Using Generative AI to Enhance Community Support for Alternative Grading Practices

Mr. Abhinav Menon, Virginia Polytechnic Institute and State University

Abhinav Menon is a graduate student in Computer Science at Virginia Tech, where his research focuses on applying generative AI to systematically improve educational resources in equitable grading practices. His thesis work explores reproducible methodologies for AI-assisted documentation quality assessment and enhancement, with emphasis on community-maintained educational materials. This work was conducted under the supervision of Dr Stephen H Edwards and Dr Kenneth "Bob" Edmison in the Department of Computer Science.

Prof. Stephen H Edwards, Virginia Polytechnic Institute and State University

Stephen H. Edwards is a Professor and the Associate Department Head for Undergraduate Studies in the Department of Computer Science at Virginia Tech, where he has been teaching since 1996. He received his B.S. in electrical engineering from Caltech, and M.S. and Ph.D. in computer and information science from The Ohio State University. His research interests include computer science education, software testing, software engineering, and programming languages. He is the project lead for Web-CAT, the most widely used open-source automated grading system in the world. Web-CAT is known for allowing instructors to grade students based on how well they test their own code. In addition, his research group has produced a number of other open-source tools used in classrooms at many other institutions. Currently, he is researching innovative methods for giving feedback to students as they work on assignments to provide a more welcoming experience for students, recognizing the effort they put in and the accomplishments they make as they work on solutions. The goals of his research are to strengthen growth mindset beliefs while encouraging deliberate practice, self-checking, and skill improvement as students work.

Bob Edmison, Virginia Polytechnic Institute and State University

Dr. Edmison is a Collegiate Associate Professor in the Department of Computer Science at Virginia Tech. His research focuses on alternative grading practices, accessibility, and the tools to support them.

Dr. Adrienne Decker, University at Buffalo, The State University of New York

Adrienne Decker is a faculty member in the newly formed Department of Engineering Education at the University at Buffalo. She has been studying computing education and teaching for over 15 years, and is interested in broadening participation, evaluating t

Dr. Manuel A. Pérez-Quiñones, University of North Carolina at Charlotte

Dr. Manuel A. Pérez Quiñones is a Professor of Software and Information Systems at UNC at Charlotte. His research interests include diversity issues in computing, CS education, and human-computer interaction. He currently serves on the Committee on Women in Science, Engineering and Medicine at the National Academies and served as Program Officer at the National Science Foundation. His efforts to diversify computing have been recognized with an ACM Distinguished Member status (2019); the CRA A. Nico Habermann award (2018); and the Richard A. Tapia Achievement Award (2017). He is originally from San Juan, Puerto Rico.

Audrey Rorrer, University of North Carolina at Charlotte

Audrey Rorrer, PhD, is a Research Associate Professor in the Computer Science Department at UNC Charlotte, where she also serves as Assistant Director of the Center for Education Innovation & Research. Dr. Rorrer's scholarship areas include the science of broadening participation in computing, SoBP, which is a recognized domain of critical importance in STEM workforce development and educational programming. Her work has focused on educational programs, outreach and collective impact activities that expand the national pipeline into STEM careers. College student development and Faculty career development are central themes across her body of work.

Using Generative AI to Systematically Improve Community Educational Resources

Abstract

Computing and engineering educators are increasingly adopting alternative grading practices to foster more supportive and successful learning environments. These practices de-emphasize the traditional focus on points, empowering students to demonstrate mastery of learning outcomes with greater clarity and reduced stress. The *Playbook of Equitable Grading Practices* is an open-source community effort to support this transition, but many contributions suffer from inconsistent structural quality and a lack of concrete implementation detail. This paper introduces a transferable, three-phase framework for using Generative AI (GenAI) to systematically evaluate, enhance, and expand structured educational documentation. The method comprises: (1) the iterative calibration of a GenAI-based rubric against expert quality standards, first on a six-play pilot set and then across all thirty community-contributed plays; (2) the enhancement of documentation through source-grounded extraction of missing methodological details; and (3) the discovery of novel grading mechanisms via an exclusion-based discovery strategy designed to overcome search bias. Experimental results demonstrate that the calibrated evaluator attained 85.3% section-level agreement (within ± 1) with expert judgments. Significant quality improvements were observed in practitioner-oriented sections, such as "How to Implement" and "Applicability," where documentation was previously most deficient. This work contributes a reproducible calibration strategy for "LLM-as-a-Judge" applications and demonstrates that GenAI can serve as a collaborative editor that lowers the barrier to high-quality documentation meeting structural quality criteria, provided it is constrained by rigorous source grounding and human review to maintain authentic practitioner voice.

1 Introduction

Alternative grading practices including specifications grading [1], standards-based grading, and ungrading [2] are increasingly being adopted in higher education. Educators also are seeking assessment methods that are more accurate, bias-resistant, and motivational [3]. However, adoption remains inconsistent, particularly in engineering and computing where traditional point-based systems are deeply entrenched.

The consequences are particularly significant in these fields. High-attrition course sequences and competitive grading cultures disproportionately affect underrepresented students [3]. Alternative grading practices that reduce the emphasis on point accumulation and increase opportunities for demonstrating mastery can directly address these equity concerns [4]. However, without

accessible, high-quality documentation, adoption depends on individual faculty members discovering and adapting scattered research findings—a barrier that reinforces the very silos identified in the literature.

Community resources like *The Playbook of Equitable Grading Practices* [5, 6] bridge this gap by consolidating practices into a shared repository. But like many community-driven projects, the *Playbook* struggles with quality consistency. Plays vary in structure and clarity, creating hurdles for the educators they aim to serve. Traditional manual reviews scale poorly; comprehensive quality checks require significant grader time and create supervision bottlenecks that persist regardless of volume [7]. Recent advances in generative AI, particularly the emergence of Large Language Models (LLMs) as scalable evaluators [8, 9], offer a plausible solution. If calibrated to assess documentation against practitioner-relevant criteria, GenAI systems can provide systematic, reproducible quality evaluation at a scale that manual review cannot achieve.

This paper presents a transferable framework for using GenAI to systematically assess and improve structured document collections. We use the *Playbook* as a case study for our three-phase approach: (1) calibrating a rubric to ensure consistent human-GenAI agreement, (2) improving weak plays by extracting implementation details from source papers, and (3) generating new plays from academic literature. This work applies the LLM-as-a-Judge paradigm [8] to structured, human-authored documentation, assessing content against practitioner usability standards rather than general text quality.

While the *Playbook* is our specific context, we attempt to frame the calibration methodology itself as the primary contribution. The iterative process, prompt architecture, and observed evaluation patterns are intended to transfer to any structured document collection where quality is defined by template criteria.

Specifically, this paper contributes:

1. A documented account of iteratively calibrating a GenAI evaluation prompt against a grader's applied quality standards, along with the parallel use of GenAI to improve existing plays and generate new ones from academic literature.
2. A set of evaluation observations from our calibration sessions, reported as context-specific findings within our rubric and model rather than general claims about LLM behavior.

2 Background

The Playbook of Equitable Grading Practices was created by an ACM SIGCSE Virtual 2024 working group [6] to document alternative grading practices used in literature. The *Playbook* contains thirty equitable grading practices, pedagogical approaches that reduce bias and focus on learning over points. These "plays" are organized across six chapters covering grading assignments, grading courses, flexible resubmission, culture change, e-learning tools, and integration strategies.

Each play follows a structured template including five primary content sections evaluated in this work: Intent, Problem, Solution, Applicability, and How to Implement. Additional sections (See Also, Source, References, Community Discussion) are administrative and not scored. A core

principle is that plays provide general practices rather than rigid lesson plans. This gives instructors enough detail to understand the approach without prescribing specific content. An instructor should understand what to create and what it should accomplish while retaining the flexibility to adapt it to their specific course.

2.1 Structural vs. Functional Quality

We focus on structural quality: whether plays follow the template and provide the information necessary for implementation. Functional quality assessment, which evaluates whether a practice actually improves student outcomes, requires deep domain expertise and involves subjective pedagogical judgments.

Structural quality assessment provides a more objective foundation. We can assess whether an Intent statement clearly describes a play's goal or if implementation guidance is sufficiently detailed. Structural quality is a prerequisite for adoption; regardless of how pedagogically sound a practice is, unclear documentation prevents it from being used. Our goal is to create high-quality documentation of community practices rather than evaluating the practices themselves.

2.2 Calibration-to-Fixed-Standard as a Methodological Choice

Calibration of a GenAI evaluator can target different reference points, each appropriate to different research goals. Common options include: (1) consensus among multiple raters, which is the target of traditional inter-rater reliability (IRR) metrics such as Cohen's kappa; (2) an external, published gold-standard rubric, which is the target when validating against an authoritative benchmark; and (3) a specific, stable quality standard applied by a single decision-maker, which is the target when the downstream use case involves that decision-maker applying their own judgment at scale [10]. Each target is methodologically distinct, and the choice depends on what the calibrated tool is intended to do.

This work adopts the third approach: calibration-to-fixed-standard. The research question is not "does the GenAI match the average human?" but "can the GenAI be calibrated to apply a consistent, pre-defined quality standard across a large corpus?" The standard here is embodied in one grader's reference scores across the full playbook, but the methodological contribution generalizes to any setting where a single, stable evaluation criterion is applied consistently. This could describe a style guide applied across documentation, a compliance checklist applied across policy artifacts, or any context where the target is fidelity to a defined standard rather than consensus across independent raters.

2.3 GenAI Selection

We used Gemini (Google) as the primary evaluation model, accessed programmatically via the Gemini CLI in headless mode to facilitate reproducible batch evaluation with consistent prompt-template handling. For improved-play evaluation, the same model and prompt version were used to ensure that before-and-after scores reflect only changes in the play text rather than differences in the evaluation instrument.

We chose deep calibration of a single model over shallow validation across multiple models. A

tightly calibrated prompt on one model serves our research question better than weak cross-model agreement on a pilot subset. During an early pilot phase, however, we used cross-model comparisons on a small benchmark set as an internal consistency check (described in Section 4.1.1). We did not repeat cross-model validation on the full 30-play corpus.

A limitation of closed-source models is opacity; providers may update underlying models without notice. Our results reflect the model versions served during our evaluation window. Some models, such as DeepSeek, were excluded due to institutional policy restrictions at Virginia Tech.

3 Literature Review

Alternative grading practices aim to address equity issues and measurement accuracy problems in traditional point-based grading [1, 2, 3, 4]. Clark and Talbert identify four pillars of these systems [4]: defined standards, helpful feedback, marks indicating progress, and penalty-free reassessment. Despite significant interest, adoption remains difficult, and Hackerson et al. observed that research on alternative grading in STEM is conducted within disciplinary silos with little consensus on theoretical foundations [11]. This fragmentation forces interested educators to piece together resources from different fields, and the community resources intended to address this problem, such as the *Playbook* [6], struggle with their own quality-consistency issues [12]. Maintaining consistency across distributed contributions is a challenge manual reviews cannot solve alone.

GenAI systems, particularly LLMs, have emerged as a plausible solution for scalable evaluation. Strong LLM evaluators can match human preferences with over 80% agreement on rated outputs, though they exhibit biases including position bias, verbosity bias, and self-enhancement bias [8]. Accuracy improves when models use chain-of-thought reasoning with explicit criteria [13] and when rubrics are calibrated against human judges [14]. Jiang et al. provide a survey of these methodologies and known challenges [9], and Ye et al. document systematic scoring patterns such as leniency on open-ended dimensions [15]. Most directly relevant to our work, Zhao et al. [10] describe an AI-supported grading pipeline in which graders iteratively refine rubric items based on disagreements between AI and grader judgments. Their central finding, that even experienced educators struggle to articulate rubric criteria that are simultaneously specific and interpretable, parallels our own experience during prompt calibration. Their work focuses on binary rubric items for student free-response questions; we extend the iterative-refinement approach to multi-point rubrics for structured documentation quality.

Generative AI is increasingly used to create educational content, from quiz questions to full lesson plans [16]. A primary concern is quality assurance; AI-generated content can include hallucinations, and human-AI collaboration is essential for verification [16]. Messer et al. highlight both the potential of AI assessment and the persistent gap between automated and human evaluation quality [17]. Existing work in this space typically focuses on evaluating AI outputs; this paper applies these evaluative capabilities to human-authored content, verifying the actionable detail and structural consistency identified as critical for successful community resource adoption [12, 18].

Our primary contribution is a transferable assessment framework for calibrating GenAI systems to evaluate structured document collections. Specifically, we adapt the LLM-as-a-Judge paradigm

Play	ChatGPT	Gemini	Grok	Variance
Play 6	19	20	19	1
Play 13	21	21	22	1
Play 16	26	27	26	1
Play 19	20	21	21	1
Play 21	20	20	21	1
Play 24	26	26	27	1

Table 1: Pilot-run cross-model scoring on six benchmark plays, used as an internal consistency check during calibration development. Reported for context.

to evaluate human-authored documentation against usability rather than general text quality, use GenAI to improve existing content through source-grounding, and document context-specific observations of GenAI behavior within educational document evaluation.

4 Methodology

Our three-phase methodology addresses the challenge of systematically improving community resources. Phase 1 develops a calibrated rubric for human-GenAI agreement. Phase 2 identifies weak plays and improves them using source-grounding. Phase 3 generates new plays from literature while avoiding search biases.

4.1 Phase 1: Evaluation Rubric Calibration

The calibration phase proceeded in two stages: an initial pilot using six benchmark plays with cross-model comparisons, and a full-corpus calibration across all thirty plays. Each stage informed the other: the pilot surfaced core calibration insights, and the full-corpus evaluation exposed systematic biases that a small sample could not reveal.

4.1.1 Pilot Run: Six Benchmark Plays with Cross-Model Comparison

The pilot used six plays chosen to represent a full quality spectrum: Play 6 (Mathematically Accurate Grading), Play 13 (Ungrading), Play 16 (Outcome Evaluation Clobbering), Play 19 (Emphasize Feedback and Reflection), Play 21 (Autograding for CS), and Play 24 (Competency/Equity Hybrid). Play 13 is conceptually dense but vague on implementation, Play 24 is highly detailed, and Play 16 is concise and well-structured. Using these plays across early iterations allowed for consistent comparison while the rubric was under active development.

During the pilot, we used cross-model comparisons across ChatGPT (OpenAI), Gemini (Google), and Grok (xAI) as an internal consistency check on the rubric. Table 1 shows that final pilot-version scores were consistent across the three models, with consistent agreement (range of approximately 1 point) across plays and models. This served as a signal that the rubric was not fitting model-specific idiosyncrasies at the pilot scale. We did not repeat cross-model validation on the full thirty-play corpus, and we do not claim cross-model generalization for the final calibrated rubric; the check was an internal benchmark used during pilot development only.

All subsequent evaluation, reporting, and statistics in this paper use the full 30-play corpus with Gemini as the sole evaluation model.

4.1.2 Full-Corpus Calibration: All Thirty Plays

We expanded calibration to all thirty plays in order to surface systematic biases that a six-play sample could not reveal. The rubric evaluates five content sections on a 0–5 scale with half-point resolution, resulting in totals from 0 to 25. The five sections are Intent (the accomplishment the play achieves), Problem (what the play addresses), Solution (the approach or practice), Applicability (when the play fits), and How to Implement (actionable guidance).

Reference Scores. A grader (the project maintainer) generated reference scores for all thirty plays across all five sections, along with written justifications for each score. This dataset served as ground truth for every iteration of the calibration. As described in Section 2.2, this aligns with the calibration-to-fixed-standard methodology: the goal is consistent application of a pre-defined quality standard, not consensus across multiple raters.

Pipeline Architecture. To enable systematic prompt iteration without disrupting output schema, we adopted a two-file architecture: a stable wrapper (`GEMINI.md`) specifying JSON output format, behavioral constraints, and perspective framing; and a swappable rubric file (`prompt.md`) containing the scoring criteria. The Gemini CLI was invoked in headless mode with batches of three plays per invocation, with randomized play ordering across runs to mitigate within-batch anchoring.

```
EVALUATION PROMPT ARCHITECTURE

1. PERSPECTIVE FRAMING
   "You are an instructor: can I use this?"

2. TASK-LEVEL PRINCIPLES
   Practices vs. lesson plans
   Structural vs. functional quality

3. SECTION CRITERIA (x5 sections)
   - Core Question
   - Score Levels (0-5)
   - Positive / Negative Examples
   - Test Questions

4. OUTPUT FORMAT
   Fixed JSON schema
   Section scores + reasoning
```

Iterative Refinement. Each prompt version was evaluated against all thirty plays, with five runs per version per play. Agreement rates were computed per section and overall. Systematic biases were diagnosed by examining plays where the GenAI and reference scores disagreed by more than one point, comparing reasoning traces against reference justifications. Targeted rubric edits addressed specific failure modes rather than adjusting the entire prompt. The calibration trajectory

Version	Key Change	Int	Prb	Sol	App	HtI
Phase A: Baseline and overcorrection recovery						
v1	Baseline, 5 sections /25	23	26	21	21	20
v2	Added negative-example traps	21	26	22	24	20
v3	Softened v2 overcorrections	21	28	26	23	23
Phase B: Section-specific cognitive reframing						
v4	Intent: cognitive extraction cost	28	26	23	25	26
v5	Applicability: no tradeoff cap	27	28	25	26	22
v6	Best-of-breed assembly	26	28	26	24	25
Phase C: Selective integration and refinement						
v7	Environment-as-approach clause	25	26	19	24	22
v8	HtI prevalence test, locked version	27	28	21	26	26

Table 2: Prompt iteration history showing per-section plays scored within ± 1 point of the reference (out of 30). Iterations are grouped into three phases: baseline stabilization, cognitive reframing, and selective refinement.

across eight versions is summarized in Table 2, grouped into three conceptual phases.

Acceptable Variance Threshold (± 1 point). We established ± 1 point as the acceptable per-section variance threshold a priori. On a 5-point scale, this represents a 20% tolerance, corresponding to roughly 80% scale-resolution agreement; the LLM-as-a-Judge literature has reported strong agreement around this same level [8]. A one-level adjacency disagreement rarely changes the feedback we would provide to a contributor; both scores flag the section as needing revision in similar ways.

Key Calibration Insights.

Practitioner framing. Initial prompts without explicit perspective framing produced scoring patterns consistent with academic peer review, rewarding philosophical depth and comprehensive coverage [8, 15]. The framing "You are an instructor asking, 'Can I understand and use this?"' substantially shifted scoring toward practitioner usability and proved essential across all subsequent versions.

Section-level cognitive criteria. Specific rubric terminology had outsized effects. The Intent section improved dramatically when reframed around how much work the reader does to identify what the play accomplishes, rather than around grammatical structure such as "no reasoning, no mechanism detail." The prior framing caused the GenAI to penalize short, clear intent statements containing purpose clauses; the new framing focused on what actually mattered to a practitioner reading for comprehension.

Prevalence over presence. In the How to Implement section, we found that a section could contain one or two named activities and still be dominated by generic filler like "explain the concept to students" or "create opportunities for reflection." Earlier versions scored these sections the same as sections where named strategies were the main content. The the final refinement added a simple test: "if you removed all the generic directives, would any meaningful strategies remain?" This corrected How to Implement overscoring without disturbing other sections.

Majority Voting. To separate stochastic variance from systematic disagreement, each play was evaluated across five runs per prompt version. Run-to-run standard deviations averaged 0.72 points, small relative to section-level disagreements. We report results using the mode across five runs per play, selecting the score the model committed to most frequently. Ties were broken by selecting the modal score closest to the run-wise mean. On a discrete 0–5 scale with half-point resolution, mode selection preserves the granularity of the scoring scheme.

4.2 Calibration-Surfaced Inconsistencies

We also observed that systematic GenAI disagreement occasionally surfaced inconsistencies in the reference scores themselves. During analysis of mid-iteration disagreements, several plays showed persistent disagreements where the GenAI's interpretation was defensibly consistent with the rubric text while the reference scores appeared to apply different implicit standards. Three of the 150 section-level reference scores were revised for internal consistency during this process; in other cases reviewed, no change to the original score was deemed necessary. The grader retained authority over contested cases.

We note this as an observation rather than a formal methodological claim: the calibration process may provide a secondary benefit by surfacing inconsistencies in reference scores that would be difficult to detect through manual review alone.

4.3 Phase 2: Play Improvement Process

Phase 2 tested whether GenAI could improve weak plays without hallucinating or changing pedagogical mechanisms.

Chained Evaluation-Enhancement Methodology. We used a chained approach: the GenAI first evaluated a play, then received the source paper and improvement instructions, all in the same context.

Five Critical Constraints. The prompt established these rules:

1. **Maintain intent:** Do not alter the core practice.
2. **Preserve voice:** Maintain the author's voice and the community-contributed feel.
3. **Source grounding:** Every update must trace back to the paper.
4. **Structural focus:** Improve clarity and completeness, do not evaluate the validity of the practice.
5. **No lesson plans:** Keep practices adaptable, not course-specific.

```
IMPROVEMENT PROMPT ARCHITECTURE

1. Evaluation results (from context)
2. Five explicit constraints
3. Source paper content
4. Output format specification
```

Scope of Improvement. The improvement process was applied to all thirty plays in the playbook. Unlike the earlier pilot work that improved only thirteen plays, we improved the full corpus to enable a paired comparison with adequate statistical power.

Evaluation of Improved Plays. Improved plays were scored using the same calibrated rubric and the same five-run mode-selection protocol as the original plays. This symmetrical design isolates the change in play text as the only variable. Phase 1 reliability addresses evaluator stability and consistency; it does not by itself address the rubric-improvement loop discussed in Section 6, which we treat as a separate threat to interpretation.

Verification Process. Each improved play was verified section-by-section against the source paper. We found that asking the same model to improve and verify led to self-preference bias [8]. Using a separate prompt and session for verification was more reliable at catching unsourced claims. The same verification process was used for the newly generated plays in Phase 3.

4.4 Phase 3: Generating New Plays from Academic Literature

Phase 3 demonstrated the potential and limits of GenAI-assisted literature search and document generation.

Phase 3a: Base Generation. We used Perplexity AI to search for highly-cited, peer-reviewed STEM papers, then generate plays. The prompt prioritized papers with high citation counts, clear implementation guidance, evidence of effectiveness, and adaptability across contexts. Critical constraints for the output mirror those used in improving prior plays: requiring source grounding, describing practices rather than lesson plans, and instructor usability framing. The prompt incorporates explicit negative examples of vague implementation guidance and emphasizes concrete examples as essential for usability.

```
PHASE 3A GENERATION PROMPT ARCHITECTURE

1. TASK FRAMING
   Search for relevant paper + create play

2. SEARCH CRITERIA
   - High citation count
   - Implementation details present
   - STEM higher education context
   - Published 2010-2025

3. TEMPLATE WITH GUIDANCE
   For each section:
   - What to include
   - Positive / Negative examples

4. CRITICAL PRINCIPLES
   - Source grounding
   - Practices not lesson plans
   - Instructor usability
```

Retrieval Issues. We encountered two problems. First, repeated searches consistently returned the same popular grading philosophies such as specifications grading, mastery grading, and token systems. This occurs because LLMs surface frequently-appearing "head" knowledge over "long-tail" information [19]. Second, plays remained too discipline-specific. For example, a chemistry-based play did not extract the general principles applicable to other STEM fields.

Phase 3b: Searching for Novelty. We developed a Phase 3b prompt to discover underexplored practices. This used an exclusion-based approach, explicitly banning common topics and instructing the GenAI to prioritize novel combinations, unusual grading logic, or cross-field practices. This prompt successfully uncovered unique practices not represented elsewhere in the *Playbook*, such as visual pictogram grade maps, grade cockpit systems, and program-level pass/no-record windows. This approach confirms that strategic prompt design can overcome search convergence through explicit exclusions, and that transferability must be explicitly required.

```
PHASE 3B NOVELTY PROMPT ARCHITECTURE

1. MISSION STATEMENT
   Discover unusual/experimental mechanisms

2. EXPLICIT EXCLUSIONS
   - Common mechanisms list
   - Theoretical papers
   - Failed experiments

3. NOVELTY CRITERIA
   - Novel combinations
   - Unusual grading logic
   - Novel scope/scale

4. EVIDENCE AND TRANSFERABILITY
   - Demonstrated success required
   - Extract general principles
```

5 Results

5.1 Initial Playbook Quality Distribution

Reference evaluation of all thirty original plays demonstrated substantial quality variance on the 0–25 scale. Reference total scores ranged from 12 to 23.5 (mean 19.2, median 19.25). Of the thirty plays, 9 scored in the "fair" range (12–17.5), 13 in "good" (18–21.5), and 8 in "excellent" (22–25).

Section-level analysis showed that How to Implement and Applicability were the weakest sections (reference means of 3.27 and 3.22 of 5), while Problem and Solution were strongest (4.45 and 4.12). This pattern confirms that structurally defined sections mapping directly to source content are of higher quality, while sections requiring greater synthesis of ideas exhibit the most significant deficiencies. Community resource templates that require contributors to

Section	± 1 Agreement	ICC(3,1)	Mean Diff.	MAE
Intent	27/30	0.577	+0.25	0.55
Problem	28/30	0.103	+0.48	0.55
Solution	21/30	0.272	+0.44	0.86
Applicability	26/30	0.723	+0.29	0.82
How to Implement	26/30	0.809	+0.40	0.63
Overall (section-level)	128/150	0.686	+0.37	0.68
Overall (excl. outliers)	122/140	0.706	+0.30	0.63

Table 3: Phase 1 calibration reliability results for the final calibrated prompt (five-run mode selection, all thirty plays). ICC(3,1) computed as two-way mixed, consistency, single measures. Mean Diff. is the signed difference (Gemini minus reference). The low Problem ICC reflects restricted variance in reference scores (range 3–5), which limits ICC's sensitivity even when raw ± 1 agreement is high; for sections with restricted reference variance, ± 1 agreement is the more interpretable metric.

transform research findings into practitioner-oriented guidance, rather than summarizing source material, may need additional scaffolding or quality review mechanisms [12].

5.2 Phase 1: Calibration Reliability

Table 3 presents the final calibration performance of the final calibrated prompt. Across all thirty plays and five sections (150 section-scores total), the GenAI agreed with reference scores within ± 1 point on 128/150 section-scores, representing 85.3% agreement. Two plays (Plays 2 and 23) showed persistent multi-section overscoring across all prompt versions and are discussed separately below as structural outliers; excluding these, agreement rises to 122/140 (87.1%).

Formal Agreement Metric. The intraclass correlation coefficient ICC(3,1) at the section level was 0.686, with a bootstrapped 95% confidence interval of [0.594, 0.767] (resampled at the section-score level over 1,000 iterations). Following Koo and Li's interpretive thresholds [20], this categorizes overall agreement as *moderate*, with the upper bound approaching the *good* threshold. Excluding the two structural outlier plays, ICC rose to 0.706 with 95% CI [0.610, 0.778]. Across all five sections, the calibrated evaluator exhibited a marginal but consistent leniency tendency (mean differences of $+0.25$ to $+0.48$), well within the ± 1 tolerance but worth noting as a directional pattern in the calibrated tool's behavior.

5.3 Phase 2: Improvement Effectiveness

Phase 2 improved all thirty plays. Table 4 presents before-and-after means for the full corpus and per-section. Mean total score increased from 21.10 (SD 1.87) to 24.47 (SD 0.95), a mean gain of +3.37 points on a 25-point scale. A paired t-test on the thirty play-level paired observations gave $t(29) = 9.55$, $p < 0.001$, with Cohen's $d = 1.74$ computed from paired differences. The 95% confidence interval on the mean improvement was [2.65, 4.09] points.

Differential Section Improvement. The per-section pattern is the more substantive finding than the aggregate. How to Implement showed the largest improvement ($d = 1.03$, $p < 0.001$),

Section	Before M	After M	Δ	t	p	d
Intent	4.43	4.90	+0.47	2.97	5.91e-03	0.54
Problem	4.93	5.00	+0.07	1.44	1.61e-01	0.26
Solution	4.56	4.97	+0.41	2.60	1.44e-02	0.48
Applicability	3.51	4.72	+1.21	5.76	3.07e-06	1.05
How to Implement	3.67	4.88	+1.22	5.62	4.60e-06	1.03
Total	21.10	24.47	+3.37	$t(29) = 9.55$	1.88e-10	1.74

Table 4: Phase 2 improvement results: per-section and total paired comparison across all thirty plays. Cohen's d computed from paired differences.

followed by Applicability ($d = 1.05$, $p < 0.001$). Intent and Solution showed moderate improvements (Intent $d = 0.54$ and Solution $d = 0.48$), while Problem showed the smallest change ($d = 0.26$, ns).

The sections that showed the largest improvements were precisely those that scored lowest before improvement, which are the practitioner-facing sections where community authors most commonly struggle to translate research into actionable guidance [12, 16]. Sections that were already strong (Problem, Solution) showed minimal change, indicating the improvement process was not uniformly inflating scores; the observed effect is concentrated in the sections that needed improvement.

5.4 Structural Outliers

Across early prompt iterations, four plays (Plays 2, 8, 12, 23) exhibited persistent GenAI overscoring by three or more total points. Score revisions during calibration (Section 4.2) and continued rubric refinement closed the gap substantially on Plays 8 and 12, where most sections fell within ± 1 of the reference in the final calibration. Two plays remained persistent outliers: Plays 2 and 23, which continued to show multi-section gaps of two or more points despite all calibration efforts.

The remaining outliers share structural characteristics distinct from the rest of the playbook: Solution sections describing how the play addresses problems rather than the mechanics of the play itself, and Applicability sections where constraints can be inferred but are not explicitly stated. The GenAI's reasoning on these plays was internally consistent with the rubric text, as was the reference reasoning, but the two applied different implicit standards.

No amount of rubric refinement closed the gap on these plays without overcorrecting on the other twenty-eight. We report this as a finding rather than a failure: practitioners applying calibration-to-fixed-standard methodology should expect a small subset of documents where rubric text cannot fully encode the grader's implicit judgment.

5.5 GenAI Evaluation Observations

We report the following as context-specific observations within our calibration setup rather than general claims about GenAI behavior.

Scoring conservatism on subjective sections. The GenAI consistently demonstrated more caution

about assigning perfect scores in subjective categories, consistent with documented LLM conservatism on open-ended evaluation dimensions [8, 15]. In the final calibration, 27% of plays received 5/5 on Applicability compared with 73% on Solution, despite Applicability being the section where improvement had the largest effect.

Section-level bias direction. The iteration process surfaced that different sections exhibited opposite biases requiring different corrections. How to Implement was systematically lenient in early versions (the GenAI credited named strategies without requiring execution detail), while Intent was systematically strict (the GenAI penalized clear intent statements containing purpose clauses). These opposing biases required separate rubric edits and could not be resolved via general reframing alone.

Convergent stochastic behavior. Run-to-run standard deviations averaged 0.72 points per section, small relative to the scoring range. Failures were reproducible across runs rather than random, indicating that remaining disagreements reflect genuine interpretive differences rather than evaluator instability. In this setting, majority voting primarily surfaces the model's systematic interpretation rather than denoising random variance.

5.6 New Play Creation

New plays generated in Phase 3 averaged 23.5 points on the 25-point scale on a preliminary evaluation, outperforming the original playbook mean, with the largest advantages in Applicability and How to Implement. We report this as preliminary; a full paired analysis of new-play quality against the original corpus is left for future work as the new-play set is expanded. Each new play was verified section-by-section against its source paper using the process described in Section 4.3.

6 Limitations

Rubric-Driven Improvement Loop. The most important limitation of this work is the circular relationship between the rubric and the improvement process. The GenAI improvements were generated against the same rubric used to evaluate them, so high post-improvement scores are partly a consequence of that alignment rather than independent evidence of quality gain. The rubric sets the yardstick and also shapes what the improvement output looks like. We partially address this by reporting differential per-section gains, which show that improvement is concentrated in the sections that scored lowest before improvement rather than distributed uniformly across all sections. However, readers should interpret Phase 2 scores as evidence that the improvement process *addresses* the rubric criteria rather than independent evidence that the plays are objectively better in ways the rubric does not capture.

Source Dependency. GenAI cannot (and should not) create details that do not exist in the source paper. Conceptual papers result in structurally weaker plays unless supplemented by other sources.

Context Window Constraints. LLMs exhibit degraded performance as context windows grow. After evaluating at most three plays, we recommend starting a new conversation window. Some systems do not explicitly surface context limits to users; instead, the model begins forgetting

earlier content while continuing to generate outputs, creating subtle quality degradation that appears normal [21]. Starting fresh conversations maintains evaluation consistency and prevents this contextual drift.

Template Uniformity. The approach described here is built on a uniform structure for all the artifacts. Real-world collections often lack this uniformity and would require an initial classification step to group documents before applying rubrics. The core framework contributions (iterative calibration, practitioner-oriented framing, negative boundary examples) transfer regardless of template uniformity, but some adaptation would be needed.

Authenticity. GenAI assistance can improve documentation quality, but over-reliance might erode the value that comes from the diverse range of perspectives authors bring to their own plays, gradually eroding trust in community resources and discouraging contributions. These plays must be subjected to community review to maintain pedagogical integrity.

Single-Reference Standard. As discussed in Section 2.2, calibration-to-fixed-standard is deliberate, but it limits external generalization: the calibrated rubric applies one particular grader's standards, not universal standards. A different maintainer applying the same methodology would produce a different calibrated rubric.

Single Model. The final reported results use only Gemini. While the pilot-run cross-model check provided some confidence in the rubric's design, we do not claim that the calibrated rubric produces equivalent agreement on other models across all thirty plays. Cross-model replication is a reasonable direction for future work.

7 Conclusion

This paper presented a framework for using calibrated GenAI to assess and improve structured documentation. Our calibration methodology is intended to transfer to other domains where quality is defined by a template, such as API documentation or tutorial libraries.

The key takeaways are that iterative prompt calibration against full-corpus reference judgments reaches 85.3% section-level agreement within ± 1 point on the *Playbook*; that GenAI improvements concentrate in the practitioner-facing sections where community authors most commonly struggle; and that a small subset of plays resist calibration through structural characteristics that exceed rubric expressiveness, and should be reported as findings rather than eliminated through increasingly specific criteria.

We also note a tension between resource quality and community legitimacy: GenAI can improve the structural clarity of plays, but communities must navigate the risk of sanitizing authentic practitioner voice. We posit that GenAI should serve as a collaborative editor that lowers the bar for meeting structural quality criteria, rather than an autonomous author that replaces human insight. Next steps include community validation, cross-grader replication of the calibration methodology, and exploring the balance between efficiency and authentic voice.

References

[1] L. B. Nilson, *Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time*. Stylus Publishing, 2015.

[2] S. D. Blum, *Ungrading: Why Rating Students Undermines Learning (and What to Do Instead)*. West Virginia University Press, 2020.

[3] J. Feldman, *Grading for Equity: What It Is, Why It Matters, and How It Can Transform Schools and Classrooms*. Corwin Press, 2018.

[4] D. Clark and R. Talbert, *Grading for Growth: A Guide to Alternative Grading Practices that Promote Authentic Learning and Student Engagement in Higher Education*. Routledge, 2023.

[5] "The playbook of equitable grading practices," https://cs-equitable-grading-practices.github.io/playbook/, 2024, accessed: 2025.

[6] S. H. Edwards, D. L. Largent, B. Schafer, D. Basu, A. C. Dalal, D. Garcia, B. Harrington, K. Lin, D. S. Myers, C. Roberson, G. Toti, and B. Wortzman, "Developing a playbook of equitable grading practices," in *Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 2 (SIGCSE Virtual 2024)*. New York, NY, USA: ACM, 2024, pp. 41–53.

[7] L. J. J. van Rensburg, "Ai-powered citation auditing: A zero-assumption protocol for systematic reference verification in academic research," 2025. [Online]. Available: https://arxiv.org/abs/2511.04683

[8] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36. NeurIPS, 2023.

[9] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo, "A survey on LLM-as-a-Judge," *arXiv preprint arXiv:2411.15594*, 2024.

[10] C. Zhao, M. Fowler, Y. Gertner, S. Poulsen, M. West, and M. Silva, "Ai-supported grading and rubric refinement for free response questions," in *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.1 (SIGCSE TS 2026)*, 2026.

[11] E. L. Hackerson, T. Slominski, N. Johnson, J. B. Buncher, M. Burnett, D. Grimm, A. K. Hoover, R. Simha, Q. Wang, H. Manzoor, R. Downey, and A. Shakibnia, "Alternative grading practices in undergraduate STEM education: a scoping review," *Disciplinary and Interdisciplinary Science Education Research*, vol. 6, no. 15, pp. 1–28, 2024.

[12] A. Decker, S. H. Edwards, B. M. McSkimming, B. Edmison, A. Rorrer, and M. A. Pérez Quiñones, "Transforming grading practices in the computing education community," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2024. New York, NY, USA: Association for Computing Machinery, 2024, p. 276–282. [Online]. Available: https://doi.org/10.1145/3626252.3630953

[13] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 2511–2522.

[14] H. Hashemi, J. Eisner, C. Rosset, B. Van Durme, and C. Kedzie, "LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 13 806–13 834.

[15] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, and X. Zhang, "Justice or prejudice? quantifying biases in LLM-as-a-Judge," in *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[16] M. Giannakos, R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriadis, D. Hernandez-Leo, S. Järvelä, M. Mavrikis, and B. Rienties, "The promise and challenges of generative ai in education," *Behaviour & Information Technology*, vol. 44, no. 11, pp. 2518–2544, 2025.

[17] M. Messer, N. C. C. Brown, M. Kölling, and M. Shi, "Automated grading and feedback tools for programming education: A systematic review," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–43, 2024.

[18] E. Wenger, *Communities of Practice: Learning, Meaning, and Identity*, ser. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press, 1998.

[19] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 15 696–15 707.

[20] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[21] Comet ML, "Context window: What it is and why it matters for LLMs," https://www.comet.com/site/blog/context-window/, 2024, accessed: 2025.