A Multi-Agent LLM Framework for Rating the Quality of Surgical Feedback

Rafal Kocielnik[1], J. Everett Knudsen[3], Steven Y. Cen[3], Jasmine Lin[2], Cherine H. Yang[2], Atharva Deo[2], Ujjwal Pasupulety[2], Peter Wager[2], Anima Anandkumar[1], Andrew J. Hung[1*]

[1]Computing + Mathematical Sciences, California Institute of Technology, 1200 E. California Blvd, Pasadena, 91125, CA, USA.
[2]Department of Urology, Cedars-Sinai, 8700 Beverly Blvd, Los Angeles, 90048, CA, USA.
[3]Keck School of Medicine, University of Southern California, 1500 San Pablo Street, Los Angeles, 90033, CA, USA.

*Corresponding author(s). E-mail(s): ajhung@gmail.com;
Contributing authors: rafal.kocielnik@gmail.com; everettknudsen@gmail.com; steven.cen@med.usc.edu; Jasmine.Lin@cshs.org; cherine.yang@cshs.org; atharva.deo@cshs.org; ujjwalpasupulety@gmail.com; peter.wager@cshs.org; anima@caltech.edu;

Abstract

Verbal feedback delivered by attending surgeons in the operating room plays a critical formative role in resident trainee skill acquisition. Yet, assessing the quality of trainer feedback and its effectiveness in influencing trainee behavior during live surgery remains a challenge. Prior studies assessed feedback content relying on extensive manual annotation by expert human raters and focused on developing broad taxonomies that overlook the qualitative aspects of feedback delivery such as clarity or urgency. Limited existing automated methods, including keyword analysis and topic modeling, also fail to capture these nuanced aspects. We introduce a two-stage LLM-based framework that discovers interpretable feedback quality criteria grounded in the context of surgical training. Our method uses multi-agent prompting and surgical domain knowledge injection to discover a small set of human interpretable scoring criteria

(e.g., *Encouraging*, *Urgent*, *Clear*). These criteria are then used to automatically score live surgical feedback via an LLM-as-a-judge approach. Evaluation on 4.2k trainer feedback instances demonstrates that our AI-discovered criteria outperform prior content-based frameworks in predicting feedback effectiveness, including observed trainee behavioral adjustments and trainer approval. This work advances scalable, human-aligned assessment of communication quality in the operating room and provides a foundation for improving surgical teaching practices.

Keywords: large language models, unsupervised discovery, surgical feedback, robot assisted surgery

1 Introduction

Formative verbal feedback to surgical trainees in the operating room (OR) plays a critical role in enhancing surgical education and outcomes [1]. High-quality feedback during surgical training is associated with improved intraoperative performance [2], faster acquisition of technical skills [3], and greater trainee autonomy [4]. Feedback in the OR is typically triggered by a trainer's observation of trainee behavior and is intended to shape future actions or decision-making. The effectiveness of a feedback utterance lies in its ability to make a trainee adjust their behavior or verbally acknowledge the feedback in a manner that elicits trainer approval. Understanding how feedback is delivered—its clarity, urgency, timeliness, actionability, and emotional tone—is essential to improving its effectiveness in surgical training. Yet, systematically quantifying these aspects in live settings remains an open challenge.

Previous research on surgical feedback has primarily focused on categorizing what instructors communicate during procedures. This includes typologies of trainer command types (e.g., guiding, questioning, chastising) [5], thematic content such as anatomy or instrument handling [6], discourse structures and planning strategies [7], and broad feedback categories (e.g., procedural, technical, praise, criticism) [8]. These frameworks have advanced understanding of instructional goals and content in surgical education. However, significantly less attention has been given to how feedback is delivered and interpreted. Systematic analysis is also difficult due to the combined need for specialized domain knowledge and labor-intensive annotation processes. Some automation methods have been proposed, including keyword-frequency techniques (e.g., LIWC [9]) and topic modeling with language model embeddings [10]. LIWC lacks sensitivity to clinical language, relies on predefined keyword dictionaries, and cannot capture delivery-focused attributes such as clarity, urgency, or instructional tone. Topic modeling approaches like BERTopic [11] cluster semantically related content (e.g., feedback on procedures such as *"sweeping and cutting"* or *"needle positioning"*) but similarly fail to distinguish meaningful differences in how feedback is delivered even in identical instructional contexts. For example, a directive like *"Move the needle to the left now"* carries a different instructional tone and urgency compared to *"Let's try adjusting the needle slightly to the left"*, despite both referring to the same

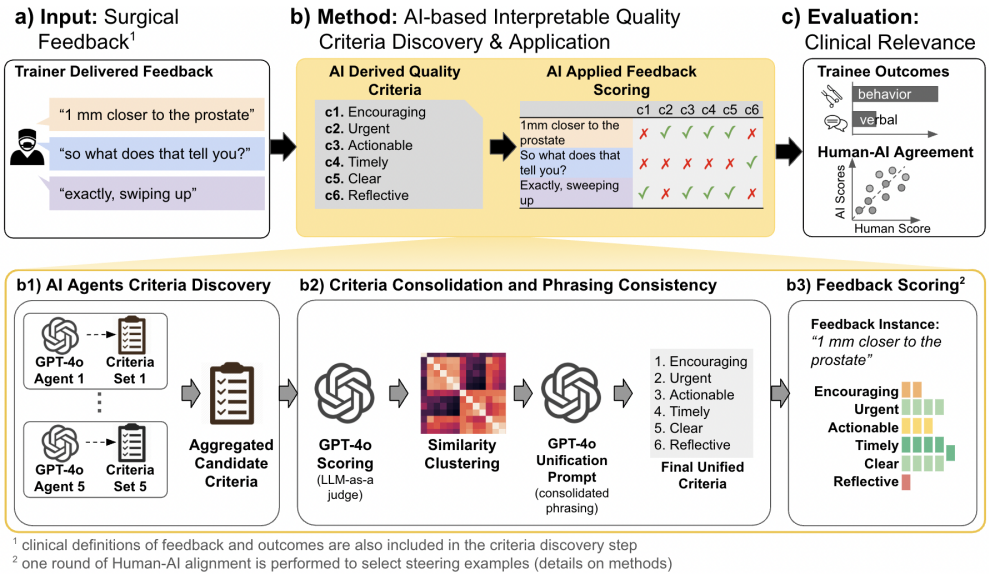

Fig. 1 Overview of our AI-based framework for interpretable discovery and evaluation of surgical feedback quality. (a) Trainer-delivered feedback during live procedures is used as input. (b) The method identifies and scores core feedback quality criteria. First, multiple GPT-4o agents generate candidate criteria sets (b1), which are then consolidated (b2) using LLM scoring, similarity clustering, and prompt-based unification into six core criteria: *Encouraging*, *Urgent*, *Actionable*, *Timely*, *Clear*, and *Reflective*. Each feedback instance is scored along these dimensions (b3) using LLM-as-a-judge approach. (c) Scored feedback is evaluated for clinical relevance through its association with observed trainee outcomes (verbal and behavioral) and agreement with human annotations.

task (*"instrument handling"*). These delivery aspects are critical for understanding feedback effectiveness from the trainee's perspective.

To address these challenges, we introduce an LLM-based framework for analyzing the delivery quality of surgical feedback (Figure 1). The framework takes as input (a) samples of raw transcripts of trainer-delivered verbal feedback during live surgical cases together with clinical domain knowledge (i.e., clinically validated definitions of feedback and effectiveness outcome criteria from [8]). This input is passed to a large language model to discover and operationalize core feedback quality criteria (b). In the **discovery phase (b1)**, we use multi-agent prompting with separate GPT-4o instances, where agents are seeded with both formal definitions and a representative subset of feedback examples to independently propose candidate evaluation criteria. In the **criteria consolidation phase (b2)**, these candidate criteria are then applied to feedback, clustered based on scoring similarity and unified into one phrasing through an additional GPT-4o consolidation prompt, producing a small set of stable, human-interpretable, and domain-grounded dimensions: *Encouraging*, *Urgent*, *Actionable*, *Timely*, *Clear*, and *Reflective*. In the **scoring phase (b3)**, the final criteria are applied to individual feedback instances using an LLM-as-a-judge setup, enabling automated scoring with human-interpretable rubric. We assess the clinical relevance of the criteria (c) by assessing their ability to predict behavioral outcomes of trainer

and trainee, in comparison to prior work; evaluating their alignment with human reasoning; and uncovering how different feedback quality properties lead to particular behavioral outcomes. This step validates the practical impact of feedback quality dimensions in real-world settings and their interpretability. Our framework enables fine-grained, scalable, and interpretable evaluation of feedback delivery, obviating the need for human annotation. Our approach departs from prior work in several key technical ways. Rather than relying on predefined taxonomies or unsupervised clustering methods such as BERTopic, we adopt a two-stage, LLM-guided strategy and combine it with prior clinical knowledge.

Our framework uncovers six interpretable feedback quality criteria—such as clarity, actionability, and urgency—that effectively predict trainee behavior change. The six quality criteria by themselves produce high AUC scores ranging from 0.71 to 0.75 for predicting trainee reaction to feedback (Table 1). Competitive analysis against existing feedback categorization frameworks from prior work, revealed that our criteria consistently improve prediction of trainee behavior by 9–12% and trainer reaction by 3-11%. In combination with content categories from prior work, our quality criteria reach AUCs ranging from 0.74 to 0.78 for trainee outcome prediction. Human scoring of feedback using the discovered criteria aligns substantially with LLM-applied scores (weighted $\kappa = 0.60$–0.70) for 5 of the 6 criteria, underscoring their interpretability and practical usability. On the methodological side, our results highlight the value of multi-agent prompting and consolidation steps for discovery of novel human-interpretable scoring criteria. This extends the prior approaches, such as LLM-as-a-judge relying on fixed rubric representing a priori provided criteria.

By enabling automated, interpretable assessment of feedback delivery grounded in real-world behavioral outcomes, our framework offers a practical tool for improving intraoperative teaching effectiveness and supporting trainer development. Beyond individual evaluations, this approach lays the groundwork for scalable integration into surgical education pipelines and clinical quality assurance systems, advancing the broader goal of optimizing communication-driven learning in high-stakes healthcare environments.

2 Results: Clinical Interpretation and Validation

Our process led to the discovery of 6 interpretable feedback quality criteria rated on a 5-point Behavioral Anchored Rating Scales (BARS). Using LLM-as-judge approach [12], where an LLM is asked to rate feedback using provided criteria and scales, we applied GPT-4o to rate 4210 lines of live surgical feedback collected in prior work [8]. We subsequently evaluated these quality criteria ratings for their ability to predict clinical effectiveness of feedback in affecting trainee behavior and leading to subsequent approval from a trainer. In this evaluation, we compared our AI discovered quality criteria to automated topic modeling approach from recent work [10] and fully manually annotated human expert proposed categories [8]. We further analyzed the statistical associations of the individual quality criteria with trainee behavioral adjustment and trainee verbal acknowledgment to understand how each quality criterion affects outcomes. Finally, we evaluated the ability of human raters with domain knowledge to

Feedback Criteria	Trainee Reaction to Feedback		Final Trainer Reaction to Trainee Reaction	
	Behavior Change	**Verbal Response**	**Approval**	**Disapproval**
AI-derived Quality Scores	$0.75_{\pm 0.01}$	$0.71_{\pm 0.02}$	$0.66_{\pm 0.02}$	$0.60_{\pm 0.06}$
Prior Topic Modeling [10]	$0.69_{\pm 0.02}$	$0.66_{\pm 0.02}$	$0.67_{\pm 0.02}$	$0.57_{\pm 0.06}$
+ AI Quality Scores	$0.77^{*}_{\pm 0.01}$	$0.73^{*}_{\pm 0.02}$	$0.69^{*}_{\pm 0.02}$	$0.63^{*}_{\pm 0.06}$
Prior Human-defined Categories [8]	$0.70_{\pm 0.02}$	$0.68_{\pm 0.02}$	$0.63_{\pm 0.02}$	$0.59_{\pm 0.06}$
+ AI Quality Scores	$0.78^{*}_{\pm 0.01}$	$0.74^{*}_{\pm 0.01}$	$0.69^{*}_{\pm 0.02}$	$0.62_{\pm 0.06}$

Table 1 Predicting trainee and trainer behavioral outcomes in reaction to feedback. Performance of LLM-driven interpretable scoring criteria discovery compared to prior approaches. We report mean AUROC ± 95% CI for predicting four behavioral outcomes from Wong et al. [8]: Behavior Adjustment, Verbal Acknowledgment, Trainer Approval, and Trainer Disapproval. AI-derived quality criteria are evaluated alone and in combination with prior topic modeling [10] and human-defined feedback categories [8]. Statistical significance (* $p < 0.05$) indicates that adding AI-derived quality scores led to a significant improvement, assessed via DeLong's test (95% CI of AUROC difference excludes 0).

use these AI-discovered quality criteria to rate the feedback instances consistently. Further details can be found in the *Methods* section.

2.1 Feedback Effectiveness Prediction

We evaluated the predictive performance of six LLM-derived feedback quality ratings across four behavioral outcomes using fivefold stratified cross-validation with Random Forest classifiers (Table 1).

Models using only the six quality ratings achieved strong performance, including AUROC=0.75, 95% CI: [0.74, 0.77] for *Trainee Behavior Change*, 0.71 [0.69, 0.72] for *Trainee Verbal Response*, 0.66 [0.64, 0.68]— for *Trainer Approval*, and 0.60 [0.54, 0.66] for *Trainer Disapproval* of Trainee Reaction. Augmenting prior topic modeling categories [10] with our AI-derived quality scores led to consistent improvements across all outcomes: AUROC=0.77 [0.76, 0.79] (+12% gain), 0.73 [0.71, 0.74] (+9%), 0.69 [0.67, 0.71] (+3%), and 0.63 [0.58, 0.69] (+11%), respectively. Similarly, augmenting prior manually annotated Human Categories [8] with our AI-derived quality ratings provided consistent gains in predictive performance, yielding AUROC=0.78 [0.77, 0.80] (+12% gain over human proposed categories) for trainee behavior adjustment, 0.74 [0.73, 0.76] (+9%) for verbal acknowledgment, 0.69 [0.67, 0.71] (+9%) for trainer approval, and 0.62 [0.56, 0.68] (+5%) for trainer disapproval.

To assess the significance of these gains, we applied DeLong's test for correlated AUROC curves (Table 6 in Methods). Compared to models using only Topic Modeling features, the addition of AI quality scores resulted in significant AUROC improvements for *Trainee Behavior Change* (Δ=+0.08, 95% CI: [0.07, 0.09]), *Trainee Verbal Response* (+0.06 [0.05, 0.07]), *Trainer Approval* (+0.02 [0.00, 0.04]), and *Trainer Disapproval* (+0.07 [0.00, 0.13]). Similar significant improvements were observed over Human Categories including gains of +0.08 [0.07, 0.10], +0.06 [0.05, 0.08], and +0.06

Quality Criterion	Human-Human		AI-AI		Human-AI	
	K	**95% CI**	**K**	**95% CI**	**K**	**95% CI**
Encouragement	0.79	(0.46, 0.91)	1.00	(1.00, 1.00)	0.72	(0.42, 0.86)
Urgency	0.72	(0.52, 0.85)	0.98	(0.94, 1.00)	0.68	(0.48, 0.82)
Actionability	0.76	(0.55, 0.88)	1.00	(1.00, 1.00)	0.79	(0.63, 0.89)
Timeliness	0.44	(0.24, 0.58)	0.94	(0.62, 1.00)	0.54	(0.32, 0.76)
Clarity	0.67	(0.47, 0.82)	0.92	(0.79, 1.00)	0.75	(0.49, 0.92)
Reflection	0.71	(0.42, 0.87)	0.98	(0.86, 1.00)	0.74	(0.41, 0.90)

Table 2 Agreement analysis across combinations of Human and AI raters.

Quadratic Weighted Kappa (**K**) scores and 95% confidence intervals (CIs) for quality scoring agreement across three rater configurations: two human raters (Human-Human), two AI runs (AI-AI), and average human vs. AI scoring (Human-AI). Score interpretation thresholds: 0.01–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (almost perfect agreement) [13].

[0.04, 0.08] for trainee behavior adjustment, trainee verbal acknowledgment, and trainer approval, respectively. The improvement for trainer disapproval was not statistically significant (95% CI includes 0). These results demonstrate that AI-derived quality dimensions offer statistically significant and additive value for predicting clinically relevant trainee and trainer reactions, and complement both automated content-based and manual expert-coded feedback aspects.

2.2 Alignment of AI scoring with Human Annotations

To evaluate alignment with human judgment, we took the AI-discovered quality rating definitions and asked two human raters with domain knowledge to apply them to 30 randomly selected feedback instances. Raters received a training session using a separate set of 30 examples. Further details can be found in the *Methods* section.

We evaluated inter-rater reliability across three configurations using quadratically weighted Cohen's kappa [14], which is appropriate for ordinal scales such as BARS. (Table 2): between two human raters (Human-Human), between two AI runs (AI-AI), and between the AI and the averaged scores of the human raters (Human-AI).

We observe substantial agreement among human raters across most dimensions (e.g., Encouragement: $K = 0.79$, 95% CI: [0.46, 0.91]; Actionability: $K = 0.76$, CI: [0.55, 0.88]; Urgency: $K = 0.72$, CI: [0.52, 0.85]). Agreement was lower for more subjective and contextual dimensions such as Timeliness ($K = 0.44$, CI: [0.24, 0.58]), suggesting inherent difficulty in consistently judging temporal aspects of feedback.

AI-generated ratings showed near-perfect internal consistency across repeated runs (AI-AI: $K = 0.92$–1.00), indicating deterministic and stable behavior. We note that these have been collected under the temperature setting of 0.0 to encourage deterministic behavior. Further details of AI setup using GPT-4o are provided in the *Methods* section. Importantly, Human-AI agreement was also substantial across most criteria (e.g., Actionability: $K = 0.79$, CI: [0.63, 0.89]; Clarity: $K = 0.75$, CI: [0.49, 0.92]),

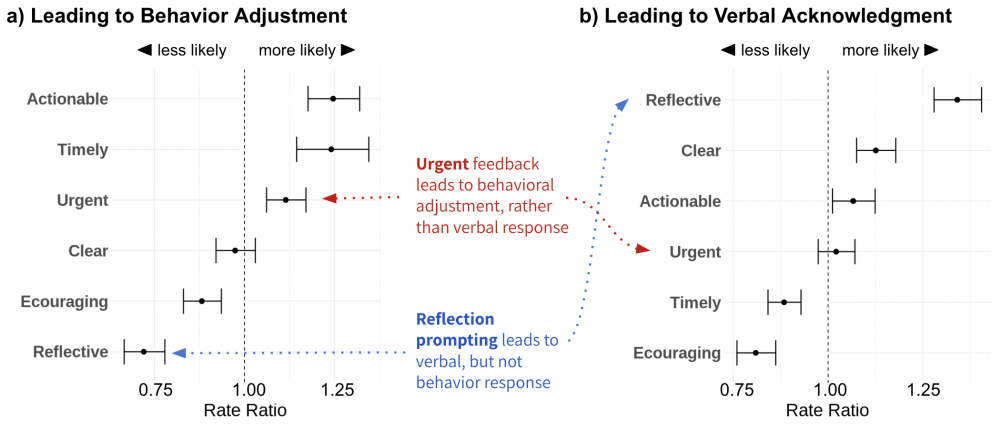

Fig. 2 Associations between discovered feedback quality criteria and trainee behavioral outcomes. Rate ratios and confidence intervals are shown for each of the six LLM-discovered feedback quality dimensions in relation to (a) *Trainee Behavioral Adjustment* and (b) *Trainee Verbal Acknowledgment* outcomes. Timely feedback was most predictive of behavioral change, while reflective and encouraging feedback were more strongly associated with verbal acknowledgment but not behavior change. These findings highlight how distinct delivery qualities of surgical feedback differentially influence trainee responses.

approaching inter-human agreement levels. This suggests that the AI model is not only consistent in its ratings but also well-aligned with expert human judgment, particularly on dimensions that are less subjective or more structurally grounded in language.

2.3 Criteria Association with Feedback Effectiveness

To understand the real-world impact of our discovered criteria, we next examine how each quality dimension relates to trainee behavioral adjustments and verbal acknowledgments following feedback.

Trainee Behavioral Adjustment was significantly associated with several feedback quality dimensions (Fig. 2a). Feedback rated as *Actionable* (Rate Ratio [RR] = 1.22, 95% CI: [1.18, 1.32]), *Timely* (RR = 1.24, CI: [1.14, 1.34]), and *Urgent* (RR = 1.11, CI: [1.06, 1.17]) was associated with higher rates of observed behavioral change. These dimensions reflect feedback that is specific, timely, and emphasizes the need for immediate action—elements that are directly conducive to real-time correction of performance. In contrast, *Encouraging* feedback (RR = 0.88, CI: [0.83, 0.94]) and *Reflective* feedback (RR = 0.72, CI: [0.67, 0.78]) were associated with reduced behavioral adjustment. This may be due to the nature of encouraging feedback, which often affirms correct behavior without requiring further adjustment, and reflective feedback, which aims to stimulate longer-term insight rather than immediate correction. *Clarity* did not show a statistically significant effect (RR = 0.97, CI: [0.92, 1.03]).

Trainee Verbal Acknowledgment showed a distinct pattern of associations with feedback quality dimensions (Fig. 2b). Feedback rated as *Reflective* (Rate Ratio [RR] = 1.34, 95% CI: [1.28, 1.40]) and *Clear* (RR = 1.13, CI: [1.08, 1.18]) was associated

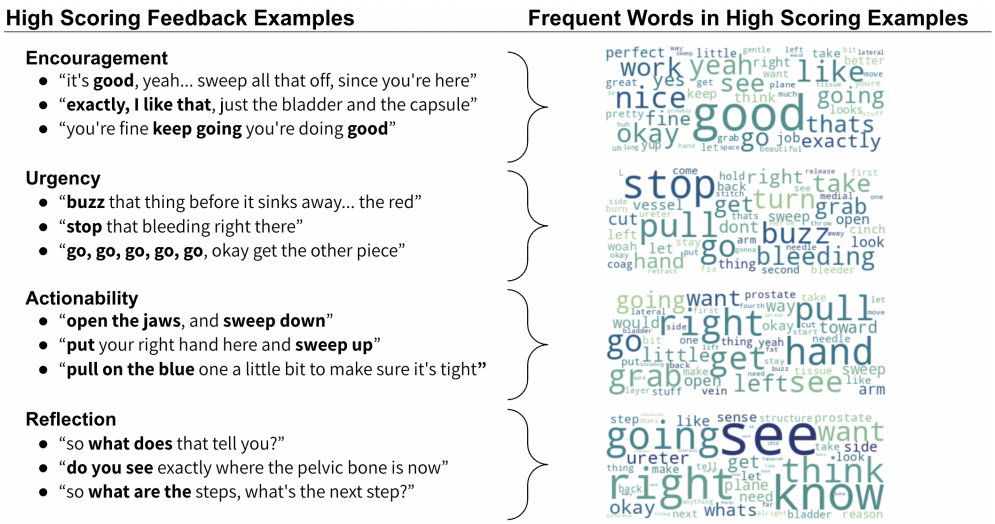

Fig. 3 Illustrative examples and linguistic patterns associated with high-scoring feedback. **Left:** Representative trainer feedback excerpts scoring highly on four of the discovered quality dimensions—*Encouragement*, *Urgency*, *Actionability*, and *Reflection*—demonstrate the linguistic structure and tone aligned with each dimension. **Right:** Word clouds visualize the most frequent terms in high-scoring instances across all criteria, highlighting key lexical patterns (Encouragement : "good", Urgency: "stop", Actionability: "pull", Reflection : "see") associated with effective feedback delivery in surgical training.

with higher rates of verbal acknowledgment. These findings suggest that verbal reactions are more likely when trainees are prompted to think or when feedback is easily understood. *Actionable* feedback showed a smaller but significant positive association (RR = 1.07, CI: [1.01, 1.12]). In contrast, *Encouraging* feedback (RR = 0.81, CI: [0.76, 0.86]) and *Timely* feedback (RR = 0.88, CI: [0.84, 0.93]) were associated with reduced acknowledgment. Again, encouraging feedback may act as affirmation, often concluding an interaction rather than prompting a response, while timely feedback may be delivered in fast-paced moments when verbal acknowledgment is less feasible. *Urgency* was not significantly associated with this outcome (RR = 1.02, CI: [0.97, 1.07]).

These results confirm distinct patterns of feedback effectiveness across outcomes. While *Actionable*, *Timely*, and *Urgent* feedback increased behavioral response rates, *Reflective* and *Clear* feedback were stronger predictors of verbal acknowledgment. *Encouraging* feedback consistently decreased the likelihood of both response types. Interestingly, *Timely* feedback had opposite effects—positively associated with behavior but negatively with acknowledgment—suggesting that different delivery styles selectively influence trainee behavior.

3 Discussion

The quality of verbal feedback in surgical training is essential for guiding real-time performance and promoting long-term skill development among surgical trainees [1–3].

Despite its importance, efforts to understand and evaluate feedback in the operating room (OR) have focused primarily on content categories—such as communication type [6], teaching behavior [5], or content themes [8], with limited focus on its quality from a trainee's perspective. These approaches also often rely on labor-intensive manual annotation [8], limited human-derived recognition of important delivery patterns [15], and lack robust validation linking feedback to behavioral outcomes [5–8].

Assessing the *quality* of surgical feedback—how it is delivered rather than just what is said—is critical for improving training outcomes. However, existing frameworks often overlook delivery dimensions like urgency, clarity, and timeliness, which are crucial for feedback effectiveness. Automated methods like topic modeling or keyword analysis fail to capture these nuances, and human-driven approaches, aside from largely omitting these aspects, are also not scalable. Our work addresses these gaps by introducing a large language model (LLM)-based framework that discovers and scores interpretable feedback quality dimensions, enabling scalable, clinically grounded evaluation aligned with trainee behavior.

Our approach surfaced a set of interpretable, behaviorally grounded feedback quality criteria that were both emergent and predictive across surgical training interactions. These criteria—*Encouragement*, *Urgency*, *Actionability*, *Clarity*, *Timeliness*, and *Reflection Prompting*—capture distinct pragmatic and pedagogical dimensions of trainer communication. Each dimension reflects a unique function: Encouragement denotes feedback that is supportive and provides positive reinforcement to boost confidence; Urgency reflects the communication that immediate action or attention is required; Actionability refers to clear, specific actions or steps the trainee can implement; Timeliness captures whether feedback is provided during or promptly after the trainee's action or decision; Clarity assesses whether the message is straightforward, unambiguous, and easily understood; and Reflection involves prompting the trainee to self-assess or reflect on their performance. The exact phrasing and scoring of behavioral anchors for each criterion with examples are presented in Tables A1 and A2

The predictive validity of these dimensions is underscored by their associations with observed trainee responses. As shown in Figure 2, Urgent feedback was strongly linked to immediate behavioral adjustments, whereas Reflective prompts were more likely to elicit verbal acknowledgment rather than action. These divergent associations suggest that different feedback types may selectively activate cognitive or behavioral processing in trainees. Some findings were counterintuitive: notably, higher levels of Encouragement were associated with a lower likelihood of both behavioral change and verbal acknowledgment. This likely reflects the role of encouragement as positive reinforcement—often used to affirm correct performance—thus requiring neither additional adjustment nor a verbal response. Similarly, Clarity was associated with increased verbal acknowledgment but showed no significant link to behavioral change. This pattern may be due to the orthogonality between clarity and actionability: a statement can be easy to understand without necessarily implying that action is needed. Clear feedback may also lower the threshold for verbal response, making it easier for trainees to affirm receipt. Finally, Figure 3 highlights representative linguistic patterns across criteria, further supporting their face validity and offering actionable insights for

feedback design. The emergence of these quality dimensions demonstrates the potential of AI-derived labels to structure, assess, and ultimately improve surgical feedback practices in real-world settings.

Our framework is the first to combine LLM-driven discovery of feedback quality dimensions with scalable rating via a behaviorally grounded rubric. Unlike prior methods, we do not rely on predefined rubrics or rigid taxonomies [12]. Instead, we task the LLM with discovering evaluative criteria from real surgical interactions and clinically grounded knowledge, enabling the model to surface delivery qualities relevant to actual training dynamics (i.e., what makes a good feedback). These criteria are defined in plain, human interpretable language, and scored on a Behaviorally Anchored Rating Scale (BARS), a well-established tool in psychometrics that enhances human interpretability and support Human-AI alignment through representative behavioral anchors [16]. This allows both humans and AI systems to consistently score feedback based on the same standards.

Technically, our approach diverges from traditional unsupervised clustering [17], topic modeling [11], or keyword-based methods [18] by focusing on delivery quality rather than content similarity.

Our method advances prior *"LLM-as-a-judge"* work [19] in two fundamental ways. First, we shift the role of the LLM from *applying* a fixed rubric to *discovering* that rubric de novo. We encourage completeness of the rubric by running five LLM *"brainstorming"* agents at a high sampling temperature (encouraging generation diversity [20]), each exposed to a different subsample of real feedback. This parallel, high-entropy generation uncovers less common yet behaviorally salient qualities—such as *Timeliness* and *Urgency*—that a single, low-temperature agent fails to surface. Second, we solve the discovery stability problem by passing the diverse candidate set through a deterministic consolidation stage. Hierarchical clustering on scoring correlations identifies semantically overlapping criteria, and the LLM then produces a finalized phrasing and BARS anchors for each cluster. Across five seeds, the final six-dimension rubric exhibits negligible lexical drift while retaining the conceptual breadth unlocked during the brainstorming step.

Expressing every criterion in plain language and anchoring it with illustrative examples enables dual use: the same rubric can be parsed reproducibly by an LLM at scale and interpreted reliably by human raters with minimal calibration. This unification of discovery, formalization and scoring distinguishes our pipeline from unsupervised topic models—whose latent factors are not scorable [21]—and from black-box classifiers, whose decision rationales remain opaque [22, 23]. By addressing *completeness*, *stability* and *interpretability* in a single workflow, we provide a principled path for deploying LLMs in safety-critical, clinician-facing settings.

We rigorously evaluated the utility of our discovered quality criteria across multiple dimensions. First, the criteria independently predicted trainee behavioral and verbal outcomes as well as follow-up trainer reactions better than prior topic-based and human-annotated baselines, showing 3–12% performance improvements. Second, we confirmed the criteria's interpretability in a human-rating study: human raters with clinical knowledge, applied the rubric to real feedback examples and achieved substantial agreement—both with each other and with an LLM on five of the six dimensions

(quadratic $\kappa = 0.60$–0.70); the remaining dimension showed moderate agreement. Third, our analysis uncovered meaningful associations between specific feedback qualities and subsequent trainee behavior. For instance, timely and actionable feedback strongly predicted behavior change, while reflective and clear feedback was more likely to prompt verbal acknowledgment. These findings confirm the practical validity of our dimensions and underscore their clinical relevance. Our framework delivers a scalable, transparent way to quantify intra-operative feedback quality—communication that shapes surgical trainee learning curves [2, 3] and ultimately impacts patient outcomes and safety [7, 24]. In practice, the rubric can underpin point-of-care dashboards that highlight especially actionable or unclear coaching, longitudinal curriculum analytics that flag faculty-wide gaps, and automated quality-assurance audits that monitor for safety-critical patterns such as high urgency coupled with low clarity. Because each criterion is expressed in plain language yet is automatically scorable by a large language model, the tool supports human-in-the-loop deployment: clinicians can inspect the rubric, contest low scores, and review exemplar anchors, ensuring both accountability and trust.

Yet, several limitations warrant mention. First, our analysis relied solely on text; prosodic cues in audio, instrument motion form video, and contextual OR events were not modeled and could modulate how feedback is interpreted by trainees. Second, our evaluation was retrospective. A prospective study that provides real-time rubric scores to trainers—and measures downstream behavioral change—will be an important next step that our lightweight, text-only pipeline readily supports. Finally, the rubric-discovery stage relied on GPT-4o, a proprietary model that may evolve over time. To reduce vendor lock-in we evaluated the application of the discovered rubric with human raters with clinical knowledge and we publish the final six-dimension rubric, BARS anchors and scoring prompt, enabling re-implementation with open-source LLMs or future clinical language models.

Although developed for the operating room, our discover-and-score pipeline can generalize to any clinical setting that relies on concise spoken guidance—such as ICU rounds, nursing shift reports, or telemedicine consultations. Because the model operates on plain text, a small speech-to-text transcript is all that must be transmitted; bulky audio or video files are unnecessary. This lightweight footprint enables remote mentorship and feedback-quality monitoring in bandwidth-constrained settings, with the potential to narrow global disparities in surgical training and, ultimately, patient outcomes. Furthermore applying our discover-and-score pipeline to other clinical domains—multidisciplinary tumor boards, emergency-department hand-offs, or virtual rehabilitation sessions—could yield domain-specific communication metrics that remain interpretable to frontline staff. Linking feedback-quality scores directly to downstream training or quality-of-care indicators would close the loop between communication analysis and measurable performance improvement, advancing the integration of transparent AI assistants into everyday digital-health workflows.

Category	Dimension	Count	Freq	Count/Case	Words/Line
Feedback	Instances	4210	100.0%	131.6 ± 77.6	8.3 ± 6.9
Trainee Behavior	Verbal Ack.	1944	46.2%	60.8 ± 31.7	10.2 ± 7.5
	Behavioral Adj.	1866	44.3%	58.3 ± 41.1	8.4 ± 6.5
Trainer Reaction	Approval	619	14.7%	19.3 ± 18.4	9.1 ± 6.9
	Disapproval	85	2.0%	2.7 ± 2.5	9.4 ± 6.5

Table 3 Statistics of behavior categories in our dataset. We report absolute counts (*Count*), relative frequency across all feedback instances (*Freq*), prevalence per surgical case (*Count/Case*), and mean words per utterance with standard deviation (*Words/Line*).

4 Methods

4.1 Ethics Approval

This study utilized datasets collected in accordance with strict ethical guidelines and approved by the Institutional Review Board (IRB) at the University of Southern California (HS-17-00113). All participants provided written informed consent prior to data collection. To ensure participant privacy and confidentiality, all datasets were de-identified before any model development or analysis was conducted.

4.2 Surgical Feedback Dataset

We used a dataset of real-world intraoperative feedback collected during robot-assisted surgeries, as introduced by Wong et al. [8]. Audio was recorded via wireless microphones worn by the surgical team, and synchronized endoscopic video was captured from the da Vinci Xi surgical system [25], providing a first-person surgical view. Using an external recording setup, audio and video streams were aligned and stored.

Utterances constituting surgical feedback—defined as trainer statements intended to modify trainee thinking or behavior—were manually identified and transcribed by surgical residents. Only utterances meeting this definition were included; other conversational content was excluded. The resulting dataset includes 4,210 feedback instances (Table 3).

Each instance was further annotated for two categories of behavioral outcomes. *Trainee Behavior* annotations captured whether the trainee responded to the feedback, including: (i) *Verbal Acknowledgment*, defined as a verbal or audible reaction confirming that the feedback was heard, and (ii) *Behavioral Adjustment*, defined as a behavioral change directly corresponding to the preceding feedback. *Trainer Reaction* annotations captured the trainer's response to the observed trainee behavior: (i) *Approval*, where the trainer verbally indicated satisfaction with the trainee's response, and (ii) *Disapproval*, where the trainer verbally demonstrated that they were not yet satisfied with the observed trainee behavioral change. The frequency, per-case prevalence, and average feedback length for each dimension are summarized in Table 3.

All annotation procedures followed standardized guidelines, and further details are available in the original dataset publication [8].

4.3 Automated Discovery of Feedback Quality Criteria

Our AI-based framework extracts interpretable feedback quality dimensions from surgical training data, designed to be scorable on a Behaviorally Anchored Rating Scale (BARS) [26, 27]. Inspired by prior tools for surgical skill assessment, such as OSATS [28], EASE [29], and DART [30], these criteria facilitate both human interpretation and automated scoring using LLMs. An overview is shown in Figure 1.

Domain-Guided Initiation

We initiate criteria discovery by injecting domain-specific knowledge into the prompt. Following in-context learning [31], GPT-4o is provided with formal definitions of key surgical feedback outcomes [8]—*Behavioral Adjustment*, *Verbal Acknowledgment*, and *Trainer Approval*—as well as a definition of feedback: *"Dialogue intended to modify trainee thinking or behavior."* These definitions are accompanied by 50 randomly selected unlabeled feedback examples (<1% of the dataset), enabling the LLM to infer relevant evaluative dimensions through analogical reasoning [32, 33]. Although LLMs are pre-trained with general language capabilities [34, 35], recent work indicates that the domain-specific information provided during prompting can help LLMs perform better by allowing them to reason appropriately within the context of the task [36]. Domain-specific information is especially important in specialized domains such as clinical natural language processing [37].

Multi-Agent Criteria Generation

We implemented a multi-agent setup where five GPT-4o instances independently propose candidate criteria. Each agent received a distinct random subset of 50 unlabeled feedback samples and was prompted to identify generalizable, abstract quality dimensions predictive of the defined outcomes. To encourage creative diversity, agents were configured with a high temperature ($T = 1.0$), which promotes variability in outputs—a key benefit in exploratory tasks like criteria discovery. Prior work has shown that higher temperatures enhance idea generation and reduce output homogenization, albeit at the cost of reduced determinism [38, 39]. Here, we prioritized discovering novel feedback qualities over reproducibility.

Critically, the prompt specified that each proposed dimension should: (1) be *definable in abstract terms*, meaning it must describe a generalizable quality applicable across feedback lines (rather than context-specific actions or examples), and (2) be feasibly scorable on a 5-point BARS scale based solely on a single transcribed feedback line, without access to surrounding dialogue or video context. Each agent was instructed to include behavioral anchors for three key levels: scores of 1, 3, and 5. These anchors served as representative examples for raters to interpret the quality being described. Each agent was asked to format its output in a structured tabular layout, listing the dimension name, its definition, and descriptions or examples of what constitutes a score of 1, 3, and 5. This structured prompting strategy ensured consistency

across generations and interpretability of the resulting criteria. The prompt wording is provided in Appendix A as *"Prompt Template for Quality-Criteria Discovery"*.

Criteria Consolidation and Definition Phrasing Finalization

We applied each agent's criteria to the full feedback dataset and computed a Spearman correlation matrix across all discovered dimensions. Hierarchical clustering (single linkage, Euclidean distance) revealed convergence patterns across agents. For cluster-number selection, we cut the dendrogram at successive values of k and computed the mean silhouette coefficient for each cut; the peak silhouette value occurred at $k = 6$, indicating six well-separated, internally cohesive clusters of semantically similar criteria [40]. For each cluster, GPT-4o (in deterministic mode, $T = 0.0$) synthesized a unified dimension definition to ensure reproducibility following [20, 38]. Prompts emphasized non-overlap, clarity, and applicability to isolated transcribed feedback, producing a refined set of six interpretable and domain-relevant feedback quality dimensions [20]. The prompt used is provided in Appendix A as *"Prompt Template for Criteria Consolidation Phrasing per Cluster"*.

Wording stability after five repetitions of the consolidation step was quantified with a cross-seed cosine-distance metric: each rubric definition was embedded using the `all-MiniLM-L6-v2` sentence-transformer [41], pair-wise cosine distances $(1-\cos\theta)$ were computed between definitions that shared the same cluster index but originated from different seeds, and the resulting values were averaged. Distances ≤ 0.05 are widely used as the near-duplicate threshold in large-scale text-deduplication pipelines [42, 43]; our mean distance of 0.02 therefore indicates negligible lexical drift. To evaluate whether conceptual breadth was retained, we tokenized each definition into uni-, bi- and tri-grams, embedded every term, and deemed a brainstorming term "covered" if its embedding showed cosine similarity ≥ 0.80 with any term in the consolidated rubric. This 0.80 cut-off is consistent with thresholds employed for near-duplicate detection and semantic-match evaluation in recent clinical-NLP studies [44, 45]. Under this criterion, the final six-dimension rubric covered 61.3 % of the vocabulary introduced during brainstorming, demonstrating that consolidation preserved the majority of the original conceptual space while standardizing phrasing.

To verify that the unified six-dimension rubric retained (or improved upon) the predictive signal discovered by individual agents, we first applied each set of criteria to every feedback instance using LLM-as-a-judge approach detailed in §4.4, yielding a vector of 5-point ordinal ratings—one score per dimension. These ratings were then used as predictors in a logistic-regression model for each behavioral outcome. We trained five agent-specific models (each using that agent's criteria vector) and one model based on the consolidated six-dimension rubric (Consolidated Criteria). Model performance was evaluated over three independent, stratified 80/20 train–test splits generated with three fixed random seeds. As summarized in Table 4, the consolidated rubric achieved the highest mean AUC across three of the four outcomes, with an above average AUC for the remaining outcome. These findings empirically support the hierarchical clustering and synthesis step, demonstrating that consolidation not only harmonizes terminology but also concentrates predictive signal for downstream modelling.

Table 4 Predictive power of quality–score criteria produced by each GPT-4o individual agent versus the final consolidated score ("Consolidated Criteria"). Values are mean AUC (± SD) across the three random-seed folds. Highest AUC for each outcome is bold-faced.

	Trainee Reaction to Feedback		**Trainer Reaction to Trainee Reaction**	
Quality Rubric source	**Behavior Adj.**	**Verbal Ack.**	**Approval**	**Disapproval**
GPT-4o Agent #1	0.66 ± 0.01	0.62 ± 0.02	0.63 ± 0.01	0.58 ± 0.06
GPT-4o Agent #2	0.67 ± 0.01	0.63 ± 0.02	0.65 ± 0.01	$\mathbf{0.64} \pm 0.02$
GPT-4o Agent #3	0.73 ± 0.02	0.68 ± 0.01	0.63 ± 0.01	0.62 ± 0.01
GPT-4o Agent #4	0.71 ± 0.01	0.67 ± 0.02	0.63 ± 0.01	0.63 ± 0.04
GPT-4o Agent #5	0.73 ± 0.03	0.64 ± 0.01	0.63 ± 0.01	0.62 ± 0.06
Consolidated Criteria	$\mathbf{0.74} \pm 0.01$	$\mathbf{0.71} \pm 0.02$	$\mathbf{0.66} \pm 0.02$	0.63 ± 0.01

4.4 Automated Feedback Scoring Based on Discovered Criteria

To systematically assess surgical feedback quality at scale, we employed a large language model (GPT-4o) to score real-world transcribed feedback instances using the rubric developed through our discovery process (Tables A1 and A2). This process follows the *LLM-as-a-judge* paradigm [19, 46], where the language model acts as a consistent evaluator applying structured criteria.

Input for Scoring

Each feedback instance was independently annotated by GPT-4o, which was prompted with (a) the full set of six scoring criteria along with their definitions and representative examples (Tables A1, A2), and (b) the specific feedback text to be rated. This structured input format enabled the model to interpret and apply the rubric definitions grounded in a behaviorally anchored rating scale (BARS). The prompt structure used for this task is detailed in Appendix A as *"Prompt Template for Multi-Criteria Feedback Scoring"*.

LLM-as-a-judge Scoring

The LLM assigned a score from 1 to 5 for each of the six criteria, returning a structured list of numerical ratings per feedback instance. All instances were scored individually in separate API calls, without batching, to minimize potential cross-instance contamination and data leakage [47]. This approach is consistent with recent methodological best practices in judgment elicitation with LLMs [12].

To ensure reproducibility and reduce variance in scoring, we set the model's temperature to 0.0, encouraging deterministic outputs. To evaluate scoring consistency, we conducted a repeated annotation of each feedback instance using the same model and prompt configuration, enabling calculation of inter-rater agreement for each quality dimension.

Human-AI Alignment Calibration

To assess the alignment between LLM-based and human judgment, we conducted a calibration study on a stratified sample of 30 LLM-scored feedback instances. For each

of the six discovered feedback quality dimensions, we selected five examples spanning the full BARS scoring spectrum: two instances with high scores (4 or 5), one with a mid-range score (3), and two with low scores (1 or 2). This stratification ensured representative coverage across the rating scale for every dimension.

Two human raters with surgical domain knowledge independently evaluated these instances, using the original LLM-discovered definitions and applying the same 5-point BARS scoring system. Raters were blinded to the LLM-generated scores to prevent bias. Following the initial rating phase, any disagreements of two or more points on the BARS scale were discussed collaboratively, allowing the raters to reconcile interpretations and establish a consensus score for each instance.

Importantly, in cases where both human raters independently agreed on a score that differed from the LLM's original rating, these instances were incorporated as new illustrative examples into the scoring rubric (up to two new examples per anchor). This iterative grounding process reinforced the behavioral anchoring of each quality dimension and aligned with rubric refinement practices recommended by prior work on human-AI collaboration in alignment tasks [48].

4.5 Comparison to Existing Automated Criteria Extraction

Most prior approaches to discovering evaluative dimensions from text rely on unsupervised topic modeling (e.g., LDA, BERTopic) or black-box LLM scoring frameworks like LLM-as-a-judge. While topic models can surface latent themes, they do not yield actionable, interpretable evaluation criteria aligned with domain-specific outcomes such as behavioral adjustment or trainer approval. Similarly, LLM-based scorers often replicate pre-defined preferences or rubrics without uncovering novel dimensions. Moreover, these systems typically do not provide explanations for their decisions unless paired with post hoc interpretability methods that generate human-understandable rationales or explanations [49]. These methods fall short in safety-critical domains like surgical education, where human-verifiable, domain-grounded criteria are essential for both evaluation and training.

Our method addresses this gap through a two-stage, LLM-assisted pipeline. In the first stage, we initiate domain-grounded criteria discovery using a multi-agent prompting strategy, where separate LLM agents generate candidate evaluation criteria informed by clinical definitions and feedback examples. These outputs are then consolidated into a set of interpretable, BARS-compatible rating dimensions. In the second stage, we leverage these discovered criteria to score new feedback instances using a structured LLM-as-a-judge approach. This enables transparent, repeatable evaluation aligned with domain outcomes, bridging the strengths of human-centered scale development and LLM-based automation. A comparison of the properties of our method against prior approaches is presented in Table 5, illustrating its unique ability to support interpretable discovery, domain-grounded scoring, and dual usability by both humans and LLMs.

Table 5 Comparison of existing methods for automated evaluation dimension discovery and scoring. Our method uniquely combines interpretable criteria discovery with scoreability and dual usability by humans and LLMs. Unlike topic modeling or direct LLM scoring approaches, it enables domain-grounded, BARS-compatible dimension induction and structured evaluation.

Method	Discovers Dimensions	Scoreable Criteria	Domain Grounded	Inter-pretable	Human + LLM Scoring
LDA / BERTopic	✓	✗	✗	Ambiguous	✗
LLM-as-a-Judge	✗	✓	Tuned	✗	LLM-only
Our Method	✓	✓	✓	✓	✓

4.6 Predictive Modeling of Behavioral Outcomes

We evaluated the predictive utility of LLM-discovered feedback quality criteria by modeling four behavioral outcomes annotated by human experts in [8]: *Trainee Behavior Change*, *Verbal Response*, *Trainer Approval*, and *Trainer Disapproval*. We implemented a fivefold *stratified* cross-validation procedure to ensure that each fold preserved the original class distribution for the respective outcome. Within each training fold, we performed nested hyperparameter tuning using an inner fivefold stratified cross-validation loop. Random Forest classifiers were tuned via grid search over the following parameter ranges: number of estimators `[200, 300, 400, 500, 1000]`, maximum number of features `[10, 25, 50]`, maximum tree depth `[20, 50]`, and minimum samples per leaf `[5, 20]`. The Gini impurity index was used as the splitting criterion, and the best hyperparameters were selected based on AUROC performance on the validation folds. To further address the effects of class imbalance during model training—particularly for less frequent outcomes such as *Trainer Disapproval*—we applied class weighting using King's method [50], which adjusts model estimation to reduce bias in rare event prediction.

We compared models using three types of features: (1) our AI-derived quality scores, (2) previously published topic modeling features [10], and (3) previously published human-defined feedback categories [8], both individually and in combination. Feature matrices were constructed accordingly, and the same cross-validation protocol was applied across all feature configurations.

Each fold's predictions were evaluated using standard classification metrics, including Area Under the Receiver Operating Characteristic Curve (AUROC), accuracy, precision, and recall. We report AUROC with 95% confidence intervals calculated using the DeLong method [51] over the pooled held-out predictions. Model performance across feature configurations and behavioral outcomes is presented in Table 1.

4.7 Model Comparison and Statistical Significance Testing

To assess the independent predictive contributions of AI-derived quality scores and previously published feedback annotations, we conducted a comparative analysis of model performance using DeLong's test for correlated receiver operating characteristic (ROC) curves. Specifically, we built three models for each behavioral outcome: (1) a *full model* incorporating both AI-derived quality scores and prior annotation features

Behavior Outcome	Full Model		Without AI Quality		Without Prior	
	AUC	95% CI	ΔAUC	95% CI	ΔAUC	95% CI
Compared to Topic Modeling [10]						
Beh. Adjustment	0.77	(0.76, 0.79)	−0.08*	(-0.09, -0.07)	−0.02*	(-0.03, -0.01)
Verb. Acknowledge	0.73	(0.71, 0.74)	−0.06*	(-0.07, -0.05)	−0.02*	(-0.03, -0.01)
Trainer Approval	0.69	(0.67, 0.71)	−0.02*	(-0.04, -0.00)	−0.03*	(-0.04, -0.01)
Trainer Disapproval	0.63	(0.58, 0.69)	−0.07*	(-0.13, -0.00)	0.03	(-0.09, 0.02)
Compared to Human Categories [8]						
Beh. Adjustment	0.78	(0.77, 0.80)	−0.08*	(-0.10, -0.07)	−0.03*	(-0.04, -0.02)
Verb. Acknowledge	0.74	(0.73, 0.76)	−0.06*	(-0.08, -0.05)	−0.03*	(-0.04, -0.02)
Trainer Approval	0.69	(0.67, 0.71)	−0.06*	(-0.08, -0.04)	−0.03*	(-0.04, -0.01)
Trainer Disapproval	0.62	(0.56, 0.68)	−0.03	(-0.09, 0.03)	−0.02	(-0.07, 0.03)

Table 6 AUROC values for models using both **AI Quality Ratings** and **Prior Work Categories**, and the respective drops in AUROC when either feature set is removed. Rows are grouped by comparison baseline. Statistically significant drops (Δ AUC 95% CI excluding 0) are underlined.

(topic modeling [10] and human-defined categories [8]), (2) a *reduced model without AI quality scores*, and (3) a *reduced model without prior annotation features*.

All models were trained using Random-Forest classifiers with class-weight balancing to account for outcome imbalance. Hyper-parameters were tuned by nested cross-validation: a fivefold stratified outer loop estimated generalization performance, while a threefold stratified inner loop executed a grid search over the number of trees `[200, 300, 400, 500, 1000]`, the number of features considered at each split `[10, 25, 50]`, the maximum tree depth `[20, 50]`, and the minimum samples per leaf `[5, 20]`. The hyper-parameter set that maximized AUROC within the inner loop was refit on the full outer-training fold and then evaluated on its corresponding held-out outer-test fold. AUROC scores were computed on these outer-test sets, and DeLong's method provided 95 % confidence intervals (CIs) both for each individual AUROC and for the difference in AUROC (ΔAUROC) between full and reduced models.

The DeLong test accounts for the covariance structure of paired predictions, enabling statistically principled comparison of correlated ROC curves. A feature set was deemed to have significant independent predictive value when the 95 % CI for ΔAUROC excluded zero. AUROC values, ΔAUROC, and their CIs are summarized in Table 6; statistically significant drops in AUROC following the removal of a feature set are interpreted as evidence for its contribution to predictive performance.

4.8 Association Between Feedback Quality and Outcomes

To assess how distinct feedback quality dimensions relate to subsequent *trainee behavioral adjustments* or *trainee verbal acknowledgment*, we fit a generalized linear mixed model (GLMM) with a Poisson distribution and log link function. The binary outcome indicated whether a feedback instance was followed by a behavioral change from the trainee. Six quality dimensions were entered as fixed effects: Encouraging, Urgent, Actionable, Timely, Clear, and Reflective. A random intercept for surgical case was

included to account for clustering across repeated observations within the same case. We report exponentiated fixed effects as rate ratios, with 95% confidence intervals.

4.9 Human-AI Alignment Evaluation

To evaluate alignment with human judgment, we took the AI-discovered quality rating definitions and asked two human raters with domain knowledge to apply them to a stratified sample of 30 feedback instances. These instances were selected to ensure coverage across the full scoring range (1–5) for each quality dimension. Specifically, for each of the six feedback quality dimensions, we sampled five examples: two with high scores (4 or 5), one with a mid-range score (3), and two with low scores (1 or 2). To avoid repetition, already sampled examples were excluded in subsequent selections using index tracking. This process ensured that the evaluation set reflected the diversity of scoring scenarios across all criteria.

Prior to annotation, raters participated in a calibration session using a separate set of 30 feedback examples. Human-human disagreements from this session were used for discussion among annotators towards reaching interpretation consensus following best practices in qualitative coding [52, 53], while human-AI disagreements were used as examples incorporated into the scoring definitions for AI behavior steering following in-context-learning principle [54], as described in Section §4.4. Final inter-rater agreement results were computed using the unseen 30-instance evaluation set.

We calculated quadratically weighted Kappa score following best practices from [16, 55]. Quadratically-weighted κ was chosen because Behaviorally Anchored Rating Scales (BARS) deliberately place the extreme anchors much farther apart—conceptually and clinically—than adjacent mid-points. Quadratic weights reflect this by squaring the category distance, heavily penalizing large misclassifications, whereas linear weights down-weight all disagreements in a strictly proportional (less discriminating) fashion [56]. Quadratic weighting is therefore recommended for ordered clinical/BARS-style rubrics and is the form originally proposed by **(author?)** [57].

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability

All GPT-4o interactions were run using standard OpenAI API [58]. The precise wording of the system and user prompts at each framework stage is provided in Supplementary Materials. Agglomerative clustering analysis has been performed using a standard scikit-learn implementation [59]. The integration code is available from the corresponding author upon reasonable request. Feedback effectiveness prediction analysis has been performed using standard Scikit-learn implementations including RandomForestClassifier [60], GridSearchCV [61], and StratifiedKFold [62]. Inter-rater reliability was assessed using standard weighted Kappa implementation from

Scikit-learn [63]. Association analysis was performed using R implementations of the generalized linear mixed-effects (`glmer`) from the `lme4` package (v1.1.37) and estimated marginal means were computed with the `emmeans` package (v1.11.2); we did not build custom code for machine learning evaluation or association analysis.

Acknowledgments

This study was supported in part by the National Cancer Institute under Award Numbers R01CA251579 and R01CA298988. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author Contributions

R.K.: conceptualization, methodology, evaluation, experimental analysis and visualization, writing original draft, review and editing. J.E.K.: conceptualization, evaluation, review and editing S.Y.C.: methodology, experimental analysis, writing review and editing. J.L., C.Y.: data curation and evaluation. A.D., U.P.: supervision, writing review and editing. P.W., A.A., J.H.: conceptualization, funding acquisition, supervision, writing review and editing.

Competing Interests

Rafal Kocielnik declares no competing financial or non-financial interests
J. Everett Knudsen declares no competing financial or non-financial interests
Steven Y. Cen declares no competing financial or non-financial interests
Jasmine Lin declares no competing financial or non-financial interests
Cherine H. Yang declares no competing financial or non-financial interests
Atharva Deo declares no competing financial or non-financial interests
Ujjwal Pasupulety declares no competing financial or non-financial interests
Peter Wager declares no competing financial or non-financial interests
Anima Anandkumar declares no competing financial or non-financial interests
Andrew J. Hung declares no competing non-financial interests, but reports financial disclosures with Intuitive Surgical, Inc. and Teleflex, Inc.

Appendix A Discovered Feedback Quality Scoring Criteria

Scale	Score 1 (None)	Score 3 (Moderate)	Score 5 (High)
Encouragement			
Definition	Neutral or negative tone; no observable encouragement.	Mild affirmations or neutral support without strong reinforcement.	Explicit, enthusiastic encouragement clearly aimed at boosting confidence.
Examples	*"That's wrong."* *"Yes, it's to define the bladder neck. . . "* *"Do you see where the pelvic bone is now?"*	*"That's fine."* *"Good."* *"Keep going, smiley face"*	*"Excellent work!"* *"Perfect."* *"Good, I love that."*
Urgency			
Definition	Calm or delayed feedback; no urgency communicated.	Prompting language or tone suggests some urgency without direct commands.	Explicit, critical urgency with immediate calls to action.
Examples	*"You might want to adjust that later."* *"This one should have been further distal."* *"Maybe next time."*	*"Watch your positioning."* *"You go a little more."* *"Let's work distally."*	*"Immediately stop!"* *"Correct your grip now!"* *"Don't coag!!"*
Actionability			
Definition	No actionable content; vague or evaluative.	Some guidance, but lacks precision.	Highly precise, step-by-step action that is immediately executable.
Examples	*"That's not right."* *"This one should have been even further distal."* *"Be more careful."*	*"Adjust your grip."* *"Keep going, smiley face."* *"Ok, open that up."*	*"Move your needle holder 2 cm forward."* *"Angle it down by 30 degrees."* *"Sweep left and then buzz."*

Table A1 AI-discovered feedback quality scoring criteria. Each dimension is rated using a 5-point Behaviorally Anchored Rating Scale (BARS); for clarity, only anchors for scores 1 (Minimal), 3 (Moderate), and 5 (High) are shown. Definitions and representative examples are provided for each anchor.

Appendix B GPT-4o prompts

Scale	Score 1 (None)	Score 3 (Moderate)	Score 5 (High)
Timeliness			
Definition	Feedback is delayed significantly, unrelated to immediate action.	Moderately prompt; refers to recent action but not immediate.	Immediate, real-time feedback during the action.
Examples	*"This one should have been even further distal."* *"Maybe next time."* *"Before you continue, let's review."*	*"So you start here."* *"This is where I want you to start."* *"You did that too fast."*	*"It's still bleeding—you gotta stop that."* *"Do you see exactly where the pelvic bone is now?"* *"You're coag'ing again!"*
Clarity			
Definition	Confusing or ambiguous; unclear what is meant.	Mostly clear but includes minor ambiguities.	Exceptionally clear, concise, and unambiguous.
Examples	*"You know what to do."* *"Good."* *"Let me see how much is bleeding."*	*"Adjust it a bit."* *"That's it, that's all you're gonna get."* *"Ok, open that up."*	*"Insert the needle at a 45-degree angle just above the marked point."* *"How many knots did you do there?"* *"Move the instrument to the left."*
Reflection			
Definition	No reflective element; purely directive or evaluative.	Occasional reflective prompt but lacks depth.	Strong, open-ended reflective guidance fostering deep evaluation.
Examples	*"Before you do the next step, clean your lens."* *"Closer to the prostate."* *"Push a little further."*	*"What went wrong there?"* *"Do you see that periurethral stuff?"* *"Do you see where the pelvic bone is now?"*	*"How could you adjust your technique to improve precision?"* *"What's the next step?"* *"Why did you choose that approach?"*

Table A2 Scoring criteria for the final two dimensions, **Clarity** and **Reflection**, discovered via AI-based analysis of surgical feedback. Anchors correspond to scores 1 (Minimal), 3 (Moderate), and 5 (High) on a 5-point Behaviorally Anchored Rating Scale (BARS).

References

[1] Agha, R. A., Fowler, A. J. & Sevdalis, N. The role of non-technical skills in surgery. *Annals of medicine and surgery* **4**, 422–427 (2015).

Prompt Template for Quality-Criteria Discovery

System Instruction:
You are working in the context of *verbal feedback* delivered by a trainer to a trainee in a live surgery. The goal of the feedback is to modify trainee thinking or behavior. There are different measures assessing feedback effectiveness, including:

- **Trainee Behavior Change** — behavioral adjustment made by the trainee that corresponds directly with the preceding feedback (e.g. trainee immediately pulls more tightly on the suture thread after receiving feedback to cinch tightly);
- **Trainee Verbal Acknowledgment** — verbal or audible confirmation by the trainee confirming that they have heard the feedback (e.g. "Okay, I see", "uh-huh, got it.");
- **Trainer Approval** — trainer verbally demonstrates that they are satisfied with the trainee behavioral change (e.g. "yes", "mhm").

Based on these descriptions, propose dimensions that would be predictive of the three outcomes above. For each dimension, supply a definition such that a rater could score a feedback instance on the Behaviorally Anchored Rating Scale (BARS) using 5 behavioral anchor levels, from **1 = feedback does *not* exhibit this quality** to **5 = feedback clearly possesses this quality**. Dimensions must be applicable to *transcribed feedback lines alone*—without preceding dialogue, video, or timing information.

User Message:
Produce an output in the format:
`No|Dimension Name|Scoring Definition|Score 1 rating|Score 3 rating|Score 5 rating`
Do **not** include this header in your reply.

Prompt Template for Criteria Consolidation Phrasing per Cluster

System Instruction:
You are given a set of *similar scoring criteria*, each with a name and definition. Combine them under one unified name and definition. Consolidate into **exactly one** refined criterion based on the list below:

Name: `[Criterion 1 Name]`, Definition: `[Criterion 1 Definition]`
Name: `[Criterion 2 Name]`, Definition: `[Criterion 2 Definition]`
...

User Message:
Based on the consolidated and refined criterion, output a Python tuple in the form:
`(No, "Consolidated Name", "Consolidated Definition")`
Only return the tuple—no additional commentary.

[2] Bonrath, E. M., Dedy, N. J., Gordon, L. E. & Grantcharov, T. P. Comprehensive surgical coaching enhances surgical skill in the operating room. *Annals of surgery* **262**, 205–212 (2015).

[3] Ma, R. *et al.* Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills—an unblinded pilot

Prompt Template for Multi-Criteria Feedback Scoring

System Instruction:
This is verbal FEEDBACK delivered during surgery by a trainer to a trainee.
Please rate it given each of the following criteria and associated scales.

Q1. [Criterion 1]
Q2. [Criterion 2]
Q3. [Criterion 3]
Q4. [Criterion 4]
Q5. [Criterion 5]
Q6. [Criterion 6]

Make the scoring concise as it needs to be parsed automatically later on; use the format of an ordered Python list, don't repeat question numbers:
Q1 score, Q2 score, Q3 score, Q4 score, Q5 score, Q6 score

User Message:
FEEDBACK: "[feedback_line]"

randomized controlled trial. *The Journal of Urology* **208**, 414–424 (2022).

[4] Haglund, M. M. *et al.* The surgical autonomy program: a pilot study of social learning theory applied to competency-based neurosurgical education. *Neurosurgery* **88**, E345–E350 (2021).

[5] Hauge, L. S., Wanzek, J. A. & Godellas, C. The reliability of an instrument for identifying and quantifying surgeons' teaching in the operating room. *The American journal of surgery* **181**, 333–337 (2001).

[6] Blom, E. *et al.* Analysis of verbal communication during teaching in the operating room and the potentials for surgical training. *Surgical endoscopy* **21**, 1560–1566 (2007).

[7] D'Angelo, A.-L. D., Ruis, A. R., Collier, W., Shaffer, D. W. & Pugh, C. M. Evaluating how residents talk and what it means for surgical performance in the simulation lab. *The American Journal of Surgery* **220**, 37–43 (2020).

[8] Wong, E. Y. *et al.* Development of a classification system for live surgical feedback. *JAMA Network Open* **6**, e2320702–e2320702 (2023).

[9] Ramprasad, A. *et al.* Language in the teaching operating room: expressing confidence versus community. *Journal of Surgical Education* **81**, 556–563 (2024).

[10] Kocielnik, R. *et al.* Human ai collaboration for unsupervised categorization of live surgical feedback. *npj Digital Medicine* **7**, 372 (2024).

[11] Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).

[12] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36**, 46595–46623 (2023). URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[13] Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *biometrics* 159–174 (1977).

[14] McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia Medica* **22**, 276–282 (2012). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/. PMID: 23092060.

[15] Quesada, S. P., Calkins, C. & Jeglic, E. L. An examination of the interrater reliability between practitioners and researchers on the static-99. *International Journal of Offender Therapy and Comparative Criminology* **58**, 1364–1375 (2014).

[16] Holland, J. R. *et al.* Reliability of the behaviorally anchored rating scale (bars) for assessing non-technical skills of medical students in simulated scenarios. *Medical Education Online* **27**, 2070940 (2022).

[17] Liu, T., Yu, H. & Blair, R. H. Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics* **14**, e1575 (2022).

[18] Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* **29**, 24–54 (2010).

[19] Gu, J. *et al.* A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[20] Patel, D. *et al.* Exploring temperature effects on large language models across various clinical tasks. *medRxiv* 2024–07 (2024).

[21] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems* **22** (2009).

[22] Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **16**, 31–57 (2018).

[23] Ribeiro, M. T., Singh, S. & Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 1135–1144 (2016).

[24] Greenberg, C. C. *et al.* Association of a statewide surgical coaching program with clinical outcomes and surgeon perceptions. *Annals of surgery* **273**, 1034–1039 (2021).

[25] Freschi, C. *et al.* Technical review of the da vinci surgical telemanipulator. *The International Journal of Medical Robotics and Computer Assisted Surgery* **9**, 396–406 (2013).

[26] Schwab, D. P., Heneman III, H. & DeCotiis, T. A. Behaviorally anchored rating scales: A review of the literature. *Academy of Management Proceedings* **1975**, 222–224 (1975).

[27] Jacobs, R., Kafry, D. & Zedeck, S. Expectations of behaviorally anchored rating scales. *Personnel psychology* **33**, 595–640 (1980).

[28] Van Hove, P., Tuijthof, G., Verdaasdonk, E., Stassen, L. & Dankelman, J. Objective assessment of technical surgical skills. *Journal of British Surgery* **97**, 972–987 (2010).

[29] Haque, T. F. *et al.* An assessment tool to provide targeted feedback to robotic surgical trainees: development and validation of the end-to-end assessment of suturing expertise (ease). *Urology practice* **9**, 532–539 (2022).

[30] Vanstrum, E. B. *et al.* Development and validation of an objective scoring tool to evaluate surgical dissection: dissection assessment for robotic technique (dart). *Urology practice* **8**, 596–604 (2021).

[31] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* **35**, 22199–22213 (2022).

[32] Ozturkler, B., Malkin, N., Wang, Z. & Jojic, N. Thinksum: Probabilistic reasoning over sets using large language models. *arXiv preprint arXiv:2210.01293* (2022).

[33] Wang, X. *et al.* Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747* (2022).

[34] Jiang, K., Mujtaba, M. M. & Bernard, G. R. Large language model as unsupervised health information retriever. *Caring is Sharing–Exploiting the Value in Data for Health and Innovation* 833–834 (2023).

[35] Wei, J. *et al.* Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[36] Maharjan, J. *et al.* Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports* **14**, 14156 (2024).

[37] Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S. & Wang, Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics* **12**, e55318 (2024).

[38] Windisch, P. *et al.* The impact of temperature on extracting information from clinical trial publications using large language models. *Cureus* **16** (2024).

[39] Anderson, B. R., Shah, J. H. & Kreminski, M. Homogenization effects of large language models on human creative ideation. *Proceedings of the 16th conference on creativity & cognition* 413–425 (2024).

[40] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).

[41] SBERT.net. sentence-transformers/all-minilm-l12-v2 · hugging face. https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2. (Accessed on 03/24/2024).

[42] Mishra, A. R., Panchal, V. & Kumar, P. Similarity search based on text embedding model for detection of near duplicates. *International Journal of Grid and Distributed Computing* **13**, 1871–1881 (2020).

[43] Rodier, S. & Carter, D. Online near-duplicate detection of news articles. *Proceedings of the Twelfth Language Resources and Evaluation Conference* 1242–1249 (2020).

[44] Tumre, S., Patil, S. & Kumar, A. Improved near-duplicate detection for aggregated and paywalled news-feeds. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)* 979–987 (2025).

[45] Zhao, K. *et al.* X-ray made simple: Lay radiology report generation and robust evaluation. *arXiv preprint arXiv:2406.17911* (2024).

[46] Li, D. *et al.* From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594* (2024).

[47] Schroeder, K. & Wood-Doughty, Z. Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509* (2024).

[48] Pan, Q. *et al.* Human-centered design recommendations for llm-as-a-judge. *Proceedings of the 1st Human-Centered Large Language Modeling Workshop* 16–29 (2024).

[49] Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D. & Groh, G. Shap-based explanation methods: a review for nlp interpretability. *Proceedings of the 29th international conference on computational linguistics* 4593–4603 (2022).

[50] King, G. & Zeng, L. Logistic regression in rare events data. *Political analysis* **9**, 137–163 (2001).

[51] Sun, X. & Xu, W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters* **21**, 1389–1393 (2014).

[52] Campbell, J. L., Quincy, C., Osserman, J. & Pedersen, O. K. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological methods & research* **42**, 294–320 (2013).

[53] Chinh, B., Zade, H., Ganji, A. & Aragon, C. Ways of qualitative coding: A case study of four strategies for resolving disagreements. *Extended abstracts of the 2019 CHI conference on human factors in computing systems* 1–6 (2019).

[54] Dong, Q. *et al.* A survey on in-context learning. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 1107–1128 (2024).

[55] Watkins, S. C., Roberts, D. A., Boulet, J. R., McEvoy, M. D. & Weinger, M. B. Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. *Simulation in Healthcare* **12**, 69–75 (2017).

[56] Viera, A. J., Garrett, J. M. *et al.* Understanding interobserver agreement: the kappa statistic. *Fam med* **37**, 360–363 (2005).

[57] Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* **70**, 213 (1968).

[58] OpenAI. OpenAI api (2023). URL https://platform.openai.com/docs/introduction. Online; accessed 07-Aug-2025.

[59] Scikit-learn. Agglomerativeclustering — scikit-learn 1.7.1 documentation (2023). URL https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html. [Online; accessed 2025-08-07].

[60] Scikit-learn. Randomforestclassifier — scikit-learn 1.7.1 documentation (2023). URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Online; accessed 2025-08-07].

[61] Scikit-learn. Gridsearchcv — scikit-learn 1.7.1 documentation (2023). URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Online; accessed 2025-08-07].

[62] Scikit-learn. Stratifiedkfold — scikit-learn 1.7.1 documentation (2023). URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html. [Online; accessed 2025-08-07].

[63] Scikit-learn. cohen_kappa_score — scikit-learn 1.7.1 documentation (2023). URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html. [Online; accessed 2025-08-07].