as it would for a written exam of the same scope. These authoring costs amortize across semesters when the course is reused.

Scaling projection. At roughly $3.40 per 30-minute exam (the above-subscription regime, Section 3.1), a 1,000-student cohort would add roughly $3,400 in inference. A comparable human-grader benchmark (two graders at $25/hour, 30 min/exam) lands at ~$25,000, so the AI-administered cohort runs roughly 7× below the human comparator under matched assumptions: cheaper, though not cleanly an order of magnitude once the voice cost is fully priced in. This supports the main-text claim that per-administration scaling is not gated by token or voice spend.

F Replicated vs Cohort-Specific Patterns

The Scope-and-contribution paragraph at the end of Section 1 distinguishes patterns that replicated across the two cohorts from numbers that are cohort-specific instances. The full split is given in Table 11 and discussed below.

What we offer as transferable. Three regularities replicated in direction across the two cohorts: (1) Gemini consistently entered Round 1 lenient on the 0–20 total; (2) Claude and GPT-5 behaved as an

Table 11. Replicated structure vs. cohort-specific magnitude across Fall 2025 and Spring 2026. Replicated rows are what we offer as the transferable Pattern 4 result; cohort-specific rows are instances rather than estimates of a population.

Property	Fall 2025	Spring 2026
Replicated (direction and shape)		
Gemini enters lenient (Round 1 Gemini − Claude on 0–20)	+3.25 pts	+1.97 pts
Gemini compresses after deliberation	+3.25 → +1.08 pts ($\sim\frac{2}{3}$ of gap absorbed)	+1.97 → +0.38 pts ($\sim\frac{4}{5}$ of gap absorbed)
Claude–GPT-5 tight pair (Round 2 pairwise α, overall)	0.97	0.99
Residual disagreement concentrates in one rater (drop-Gemini leave-one-out α)	0.97	0.99
Cohort-specific (magnitude)		
Round 2 Krippendorff's α (dim-pooled, ordinal)	0.86	0.90
Round 2 Krippendorff's α (overall, ordinal)	0.83	0.95
Lowest-agreement rubric dimension	Experimentation (67% perfect)	Risk & Ethics (78% perfect)
Magnitude of residual Gemini offset (Round 2, signed mean on 0–20)	+1.08 pts	+0.38 pts
Marginal grading-LLM cost per exam	~$0.29	~$0.96
Marginal cost per student-minute (within subscription)	~$0.01	~$0.03

interchangeable tight pair at pairwise $\alpha > 0.96$; (3) residual disagreement concentrated in a single rater with a directional offset rather than distributing as random noise. These are the empirical basis for Pattern 4 (use a multi-model council with raters chosen for argumentative asymmetry).

What we do not claim transfers. The absolute α values, the identity of the lowest-agreement rubric dimension, the magnitude of the residual Gemini offset, and the per-student dollar figures are instances from two cohorts of one course at one institution. Two iterations cannot distinguish a two-cohort coincidence from a general property; we report the per-cohort numbers as instances rather than estimates of a population. The patterns themselves are stated in terms of LLM properties rather than classroom properties, which is why we expect them to generalize, but that expectation is a conjecture rather than a demonstrated result.

Adaptation cost as a scope boundary inside higher education. Even within higher education our deployment is a single course format (a capstone project plus case discussion) in a single discipline (AI/ML product management). The adaptation cost flagged in Appendix E (rubric, case library, phase-specific agent prompts, pilot runs) would be paid again before any of the numbers above could be re-measured on a new population. Nothing in our experience predicts how that adaptation cost scales with disciplinary distance from the original deployment.

Plausible next test sites. Neighboring domains where analogous structured spoken evaluation exists are the most plausible next test sites for the patterns: technical hiring phone screens, where Jabarian and Henkel [10] report AI interviews increasing job offers by 12% and retention by 16–18% across 70,000 applicants; and professional certification formats (medical boards, legal competency, regulatory compliance). We have not deployed in any of them, and we do not claim the specific numbers would survive the move.

G What AI Administration Buys

Human audit is a core step of the protocol rather than an optional overlay; the main paper concedes as much (Section 3.2). A fair question then follows: *if scoring still requires instructor validation, what is gained?* Three things, each scaling differently with cohort size.

Automation of the time-dominant interview phase. At our per-exam durations (22-minute mean in Fall 2025, nominal 30 minutes in Spring 2026) and cohort sizes, the single-examiner interview time alone would consume roughly 13 hours in Fall 2025 and 19 in Spring 2026 ($36 \times 22/60$ and $37 \times 30/60$); doubling for a two-grader configuration at the same actual durations gives about 26 and 37 hours. Under the nominal 25-minute, two-grader comparator of Appendix E, the same cohorts would land at roughly 30 and 31 hours. Either framing is before scheduling, room booking, and between-exam overhead, and would serialize along the examiners' calendars. The AI absorbs those hours concurrently. The interview phase is the time-dominant cost of traditional orals, and it is the phase the AI fully replaces.

Audit load scales with the flagged fraction rather than requiring universal human review. The decomposition-informed audit rule fired on roughly 3% of Spring 2026 exams: one student out of 37, and on the order of 30 flagged exams at 1,000 students if the rate holds (the absolute count still grows with the cohort; the operational saving is that the human reviews a small fraction rather than every exam). We have not measured audit-time-per-case directly and do not claim audit is free; the operational property is the fraction (~3%), not the per-case duration. AI handles the volume; the human handles the flagged minority.

Scale. At 200, 500, or 1,000 students a faculty-administered oral is infeasible for most courses without a dedicated corps of trained graders. So the realistic counterfactual at those sizes is not AI-vs-human but AI-vs-written-only, with the integrity weaknesses that motivated this paper. The point of comparison shifts from "can the AI match a faculty oral" to "does an AI oral provide assessment properties that the only feasible alternative (a written exam) cannot."

H Academic Dishonesty Vectors

Three forms of academic dishonesty deserve explicit treatment given the unsupervised home environment in which the exams ran. The main paper gives the short version (Section 6); the full vector-by-vector treatment is below.

Pre-rehearsed answers. A student who builds a copy of the examiner prompt and drills against it is practicing oral explanation of course material, which is the learning behavior the format rewards (Section 6 of the main paper, transparency-and-incentive-alignment paragraph). Dynamic follow-up questioning exposes the limits of rehearsal: questions are calibrated to what the student just said, so a scripted answer that does not reflect actual understanding breaks down within one or two follow-ups. The adversarial edge case (a student who reverse-engineers the examiner prompt and rehearses exhaustively) collapses into the best-case scenario: a student who has practiced explaining the material out loud until they can answer arbitrary follow-up questions has, for all practical purposes, learned it.

Live AI coaching. A second AI system feeding answers in real time via headset is not fully preventable: a fluent student repeating coached answers naturally would not be detectable from the transcript or video alone. This is a broader societal problem, not specific to this system: the same concern applies to human job interviews and is already surfacing with audio earpieces and camera-equipped glasses. No mitigation we are aware of is absolute.

Confederate in the room. A second person whispering or typing answers off-camera is deterred by the self-recorded video requirement, which instructors can review. Detection is not guaranteed (camera placement, off-screen displays, and lip-read coaching all remain possible), but the marginal effort and risk of detection are higher than in the unproctored written setting.

Net. None of these mitigations are absolute. The format is more integrity-preserving than unproctored written work and does not overclaim: we do not claim to prevent all cheating, only to raise the effort required above the level where AI assistance provides a free ride. The transparency framing in Section 6 of the main paper compounds this: practicing against a public clone of the examiner prompt is functionally indistinguishable from learning the material.

I Council Decomposition: Per-Rater Leniency Analysis

Where does the residual disagreement in the council live, in what shape, and why do we retain Gemini despite the apparent reliability gain from dropping it? Section 3.2 of the main paper answers in brief; the full analysis follows.

Pooled α replication. A second deployment in Spring 2026 (37 students, same council and rubric) replicated the Fall 2025 deliberation gain. Dimension-pooled α rose 0.65 (Round 1) $\rightarrow$ 0.90 (Round 2); overall-score α reached 0.95 (vs. 0.83 in Fall 2025). Pooled α above the conventional 0.80 threshold for "good reliability" is necessary but not sufficient: it does not say *where* the residual disagreement lives.

Single-rater concentration with a directional offset. Decomposing the council shows a single rater accounts for nearly all of the residual disagreement, and the bias is directional rather than random. In both cohorts, Gemini grades are systematically more lenient than the other two models on the 0–20 total. Signed mean differences (Gemini − Claude, Round 2): +1.08 points in Fall 2025, +0.38 in Spring 2026. Claude and GPT-5 by contrast agree almost perfectly: pairwise $\alpha = 0.97$ in Fall 2025 and 0.99 in Spring 2026; mean absolute total-score difference ≤ 0.4 points in both.

Leave-one-out. Dropping Gemini from the council would lift overall-score α to 0.97 (Fall 2025) and 0.99 (Spring 2026), not by reducing noise (the Claude–GPT-5 pair was already tight), but by removing the directional offset.

Deliberation compression. Round 1 Gemini entered the council roughly +2 to +3.3 points lenient on the 0–20 total, depending on cohort. Round 2 Gemini compressed toward consensus, but did not converge: roughly one third of the initial gap survives in Fall 2025 (+3.25 $\rightarrow$ +1.08 points, ∼67% of the gap absorbed) and roughly one fifth in Spring 2026 (+1.97 $\rightarrow$ +0.38, ∼81% absorbed). Per-model biases survive deliberation as stable offsets rather than average out.

Why we retain Gemini. Despite the apparent reliability gain from dropping Gemini, we retain it deliberately. A persistent lenient voice forces the stricter raters (Claude and GPT-5) to cite specific transcript evidence when defending deductions rather than reaffirm low scores by default. The practical lesson for LLM-as-judge ensembles: expect per-model biases to survive deliberation as stable offsets rather than average out, and choose raters for argumentative asymmetry rather than interchangeability. The decomposition also motivates the audit rule below (Appendix J): rather than triggering on raw three-way spread, the operative criterion is disagreement between the two tightly-coupled raters.

Chair-strictness caveat. A fair objection to the "lenient voice forces stricter raters to justify deductions" argument is that the chair synthesizer is Claude, one of the two strict raters. If a

strict chair can discard the lenient input, retaining Gemini may add less than the argumentative-asymmetry story implies. We do not have a fully controlled chair-swap experiment, but the chair output is closer to the Round-2 mean than to either pole: the chair-vs-Round-2 absolute difference is small in both cohorts (≤ 0.5 points on the 0–20 total on average, with no systematic strict-side or lenient-side offset). That is consistent with the chair acting as a weighted synthesizer of all three Round-2 votes rather than a strict-side selector, but a controlled alternative-chair comparison (Gemini or GPT-5 as chair, or a non-LLM aggregation rule) is left to future work.

J Audit Rule Derivation

This appendix expands two passages that appear only briefly in the main paper: the audit-rule definition (Section 3.2) and the partial-validity check (Section 3.3).

Partial validity check (Fall 2025). The instructor and teaching assistant independently graded all 36 Fall 2025 examinations and compared their scores to the council output. The instructor's holistic grades aligned with the most lenient model (Gemini), but after reviewing the stricter models' evidence (verbatim transcript excerpts supporting each deduction) the instructor conceded most of the stricter scores. The strict grading surfaced teaching deficiencies (particularly in Experimentation, Section 3.3 of the main paper) that the instructor's own grading had obscured: when you grade your own students, it is easy to give credit for what you *know* they understood rather than what they *demonstrated*. Only two students requested regrading, and neither challenge succeeded; the transcript evidence backing each grade was too specific to dispute.

What this comparison is and is not. The comparison was informal, not a controlled validation study. The instructor's leniency was partly driven by awareness of teaching gaps, a confound we did not control for. Rigorous validation would require blinded expert re-grading of a random subset with pre-defined scoring criteria. We therefore treat human audit not as a post-hoc convenience but as a *core* methodological step of the grading protocol: reliability gives us consistency among raters; the human auditor is the operational safeguard against validity failures, but formal validity (against blinded expert re-grading on pre-defined criteria) remains future work.

Decomposition-informed audit rule. The council decomposition (Appendix I) yields a tighter audit trigger than raw three-way spread. Because residual disagreement concentrates in one rater (Gemini) with a stable directional leniency offset, the most informative criterion is *disagreement between the two tightly-coupled raters*. Specifically, we flag an exam for human audit when:

- Claude and GPT-5 themselves diverge by more than one point on the 0–20 scale, *or*
- Gemini sits more than two points away from the midpoint of Claude and GPT-5 (in either direction).

The second clause is symmetric by design: although in our two cohorts Gemini was consistently the lenient outlier, an unexpected reversal (Gemini scoring well below the strict pair on some future cohort) would be at least as informative and should fire the same rule. Either condition signals disagreement that normal deliberation cannot absorb. In Spring 2026 this combined rule fires on roughly 3% of exams (one student out of 37).

Operating regime. AI handles the volume; the human handles the flagged minority and the cases the council itself surfaces as ambiguous. We did not measure audit-time-per-case; the operational property the rule guarantees is the fraction (~3% at the disagreement structure observed across two cohorts), not a per-case cost.

K Platform Portability Reimplementation Detail

The portability summary in Section 2.1 of the main paper names which capabilities are commodity and which two (multi-agent routing and per-session dynamic variables) an adopter would reimplement at the application layer. The full per-capability assessment is below: we split the four capabilities enumerated in Section 2.1 into four commodity sub-capabilities (speech-to-text, text-to-speech, turn-taking, transcript export) plus the multi-agent-workflow capability, and treat per-session dynamic variables as a sixth platform feature.

Commodity capabilities. Four sub-capabilities are commodity across contemporary voice-AI platforms: speech-to-text, text-to-speech, turn-taking, and transcript export all appear in Retell, Vapi, Ultravox, LiveKit Agents, and (at the model layer) OpenAI's Realtime API. A port is straightforward at this layer.

ElevenLabs-specific: multi-agent workflows. The multi-agent-workflow capability (the inter-phase orchestrator that hands off between Authentication, Project Discussion, and Case Discussion agents) is implemented through ElevenLabs' native workflow primitive. On platforms without a native workflow primitive, this becomes the adopter's code: not a large amount of code, but a non-trivial source of subtle bugs around phase transitions, state leakage between agents, and early termination on silence timeouts. Phase 2→3 transitions in particular were fragile during early Fall 2025 deployment, and the workflow primitive's built-in semantics absorbed several of those bugs at no engineering cost.

ElevenLabs-specific: dynamic variables. Per-session dynamic variables (string parameters such as student name, NetID, and capstone-project context, populated at conversation start and interpolated into agent prompts) are implemented through ElevenLabs' first-class dynamic-variable system. On platforms without that primitive, per-session context is typically string-interpolated directly into the initial system prompt; this works but blurs the distinction between per-session inputs (one student's project context) and per-cohort inputs (the rubric, the case library), making per-cohort updates riskier.

Voice cloning. The cloned "Professor Provost" voice used in Fall 2025 (Section 4 of the main paper) is ElevenLabs-specific at the quality tier we used. Contemporary alternatives (e.g., open-source XTTS, OpenAI's voice models) exist but were not evaluated. Spring 2026 used a neutral ElevenLabs preset ("Alice") rather than a cloned voice, so the port-friction question for voice cloning specifically does not apply to the current production configuration.

Turn-taking tuning. The 15-second silence threshold (Section 4 of the main paper) transfers across platforms but with different default values: Retell defaults to 2 seconds, OpenAI Realtime to a server-side voice activity detection model that does not expose a numeric threshold. The separate prompt-level 10-second patience rule is a distinct knob, described in Appendix A. Re-tuning is cheap; the platform-specific configuration knob is not.

Net assessment. A port to Retell or Vapi is mostly orchestration work: the platform difference is whether you write the inter-phase state machine yourself or inherit it from the platform. A port to OpenAI Realtime is closer to a ground-up rebuild, because the workflow and dynamic-variable primitives both have to be reimplemented at the application layer.

Failure modes are LLM-layer, not platform-layer. What does *not* change on porting is the shape of the failure modes in Section 4 of the main paper: question stacking, paraphrase drift, non-random case selection, and silence interpretation all originate in the underlying LLM, not the voice platform. Four of the five engineering patterns (decompose into single-purpose modules; constrain

in code or configuration, not prompts; never delegate randomization; multi-model deliberation with argumentative-asymmetric raters) surface on any stack that delegates conversation management to a language model. The fifth (voice selection) is the one that travels least easily, because high-fidelity voice cloning and preset libraries differ across platforms. We have not evaluated any alternative platforms in deployment, so this assessment is conservative rather than empirical.

L Supplementary Figures

The main paper carries a pooled (cross-dimension) rendering of grading agreement; the per-dimension breakdown below is the companion figure for readers who want to see how the deliberation gain distributes across the five rubric dimensions.

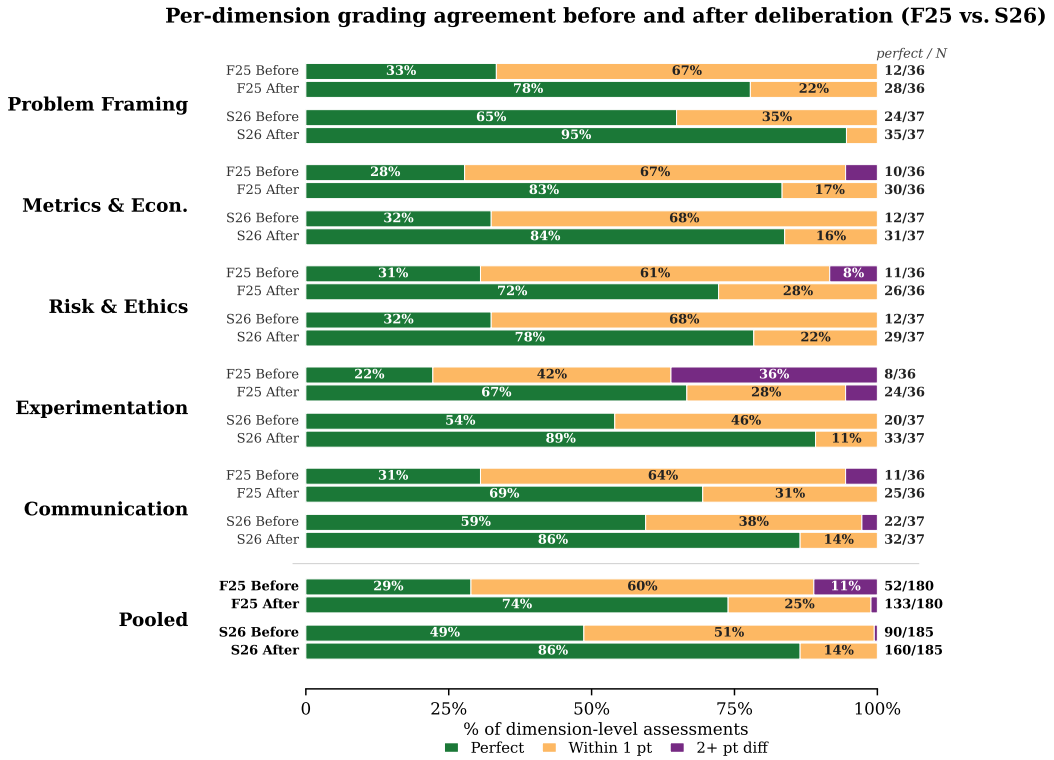

Fig. 6. Per-dimension grading agreement before and after deliberation, both cohorts (Fall 2025: $n = 36$ students, 180 dimension-level assessments; Spring 2026: $n = 37$ students, 185 assessments). Each row is the fraction of that cohort's assessments where the three council models agreed perfectly (dark green), landed within one point (amber), or diverged by 2+ points (dark purple). For each dimension, four bars are shown: F25 Before, F25 After, S26 Before, S26 After (the bottom block pools across all five dimensions). The right-hand column gives perfect-agreement count over total. The Before/After pair makes the deliberation gain visible directly: in F25, pooled perfect agreement rises 29% → 74% and 2+ point spreads collapse 11% → 1%; in S26 the same compression carries higher absolute agreement (49% → 86% perfect, with the single Round-1 2+ point spread also eliminated after deliberation). The dimension that ranks lowest on after-deliberation agreement shifts from Experimentation (F25, 67% perfect) to Risk & Ethics (S26, 78%): noise patterns are not fixed properties of the rubric.

M Discussion: Feedback and What Is Lost

The three subsections below were main-text in earlier drafts. They are moved here to keep the main paper at its target length; the substance is unchanged. (The transparency-and-incentive-alignment framing that originally led this section has been promoted back to the main paper, Section 6.)

Feedback Quality. The grading council produced feedback exceeding what human graders typically deliver in both structure and specificity. Each evaluation included strengths, weaknesses, and action items with verbatim evidence from the transcript. For the highest-scoring student: "Your understanding of metric trade-offs and Goodhart's Law risks was exceptional; the hot tub example perfectly illustrated how optimizing for one metric can corrupt another." For a lower-performing student: "Practice articulating complete A/B testing designs: state a hypothesis, define randomization unit, specify guardrail metrics, and establish decision criteria for shipping or rolling back." Human graders rarely generate such detailed, evidence-linked feedback for every student. The automated system produces it as a byproduct.

What Is Lost. Automating assessment is not a free lunch. A human examiner develops real-time intuition about student understanding, noticing by the third or fourth exam which topics the class systematically struggles with. Our Experimentation gap (Section 3 of the main paper) was only visible *after* analyzing aggregate scores from the AI system; a human examiner might have noticed it sooner. But this concern can be mitigated by other feedback channels. In our course, students complete a brief exit survey after every session (what they learned, already knew, and found unclear), and an LLM analyzes responses to generate pedagogical recommendations within minutes. We apply the same approach to assignments before grading, identifying class-wide patterns early enough to address in subsequent sessions. These mechanisms are orthogonal to the exam format and provide the continuous pedagogical feedback that automated assessment alone lacks. There is also value in the human relationship between examiner and student. A professor can read body language, offer encouragement, and exercise contextual judgment that no rubric fully captures. We do not claim AI-administered assessment is superior to well-resourced human assessment. The claim is narrower: it is superior to *no* oral assessment, the realistic alternative for most courses.

Accessibility (extended). Voice-based assessment introduces accessibility challenges that institutions must address before deployment, not after. Students with speech-related disabilities or conditions affecting verbal fluency face barriers that written formats avoid. For non-native speakers, the picture is more complicated than it might appear: because speech-to-text output feeds an LLM rather than a keyword matcher, the system can in principle apply domain-aware contextual repair that standalone ASR systems cannot (we did not benchmark this against a word error rate (WER) baseline, so we report it as a mechanism that may partly offset accented-speech errors, not as a measured reduction). Pre-exam, several Fall 2025 students worried the system would not understand their accent; post-exam, the barrier proved cognitive rather than acoustic, as one international student reported, "At times, the difficulty wasn't in knowing the material but in articulating my thoughts under pressure in English." This points to a distinct fairness risk: a live oral rewards fast verbal production, so a student who understands the material but formulates answers more slowly (because of a processing-speed difference, a stutter, or simply thinking in a second language) can be underserved by a format that a written exam would not penalize. The silence-timeout fix (Section 4) partly addresses this by not punishing thinking pauses, but pacing pressure remains. The accommodations we consider necessary before wider deployment therefore include configurable (not merely extended) time limits and pacing, a choice of agent voice and speaking rate, an ungraded practice run so the interface is familiar before stakes are real, and a text-based alternative for students who cannot use the voice format; accented speech in particular should be spot-checked

against gold transcripts rather than assumed handled. Spring 2026 added a live transcript display, a direct response to Fall 2025 exit-survey requests, giving students real-time confirmation of what the system understood and allowing immediate rephrasing. Students should be informed in advance about how the system works, what is recorded, and who accesses their transcripts. The format is not uniformly exclusionary: students with anxiety-related accommodations found the AI examiner *less* stressful than facing a senior professor, and the live transcript particularly benefits non-native speakers and those using technical vocabulary, suggesting voice AI may simultaneously create and remove accessibility barriers depending on the population.

N What's Next

Four extensions seem most immediately valuable.

Retrieval-augmented examination. In both Fall 2025 and Spring 2026 the agent received project summaries as parameters but could not access student-submitted artifacts. If it could reference specific slides or code, it could ask "On slide 7, you claim 94% accuracy. Walk me through how you validated that number." This would further close the gap between artifact quality and demonstrated understanding.

Formative deployment. The same infrastructure could support low-stakes practice examinations throughout the semester, letting students test understanding before high-stakes assessment. At a marginal cost on the order of $1 per 10 minutes of graded exam time (Section 3 of the main paper), a more immediate application is attaching a short oral comprehension check to every written assignment, a pass/fail conversation verifying that the student can explain what they submitted. This is the continuous authenticity layer the conclusion of the main paper points to: the system stops being a semester-end assessment tool and becomes a standing check attached to every assignment.

Validated expert re-grading of the flagged subset. The decomposition-informed audit rule already in place fires on roughly 3% of Spring 2026 exams (Section 3 of the main paper), and instructor + TA review of those flagged cases is now treated as a core step of the protocol. The next step is a controlled validation study (blinded expert re-grading of a random subset with pre-defined scoring criteria), so the gap from inter-rater reliability to validity against expert judgment can be measured rather than asserted.

Accessibility by design. Spring 2026 implemented live transcript display (a direct response to Fall 2025 exit-survey requests), so students see exactly what the system understood in real time and can rephrase if they spot a mistranscription. Remaining work includes ungraded practice examinations so students can familiarize themselves with the voice interface before stakes are real, configurable time limits for students with accommodations, a choice of agent voice and speaking pace, and a text-based alternative for students unable to use the voice format. Our experience is anecdotal and observational rather than measured: a handful of students with anxiety-related accommodations volunteered in exit-survey free responses that the AI format felt less stressful than a human examiner, and pre-exam concerns about accented speech did not surface as post-exam grading failures in either cohort. We did not measure word error rate against gold references and did not survey students with disabilities systematically. With those caveats, the observation is suggestive (accessibility design could make AI-administered orals more inclusive than human-administered ones for some populations), but the claim is preliminary, not validated.