Scalable and Personalized Oral Assessments Using Voice AI

PANOS IPEIROTIS, New York University, USA
KONSTANTINOS RIZAKOS, New York University, USA

CCS Concepts: • **Social and professional topics** → **Student assessment**; • **Human-centered computing** → *Natural language interfaces*; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: oral examination, voice agents, educational assessment, large language models, automated grading, academic integrity, LLM-as-judge

ACM Reference Format:
Panos Ipeirotis and Konstantinos Rizakos. 2026. Scalable and Personalized Oral Assessments Using Voice AI. *J. ACM* 0, 0, Article 0 (2026), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 The Problem

Students in our AI/ML Product Management course at NYU Stern submitted thoughtful project analyses and case-study discussions. The writing was often good: clear arguments, reasonable tradeoffs, appropriate citations. Yet when we questioned them live in class, many could not walk through a single decision from their own work. The gap between what students submitted and what they could discuss live was too consistent to be nerves.

The temptation is to call this a cheating problem [5, 11], but that framing misleads. AI-generated text is now difficult to distinguish from human writing: in a controlled study, untrained readers identified GPT-3-authored text at chance, and brief training raised accuracy to only about 55% [3]. Detection classifiers fail often enough that responsible institutions have largely stopped using them [18, 20]. The root issue is structural: written submissions once served as evidence of thinking, and that evidentiary link has broken. Proctored, closed-book exams restore something, but the open-ended, project-specific questions that make a course like ours worth taking cannot survive a timed room with no materials.

The same gap opens far beyond the classroom. A job interview, a professional licensing board, even the legal question of whether two parties reached a "meeting of the minds": anywhere a polished artifact is taken as proof of understanding that ought to be tested directly, the problem is the same, and so is the remedy. Put the person in a position where they must reason aloud under follow-up, and separate those who did the work from those who can only gesture at it. We develop the idea inside a university course because that is where we deployed it, but nothing about the method is specific to education.

Oral examinations sidestep the problem [7, 8]. A student who can reason aloud through a follow-up question, defend a specific design decision, and shift when challenged has demonstrated understanding in a way no artifact can fake. Structured oral formats are also more reliable than their reputation suggests: reliability concerns fall when examiners follow a shared rubric and

Authors' Contact Information: Panos Ipeirotis, panos@stern.nyu.edu, New York University, New York, New York, USA; Konstantinos Rizakos, konstantinos.rizakos@nyu.edu, New York University, New York, New York, USA.

ACM 1557-735X/2026/0-ART0
https://doi.org/XXXXXXX.XXXXXXX

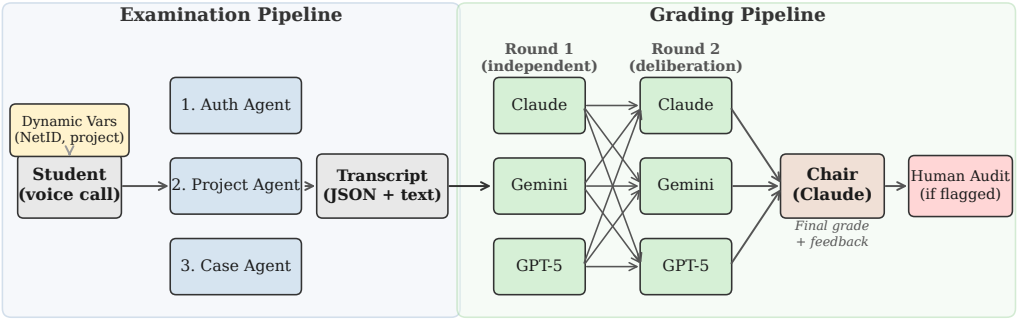

Fig. 1. End-to-end system architecture. **Left:** the examination pipeline runs three agent phases, with per-student context injected via dynamic variables. **Right:** the grading pipeline scores each transcript across two multi-model rounds, then chair synthesis; high-disagreement cases route to human audit.

pre-set criteria [15, 16]. Oral exams fell out of widespread use not because they fail but because administering and grading them at scale cost too much [1]. The same study found a partial solution by having 600 students record video responses through a learning management system, but grading remained human and the examination lacked the interactivity of human-driven exams. The scaling problem stayed unsolved.

Voice AI and automated grading change that arithmetic. Both expensive parts of an oral exam can now be automated: conducting it and grading the result. Contemporary conversational AI platforms combine speech recognition, natural-sounding synthesis, and turn-taking that tolerates thinking pauses [6]. Large language models (LLMs) can grade the resulting transcript reliably with the right design [22]. Jabarian and Henkel [10] found that AI-conducted voice interviews in hiring can match or exceed human interviewers; Nitze [17] explored LLM-driven oral simulations in text. Hung et al.'s Socratic Mind piloted an LLM-led Socratic oral-assessment assignment with 600 students, taking spoken or typed answers, with grading still manual [9]. Each piece already exists.

What no one had shown is how to assemble them into a live voice examination of record, conducted and graded end to end, that holds up under real exam conditions, and what breaks when you try. We built *Viva by NYU Stern*,[1] which conducts a personalized oral exam and then grades the transcript with a council of three LLMs that score independently, read each other's assessments, and revise (Figure 1).

Scope and contribution. This is a field report, not a controlled study. Two consecutive cohorts, one course, one institution: 36 students in Fall 2025, 37 in Spring 2026. The numbers we report (cost per student, how closely the grading models agree, rates of student stress and preference) are observations from those deployments, not population estimates. Whether the lessons that follow transfer beyond this one deployment is an empirical question we cannot settle from two cohorts. The replicated-vs-cohort-specific split appears in Appendix F.[2]

What we offer are five practical engineering lessons that surfaced in our deployments and are likely to matter in many conversational systems that hand behavior to a language model. We call them *patterns* in the design-pattern sense a programmer already knows: each pairs a recurring

[1]Available at http://viva.stern.nyu.edu/. The name nods to "viva voce," the Latin "living voice" phrase for oral examinations.
[2]Pointers of the form "Appendix X," here and throughout, refer to this article's online supplement in the ACM Digital Library, as do mentions of "the supplementary materials." The supplement collects the full voice-agent and grading-council prompts, the survey instrument and per-cohort results, the cost decomposition, the grading-council reliability decomposition, the human-audit rule, the platform-portability assessment, supplementary figures, and an extended discussion, together with a sanitized snapshot of the orchestration and grading code. The anonymized cohort data and the analysis scripts that reproduce the reported statistics are archived at Zenodo: https://doi.org/10.5281/zenodo.21522314.

failure with a reusable fix. The five: (1) *Decompose into single-purpose modules, not a single do-everything agent*; (2) *Constrain LLM behavior with code or configuration, not prompts*; (3) *Never delegate randomization to an LLM*; (4) *Use a multi-model council with a deliberation round, and choose models that disagree*; and (5) *Choose voice characteristics with the same care as question design.* The first pattern was an early architectural choice; the others surfaced from deployment failures. These patterns organize the rest of this paper. Education is where we found them.

2 System Design

2.1 Platform Requirements and Architecture

We could have run the whole examination through a single agent. We split it across three instead: isolating each phase prevents conversational drift, keeps a failure in Phase 2 from contaminating Phase 3, and simplifies debugging. That design choice shaped what the platform had to provide.

A voice examination system needs four capabilities: (1) accurate speech-to-text and text-to-speech with natural prosody, (2) turn-taking that tolerates thinking pauses, (3) multi-agent workflows to decompose the exam into phases, and (4) transcript export for grading. These can be assembled from components or bought as an integrated platform. We used ElevenLabs Conversational AI [6], but the first, second, and fourth capabilities are commodity across contemporary voice-AI platforms (Retell, Vapi, Ultravox, LiveKit Agents, OpenAI's Realtime API); only multi-agent routing and per-session dynamic variables would require application-layer reimplementation elsewhere (Appendix K).

Figure 1 shows the flow, and the full prompt scaffold is in Appendix A. Each of the three agents owns one phase of the examination:

(1) **Authentication Agent:** In Fall 2025 this agent opened the session, captured the student's name and NetID through spoken dialogue, and validated them against the course roster before handing off to Phase 2. In Spring 2026 we made it largely redundant by issuing each student a personalized examination URL ahead of time: the link establishes identity, so the phase becomes an extra security check that a brief greeting could replace.
(2) **Project Discussion Agent:** Receives the student's capstone project context via per-session dynamic variables and asks structured questions about their specific work: goals, data choices, modeling decisions, evaluation approaches, and failure modes. This phase targets the gap between LLM-generated written work and actual understanding: students cannot improvise consistent answers about decisions in work they did not personally complete.
(3) **Case Discussion Agent:** Discusses one case study from a library of real-world AI/ML product cases covered in the semester, asking questions that span course topics. In Fall 2025 the agent itself selected the case from a prompt-supplied library; in Spring 2026 the case is drawn in code from a per-session random seed and passed to the agent as a dynamic variable (Section 4). The agent generates its questions dynamically during the examination.

Only per-session dynamic variables flow across handoffs, not prior-phase transcript text; the downstream grading council (Section 2.2) still sees the full concatenated transcript. The failure modes in Section 4 originate in the underlying LLM, not the voice platform, so the patterns that address them surface on any stack that delegates conversation management to a language model (voice selection travels less easily; Appendix K).

2.2 Grading: A Council of LLMs

We implemented a "council of LLMs": three models from different families independently score the same transcript, then revise after seeing each other's reasoning. We used Claude, Gemini, and GPT-5, one from each of the three major frontier-model families available at deployment rather than

a tuned panel size.[3] This "LLM-as-judge" approach extends a long history of automated scoring [19] to oral examination transcripts: language models evaluate open-ended work in place of human raters [21, 22] and can debate their judgments in multi-agent exchanges [2]. The process has three steps:

(1) **Round 1 (Independent):** Each model receives the full examination transcript and the grading rubric, then independently scores it on the rubric's dimensions (in our case, five dimensions, each scored 0–4), with verbatim transcript evidence for each score: *Problem Framing* (business problem → ML spec), *Metrics & Economics* (evaluation metrics, trade-offs, counter-metrics), *Risk & Ethics* (failure modes; fairness, accountability, transparency, privacy, security), *Experimentation* (A/B design, hypotheses, controls), and *Communication* (structured answers under follow-up pushback). The three models run in parallel with no shared context.
(2) **Round 2 (Deliberation):** Each model receives the other two models' *complete* Round 1 output (their per-dimension scores, written justifications, and the verbatim transcript excerpts each one cited) rather than a paraphrased digest, and is asked to revise its own evaluation in light of that reasoning. We label peers neutrally ("Peer 1," "Peer 2") and withhold model identities, so a model cannot defer to a brand rather than an argument.[4] Models may change scores or reaffirm them, but must justify either decision with transcript evidence. On one Spring 2026 exam (a DocuSign "Contract Risk Analyzer" project), a grader that had scored Problem Framing higher revised it down to 3 once a peer noted the student had never worked through the project's second-order effects, arguing the change from what the transcript showed rather than from the peer's number.
(3) **Chair Synthesis:** A designated chair model (Claude) receives all Round 1 and Round 2 outputs and produces the final grade: one score per dimension, a total score (0–20), and a structured feedback report with strengths, weaknesses, and verbatim evidence.

Why not use a single model? Because initial independent assessments showed poor agreement; the deliberation phase was necessary. A known risk is *anchoring bias*: when models see each other's Round 1 scores, they may converge toward a shared (possibly wrong) answer rather than genuinely revising. We cannot fully rule this out, but deliberation transcripts show substantive engagement (models cite specific transcript evidence to argue for score changes rather than splitting differences), suggesting the convergence reflects reasoning. More on this in Section 3. The full grading prompt and a worked deliberation example are in Appendix B.

3 Deployment and Results

We deployed the system as the final examination for an undergraduate AI/ML Product Management course at NYU Stern across two consecutive cohorts, Fall 2025 (36 students) and Spring 2026 (37 students), with fixes shipped between them (Section 4). We use the semester labels throughout to line up with the dated tables and figures.

The system eliminated scheduling as a constraint; students chose any time during a nine-day window in each cohort (Table 1 lists the opening dates). Students recorded themselves (webcam and audio) during the examination as a course requirement, both to deter outsourcing or real-time

[3]Exact model identifiers, from our deployment configuration (Fall 2025) and the stored grade records (Spring 2026): Fall 2025 used `claude-opus-4-6`, `gemini-2.5-pro`, and `gpt-5.4`; Spring 2026 used `claude-opus-4-6`, `gemini-3.1-pro-preview`, and `gpt-5.4` (Gemini was the only model upgraded between cohorts). We record the exact versions because model version affects reproducibility; the chair-model tier, which differed between cohorts, is discussed in Appendix E.
[4]We concatenate the two peer assessments in a fixed order for every exam; we did not randomize presentation order, so a small position effect cannot be excluded, though because both peers are always present the room for order to matter is limited.

Table 1. Examination Statistics, both cohorts. "Messages" counts both agent prompts and student replies.

Metric	Fall 2025	Spring 2026
Students examined	36	37
Exam window opens	Dec 12	Mar 15
Average duration	22 min	30 min
Duration range	9–41 min	19–43 min
Average number of messages	58	84

assistance and as a backup if the conversational pipeline failed mid-exam. This exam-integrity obligation is separate from research consent: under IRB protocol NYU IRB-FY2023-7595, de-identified transcripts and survey responses enter the analysis only for consenting students; a student who declined would still have taken the exam for course assessment, with their data excluded from the study. No student in either cohort declined.

3.1 Cost

Two regimes matter. *Within* our $99/month ElevenLabs subscription, which covered all voice minutes in both cohorts, the per-exam marginal cost reduces to grading-LLM API calls alone: ~$0.29 per Fall 2025 exam (billed) and ~$0.96 per Spring 2026 exam (estimated; Appendix E), or $0.01–$0.03 per student-minute. The Spring jump is a configuration choice, not a scaling signal: a Claude Opus 4.6 chair accounted for roughly two-thirds of grading spend; reverting to a cheaper chair tier would halve the total. *Above* an equivalent subscription credit pool, voice bills at the ElevenLabs additional-call overage rate of $0.08 per agent-minute, so the planning figure is roughly $0.11 per student-minute, or about $3.40 for a 30-minute graded exam (before retry/setup overhead).

These per-exam figures are the *marginal* cost. The larger cost is a one-time *adaptation*: writing the rubric, curating the case library, and authoring the per-phase questions. The orchestration and grading platform already exists (we now run it as a hosted service), so this is exam-authoring effort rather than a systems-engineering burden, the same rubric-and-question work a written or in-person exam requires. In our deployments, authoring a new exam and rubric took roughly four to eight hours, and we would not expect more than a day or two even for a wholly new discipline. Much of that authoring is itself AI-assistable, shared across a course's other assessments, and amortized over every exam that reuses it (Appendix E).

3.2 Grading Reliability

Initial independent grading in Fall 2025 revealed an inconsistency: Gemini averaged 16.3/20 while Claude averaged 13.0/20, a 3.3-point gap (GPT-5 fell between at 13.7/20, within one point of Claude on 64% of students). Without deliberation, the council was useless. After deliberation, convergence was asymmetric: Gemini dropped its mean by 1.3 points after seeing stricter peer assessments, Claude rose by 0.9, GPT-5 by 0.3, all three converging toward the 14–15 range. That the largest shift came from the most lenient model suggests substantive revision rather than score-splitting.

Deliberation improves agreement, with the largest relative gain in the noisier Fall 2025 cohort (Figure 2; per-dimension breakdown in Appendix L, Figure 6). Perfect three-way agreement rose from 29% to 74% in Fall 2025 and from 49% to 86% in Spring 2026 (pooled across the five rubric dimensions); 2+ point spreads fell from 11% to 1% (Fall 2025) and from 1% to 0% (Spring 2026). We summarize agreement with Krippendorff's α, a chance-corrected agreement score on a 0–1 scale: it discounts the agreement three raters would reach just by guessing, so it is a stricter bar than the raw percentages above. On that scale $\alpha \geq 0.80$ is the conventional threshold for good agreement [13]; we compute α following [14]. Pooled the same way, α rose from $0.52 \rightarrow 0.86$ in

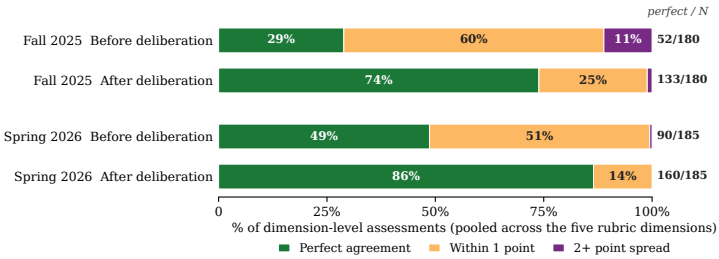

Fig. 2. Pooled grading agreement before and after deliberation. Each bar is the share of dimension-level assessments (pooled across the five rubric dimensions) where the three council models agreed perfectly (green), within one point (amber), or by 2+ points (purple). Fall 2025: $n = 36$ (180 assessments); Spring 2026: $n = 37$ (185). Per-dimension breakdown in Appendix L, Figure 6.

Fall 2025 and 0.65 → 0.90 in Spring 2026. On the 0–20 overall score it reached 0.83 and 0.95 after deliberation.

Pooled α above threshold does not say *where* the residual disagreement lives. Decomposing the council shows that a single rater accounts for nearly all of it, as a directional offset: in both cohorts Gemini is more lenient on the 0–20 total (Gemini − Claude signed mean: +1.08 Fall 2025, +0.38 Spring 2026), while Claude and GPT-5 agree almost perfectly (pairwise $\alpha = 0.97$ and 0.99). Dropping Gemini would lift overall-score α to 0.97 and 0.99, yet we retain Gemini deliberately: a persistently lenient voice forces the stricter raters to cite specific transcript evidence rather than reaffirm low scores by default (Appendix I). We added human audit as a *core* step: flag an exam when Claude and GPT-5 diverge by more than one point on the 0–20 scale, or when Gemini sits more than two points from the midpoint of Claude and GPT-5. In Spring 2026 this rule fires on roughly 3% of exams (Appendix J). The practical lesson: expect per-model biases to survive deliberation as stable offsets, and choose raters for argumentative asymmetry rather than interchangeability.

3.3 Diagnostic Power for Teaching Quality

Reliability ("do raters agree?") is necessary but not sufficient; *validity* ("do grades reflect actual student competence?") requires expert human judgment. As a partial validity check, the instructor and TA graded all 36 Fall 2025 exams independently; the instructor's initial grades sat close to Gemini's, but on review the instructor conceded most of the stricter council scores once cited transcript evidence was visible. The same evidence trail benefits students directly: every exam yields a structured, evidence-linked feedback report as a byproduct of grading, something human graders rarely produce for every student (Appendix M). External grading has a second side benefit over instructor self-grading: it surfaces teaching deficiencies that would otherwise remain hidden.

Scores by topic varied (Figure 3): Fall 2025 Experimentation (A/B testing) averaged 1.94/4 with a left tail (8% scored 0, 19% scored 1, zero demonstrated mastery), against Problem Framing at 3.39/4. This was a mirror held up to the instructors, not just the students: A/B testing had received insufficient class time, easy to overlook when instructors grade their own students leniently on topics they know were covered superficially. The gap closed in Spring 2026 after we rebalanced the syllabus: Experimentation rose to 2.70, with no student scoring below 2.

Duration was a poor proxy for score in both cohorts (Figure 4): the shortest Fall 2025 exam (9 minutes) earned the highest score (19/20), while the longest (41 minutes) earned one of the lowest (12/20). In Spring 2026 the trend was moderately negative, though at this sample size that is not strong evidence of a population-level effect.

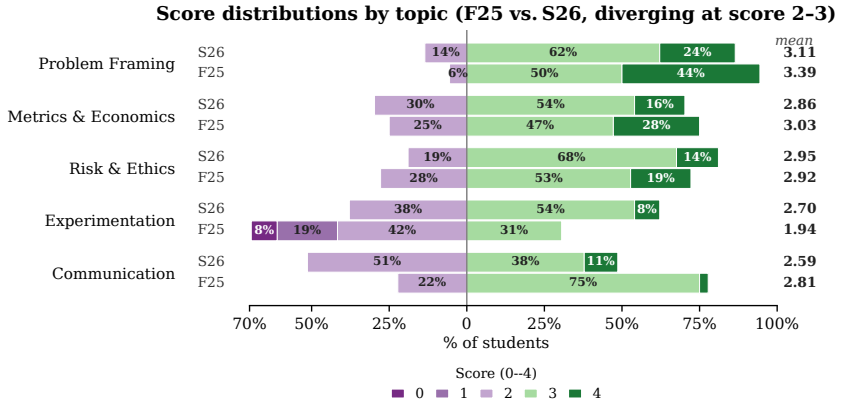

Fig. 3. Per-dimension score distributions across both cohorts (Fall 2025: $n = 36$; Spring 2026: $n = 37$). Scores 0–2 extend left of the centerline, 3–4 extend right; bold numbers are means. Fall 2025 Experimentation had a left tail (mean 1.94); Spring 2026 Experimentation shifted right (mean 2.70, no scores below 2) after we revised the syllabus to give A/B testing more class time.

4 What Broke: What We Wish We Knew Before Deploying

Fall 2025 surfaced five failure modes: the agent stacked several questions into a single turn; it paraphrased questions when asked to repeat them; it treated a student's thinking pause as a dropped call; it selected case studies far from randomly; and, in a cloned professor's voice, it came across as harsh. Some we fixed in code for Spring 2026, some we could only mitigate at the prompt level, and one (silence as thinking) survived its fix. Each part below gives the problem, then what we changed, and for the surviving one, why it resisted.

Question Stacking. The most damaging issue. The agent regularly fired off compound questions. From an actual transcript: *"Tighten it up for me: who is the user, and what decision do they make differently because of your product? And what is your North Star metric, plus one counter metric that might get worse if you over-optimize?"* A Fall 2025 student summed it up: *"It kept asking 3 to 4 questions at the same time, despite me asking it to slow down multiple times."* Over half of the student improvement suggestions on the exit survey mentioned this problem. Prompt-level instructions helped but did not eliminate the behavior: for constraining LLM behavior in extended conversations, they are necessary but insufficient. The failure is not unique to our stack: Socratic Mind's authors observed the same multi-question drift and planned a prompt-level fix [9]. We also built an "interference protocol" into the grading rubric, instructing graders to assess only the answered parts when the agent stacked questions and the student answered only some.

The lesson: *enforce behavioral constraints through system architecture, not just prompting.* Spring 2026 added a programmatic turn validator that detects a multi-question turn and asks the model to re-emit a single-question version. The rate of stacked turns (a single agent turn packing in two or more separate questions) fell from 31% in Fall 2025 to 11% in Spring 2026.[5] The turns that remained also got simpler: three-or-more-question turns nearly vanished, falling from 2.9% to 0.3%; questions spanning two rubric dimensions roughly halved, from 10.6% to 4.6%; and the median stacked turn shrank from 52 to 38 words. A few three- and four-question turns still surface, so this is a large reduction, not elimination.

[5]Precisely, from 31.2% (191 of 612 agent question-turns) to 10.9% (122 of 1,115), after excluding scripted case-briefing preambles (see the online supplementary materials); we report a descriptive reduction rather than a significance test, since turns are nested within exams. The same drop holds exam by exam: the most-stacked exam fell from 85% to 39%.

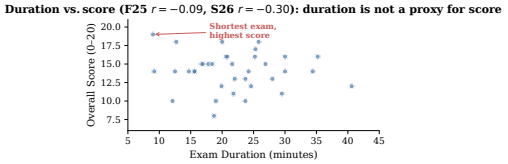

Fig. 4. Examination duration vs. overall score (Fall 2025, $n = 36$). The shortest exam (9 minutes) received the highest score (19/20); the longest (41 minutes) received 12/20. Pearson $r = -0.09$ for Fall 2025 and $r = -0.30$ for Spring 2026 ($n = 37$; not shown). Duration is not a reliable proxy for score in either cohort: short exams were not necessarily shallow, and long exams were not necessarily stronger.

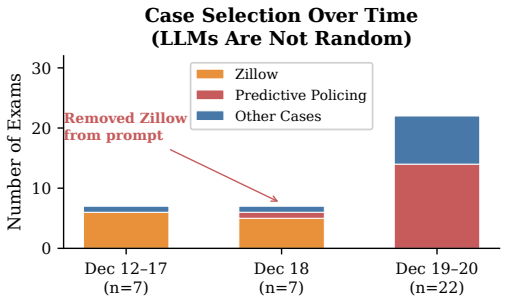

Fig. 5. Fall 2025 case selection over time. Despite instructions to "randomly select" a case from a library of eight, the agent chose Zillow in 6 of 7 examinations during December 12–17 (86%). After Zillow was removed from the prompt at the end of December 18, the agent fixated on Predictive Policing in 14 of 22 examinations on December 19–20 (64%). Spring 2026 (not shown): a deterministic seed-to-case mapping in code produced a 22% maximum share with all eight cases used.

Paraphrasing During Clarification. When students asked the agent to repeat a question, it would paraphrase rather than repeat verbatim, inadvertently changing the question. The LLM was doing what it was trained to do (being helpful by rephrasing) but exactly the wrong reflex in an examination. Explicit prompt instructions ("repeat the exact previous question") reduced but did not eliminate the behavior; this remains a prompt-level mitigation, not a code-enforced one.

Silence Handling. By default the platform treated a silence as a sign that something had gone wrong (a dropped call) rather than as a student thinking. Fall 2025 ran on the 5-second default, so the agent would cut in with "Are you there?" exactly when a student was formulating an answer. Spring 2026 raised the threshold to 15 seconds, but the interruptions persisted, and students reported that the check-ins spiked their stress precisely when they were trying to think. The lesson: silence is thinking time. Our default for future semesters is 30 seconds.

Non-Random Case Selection. This was the most instructive failure (Figure 5). The examination prompt listed eight case studies and instructed the agent to "randomly select ONE case." From December 12–17, when the Zillow case study (Zillow's iBuyer home-buying program, used to discuss concept drift and adverse selection) was on the list, the agent selected it in 6 of 7 examinations (86%). At the end of December 18, we removed this case study from the prompt. The agent immediately shifted to the Predictive Policing case study, selecting it for 14 of 22 examinations on December 19–20 (64%), better than the level of fixation on the Zillow case study but still far from the ∼14% expected under uniform selection over the remaining seven cases. A predictable case is also a leakable one: a student who can anticipate which case they will receive can prepare that scenario in advance or tip off the next student, undercutting the leak-resistance that the dynamically generated questions otherwise give the format (Section 6).

The fundamental issue: LLMs should not be trusted to implement uniform randomization by instruction. The same deterministic bias appears when models are asked simply to generate random numbers [4]. They carry ordering biases that survive every prompt-level mitigation we tried, and the bias is invisible until you examine the output distribution. Kelly [12] makes the broader case that randomization is routinely botched even when practitioners are explicitly trying to implement

Table 2. Design Patterns for Voice AI Assessment Systems.

Pattern	Failure Addressed	Implementation
1. Decompose into single-purpose modules	A monolithic do-everything agent; conversational drift across exam phases; cascading prompt failures	Separate agents per phase (authentication, project, case); independent grading and chair-synthesis steps; phase transitions enforced by orchestrator, not LLM discretion
2. Constrain in code or configuration, not prompts	Question stacking (code); silence handling (platform configuration)	Programmatic turn validator rejects multi-question responses; voice-activity-detection timeout raised from 5 s to 15 s (itself insufficient; 30 s is the future default)
3. Never delegate randomization to the LLM	Non-random case selection (ordering bias survives every prompt)	Deterministic case mapping from a per-session random seed passed as a dynamic variable
4. Multi-model council with deliberation	Single-model grading bias; silent miscalibration	Independent scoring (Round 1), cross-review with cited evidence (Round 2), chair synthesis; flag high-disagreement cases for human audit
5. Voice selection	Perception of aggressiveness	Selected a calm, education-oriented voice

it; asking an LLM to randomize by instruction is a textbook instance. The fix is to draw a random number in code and map it to a case deterministically, which is what Spring 2026 shipped. Across the 37 Spring 2026 examinations, all eight cases were used and the maximum single-case share was 22% (8/37), in contrast to the Fall 2025 single-week 86% peak.[6] The principle generalizes: any behavior requiring uniformity or randomness belongs in code, not in the LLM.

Voice Characteristics. We cloned the voice of a course professor from a conference keynote, reasoning that a familiar voice would feel more natural than the defaults. The opposite happened: students described the agent as "aggressive," "condescending," "shouting," a "sports coach" style "absolutely not appropriate for classroom use." One free-response answer reduced it to four words: *"Make it less mean."* Keynote delivery (fast-paced, projected, designed to talk *at* a large audience) is the wrong register for a one-on-one exam; without the softening cues of face-to-face interaction, a high-fidelity clone of an authority figure read as intimidating rather than familiar. Spring 2026 switched to "Alice," an ElevenLabs preset with a calm, education-oriented register; the change removed the complaints. Voice characteristics warrant the same attention as question design.

Table 2 collects the five patterns, pairing each with the failure it addresses and the implementation that enforced the fix; paraphrasing remains the one failure without an architectural pattern, mitigated only at the prompt level.

5 Student Experience

The central tension held in both cohorts: what the exam measured well versus how it felt to take. Fall 2025 students rated "tested my actual understanding" highest (70% agreement), but only 33% found questions clear and 83% found the exam more stressful than written exams. Spring 2026 replicated the student-perceived-understanding finding (66%) while clarity rose to 78% and "more stressful" fell to 63%; the timing coincides with the engineering fixes between cohorts, though with two cohorts and several simultaneous changes we cannot identify which fix moved which

[6]With $n = 37$ across 8 cases, expected cell counts under uniformity are 4.6, below the conventional ≥ 5 rule for the χ^2 approximation, so we report the observed distribution descriptively rather than as a goodness-of-fit test against uniformity.

response. "Tested my actual understanding" is a student-perceived measure, not a psychometric validity claim; the reliability-vs-validity distinction is in Section 3.3.

These rates come from voluntary surveys run after the exam but before grade release (30 of 36 responding in Fall 2025, 32 of 37 in Spring 2026; the instrument and full response distributions are in Appendices C and D). Such surveys can over-represent polarized reactions, so we checked the one thing we observe for non-respondents, their final grades: in both cohorts they were statistically indistinguishable from respondents' (Appendix D.4). Prior oral-exam experience does not appear to explain the stress response: the share who had never taken an oral exam held flat across cohorts while the stress rate fell (cross-tabulations in Appendix D.2). Format preference was more consistent with perceived stress than with prior oral-exam experience: the share preferring an AI oral exam rose from 13% to 25%. Comfort and fairness ratings also rose, so a concurrent shift in trust in AI cannot be ruled out.

The most praised feature in both cohorts was scheduling flexibility. The most common Fall 2025 complaint was question stacking, matching the 31% stacked-turn rate (Section 4); the complaint receded in Spring 2026 free responses as the rate dropped to roughly 11%.

Accessibility cut both ways. Some Fall 2025 students with anxiety-related accommodations volunteered that they found the AI examiner less stressful than typical exams, suggesting the format could serve as a test-anxiety accommodation in some populations. For international students, the format cut the other way: "At times, the difficulty wasn't in knowing the material but in articulating my thoughts under pressure in English." We added a live transcript display for Spring 2026 in response to Fall 2025 requests; it lets students catch mistranscriptions in real time, but cognitive load under pressure remains (caveats and a fuller accessibility discussion in Appendix M).

One complaint increased in Spring 2026: the adaptive nature of questioning was perceived as "unfair"; students want exam difficulty to be the same for everyone. Asked to agree or disagree that the examiner "felt like a fair evaluator," 33% agreed in Fall 2025 and 56% in Spring 2026. Yet after grades were released, only two regrading requests came in, both Fall 2025, and no points were conceded in either case. The complaint was about process, not outcome.

6 Discussion

For anyone considering adoption, three gains hold even with audit overhead: the system automates the time-dominant interview phase, concentrates human review on high-disagreement cases, and makes oral assessment operationally feasible at cohort sizes where a faculty-administered oral is not. At those sizes, the realistic counterfactual is not AI-vs-human but AI-vs-written-only; the fuller argument is in Appendix G. The caveats start with what the grades mean. The grading council's high agreement establishes reliability (consistency among AI raters), not validity (correctness against expert judgment); the human auditor on roughly 3% of Spring 2026 exams is a guardrail, not a full fix. What we can claim is operational feasibility and promising reliability. What we cannot claim is that these grades carry the validity of a well-run faculty oral, or that the format is demonstrably fair to every student: the first would take blinded expert re-grading, the second a subgroup fairness analysis, and we have performed neither.

The format inverts the traditional cheating incentive. The LLM generates questions dynamically against published guidelines rather than from a fixed question bank, so there is no exam to leak. Practicing against the published structure is itself the learning behavior the format rewards. Two academic-dishonesty vectors still deserve treatment in the unsupervised home environment: *live AI coaching* via a headset or phone (not fully preventable; a broader societal concern), and a *confederate in the room* (deterred by the self-recorded video requirement). The vector-by-vector detail is in Appendix H.

Voice-based assessment introduces accessibility tradeoffs that a written format avoids. Speaking under time pressure adds cognitive load and can disadvantage students who articulate or process more slowly, or whose accented speech is transcribed less accurately; students with speech-related disabilities face barriers a keyboard would not impose. In practice, speech recognition handled the technical course vocabulary well, and because the speech-to-text output feeds an LLM that already has the course context, the system can in principle repair some accented-speech errors. But we did not measure word error rate against gold references and cannot quantify the residual. These tradeoffs call for explicit accommodations: extended or configurable time limits, a choice of voice and speaking pace, an ungraded practice run to defuse unfamiliarity, and a text-based alternative for students who cannot use voice. We detail these, alongside Spring 2026's live-transcript mitigation, in Appendix M.

Finally, data governance. The system transmits voice audio and transcripts to ElevenLabs and graded-transcript text to Anthropic, Google, and OpenAI for council grading, governed by each provider's API privacy policy and (where available) by enterprise/zero-retention agreements; we did not run an independent audit. Locally recorded webcam and audio (a course-integrity requirement, Section 3) are not transmitted to any vendor. Appendix M covers feedback quality and what a human examination provides that this system loses.

7 Conclusion

The spread of LLMs has disrupted traditional assessment. Rather than pursuing a detection arms race, we can adopt formats that inherently reward actual understanding. Oral examinations have always had this property but never scaled; voice AI restores scalability. Our deployment demonstrates operational feasibility across two cohorts (73 students). Agreement within the grading council is high. That is encouraging, but it does *not* by itself settle validity against double-blind expert human graders. Student experience was mixed. Students felt the exams tested their understanding but found them more stressful than written exams; the "more stressful" rate fell from the first cohort to the second as engineering fixes shipped, and question stacking, the most-cited culprit, fell with it. They dislike the adaptive nature of questioning as procedurally unfair; yet once grades were released, almost no one disputed them (Section 5).

The practical contribution is the five patterns from Section 1. Two iterations were enough to surface them; how far they carry is what we hope future deployments will show. The same infrastructure has an application beyond the final exam: attaching a short oral comprehension check to every written assignment would convert the system from a semester-end assessment tool into a continuous authenticity layer (Appendix N).

The online supplement includes the voice-agent and grading-council prompts and a sanitized snapshot of the code; the anonymized data and analysis scripts are archived at Zenodo (https://doi.org/10.5281/zenodo.21522314). Deploy it, use it, and report back what works and what breaks.

Acknowledgments

We thank Brian Jabarian for discussions on AI-administered interviews, Foster Provost for voice contribution, and Andrej Karpathy for the council-of-LLMs concept. Approved under NYU IRB-FY2023-7595.

References

[1] Tiffany Bayley, Kyle D. S. Maclean, and Tessa Weidner. 2026. Back to the Future: Implementing Large-Scale Oral Exams. *Management Teaching Review* 11, 1 (2026), 159–170. doi:10.1177/23792981241267744

[2] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *The Twelfth International Conference*

on Learning Representations (ICLR). https://arxiv.org/abs/2308.07201

[3] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7282–7296. doi:10.18653/v1/2021.acl-long.565

[4] Javier Coronado-Blázquez. 2025. Deterministic or probabilistic? The psychology of LLMs as random number generators. arXiv:2502.19965 [cs.CL] https://arxiv.org/abs/2502.19965

[5] Paul Denny, James Prather, Brett A. Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N. Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing Education in the Era of Generative AI. *Commun. ACM* 67, 2 (Jan. 2024), 56–67. doi:10.1145/3624720

[6] ElevenLabs. 2024. Conversational AI Platform. https://elevenlabs.io/docs/eleven-agents/overview; pricing reference: https://elevenlabs.io/pricing/agents. Agents documentation accessed July 2026; pricing page accessed May 2026 and re-verified July 2026 for the $0.08/minute additional-call overage rate cited in Section 3 and Appendix E.

[7] Andrea Fenton. 2025. Reconsidering the Use of Oral Exams and Assessments: An Old Way to Move Into a New Future. *Educational Researcher* 54, 7 (2025), 430–436. doi:10.3102/0013189X251333638

[8] Catherine Hartmann. 2025. Oral Exams for a Generative AI World: Managing Concerns and Logistics for Undergraduate Humanities Instruction. *College Teaching* (2025). doi:10.1080/87567555.2025.2558563

[9] Jui-Tse Hung, Christopher Cui, Diana M. Popescu, Saurabh Chatterjee, and Thad Starner. 2024. Socratic Mind: Scalable Oral Assessment Powered By AI. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. Atlanta, GA, 340–345. doi:10.1145/3657604.3664661

[10] Brian Jabarian and Luca Henkel. 2025. Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews. doi:10.2139/ssrn.5395709 SSRN 5395709.

[11] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. doi:10.1016/j.lindif.2023.102274

[12] Terence Kelly. 2025. What Every Experimenter Must Know About Randomization. *ACM Queue* 23, 6 (November/December 2025). doi:10.1145/3778029 Drill Bits column.

[13] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage, Thousand Oaks, CA.

[14] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. *Departmental Papers (ASC)* (2011). https://repository.upenn.edu/asc_papers/43/ University of Pennsylvania.

[15] Muhammed Ashraf Memon, Gordon Rowland Joughin, and Breda Memon. 2010. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. *Advances in Health Sciences Education* 15, 2 (2010), 277–289. doi:10.1007/s10459-008-9111-9

[16] Shashi Nallaya, Sheridan Gentili, Scott Weeks, and Katherine Baldock. 2024. The validity, reliability, academic integrity and integration of oral assessments in higher education: A systematic review. *Issues in Educational Research* 34, 2 (2024), 629–646. http://www.iier.org.au/iier34/nallaya.pdf

[17] André Nitze. 2024. Future-proofing Education: A Prototype for Simulating Oral Examinations Using Large Language Models. arXiv:2401.06160 [cs.CY] https://arxiv.org/abs/2401.06160

[18] OpenAI. 2023. New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/. Blog post, January 31, 2023; updated July 20, 2023, announcing the classifier's withdrawal "due to its low rate of accuracy".

[19] Mark D. Shermis and Jill Burstein (Eds.). 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, New York. doi:10.4324/9780203122761

[20] Vanderbilt University. 2023. Guidance on AI detection and why we're disabling Turnitin's AI detector. https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/. Brightspace guidance, August 16, 2023.

[21] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796 [cs.CL] https://arxiv.org/abs/2404.18796

[22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685