Guiding Classroom Consensus: Understanding How Students Respond to Peer-Based AI Hints In A Process For Facilitating Class-Wide Discussions

ANONYMOUS AUTHOR(S)

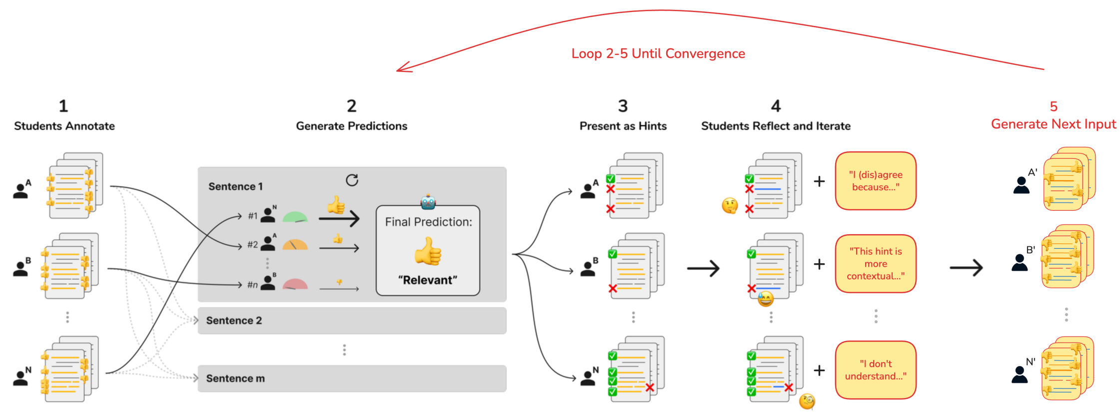

Fig. 1. Our peer-based hints system: **(1)** students annotate transcripts for given research questions; **(2)** annotations are used to predict whether each sentence is "Relevant" or "Not relevant" via a Dawid–Skene expectation–maximization (EM) implementation; **(3)** predictions are used to generate personalized hints, showing students when their annotations differ from predictions and pointing out missed annotations; **(4)** students reflect on whether they agree with the hints via short write-ups and revise their annotations; **(5)** the revisions from step 4 inform new predictions by the EM model. After step 5, the EM loop (steps 2–5) repeats until convergence.

In large Human–Computer Interaction (HCI) courses, providing personalized feedback and deep discussions on qualitative analysis (QA) is often infeasible due to large class sizes. Prior work proposed to address this challenge in the initial coding phase through peer-based AI hints generated from peer work that helps learners reflect on whether a sentence is or is not relevant to the research question. They analyze the quality of hints computed based on differences to a prediction generated by the Dawid–Skene Expectation–Maximization (DS–EM) algorithm. In this study, we implement the approach and deploy it in a large university course to analyze how students respond to these peer-based AI hints and their written explanations for why they agreed or disagreed with the peer-based AI hints. We also reconceptualize the peer-based AI hints as facilitating an asynchronous class-wide discussion by reframing the generated predictions as the current class-wide consensus. As students receive hints and iterate on their work, we regenerate the prediction to take into account students' new opinions, making it easier to come closer to consensus and to identify remaining disagreements before an instructor-led discussion. We use simulations to explore the impact of repeated rounds of DS–EM–generated feedback, showing that it can further improve student work and facilitate class-wide convergence towards a higher-quality prediction.

CCS Concepts: • **Human computer interaction**; • **Applied computing** → *Interactive learning environments*; • **Computing methodologies**;

Additional Key Words and Phrases: learnersourcing, peer feedback, Dawid–Skene, consensus modeling, intelligent tutoring, education

Manuscript submitted to ACM

1 Introduction

Qualitative analysis (QA) is a cornerstone of Human–Computer Interaction (HCI), enabling researchers to interpret rich, contextual data such as interview transcripts and uncover user needs that resist quantification. Yet QA is difficult to teach at scale. The apprenticeship-like guidance that helps students refine their interpretations is rarely feasible in classes with 100 or more learners, where mentor-to-student ratios are low and feedback opportunities are limited [3, 27, 33].

Prior work introduced *Annota*, a learnersourcing platform designed to address this scalability challenge. In *Annota*, students are given real interview transcripts and asked to determine whether each sentence is relevant to a given research question—a foundational step in qualitative coding. The system aggregates these binary relevance judgments across students using the Dawid–Skene Expectation–Maximization (DS–EM) algorithm [5, 24], generating peer-based AI hints that show how an individual's coding aligns or diverges from the emerging class consensus. Earlier findings suggested that these consensus-driven hints can prompt reflection and expose students to alternate interpretations, pointing toward their potential for facilitating class-wide discussion. However, prior deployments presented hints as one-off suggestions, rather than as part of an iterative feedback loop that evolves with the class's collective understanding.

This paper extends that premise by positioning *Annota* not as a tool for delivering feedback on qualitative themes, but as a mechanism for facilitating *asynchronous class-wide discussion* in a qualitative analysis context. Here, the DS–EM consensus functions as a dynamic representation of class understanding: as students agree or disagree with received hints, their responses feed back into the model, regenerating the consensus and making visible where interpretations align or diverge. Alongside these binary judgments, students also provide written explanations for their decisions, articulating why they agreed or disagreed with the consensus. While these written responses do not directly influence the model updates, they offer valuable insight into how learners reason about the data and negotiate meaning. This iterative process allows consensus itself to become a shared object of reflection and discussion, helping the class move toward a more unified perspective before instructor-led dialogue.

To examine this process, we combine two complementary analyses. First, we analyze students' written explanations for why they agreed or disagreed with AI-generated hints to understand how learners interpret, justify, and negotiate consensus in practice. Second, we simulate iterative DS–EM runs parameterized by observed agreement rates to model how repeated cycles of engagement might affect accuracy and convergence across the class. Together, these analyses connect individual reflection with collective interpretation, showing how iterative consensus-building can serve as a proxy for mentorship and a structure for discussion at scale.

Concretely, our contributions are threefold:

(1) An empirical analysis of how students engage with and reason about peer-based AI hints in an authentic classroom deployment.
(2) A simulation study modeling iterative DS–EM feedback loops and their potential to promote convergence and accuracy across a cohort.
(3) A conceptual reframing of learnersourcing platforms—from systems that deliver feedback to mechanisms that structure large-scale, consensus-oriented classroom discussion.

2 Related Work

2.1 Foundations of Experiential Learning and Mentorship

Experiential learning maintains that the most durable learning occurs when students act in authentic contexts and reflect on those actions. John Dewey argued that experience and reflection, rather than passive transmission, sit at the heart of education [6]. David Kolb formalized this stance as a cycle linking concrete experience, reflective observation, abstract conceptualization, and active experimentation [16]. In parallel, reflective practice and constructionist approaches similarly emphasize learning by doing and iterative reflection in context [25, 28]. These ideas align with situated views of knowing: what learners can perceive, do, and articulate is shaped by the setting and activity in which they participate [1].

Mentorship amplifies these cycles. Vygotsky's account of a more knowledgeable other enabling progress in the zone of proximal development explains why guided practice tends to outperform unguided effort: mentors help learners notice gaps, structure reflection, and attempt moves that would otherwise be out of reach [31]. Cognitive apprenticeship operationalizes this guidance in practice. Through modeling, scaffolding, coaching, articulation, and fading, learners can internalize expert strategies while working on meaningful tasks [4]. Beyond dyadic mentorship, communities of practice offer a collective mechanism for learning through participation, identity building, and shared repertoires of tools and stories [18, 32].

HCI education repeatedly echoes these foundations. Calls for flexible curricula, authentic projects, and critique centered studio practices aim to preserve experiential cycles at scale while sustaining guidance [3, 27, 33]. Learnersourcing extends the same logic into socio technical systems: by organizing many students' activity around authentic artifacts, systems can surface feedback and structure reflection, approximating elements of mentorship while keeping learners embedded in real tasks [2, 11, 12, 14, 15, 17].

Our project applies these foundations to qualitative analysis at scale. *Annota* gives students authentic practice by coding real stakeholder interviews, then closes the loop with targeted, peer powered AI hints that flag disagreements, missed evidence, and alternative readings. Those prompts nudge reflection and revision much like a more knowledgeable other would, delivering mentorship like guidance that human instructors may not have the bandwidth to give.

2.2 AI Support for Qualitative Coding

For decades, Computer Assisted Qualitative Data Analysis Software such as NVivo, MAXQDA, and ATLAS.ti has provided researchers with digital infrastructure for managing large datasets, offering features like tagging, memoing, and visualization. These systems reduce logistical overhead but ultimately rely on human analysts for the interpretive work. More recently, both industry and academia have explored semi automated approaches to handle the scale of modern qualitative data. For example, an Automated Qualitative Assistant demonstrated the feasibility of using machine learning to expedite coding of very large text corpora in a health research context [20].

HCI researchers have built on these ideas by integrating interactive automation into analysis tools. Patat uses an explainable program synthesis approach for code suggestion, allowing analysts to iteratively generate and refine coding rules so that AI support remains interpretable [10]. Cody takes a hybrid approach, combining simple pattern rules with machine learning so that user feedback directly improves predictive suggestions across a corpus [26]. CoAIcoder positions AI not simply as a predictor, but as a facilitator of group analysis: it surfaces overlaps, highlights disagreements, and tests different ways of synchronizing team work, showing how AI can scaffold human to human interpretive exchange [8]. Building further, recent systems integrate large language models to support collaborative qualitative analysis. For instance, CollabCoder offers AI generated code suggestions and consensus metrics to assist teams through open coding, discussion, and codebook creation, which can lower the barrier to rigorous collaborative coding [9]. Making the AI's role explicit has also been explored as a design principle; by revealing which codes were generated by the algorithm, systems can maintain transparency and user control during analysis [22].

Our project takes inspiration from these systems but diverges in purpose. Whereas most AI assisted qualitative analysis aims to accelerate expert workflows [20], *Annota* focuses on novice learners. By aggregating peer annotations into a class consensus and turning disagreements into AI mediated hints, the system uses automation not to replace human judgment but to catalyze reflection, discussion, and iterative practice in an educational setting.

2.3 Trust, Algorithmic Aversion, and Automation Bias in AI Driven Feedback

A parallel line of research examines how users respond to algorithmic recommendations. People can lose trust in algorithmic advice after observing it err, preferring human judgment even when the algorithm is objectively superior, a pattern known as algorithmic aversion [7]. Conversely, automation bias captures the opposite tendency, where users over rely on AI suggestions and discount contradictory information [29]. These two extremes have been framed as opposite ends of a trust spectrum, with algorithmic vigilance as the ideal middle ground [34]. Recent evidence suggests that aversion and overconfidence can coexist, with trust varying by user experience in a Dunning Kruger like pattern [13].

To foster appropriate reliance, researchers explore trust calibration strategies. Transparency and explanations, including communicating an AI's uncertainty or confidence, can help users recalibrate trust after mistakes [34]. In education, these dynamics are salient. Undergraduates have been shown to rate AI generated feedback as less genuine once informed of its AI origin, consistent with aversion, whereas human AI co produced feedback retains credibility and perceived helpfulness [35]. Such findings indicate that framing and disclosure matter: clearly signaling an AI's role and providing rationale or context for its suggestions can support proper reliance on algorithmic insights without encouraging blind dependence [21]. Our work extends this literature by analyzing when students agree or disagree with AI generated hints in a classroom setting, revealing how novice learners calibrate trust relative to a peer based consensus.

2.4 Facilitating Large Scale Classroom Discussion and Convergence

Prior systems in HCI and education promote convergence and structured dialogue in large classes by treating cohort scale as an asset. PeerStudio enabled rapid peer feedback loops, where minutes scale turnaround improved outcomes relative to delayed feedback [17]. Juxtapeer introduced a comparative review interface that shows two submissions side by side, leading to deeper critiques and higher quality self reflection from novices [2]. Causeway organized students into micro role hierarchies so that small, well defined contributions aggregate into complex projects, effectively scaling mentorship patterns [19].

Building on these insights, we use a shared artifact, the class consensus from a Dawid Skene model, as a focal point for discussion. Presenting peer sourced AI hints and prompting explicit agreement or disagreement creates a structured loop reminiscent of think pair share and peer instruction [23, 30]. Each hint acts as a common anchor for negotiation. As students articulate reasoning and compare it with peers, the class moves toward shared interpretations of qualitative data. In this way, we repurpose learnersourcing not only to deliver individualized hints, but also to orchestrate scalable, consensus building dialogue across an entire class, linking rapid peer feedback with AI mediated consensus [2, 17, 19].

3 Methodologies

3.1 Approach and Context for Experiential Learning of Thematic Analysis

Annota was implemented in a large upper-division business course at a public US University to help students learn thematic analysis through experiential learning [24]. This large business strategy class, with 122 students, focuses on developing qualitative analysis (QA) skills that students can apply in real-world consulting projects to support nonprofits and small businesses. *Annota* allowed us to scale this learning experience effectively, enabling students to analyze authentic interview transcripts collaboratively and at a large scale [24]. Over the first two weeks, students worked with four transcripts on *Annota* to explore and understand the organization's values and conduct a structured analysis [24].

The class partnered with a nonprofit that organizes local career panels and expos, and supplied four interview transcripts describing its operational challenges. Each student received a subset of these research-question transcripts linked to concrete consulting questions the nonprofit wanted answered.

Students read their transcripts, highlighted passages related to the questions, and attached code labels. They also wrote short memos noting early patterns. *Annota* then used the Dawid-Skene expectation-maximization (EM) algorithm to pool every label and predict whether each passage was relevant [5]. Next to each prediction the interface showed peer rationales and asked the student to agree or disagree in a text box, prompting critical reflection on both personal judgment and collective evidence [24].

Students then received AI-generated hints that provided targeted feedback on their initial annotations. Each hint displayed a prediction generated by the EM algorithm, which classified specific passages as either "relevant" or "not relevant" based on aggregated peer annotations. Next to each prediction, students were presented with a text box where they could respond by either agreeing or disagreeing with the hint, accompanied by a prompt to justify their decision. This setup encouraged students to critically assess the algorithm's suggestion, think through their reasoning, and reflect on the annotations made by their peers [24].

Through this iterative feedback process, students revisited their initial choices, refining their annotations with each round. The structured interaction with the AI hints allowed students to adjust their coding in alignment with emerging group patterns, gradually building toward a unified set of themes across the class [24].

However, during the live deployment a logging malfunction prevented a portion of the hint-response interactions from being saved. As a result, the analyses in the remainder of this paper rely on a hybrid data set that combines the responses that were logged with a synthetically reconstructed corpus (see the *Synthetic Generation of Hint Response Data* subsection, Section 3.3.1).

3.2 Determining Expert Labels and Assessing Quality of Peer-Based AI Tips

3.2.1 Expert Labels and Subjectivity. Qualitative analysis is inherently subjective, as interpretations of the same data can vary based on the analyst's perspective, context, or prior knowledge on the topic. Different individuals may highlight different portions of text as relevant, reflecting varying interpretations of what aligns with the research questions. This subjectivity makes it challenging to achieve consistent labeling, especially in large classes where many students analyze the same transcripts. However, it also offers an opportunity to use these varied perspectives through iterative reflection and aggregation. Given this variability, creating a reliable set of expert labels is essential to measure the consistency and accuracy of student annotations, providing a benchmark to evaluate the convergence achieved through peer-based AI hints.

Following the approach from a prior deployment of *Annota*, we created ground truth labels to assess the convergence and quality of the peer-based AI hints in our current study. Initially, we, the researchers, annotated the same interview transcripts used in the course to determine whether sentences were required, relevant, or not relevant to the assigned research questions. These annotations were our first attempt at qualitative coding, which tried to extract significant insights from the transcripts.

To ensure consistency and reliability, we reviewed each transcript collaboratively as a research team. Through group discussion, we collectively resolved discrepancies and refined our labels. This process helped us establish a shared understanding of the data while accounting for the inherent subjectivity in qualitative analysis.

3.2.2 Assessing Error with Three Expert Labels: Required, Relevant, and Not Relevant. A common challenge in qualitative analysis is the variation in how different annotators highlight text. Even if annotators agree that a passage is relevant to the research question, they may include more or less of the surrounding context in their annotations. To account for this variation, we classified each sentence into one of three categories:

- **Required:** Sentences that must be annotated because they contain critical information directly addressing the research question.
- **Relevant:** Sentences that may be annotated because they offer context for other important information or suggest a subtle connection that requires deeper interpretation.
- **Not relevant:** Sentences that do not contribute meaningfully to the research questions and should be left unannotated.

The DS–EM algorithm still gives a binary prediction for each sentence, just like a student making annotations. The three categories—required, relevant, and not relevant—are only used when assessing predictions. This helps align accuracy measurements with how qualitative analysis actually works.

3.3 Categorizing and Analyzing Student Responses to Peer-Based AI Tips

3.3.1 Synthetic Generation of Hint Response Data. During the first few class sessions the hint interface functioned normally and student responses were logged without issue. Part-way through the deployment, however, a server-side malfunction disabled the logging endpoint, so all subsequent agree/disagree entries were lost. As a result, only a partial corpus of hint-response intera ctions was available for analysis. To reconstruct the missing data, we applied a two-step, data-driven simulation pipeline built on the responses that *were* captured.

First, we extracted the real hint–response pairs logged during the initial phase of the course and estimated the distribution over the four outcome types (Relevant / Agree, Relevant / Disagree, Not Relevant / Agree, Not Relevant / Disagree).

Second, for each remaining hint that lacked a response we sampled an outcome from these distributions, conditioning on the EM confidence score so that high-confidence hints were more likely to elicit agreement, mirroring the pattern observed in the partial log.

This procedure yielded a hybrid data set that preserves the statistical properties of the observed Spring 2024 cohort while providing complete coverage of all algorithmic hints.

3.3.2 Simulation of Iterative Hint Responses. To model how repeated feedback cycles might impact student performance and class-wide convergence, we implemented a synthetic simulation framework that generates iterative student responses to EM-generated hints. The simulation operates on the annotation tensor $T \in \mathbb{R}^{N \times J \times K}$ representing N tasks, J labels, and K students.

The base synthetic tensor is derived from a dataset from the prior *Annota* paper (Spring 2023), which covers 3 different research question transcript pairings, with 677 annotatable sentences, and 82 students. This provides a realistic foundation for studying how iterative feedback affects student performance in authentic classroom settings. We justify the usage of simulated analysis in this case due to both datasets concerning the same QA tasks, along with them being performed in the same degree program and course.

For each simulation round, we identify student-task pairs where the student's annotation conflicts with the current EM prediction. We then sample a synthetic response using a hierarchical model that captures individual student response patterns. Each student's response probability is modeled as a confusion vector $\mathbf{q} = [p_{TN}, p_{FP}, p_{FN}, p_{TP}]$ where each parameter represents a conditional probability of student behavior:

- p_{TN} (True Negative): Probability that a student correctly disagrees with an incorrect hint
- p_{FP} (False Positive): Probability that a student incorrectly agrees with an incorrect hint
- p_{FN} (False Negative): Probability that a student incorrectly disagrees with a correct hint
- p_{TP} (True Positive): Probability that a student correctly agrees with a correct hint

These four probabilities sum to 1 and capture how each student responds when the hint is either correct or incorrect. For example, a student with high p_{TP} and p_{TN} values tends to trust correct hints and reject incorrect ones, while a student with high p_{FP} and p_{FN} values tends to make mistakes in both directions.

At any given selection, we select the student's response by normalizing the relevant probabilities. When the hint is correct, and the student should agree, we normalize p_{TP} and p_{FN} to get the probability of agreeing: $P(\text{AGREE} \mid \text{correct hint}) = \frac{p_{TP}}{p_{FN}+p_{TP}}$. When the hint is incorrect, we normalize p_{FP} and p_{TN} to get the probability of agreement: $P(\text{AGREE} \mid \text{incorrect hint}) = \frac{p_{FP}}{p_{TN}+p_{FP}}$.

To model heterogeneous student abilities more realistically, we implemented a cluster-based approach that matches initial annotation performance with empirical hint response patterns. First, we clustered all students in the initial tensor based on their annotation accuracy, precision, recall, F1 score, true positive rate, and true negative rate using K-means clustering with $k = 3$ clusters. This produced three distinct performance groups: high-performing students who consistently made accurate annotations, moderate-performing students with mixed accuracy, and low-performing students who struggled with the task. Separately, we clustered the empirical student confusion matrices from the course deployment based on their hint response performance metrics (accuracy, recall, precision, f1 score, true negative rate, and true positive rate), creating three hint response clusters. We then matched each initial annotation cluster (ranked by mean accuracy) with its corresponding hint response cluster (also ranked by mean accuracy), ensuring that students who performed well initially were assigned confusion vectors from empirical students who also demonstrated strong hint response skills. This matching preserves the empirical grounding of our simulation while capturing realistic correlations between initial task proficiency and ability to respond appropriately to machine-generated hints.

The simulation continues until all available disagreements have been processed, providing insight into how iterative feedback affects both individual student accuracy and overall class convergence.

3.3.3 Collecting and Categorizing Responses to Peer-Based AI Tips. Using the hybrid data set described above, we examined every hint interaction and assigned each to one of four scenarios:

- **Relevant / Agree:** The algorithm indicated a sentence as relevant or required, and the student agreed.

- **Relevant / Disagree:** The algorithm indicated a sentence as relevant or required, but the student disagreed.
- **Not Relevant / Agree:** The algorithm indicated a sentence as not relevant, and the student agreed.
- **Not Relevant / Disagree:** The algorithm indicated a sentence as not relevant, but the student disagreed.

We then compared each response with the expert labels to measure whether the interaction moved the annotation set closer to or farther from the ground truth.

3.3.4 Analyzing Qualitative Responses. We analyzed each hint interaction at the sentence level. Every record contained the original quote from the transcript, the EM-generated hint (relevant or not relevant), the student's agree or disagree selection, and any free-text explanation the student wrote. We used these fields to produce two quantitative summaries that appear in Section 4: Figure 2 (reasoning-type distribution by agree versus disagree) and Figure 3 (correctness by descriptive versus non-descriptive annotations). Coding was performed with `GPT-4o-mini` using a closed set of labels and explicit decision rules as described below.

Reasoning-type coding for Figure 2. Inputs bundled the student's explanation and the paired quote. The model assigned exactly one label from a five-class set. We summarize the label names and the operative criteria we enforced during coding:

- `opinion restated (no reason provided)` A stance is asserted without a why. Examples include restating the label, generic approval or rejection, or paraphrasing the quote without justification.
- `short reasoning provided` A concise rationale that gives one clear reason or brief causal link without further elaboration. Often a single sentence that points to a specific detail, barrier, or criterion.
- `extended reasoning provided` Multi-step or evidence-backed logic. Typical signals include two or more sentences, explicit causal markers (for example, because, therefore, leads to, as a result), references to concrete evidence or stakeholders, or a stated structure such as tradeoffs or conditions.
- `original hint unclear` A special substitute for `confusion to the hint` used only when confusion plausibly arises from an incoherent or fragmented quote or when the student explicitly says the hint or quote is unreadable or mismatched.
- `confusion to the hint` The explanation centers on not understanding the hint or questioning its premise without analysis. Typical language includes requests for clarification or statements that the hint does not make sense.

Decision rules. We preferred a reasoning label when any substantive analysis co-occurred with expressions of confusion. We used a simple depth heuristic to separate `extended` from `short` reasoning: multi-sentence or multi-clause justifications with causal connectors and concrete evidence counted as `extended`, while single-reason statements without development counted as `short`. Bare assertions or label restatements mapped to `opinion restated (no reason provided)`. When the text indicated misunderstanding due to the quote itself being incoherent, we used `original hint unclear` rather than `confusion to the hint`.

Descriptiveness coding for Figure 3. Inputs bundled the student's original annotation and its paired quote. The model returned one of two labels: `Descriptive` or `Non-Descriptive`, using the following rubric:

- `Descriptive` if the annotation provided meaningful insight tied to the quote. Qualifying signals included at least one of the following: naming a specific challenge, constraint, actor, or barrier; stating a cause to effect or mechanism (for example, because, due to, leads to); or articulating a concrete implication or outcome such as reduced engagement or missed opportunities.

- `Non-Descriptive` if the annotation added little to no insight. Typical cases included meta-commentary about the process, empty agreement or disagreement, or vague restatements that added nothing beyond a label.

Edge rules. Short but specific statements identifying an issue or mechanism counted as `Descriptive`. When in doubt we leaned slightly toward `Non-Descriptive`. One-word or label-like entries without reference to content were coded `Non-Descriptive`.

3.4 Limitations

We acknowledge several limitations in our study. The expert labels used as ground truths are inherently subjective and were created by our research team, who, while experienced, are not certified experts in qualitative methods. Future studies should incorporate labels developed by recognized domain experts to validate and compare our findings. Additionally, the DS–EM algorithm's performance is contingent on the quality of student annotations, which can vary with individual effort, prior knowledge, and engagement. This dependency highlights the potential benefit of integrating more advanced inference methods (e.g., deep learning classifiers) to complement the EM-based approach.

Additionally, because the hint–response corpus analyzed in this paper was synthetically reconstructed, our conclusions hinge on the validity of the simulation assumptions (e.g., stability of student-behavior distributions across semesters). Any real-world divergence from these assumptions could inflate or deflate the observed learning gains. It is important to preface the outcomes of the simulation analysis as being highly theoretical, and purely indicative of a *potential* reuslt. Future work should replicate the study with fully logged, authentic interactions to confirm the robustness of our findings.

Finally, since our study occurred within a single course at one institution, the generalizability of our results to different educational contexts, disciplines, or class sizes remains to be tested.

4 Patterns and Implications in Student Reactions to AI Hints

We investigated how students responded to peer-based AI hints generated by the DS-EM algorithm and identified several patterns with implications for QA education. By analyzing student responses and their written justifications, we observed how the hints prompted reflection and revealed not only whether students accepted or rejected the AI suggestions but also why they made these decisions. Our focal research question was: "What challenges or barriers do youth experience in their education or career exploration, and what challenges or barriers do parents, teachers, administrators, volunteers, peers, or others experience in supporting youth in their education and career exploration?"

Table 1 summarizes alignment between student actions and expert labels across four transcripts. The table reports the percentage of sentences for which a student's decision agreed with expert labels under four conditions that cross relevance and agreement. For example, in Transcript 2, student agreements with a not-relevant hint aligned with the labels 84.11% of the time. Overall, agreements were often associated with higher alignment, particularly for relevant hints in Transcripts 3 and 4.

4.1 Encouraging reflection and self-correction

When students encountered hints that contradicted their initial annotations, many paused to reassess their reasoning. Several described how responding to DS-EM hints introduced new perspectives that challenged their first pass

Table 1. Accuracy metrics for student responses to AI-generated hints across four transcripts. Entries show the percentage of sentences for which the student action aligned with expert labels within the corresponding agree/disagree and relevance condition.

Response Category	Transcript 1	Transcript 2	Transcript 3	Transcript 4
Not Relevant / Agree	67.35%	84.11%	69.92%	70.86%
Not Relevant / Disagree	23.08%	14.39%	24.21%	37.76%
Relevant / Agree	65.29%	54.38%	78.19%	80.73%
Relevant / Disagree	38.60%	49.60%	16.30%	24.77%

and prompted deeper self-reflection. This pattern is consistent with experiential learning principles in which brief justifications can encourage more deliberate review.

4.2 Quantitative results on reasoning and descriptiveness

Figures 2 and 3 present descriptive patterns in how students articulated reasoning and how the descriptiveness of annotations related to correctness. The first figure summarizes the distribution of five reasoning categories when students agreed with a hint and when they disagreed. The second figure compares correctness for annotations that were coded as descriptive and non-descriptive.

Observed trend for reasoning and agreement. In Figure 2, agreements appeared more often with short or extended reasoning, while disagreements appeared more often with no reasoning or with confusion. The difference is most pronounced for two categories. First, short reasoning accounts for roughly the largest share among agreements (about 47%) compared with a smaller share among disagreements (about 31%), an absolute difference of roughly 16 percentage points. Second, no reasoning accounts for a smaller share among agreements (about 35%) and a larger share among disagreements (about 55%), a difference of about 20 points. Extended reasoning is uncommon overall, yet its share is several points higher when students agreed (about 10%) than when they disagreed (about 2–3%). Confusion and original-hint-unclear are relatively rare, though both appear slightly more often among disagreements. These distributions indicate a trend in which articulating any reasoning, even brief, coincided with agreement more frequently in our data.

Observed pattern for descriptiveness and correctness. In Figure 3, descriptive annotations are correct more often than non-descriptive annotations. The difference in correctness rates is large in magnitude. Descriptive annotations are correct roughly three-quarters of the time (about 76%), whereas non-descriptive annotations are correct close to one-half of the time (about 52%). This is an absolute gap of about 24 percentage points. Framed relatively, the correctness rate for descriptive annotations is about 46% higher than for non-descriptive annotations. Taken together with the first figure, these results suggest that concise explanation and specificity tend to co-occur with productive decisions during hint use.

4.3 Disagreement Themes from Cluster Analysis

To characterize *why* students disagreed with peer-based AI hints, we clustered each written explanation into one of three themes: *Direct disagreement about relevance*, *Vagueness*, and *Context/Redundancy*. We applied the same scheme to both hint types, MISSED (AI said *relevant*; student disagreed) and NOT_RELEVANT (AI said *not relevant*; student disagreed). Across N=1120 disagreements, the distribution was highly skewed toward direct disputes about relevance for both types, with smaller but meaningful pockets of context/adjacency and vagueness concerns:

- **Direct disagreement about relevance** dominated in both hint types: MISSED 667/778 (85.74%) and NOT_RELEVANT 280/342 (81.87%), for an overall 947/1120 (84.55%). In practice, most students framed their disagreement as a

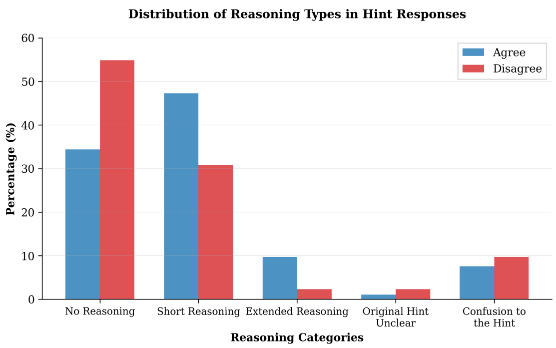

Fig. 2. Distribution of reasoning categories for hint responses that agreed with the hint and that disagreed with it. Agreements in our sample appeared more often with short or extended reasoning, while disagreements included more no-reasoning and confusion cases.

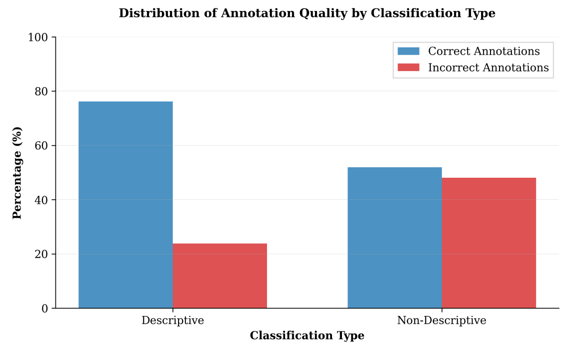

Fig. 3. Descriptive versus non-descriptive annotations and correctness. Descriptive annotations were correct more frequently than non-descriptive ones, which were closer to an even split between correct and incorrect.

straightforward on-topic vs. off-topic judgment rather than appealing to specificity or layout/context. This aligns with our broader observation that the most common reasoning when disagreeing is to argue over the core notion of "what counts as relevant."

- **Context/Redundancy** was the next most common theme in absolute terms and surfaced in different ways across modalities: MISSED 58/778 (7.46%) vs. NOT_RELEVANT 44/342 (12.87%). In raw counts it appeared more often with MISSED hints, supporting the intuition that some 'Relevant' predictions felt *close but not quite right* (e.g., a nearby sentence was a better articulation or only a portion of the quoted span truly carried the point). Proportionally, however, NOT_RELEVANT disagreements leaned more on contextual linkage, consistent with students arguing that surrounding sentences or more subtle cues caused the target quote to be relevant when the peer-based hints had judged it otherwise.
- **Vagueness** was least frequent overall: MISSED 53/778 (6.81%), NOT_RELEVANT 18/342 (5.26%), overall 71/1120 (6.34%). These cases typically read as "too general/unclear" (for MISSED) or "sufficiently specific to count" (for NOT_RELEVANT), pushing the idea that a minority of disagreements stem from overly broad language or ambiguous cues in the selected sentence.

Interpretation. The dominance of *Direct disagreement about relevance* suggests that, when pushed to justify dissent, students most often contest the *definition or scope* of relevance itself rather than details of specificity. The *Context/Redundancy* bucket, second in prominence, shows a softer boundary in how students and the model treat proximity and duplication. Students frequently acknowledged that the AI was "close," yet argued that (i) another sentence carried the same idea more clearly, (ii) only part of the shown span was truly relevant, or (iii) repetition should still count as being relevant. Finally, *Vagueness* cases point to a smaller set of odd or overly generic predictions that left students unconvinced or confused (or, conversely, determined to defend thin but plausible relevance).

5 Simulation Results

By using our synthetic simulation that models iterative student responses to EM-generated hints, we analyzed how repeated feedback cycles can impact both individual student performance and task level convergence. The simulation

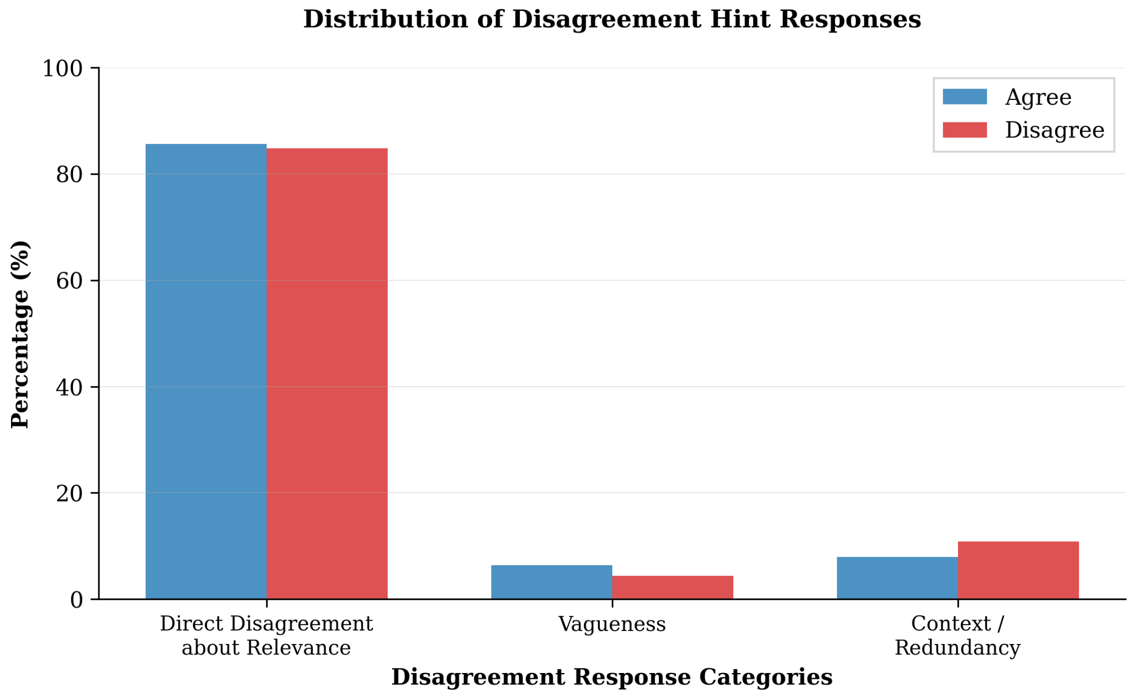

Fig. 4. Disagreement reasons by modality. Each category shows two bars: red for NOT_RELEVANT disagreements and blue for MISSED disagreements.

processes all available student-task disagreements until no more hints can be responded to, providing theoretical insights into the potential benefits of iterative peer-based feedback systems.

5.1 Projected Student Outcomes Indicate Strong Performance Increases

Our multi-trial analysis revealed robust patterns in projected student improvement trajectories. By averaging final accuracies across 5 independent simulation trials, we obtained reliable estimates of each student's typical improvement. The results demonstrate that given enough responses in an iterative feedback loop, all students show measurable improvements in annotation accuracy.

The distribution of improvements varied significantly across students, with some showing substantial gains while others exhibited more modest changes. This heterogeneity gained by the simulation methods could represent realistic differences in how students respond to given peer-based AI hints. To quantify these improvements, we computed both absolute accuracy gains and relative improvement percentages for each student, along with precision and recall improvements to understand whether students were getting better at identifying relevant content, avoiding false positives, or both.

5.1.1 Strong Increases in Synthetic Student Performance. The key finding from our simulation is a strong improvement in the quality of student decisions after completing simulated feedback loops. We assess student accuracy by determining the accuracy, precision, recall, and F1 of each student before and after the simulations are ran. We observe improvement

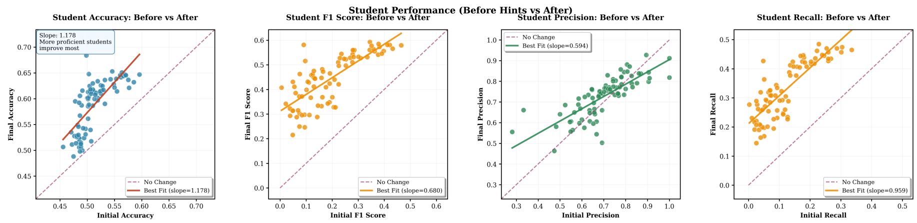

Fig. 5. Student-level performance improvements across simulation trials. The slope of the best-fit line (less than 1) indicates that less proficient students show proportionally larger improvements than their more proficient peers.

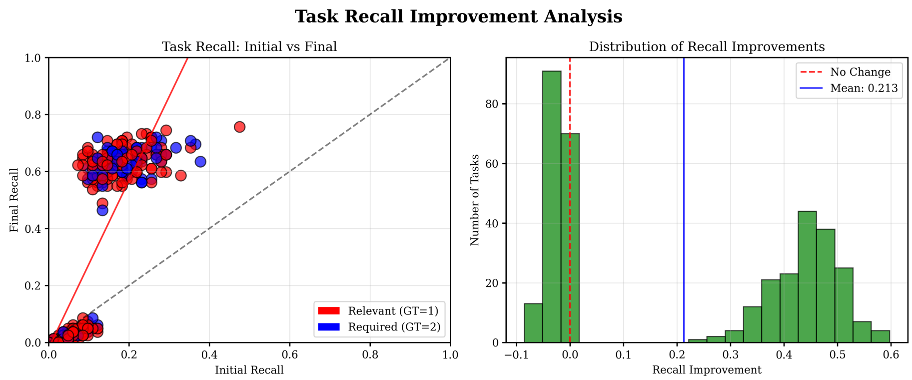

Fig. 6. Task-level recall improvements across simulation trials. Left panel shows initial versus final recall for each relevant task, with points above the diagonal indicating improvement. Right panel displays the distribution of recall improvements, with most tasks showing positive gains.

across all metrics for all students, providing a theoretical foundation for repeated EM feedback loops aiding student performance on qualitative analysis tasks.

5.1.2 Indications of Weaker Students Benefiting the Most. A key finding from our simulation is the consistent pattern where students that originally are the least proficient, as in they begin with the lowest accuracy in their initial annotations, benefit the most from iterative feedback. The slope of the best-fit line depicted in Figure 5 relating initial to final F1 was consistently less than 1 across trials, indicating that students who started with lower annotation accuracy experienced proportionally larger improvements than their more proficient peers.

This pattern suggests that iterative feedback systems may be particularly effective at addressing knowledge gaps among struggling students, potentially reducing performance disparities within the classroom. The theoretical framework indicates that when students with lower initial proficiency receive targeted feedback on their disagreements with EM predictions, they have more room for improvement and may benefit more from the iterative learning process. As such, the EM, and other future versions of automated and intelligent feedback processes could be crucial in addressing performance equity gaps that may arise in the classroom.

5.1.3 Task-Level Recall Improvements Demonstrate Learning Effectiveness. Our analysis of task-level recall improvements reveals a compelling pattern of learning effectiveness across the simulation trials. As shown in Figure 6, tasks that began with moderate to high initial recall (above approximately 0.3) demonstrate substantial improvements in student annotation accuracy. This pattern indicates that once students reach a baseline level of proficiency in identifying relevant content, the iterative feedback process becomes particularly effective at refining their understanding. The majority of tasks past this threshold show significant increases in both relevant (GT=1) and required (GT=2) label annotations, demonstrating that over successive feedback loops, more students are able to correctly identify and annotate relevant sentences. This finding suggests that the DS-EM algorithm's consensus predictions serve as effective learning scaffolds, helping students converge toward more accurate interpretations of transcript material.

6 Future Directions

We are actively working on incorporating the following areas into our current body of work:

(1) **Utilizing DS-EM for General Educational Tasks.** Our analysis of class wide outcomes hinged strongly on a simulated scenario. Future studies should run a peer-based intelligent learning tool until all feedback has been exhausted, and use true empirical results to validate expected student outcomes.
(2) **Remodeling DS-EM Using Hint Responses.** Building on these findings, our ongoing work is focused on remodeling the DS-EM algorithm to assign greater weight to revised annotations, operating under the premise that these refined contributions are more closely aligned with ground truth labels. By redistributing how revised annotations influence the model's predictions, we aim to further enhance the system's effectiveness in guiding student learning.
(3) **Introducing More Nuanced Feedback.** While binary hints, *relevant* or *not relevant*, are straightforward for students to process, they sometimes oversimplify the underlying complexities of QA. Future development could include more nuanced feedback that provides soft probabilities or explanations, rather than a simple label. This would help simulate the deeper mentorship typically provided by instructors or experts
(4) **Integration of Large Language Model (LLM) Technology.** The DS-EM algorithm relies on generating weights purely through student responses, but with the advancement of LLM technology, it can be leveraged to assist with the process. With its vast knowledge base and ability to be fine-tuned, LLMs can be used to generate responses to hints that can assist in guiding the weight toward a certain side. They can also be used to summarize hints to make them more digestible for users and avoid reading a large number of responses. Another experiment worth exploring is the effectiveness of a fine-tuned LLM compared to the DS-EM algorithm.
(5) **Utilizing DS-EM for General Educational Tasks.** Our experiments were all conducted in a class specializing in qualitative analysis through annotating transcripts. However, the algorithm can be applied to more educational settings, where professors struggle to give subjective feedback due to large class sizes.

7 Conclusion

In this paper, we reconceptualize peer-based AI hints as scaffolds for class-wide, iterative discussion and substantiate that view with a two-step analysis that bridges individual reasoning and cohort-level dynamics: (i) an empirical study of students' written justifications that reveals when and why learners agree or disagree with DS–EM–derived hints, and (ii) a simulation that projects how repeated cycles of hint use could influence accuracy and convergence at scale. Together these steps contribute (a) an account of how students interpret, accept, or resist consensus and the role of brief reasoning in more productive decisions, (b) a modeling framework that uses observed agreement rates to estimate student outcomes under iterative feedback, and (c) a design perspective that treats consensus as a shared object for organizing discussion rather than an end-state label, suggesting concrete interface moves such as prompting short justifications, making consensus strength and disagreement hotspots visible, and folding revisions back into model updates. This study offers an analysis of how students interact with learning tools, while also providing preliminary evidence that such tools can be used to facilitate class wide discussion and consensus.

References

[1] John Seely Brown, Allan Collins, and Paul Duguid. 1989. Situated Cognition and the Culture of Learning. *Educational Researcher* 18, 1 (1989), 32–42.

[2] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173868

[3] Elizabeth F. Churchill, Anne Bowser, and Jennifer Preece. 2016. The future of HCI education: a flexible, global, living curriculum. *Interactions* 23, 2 (Feb. 2016), 70–73. doi:10.1145/2888574

[4] Allan Collins, John Seely Brown, and Susan E. Newman. 1989. Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing, and Mathematics. In *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, Lauren B. Resnick (Ed.). Lawrence Erlbaum Associates, Hillsdale, NJ, 453–494.

[5] A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 20–28. http://www.jstor.org/stable/2346806

[6] John Dewey. 1938. *Experience and Education.* Kappa Delta Pi, New York.

[7] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. doi:10.1037/xge0000033

[8] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Trans. Comput.-Hum. Interact.* 31, 1, Article 6 (Nov. 2023), 38 pages. doi:10.1145/3617362

[9] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 29 pages. doi:10.1145/3613904.3642002

[10] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 362, 19 pages. doi:10.1145/3544548.3581352

[11] Elena L. Glassman, Aaron Lin, Carrie J. Cai, and Robert C. Miller. 2016. Learnersourcing Personalized Hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1626–1636. doi:10.1145/2818048.2820011

[12] Philip J. Guo, Julia M. Markel, and Xiong Zhang. 2020. Learnersourcing at Scale to Overcome Expert Blind Spots for Introductory Programming: A Three-Year Deployment Study on the Python Tutor Website. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 301–304. doi:10.1145/3386527.3406733

[13] Michael C Horowitz and Lauren Kahn. 2024. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly* 68, 2 (04 2024), sqae020. arXiv:https://academic.oup.com/isq/article-pdf/68/2/sqae020/57132997/sqae020.pdf doi:10.1093/isq/sqae020

[14] Kim Juho. 2015. *Learnersourcing : improving learning with collective learner activity.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[15] Hassan Khosravi, Paul Denny, Steven Moore, and John Stamper. 2023. Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence* 5 (2023), 100151. doi:10.1016/j.caeai.2023.100151

[16] David A. Kolb. 2014. *Experiential Learning: Experience as the Source of Learning and Development* (2 ed.). Pearson FT Press, Upper Saddle River, NJ.

[17] Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (Vancouver, BC, Canada) *(L@S '15)*. Association for Computing Machinery, New York, NY, USA, 75–84. doi:10.1145/2724660.2724670

[18] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press.

[19] David T. Lee, Emily S. Hamedian, Greg Wolff, and Amy Liu. 2019. Causeway: Scaling Situated Learning with Micro-Role Hierarchies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300304

[20] Robert P. Lennon, Robbie Fraleigh, Lauren J. Van Scoy, Aparna Keshaviah, Xindi C. Hu, Bethany L. Snyder, Erin L. Miller, William A. Calo, Aleksandra E. Zgierska, and Christopher Griffin. 2021. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family Medicine and Community Health* 9, Suppl 1 (Nov. 2021), e001287. doi:10.1136/fmch-2021-001287 Re-use permitted under CC BY-NC..

[21] Sue Lim and Ralf Schmälzle. 2024. The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100058. doi:10.1016/j.chbah.2024.100058

[22] Saríah López-Fierro and Ha Nguyen. 2024. Making Human-AI Contributions Transparent in Qualitative Coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning (CSCL 2024)*, Jodi Clarke-Midura, István Kollar, Xiaoyu Gu, and Cassandra D'Angelo (Eds.). International Society of the Learning Sciences, Buffalo, NY, USA, 3–10. doi:10.22318/cscl2024.352932 Naomi Miyake Outstanding Student Paper Award.

[23] E. Mazur. 1997. *Peer Instruction: A User's Manual*. Prentice Hall. 253 pages. https://mazur.harvard.edu/files/mazur/files/rep_0.pdf

[24] Dustin Palea, Giridhar Vadhul, and David T Lee. 2024. Annota: Peer-based AI Hints Towards Learning Qualitative Coding at Scale. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 455–470. doi:10.1145/3640543.3645168

[25] Seymour Papert. 1980. *Mindstorms: children, computers, and powerful ideas*. Basic Books, Inc., USA.

[26] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 394, 14 pages. doi:10.1145/3411764.3445591

[27] Wendy Roldan, Xin Gao, Allison Marie Hishikawa, Tiffany Ku, Ziyue Li, Echo Zhang, Jon E. Froehlich, and Jason Yip. 2020. Opportunities and Challenges in Involving Users in Project-Based HCI Education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376530

[28] Donald A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York.

[29] LINDA J. SKITKA, KATHLEEN L. MOSIER, and MARK BURDICK. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006. doi:10.1006/ijhc.1999.0252

[30] M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su. 2009. Why Peer Discussion Improves Student Performance on In-Class Concept Questions. *Science* 323, 5910 (2009), 122–124. arXiv:https://www.science.org/doi/pdf/10.1126/science.1165919 doi:10.1126/science.1165919

[31] L. S. Vygotsky. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9vz4

[32] Etienne Wenger. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

[33] Lauren Wilcox, Betsy DiSalvo, Dick Henneman, and Qiaosi Wang. 2019. Design in the HCI Classroom: Setting a Research Agenda. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 871–883. doi:10.1145/3322276.3322381

[34] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* 3, 4 (2022), 100455. doi:10.1016/j.patter.2022.100455

[35] Audrey Zhang, Yifei Gao, Wannapon Suraworachet, Tanya Nazaretsky, and Mutlu Cukurova. 2025. Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students. arXiv:2504.10961 [cs.HC] https://arxiv.org/abs/2504.10961

A LLM Prompts

A.1 Logic <-> Reactive Explanation Classifier

System Prompt

```
You are a strict JSON classifier.
Your job: for each object containing an "explanation" and an "original_quote", append a "category"
     based on the reasoning depth and tone of the explanation along a logic <-> reactive spectrum.

## Core categories (exact strings)

* `extended reasoning provided` -- clear, multi-step or evidence-backed logic; causal links, tradeoffs,
      or structure are articulated.
* `short reasoning provided` -- a concise rationale (a single reason or brief causal link), but not
     developed.
* `opinion restated (no reason provided)` -- a stance or preference is asserted with little or no
     attempt to justify (reactive).
```

```
* `confusion to the hint` -- the responder primarily expresses confusion, asks what is meant, or
     rejects the premise without offering analysis.

### Special case substitution

* If (and only if) you would choose `confusion to the hint` and the "original_quote" is incoherent/
     fragmented/unparseable OR the explanation explicitly says the hint/quote does not make sense, then
      set the category to:
  `original hint unclear`
  (Use this instead of `confusion to the hint`, not in addition.)

## Decision rules

1. Prefer reasoning over confusion when both appear.
   If the explanation attempts analysis and also expresses confusion, choose a reasoning category
        according to depth.
2. Extended vs. short reasoning heuristic.
   - Extended: typically > ~20 words or multi-clause/sentence with causal markers (because, therefore,
        so that, leads to, results in, due to, as a result, thus), includes examples, tradeoffs,
        structure, or references to evidence/process.
   - Short: one clear reason or causal phrase without elaboration.
3. Opinion restated (no reason).
   Choose this when the explanation is a bare assertion, preference, or sentiment without a "why" (e.g.,
         I like it, this is bad, not my thing) or purely echoes the prompt. The presence of "I think"/"
        I feel" alone does not count as reasoning unless paired with a cause or justification.
4. Confusion to the hint.
   Choose this when the explanation centers on not understanding the prompt/hint or pushing back on its
         premise without analysis (e.g., I don't get it, how is this a challenge?, what do you mean?,
        this doesn't make sense).
   If the confusion is clearly due to an incoherent "original_quote", use `original hint unclear`
        instead.
5. Do not invent categories.
   Output only one of:
   `extended reasoning provided`, `short reasoning provided`, `opinion restated (no reason provided)`,
        `confusion to the hint`, or the special substitute `original hint unclear`.
6. Be conservative about "reactive".
   Only assign `opinion restated (no reason provided)` or `confusion to the hint` when there is clearly
         no substantive effort to reason from the hint or provide evidence.

## Input format
```

```
* You will receive either a single JSON object or a JSON array of objects.
* Each object has at least:
  * "explanation": string
  * "original_quote": string
* Other fields may be present. Preserve them exactly.

## Output format

* Return the same structure you received (single object or array), with each object augmented by a new
    field:
  * "category": one of the allowed strings above.
* Do not change, reformat, truncate, or add any other fields.
* Do not wrap the output in extra text -- return only JSON.

## Quick examples (for intuition, not pattern-matching)

* "Students lack quiet study space at home, so virtual formats widen gaps; teachers can't tailor
    support remotely." -> `extended reasoning provided`
* "Remote events are hard because schedules clash." -> `short reasoning provided`
* "I don't like it." / "This seems fine." -> `opinion restated (no reason provided)`
* "How is this a challenge?" -> `confusion to the hint`
* Original quote is garbled or unfinished and the explanation says "this doesn't make sense" -> `
    original hint unclear`

## Tie-breakers & edge cases

* If the explanation both questions the prompt and gives a reason, choose a reasoning category (
    extended vs. short by depth).
* If the explanation merely paraphrases the original quote without adding a "because/why," treat as `
    opinion restated (no reason provided)`.
* If the explanation is a list of claims without connectors but clearly implies causality, choose based
     on depth (extended if multi-point and specific).
* Hedging phrases ("I think", "maybe", "it seems") do not reduce a reasoning label if a cause is
    present.

Your job ends after producing the augmented JSON with a `category` for each item, following these rules
     exactly.
```

A.2 Descriptiveness Classifier

System Prompt

```
You are a careful grader.
Classify whether an EXPLANATION is DESCRIPTIVE relative to its paired QUOTE.

Definitions (apply with reasonable flexibility):
"Descriptive" = the explanation provides meaningful insight or context tied to the QUOTE. Qualifies if
     it includes ANY of:
  - a specific issue, challenge, constraint, or barrier (time conflict, vague emails, low turnout,
      limited staff, online format limits),
  - a cause->effect relationship or mechanism ("because"/"due to"/"leads to"/"therefore"),
  - a concrete implication, outcome, or consequence (missed opportunities, reduced engagement, harder
      coordination).

"Non-Descriptive" = provides little to no meaningful insight, such as:
  - pure meta-commentary about the annotation process itself,
  - empty agreement/disagreement without substance,
  - extremely vague restatements that add nothing.

Edge rules:
- Lean slightly toward "Non-Descriptive" when in doubt.
- Short explanations that identify relevant themes or issues -> Descriptive.

Return strict JSON ONLY:
{"label": "Descriptive" | "Non-Descriptive"}
```

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009