Annota: Peer-based AI Hints Towards Learning Qualitative Coding at Scale

DUSTIN PALEA, UC Santa Cruz, USA
GIRIDHAR VADHUL, UC Santa Cruz, USA
DAVID T. LEE, UC Santa Cruz, USA

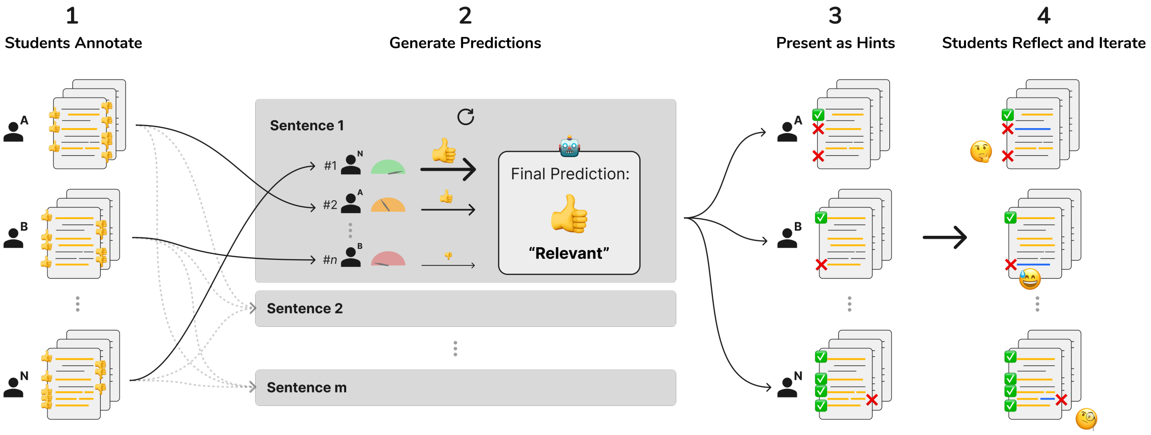


Fig. 1. Our peer-based hints system: **1)** students annotate transcripts for given research questions, **2)** annotations are used to predict whether each sentence is "Relevant" or "Not relevant" to the research question, via an implementation of the Dawid-Skene expectation maximization algorithm, **3)** 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 or not and iterate.

Learning qualitative analysis requires personalized feedback and in-depth discussion not possible for educators to provide in a large course, resulting in many students obtaining only a shallow exposure to qualitative user research and interpretative skills. To overcome this challenge, we introduce a learnersourcing method that builds on the Dawid-Skene expectation maximization (EM) algorithm to generate peer-based AI hints that support students in one aspect of qualitative analysis: determining what sentences are relevant to the research question. After one annotation round, class-wide annotations are used to predict relevant sentences and to generate hints prompting students to revisit missed or incorrectly annotated sentences. An in-the-wild deployment within a large course (N=122) showed that our algorithm converged to comparatively high accuracy despite noisy student labels, and after only ~20 students. An analysis of student interviews found that peer-based AI hints helped improve understanding of research questions, led to more careful examination of transcript annotations, and improved understanding of when they were over-annotating or under-annotating.

CCS Concepts: • **Applied computing** → **Collaborative learning**; • **Human-centered computing** → **Collaborative and social computing**; *Computer supported cooperative work*;

Additional Key Words and Phrases: AI-assisted human-to-human collaboration, learnersourcing, learning qualitative analysis

Manuscript submitted to ACM

ACM Reference Format:
Dustin Palea, Giridhar Vadhul, and David T. Lee. 2023. Annota: Peer-based AI Hints Towards Learning Qualitative Coding at Scale. In . ACM, New York, NY, USA, 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Qualitative analysis (QA) of stakeholder interviews is a fundamental method in Human-Computer Interaction (HCI) that helps designers deeply understand their users, ultimately guiding them toward creating solutions better attuned to the complex context of user needs. However, despite such qualitative research methods being recognized as a "very important" part of HCI education [14], especially for developing human-centered technologies for the public interest [46], the growing demand for HCI education has made it increasingly challenging to engage students deeply and experientially in learning user research and other design and research methods [50]. This is particularly evident in large project-based courses, where stakeholder interviews and analyses can sometimes be cursory rather than in-depth and rigorous, leading to various efforts aimed at addressing different challenges in the process ranging from supporting better student recruitment of interviewees [41] to better integrating theory in the process of framing problems and solutions [8].

One important challenge is that qualitative analysis skills are typically developed experientially through in-depth discussions and feedback with experienced mentors or collaborators (a "more knowledgeable other" in Vygotsky's Zone of Proximal Development [48]), through which students hone interpretative skills and develop the intuition and patience for detail-oriented annotation and reflection. Providing such an experience is next to impossible in the 100+ student design courses that are now common, and even more so in under-resourced institutions unable to fund large numbers of graduate or undergraduate student teaching assistants, leading us to the motivating question for this paper: *How might we better support experiential, apprenticeship-like learning of qualitative analysis in large classes?*

In this paper, we explore whether a learnersourcing approach [29] might be able to turn the large size of class into an advantage rather than a disadvantage, through leveraging the fact that there are many students all working on annotating the same transcripts. Specifically, we introduce peer-based AI hints for qualitative analysis, an approach that builds on Dawid and Skene's Expectation Maximization algorithm (DS-EM) [18] to support students in one aspect of qualitative analysis: determining what sentences are relevant to the research question. We implement peer-based AI hints within Annota, a web-based platform we developed for students to conduct and learn qualitative coding of interview transcripts. After students annotate transcripts individually, the platform aggregates all student annotations using the DS-EM algorithm to predict relevant sentences and then provides students with personalized hints on potential passages they missed or incorrect annotations they made. Students reflect on these hints by agreeing or disagreeing with the hints and justifying their stance with a short explanation. In this way, peer-based AI hints seek to play the role of a "more knowledgeable other" in supporting students to work experientially in their Zone of Proximal Development [48] and to engage in reflection on their work towards evolving schemas of ideal practice [38]. See **Figure 1** for an overview.

We conducted an in-the-wild deployment of Annota in a large course ($N = 122$) to collect authentic annotations from learners engaged in learning and doing qualitative analysis in the context of a real-world consulting project for a local non-profit (part of which involved analyzing interviews of the organization's stakeholders). We then used all student annotations to generate peer-based AI hints on the platform which students could optionally engage in to improve their work. In our assessments of the convergence properties of our approach for generating AI tips, we found that with 2-3 interview transcripts, $\sim$ 20 annotators is sufficient for the algorithm to converge to comparably high-quality predictions. Specifically, while only 34% of students on average are able to identify any given 'required' (highly relevant) sentence, our approach is able to achieve high performance, at times performing better than even the best students. To understand the impact on student learning, we conducted semi-structured interviews to understand the experience of students who used the peer-based AI hints. We found that our approach led students to improve their understanding of research questions and transcripts, to more carefully examine their justifications, and to improve their understanding of when they were over-annotating or under-annotating the transcript. We conclude by discussing directions for extending the work to supporting other aspects of qualitative analysis and broader implications for the design of platforms that leverage large classrooms of learners to provide apprenticeship-like support for learning. Our paper contributes:

- A system for generating peer-based hints which demonstrates the novel application of the crowdsourcing DS-EM algorithm to augmenting experiential learning of qualitative coding,
- A quantitative evaluation of the system showing that our approach is able to achieve high accuracy without requiring a large number of students or transcripts,
- A qualitative evaluation of student experiences showing that the resulting hints provides value for student learning in diverse ways,
- A discussion around AI-augmented crowd-based collaborative learning and implications for real-world learning, complex crowd work, and qualitative analysis,

2 RELATED WORK

2.1 Experiential Learning and Mentorship

Experiential learning under guidance of a mentor is a gold-standard in education. Experiential learning can be traced back to John Dewey who believed that genuine learning occurs when students actively engage in their environment [19] i.e. they learn by doing. Later, educational theorist David Kolb formalized and expanded upon these ideas, notably with his iterative Experiential Learning Cycle [31] which consists of: 1) a concrete experience, 2) reflective observation, 3) abstract conceptualization, and 4) active experimentation. This dynamic is greatly enhanced by mentors who bring their expertise to provide personalized feedback to coach and support learners in reflecting on and identifying areas for improvement. In Vygotsky's theory, this More-Knowledgeable Other (MKO) supports learners in going beyond tasks they can do with their current abilities to ones that they can do with support, a region he termed the Zone of Proximal Development (ZPD) and where he believed learning occurs most effectively [48].

This can be challenging, however, due to the limited availability of authentic work contexts for learners to practice and limited availability of mentorship from a MKO [9, 27]. The implications are significant, ranging from making work-integrated apprenticeship programs unsustainably costly to run and dependent on large government subsidies [4, 5] to limiting access to undergraduate research experiences [6, 45]. Our project explores how one might scale experiential learning and mentorship-like dynamics in the specific context of teaching and learning QA. Annota enables large numbers of students to actively practice qualitative analysis in authentic contexts in which students analyze real stakeholder interviews. At the same time, it provides a light form of mentorship via peer-based AI hints, providing a form of critique and prompted reflection that a more knowledgeable other might provide in traditional settings.

2.2 Crowdsourcing and Learning

A large literature exists around learnersourcing, "a form of crowdsourcing in which learners collectively contribute novel content for future learners while engaging in a meaningful learning experience themselves" [29]. As a form of peer learning, learnersourcing has been used to help learners better navigate [30] and interpret [49] educational materials, understand concepts through generated peer examples [39, 44] and explanations [23–25, 51], and receive individualized critique [7, 10, 32, 47]. There are many ways in which this is achieved, e.g. through 1) aggregating usage data, such as in LectureScape [30] which uses learners' watch activity on videos to surface potential points of interest, 2) supporting communities of practice where learners create, share, and remix learning content [39, 44], such as how the Scratch [44] community shares project examples with each other for learners to remix [17], 3) eliciting, aggregating, and recommending learning content [23–25], such as AXIS [51] which collects explanations from learners and then uses a multi-armed bandit algorithm to select and present the best explanations to future learners, 4) providing variations of 1-on-1 feedback [7, 10, 32, 47], such as PeerStudio [32] where learners use rubrics to provide feedback on draft writing, and finally, 5) coordinating smaller individual contributions from multiple learners in workflows to aggregate them into one cohesive output [13, 49], such as Crowdy [49] which collects subgoal labels from learners watching a how-to video and then uses voting to aggregate them into a final cohesive outline of groups of steps for the video.

We introduce a new technique in which crowdsourcing label aggregation algorithms are used (specifically, the Dawid-Skene EM algorithm [18]), not to aggregate micro-task work from individuals, but to support learners in complex work. We view a complex task (thematic analysis) as partially consisting of many smaller tasks (determining which passages are relevant to the research question). But rather than dividing these up among many individuals, we have learners work on the full complex task, and then use crowdsourcing algorithms to aggregate the work of the crowd towards providing peer-based AI hints. As will be discussed further in **Section 7**, we see this as a very promising direction for experiential learning and crowdsourcing algorithms.

2.3 Designing to Support Qualitative Analysis

Many computing systems exist for supporting QA, also known as Computer-assisted Qualitative Data Analysis Software (CAQDAS). Widely used examples include NVivo [3], MAXQDA [2], and ATLAS.ti [1] (though there are a breadth of other options fulfilling different niche needs [34, 36, 43]). Traditionally, these systems are designed to support researchers in *manually* managing and analyzing qualitative data towards making sense of the data. They do this by providing tools for tagging data with codes, capturing thoughts and insights with memos, querying data for keywords, generating visualizations of the data, and more. In regard to these aspects of supporting QA, the Annota system we build our peer-based AI hints on is similar in the tools it provides for coding and memoing.

Recent research has considered how to extend traditional systems through AI for *semi-automated* coding. The focus in these AI systems is to make the laborious process of coding large corpuses of data more efficient by intelligently applying codes to unseen data. This can be done through explicitly defining rules that match certain text [15, 22, 37, 40] or ML powered text-classification [15, 21, 28, 40, 53, 54].

For example, Cody [40] semi-automates coding through a hybrid approach involving both pattern rules and supervised ML. Pattern rules are used to quickly identify and suggest matching portions of text. User responses to these suggestions then kickstart the training of an ML model that generates further suggestions across the data corpus. Researchers found that Cody increased coding quality and user understanding of transcript data. This was followed by other systems like PaTAT which take an interactive program synthesis approach to learn pattern rules, thus lessening the need for black box models to increase the interpretability and transparency in model recommendations. These features enabled users to better learn about their datasets, discover suggestion errors, repair and modify what the model was learning, and discover new themes [22].

Our work builds on this thread of work, but has important differences. First, our setting and goal are distinct due to our focus on supporting learning. Existing systems mostly assume *expert-annotator codes are truthful*, and aim to *generalize them to unseen data*, for the purpose of *more efficient annotation*. In contrast, our system assumes *novice-annotator codes are noisy*, but that *collective expertise exists within student work* on data already seen by many annotators, for the purpose of providing *educationally valuable feedback*. Second, existing systems may rely on human codes as initial input, but then make human-independent inferences to generalize them to unseen data, making them comparatively complex to interpret for novices. In contrast, peer-based AI hints use AI to intelligently aggregate human codes for providing recommendations/hints, which means that hints are always a result of direct human interpretation, never human-independent inferences. Consequently, our hints are highly transparent: with each hint, we show the codes created by individual peer-annotators that led to the hint, conveying a human's interpretation of the data in natural language that is intuitive for even novices to interpret (as opposed to, for example, an explanation that is abstracted behind a pattern rule).

Beyond *individual* AI-assisted QA, our work also has connections to the nascent space of AI-assisted *collaborative* QA. In one of the few papers on the topic, researchers introduce CoAIcoder [21], an AI-assisted collaborative qualitative coding tool that leverages AI to enhance human-to-human collaboration by providing code suggestions based on users' coding history. Through a controlled study they compare multiple ways of using the system along three dimensions: 1) AI – whether or not the predictive model is applied to provide code suggestions, 2) Synchrony – whether or not annotators create codes at the same time, and 3) Shared Model – whether or not annotators use a shared prediction model based on all annotators' codes or just based on the individuals' codes. Our work extends this thread of work to consider large groups (or classes) of annotators, and how collaborative coding in these contexts might enable one to leverage the large size of the class to support learning.

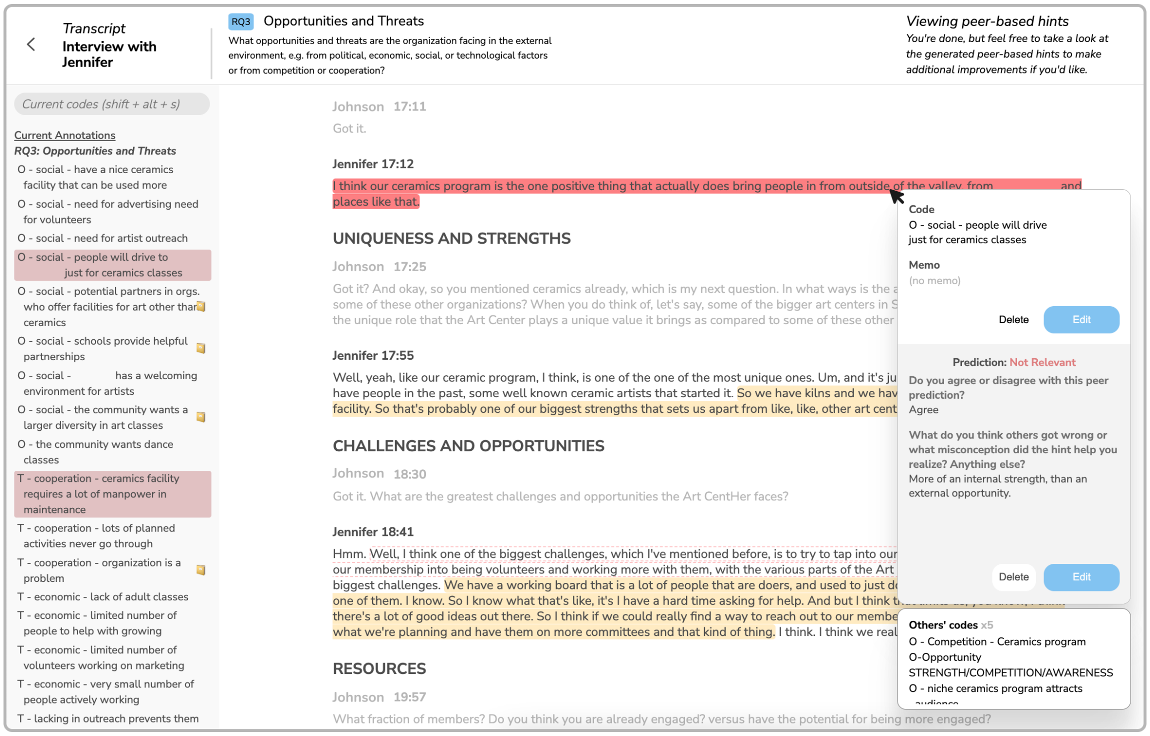
3 SYSTEM: PEER-BASED AI TIPS FOR EXPERIENTIAL LEARNING OF QUALITATIVE ANALYSIS AT SCALE

Peer-based AI tips are designed to better enable and support experiential learning of qualitative coding at scale. As described in Related Work (**Section 2**), experiential learning requires opportunities for authentic real-world learning and reflective incorporation of those experiences into schemas of ideal practice. This often involves a "more knowledgeable other" who might use various apprenticeship learning methods such as coaching, scaffolding, modeling, reflection, and articulation to support learners in successfully making real-world contributions (within their Zone of Proximal Development) and to facilitate reflection on their work and process. In the following, we describe our conceptualization and implementation of peer-based AI tips in Annota and how its design enables experiential learning of qualitative analysis at scale. As a reminder, for the scope of this paper, we are focused specifically on supporting students in one aspect of qualitative analysis as a starting point: determining what sentences are relevant to the research question.

3.1 Authentic experiential learning of qualitative analysis on Annota

We developed and evaluated peer-based AI tips within Annota, a platform for learning qualitative analysis that aims to make it possible to organize large numbers of students (e.g. in a large class) to learn in the context of doing qualitative analysis in real-world projects. For example, as will be described further in **Section 4**, our study took place in the context of organizing students to conduct and analyze interviews as part of a consulting project for a local non-profit which involved eliciting and synthesizing internal and external perspectives on various aspects of the organization.

On Annota, students are assigned some number of interview transcripts as well as some number of research questions per transcript. In the remainder of the paper, we refer to each transcript and research question pair as an RQT. For



Transcript
Interview with Jennifer

RQ3 Opportunities and Threats
What opportunities and threats are the organization facing in the external environment, e.g. from political, economic, social, or technological factors or from competition or cooperation?

Viewing peer-based hints
You're done, but feel free to take a look at the generated peer-based hints to make additional improvements if you'd like.

Current codes (shift + alt + s)

Current Annotations
RQ3: Opportunities and Threats
- O - social - have a nice ceramics facility that can be used more
- O - social - need for advertising need for volunteers
- O - social - need for artist outreach
- O - social - people will drive to just for ceramics classes
- O - social - potential partners in orgs. who offer facilities for art other than ceramics
- O - social - schools provide helpful partnerships
- O - social - has a welcoming environment for artists
- O - social - the community wants a larger diversity in art classes
- O - the community wants dance classes
- T - cooperation - ceramics facility requires a lot of manpower in maintenance
- T - cooperation - lots of planned activities never go through
- T - cooperation - organization is a problem
- T - economic - lack of adult classes
- T - economic - limited number of people to help with growing
- T - economic - limited number of volunteers working on marketing
- T - economic - very small number of people actively working
- T - lacking in outreach prevents them

Johnson 17:11
Got it.

Jennifer 17:12
I think our ceramics program is the one positive thing that actually does bring people in from outside of the valley, from and places like that.

UNIQUENESS AND STRENGTHS

Johnson 17:25
Got it? And okay, so you mentioned ceramics already, which is my next question. In what ways is the a
some of these other organizations? When you do think of, let's say, some of the bigger art centers in S
the unique role that the Art Center plays a unique value it brings as compared to some of these other

Jennifer 17:55
Well, yeah, like our ceramic program, I think, is one of the one of the most unique ones. Um, and it's ju
have people in the past, some well known ceramic artists that started it. So we have kilns and we hav
facility. So that's probably one of our biggest strengths that sets us apart from like, like, other art cent

CHALLENGES AND OPPORTUNITIES

Johnson 18:30
Got it. What are the greatest challenges and opportunities the Art CentHer faces?

Jennifer 18:41
Hmm. Well, I think one of the biggest challenges, which I've mentioned before, is to try to tap into our
our membership into being volunteers and working more with them, with the various parts of the Art
biggest challenges. We have a working board that is a lot of people that are doers, and used to just do
one of them. I know. So I know what that's like, it's I have a hard time asking for help. And but I think t
there's a lot of good ideas out there. So I think if we could really find a way to reach out to our membe
what we're planning and have them on more committees and that kind of thing. I think. I think we real

RESOURCES

Johnson 19:57
What fraction of members? Do you think you are already engaged? versus have the potential for being more engaged?

Code
O - social - people will drive just for ceramics classes

Memo
(no memo)

Delete Edit

Prediction: Not Relevant
Do you agree or disagree with this peer prediction?
Agree

What do you think others got wrong or what misconception did the hint help you realize? Anything else?
More of an internal strength, than an external opportunity.

Delete Edit

Others' codes x5
O - Competition - Ceramics program
O-Opportunity STRENGTH/COMPETITION/AWARENESS
O - niche ceramics program attracts audience

Fig. 2. A student reviewing their hints in the annotation view of Annota. Red highlights indicate annotations predicted as "Not Relevant" and yellow as "Relevant", dashed red borders indicate sentences predicted as "Missed". On the right hand side, the student has hovered over a "Not Relevant" annotation and has responded to the hint, indicating that they agree they have incorrectly annotated the sentence and explaining that they now realize it is better suited for a different research question.

each RQT, students read and code the assigned interview transcripts by highlighting passages and writing labels (i.e. codes) that denote how the coded passages relate to the given research question. They can also write analytical memos containing reflections on their codes and on potential patterns or higher-level themes they are seeing. Once they submit their annotations for all assigned RQTs, they can enter a separate view in which they can see the annotations they made across all interviews (per research question) and organize them into higher-level themes and subthemes.

3.2 Coaching, scaffolding, and modeling through peer-based AI tips

After each student submits their initial annotations for an RQT, a Firebase cloud function is triggered that checks how many students have submitted annotations so far, and after a certain threshold is reached (e.g. 20 students), runs an algorithm to generate predictions of relevant sentences based on all submitted annotations (details in **Section 3.4**). These predictions are then used to generate hints directing students to re-examine: (1) annotations they made that are predicted to be not relevant or (2) passages they did not annotate that are predicted to be relevant.

Specifically, once hints have been generated, students see an updated ‘Hints Generated!’ status icon for the given RQT in the Annota dashboard. Upon re-entering the annotation view for a given RQT, students are greeted with a summary of the personalized hints they will be provided for that RQT, i.e. the number of annotations predicted as not relevant and number of not-annotated passages predicted as relevant. Hints are presented not as definitely correct, but as derived from peer submissions that may indicate that either the student got something wrong or that they got something right that many others got wrong. Once they click past the summary page, students see the transcript enhanced with the peer-based AI tips. Specifically, unannotated passages that the hints predict as relevant - and thus 'missed' - are outlined in red, while annotations that are predicted to be irrelevant are shaded in red (see **Figure 2**). Annotations predicted to be correct maintain their original yellow shading. When students hover over a given annotated or recommended passage, they are shown the codes that other students made for the given passage so that they can see how their peers described the annotated passage and its relevance to the research question[1]. This can help with student sensemaking when evaluating whether to accept or reject the peer-based AI tip.

Peer-based AI tips support experiential learning by providing a light form of coaching and scaffolding, pointing students potential errors or areas of improvement, and by providing a light form of modeling, showing students work of other peers to compare with. They can help to surface misunderstandings of the research question or transcript and misconceptions around the detail-oriented nature of annotation (see **Section 6**).

3.3 Reflection and articulation through interactions with peer-based AI tips

When hovering over hints, users are prompted to either agree or disagree with the hints. If they agreed with the hint, they are asked to describe what misconceptions the hint helped them to realize. If they believe the hint to be wrong, they are asked to describe what they believe other students got wrong. In this way, hints not only point students to potential errors, but also provide a context for students to reflect on their work and to articulate their reasons and conclusions. Depending on the instructor, students could be required to respond to these prompts as part of the assignment or they can be optional. However, in either case, students need to describe their rationale in order to update a given annotation to match hint suggestions.

3.4 Surfacing latent collective student expertise as the "more knowledgeable other"

In general, for any given relevant sentence, far less than 50% of students identify it in their annotations (35.8%), which means that naive approaches to aggregating student annotations will perform poorly. The core idea in our implementation of peer-based AI hints is to connect the qualitative analysis process to a crowdsourced label aggregation problem, allowing us to use Dawid and Skene's EM algorithm as a way to surface latent collective student expertise as a "more knowledgeable other" that can be used to predict relevant passages and support student work.

In the crowdsourcing label aggregation problem, one has a set of workers and tasks. Workers are assigned to a subset of the tasks, each requiring them to determine a label for that task. Based on the labels provided by each worker for their assigned tasks, one would like to then determine the most likely true labels for each task. Dawid and Skene's expectation-maximization (DS-EM) algorithm considers an underlying model in which each worker has a confusion matrix specifying how likely they are to label a task of a particular class with any of the possible classes and in which there is a prior probability that any given task belongs to one of the possible classes. The DS-EM algorithm then seeks to jointly calculate these parameters along with predictions of labels for each task through an iterative process.

This can be used within the qualitative learning process by considering each sentence in the transcript as a crowdsourcing task to be labeled as relevant or not relevant to the research question (a binary classification problem) and each student as a crowd worker. Task assignments correspond directly to the RQTs assigned to each student, with worker labels for a given task corresponding to the presence or absence of an annotation by that student overlapping

[1] During the time of the study, this feature was only available for annotations they made that were predicted to be not relevant.

with the corresponding sentence (i.e. an indication that the student believes there is something in this sentence relevant to the RQ). DS-EM can be directly applied to the dataset to produce predictions per sentence.

We then use the resulting predictions conservatively to generate peer-based AI tips. Specifically, our tips point to an annotation as potentially incorrect if it does not overlap with any sentence predicted to be relevant by DS-EM. Similarly, we determine missed passages by identifying continuous sets of sentences all predicted as relevant (a sentence chunk) and noting it as something the student missed if they did not make any annotation overlapping with that sentence chunk. We use this conservative approach to generate hints because there are equally legitimate choices students could make regarding how much of the background/context sentences they want to include in their annotation when identifying something relevant to the research question.

Understanding the intuitions underlying DS-EM helps to understand why the hints produced can be thought of as playing the role of a "more knowledgeable other" in supporting a peer-based form of apprenticeship learning. In our case, each task has two potential labels: relevant ("1") and not relevant ("0"). The confusion matrix for a worker u is a 2x2 matrix π_u with $\pi_u[x, y]$ denoting the probability the worker will produce a y label when the true label is x. In each round of iterative parameter/prediction updates, after worker confusion matrices are updated based on a maximum likelihood estimation of parameter values, new label predictions are generated by computing the probability of the task being a "0" or "1" and choosing the class that maximizes this probability. This turns out to be equivalent to a weighted majority vote where "1" votes from student u are weighted $\log \frac{\pi_u[1,1]}{\pi_u[0,1]}$ and "0" votes from student u are weighted $\log \frac{\pi_u[0,0]}{\pi_u[1,0]}$. In other words, the weight of a student's vote is based on the specificity and sensitivity of their labels as defined in positive/negative likelihood ratios. A student annotation of a sentence as "relevant" is only a positive vote if the fraction of times they label a truly relevant sentence as relevant is greater than the fraction of times they label an irrelevant sentence as relevant. The extent to which they are better at identifying a relevant sentence when it is truly relevant (in comparison to when they mark an irrelevant sentence as relevant), the more their vote counts. From an apprenticeship learning perspective, the EM-algorithm is identifying the strengths of each student and using these to aggregate predictions for each sentence across the entire class. By doing so, it turns the large size of a class into an advantage rather than a disadvantage for providing a peer-based version of a "more knowledgeable other".

4 METHODS

4.1 Approach and Context for Experiential Learning of Thematic Analysis

In order to study the effectiveness and impact of peer-based AI hints, we ran an in-the-wild deployment of Annota in a large business strategy class context ($N = 122$) in which students learned qualitative analysis while analyzing real interviews that were part of a larger consulting project.

This paper centers on student work in the first two weeks of the class when they are analyzing interview transcripts to understand the organization's values and carrying out a SWOT analysis, i.e. mapping the organization's internal strengths and weaknesses, and external opportunities and threats. Each student was assigned to analyze 8 transcripts over a two-week period (4 transcripts per week), and was assigned *one* of the following three research questions (roughly 40 students per question):

- *Values:* What are the core values or qualities of the organization that members want to achieve or maintain regardless of the strategy?

- *Strengths and Weaknesses*: What are the organization's unique resources, unfair advantages/strengths, and ways in which its products, services, and brand are differentiated? What are the weaknesses of the organization? (S&W)
- *Opportunities and Threats:* What opportunities and threats are the organization facing in the external environment, e.g. from political, economic, social, or technological factors or from competition or cooperation? (O&T)

Deploying Annota in this setting allowed us to collect a dataset of authentic student annotations produced in a real experiential learning context, where most students are learning thematic analysis for the first time while contributing to a real-world project.

We then ran the EM algorithm on this dataset to generate hints, which we evaluated in two ways. First, we ran simulations to assess the quality and convergence properties of peer-based AI tips. Second, students could do an extra credit assignment in which they could read, reflect on, and respond to each generated peer-based AI tip on Annota followed by filling out a survey and taking part in a 20-30 minute semi-structured interview. We then analyzed the survey and interview responses to understand the different ways in which peer-based AI tips supported student learning. In what follows, we describe each of these in more detail.

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

Our first question centers on understanding the quality of the EM-based predictions and how many students or transcripts are needed to obtain high-quality results.

4.2.1 Expert labels and subjectivity. At first glance, it is not clear how this can be done since qualitative analysis is inherently subjective in nature. Some people may view a passage to be relevant while others may view it to be not relevant due to different contexts and worldviews from which they are interpreting the passage. A particular statement may be evidence for an organizational strength only when viewed from a certain light or in connection to other information. At the same time, for simpler research questions, these are more the exceptions rather than the norm, with the vast majority of passages being easy to agree on with discussion. Thus, for our evaluation, we used the following two simpler research questions: the internal strengths and weaknesses and the external opportunities and threats. We defined expert "ground truth" labels by having the two first authors code what sentences were relevant to the research questions. For the first two transcripts, they coded and discussed their annotations together to ensure they had a common understanding of what was relevant or not relevant, and then split up the remaining transcripts, raising and discussing any cases they encountered that they were unsure of. One of the first authors was a TA for the course over multiple years and has published qualitative research papers, so has deep understanding of both the course context and the qualitative research process. Thus, using the labels they produced as our ground truth for quality and convergence evaluation is equivalent to assessing the ability of DS-EM to match the performance of feedback from course staff with the time to deeply evaluate and discuss all annotations. We discuss the issue of subjectivity for harder research questions in the discussion section.

4.2.2 Assessing error with three expert labels: required, relevant, and not relevant. A second challenge is that even if there is agreement that a particular passage is relevant to the research question, different people might choose to highlight different amounts of the passage. Beyond the central parts of the passage, people may legitimately choose to include more or less of the bordering context in their annotation. To account for this in our assessment of quality, our expert labels classified each sentence into one of three options:

- *Required*: these are sentences that must be included in annotations since they are clearly relevant,
- *Relevant*: these are sentences that could be annotated, e.g. due to providing context for a required sentence or due to a potential but more tenuous connection requiring reading between the lines,
- *Not relevant*: these are sentences that should not be annotated,

Of the 1700 sentences present in the 8 graded transcripts, we found 842 not relevant, 660 relevant, and 198 required sentences for the 'Strengths and Weaknesses' research question. We found 1253 not relevant, 288 relevant, and 159 required sentences for the 'Opportunities and Threats' research question.

We note that the DS-EM algorithm still produces a binary prediction for each sentence just like any student making annotations. The three cases (required, relevant, or not relevant) are only used in our assessment of the predictions to make sure that we are assessing accuracy in ways that better match the reality of the qualitative analysis process.

4.2.3 Metrics: positive and negative precision and recall, and macro-F1. Since the distribution of labels can be imbalanced (e.g. most sentences are not relevant to Opportunities and Threats), we compute the *Macro-F1* score as our quality metric, which takes the unweighted mean of class-wise F1 scores calculated based on the below definitions of Positive and Negative Precision and Recall. Here, 'Positive' predictions refer to predictions of when something *should* be annotated and 'Negative' predictions refer to predictions of when something *should not* be annotated:

- *Precision*: the fraction of sentences predicted by DS-EM to be relevant that were marked as relevant or required in the expert labels. This gives us a sense of how often positive predictions are correct,
- *Negative Precision*: the fraction of sentences predicted by DS-EM to be not relevant that were marked as not relevant or relevant in the expert labels (since it isn't necessary to highlight a relevant but not required sentence). This gives us a sense of how often negative predictions are correct,
- *Recall*: the fraction of sentences marked as required in the expert labels that were predicted by DS-EM to be relevant. This gives us a sense of how well positive predictions cover the required sentences.
- *Negative Recall*: the fraction of sentences marked as not relevant in the expert labels that were predicted by DS-EM to be not relevant. This gives us a sense of how well negative predictions cover the not relevant sentences.

4.2.4 Evaluation comparisons. To assess the quality of our predictions for students, we compare our predictions against the average score of all student annotators and the score of the very best annotator, defined to be the annotator that maximizes macro-F1. To assess the quality of our predictions compared to a naive algorithm, we also compare our predictions against the prediction produced by a naive majority vote aggregation of submitted annotations, in which a sentence is predicted as relevant if a majority of annotators mark it as relevant. In later figures, we refer to these as 'Student Average', 'Best Student', and 'Majority Vote' respectively, and refer to our own predictions as 'Annota'. All baselines (i.e. Student Average, Best Student and Majority Vote) in **Section 5.1** and all algorithms used in our performance simulations and evaluations only use annotations made by students *before* seeing any peer-based hints.

4.3 Collecting and Analyzing Qualitative Data to Understand the Value for Student Learning

Our second question centers on understanding the value of peer-based AI hints for student learning.

4.3.1 Study context and participants. To understand this, we used the generated predictions to deploy peer-based AI hints on the Annota platform[2] and then conducted a qualitative study to assess student experiences. Specifically, students could do an extra credit assignment[3] towards the end of the class in which they could read, reflect on, and respond to each generated peer-based AI hint on Annota followed by filling out a survey and taking part in a 20-30 minute semi-structured interview. As described earlier in the Systems section, responding to each hint involved agreeing or disagreeing with it and articulating their reason. A total of 14 students filled out the survey (7 male, 7 female), and 10 continued on to the interview (7 Seniors, 3 Juniors; 3 Males, 6 Females and 1 Undisclosed; 7 studying "Technology and Information Management", and one each studying Cognitive Science, Business Management and Economics, and Global Economics), where 2 were assigned to the Strengths and Weaknesses research question, 6 to Opportunities and Threats, and 2 to Values. Survey responses and hint responses were reviewed prior to each interview to inform follow up questions. However, our qualitative analysis only centered on the survey and interview responses (**Section 6**). We also note that there was a bug in our code which in rare occasions showed annotations as wrong even when they were predicted to be right. However, this does not affect any of the simulations described beforehand nor does it affect any of the qualitative analysis described later because our analysis centers on understanding experiences in which students derived value from the hints for learning.

4.3.2 Interviews and qualitative analysis. Our semi-structured interviews sought to understand their experience of the peer-based AI hints, as well as their experience using non-hint related features of Annota and their experience in the course overall (both not within the scope of this paper). In the portion dedicated to the AI hints, we had students log into Annota and share their screen if they were able. They then walked through their process of using Annota from their initial analysis to their use of the peer-based hints, followed by questions aimed at eliciting concrete stories rather than vague recollections. Questions related to the peer-based hints included, for example, *"Did you find the feedback useful? Why or why not?"* and *"Did this affect your learning of qualitative analysis? If so, in what ways?"*. Our interviews were recorded, transcribed, and analyzed through an inductive thematic analysis process.

We started by having the first two authors open-code the transcripts and discuss discrepancies and potential themes through an affinity diagramming process. These were discussed together with the last author to finalize the themes after which the first author conducted a final round of annotation.

4.4 Limitations

We acknowledge several limitations of our study. First, our findings that resulted from the analysis of our surveys and interviews are inherently limited by their qualitative nature. Therefore, further studies are needed to quantitatively measure the effects of peer-based hints on learning outcomes. Second, we note that a limitation of our peer-based hints system is that it supports students in just one aspect of qualitative analysis: identifying which sentences are relevant to the given research question or not. Further work needs to be done to extend these ideas to creating themes and the other more complex interpretive aspects of qualitative analysis. The class involved students in an inductive process that drew on elements of grounded theory [11], with students engaging in techniques such as open coding, codeweaving, memoing [42], and discussion in small student teams to reflect on, challenge, and iteratively build their understanding of the data. However, because the research question we gave them was fixed, it did not capture a fully inductive process

[2] Note that while our study of peer-based AI hints only focused on the research questions described above, there were additional research questions in subsequent homework assignments, and the live EM algorithm is run on the full set of tasks in the database. Simulations for evaluating convergence properties do not include these other research questions.

[3] We had initially intended to deploy the hints to all students, but the course assignments did not have space for students to iterate on their annotations so there was no reason for students to use the hints. We also discovered a bug partway through the deployment that was fixed by the extra credit assignment.

in which the research question itself may evolve in the analysis process. In such a context, the peer-based AI hints we defined would come into play in later rounds of analysis when discussing and converging to a common set of themes.

5 QUALITY AND CONVERGENCE RESULTS

5.1 Prediction quality is comparatively high

We find that Annota predictions are similar to the best students and far better than the average student or majority vote in macro-F1 scores, with quality gains primarily coming from strong positive recall and negative precision.

5.1.1 Annota hints do better than the average student and similar to the best student. As can be seen in **Figure 3, row 1**, Annota predictions do far better the average student, with a macro-F1 score of 0.88 and 0.72 for the Strengths and Weaknesses (S&W) and the Opportunities and Threats (O&T) research questions respectively, which is a 20% increase or more on the macro-F1 scores for the average student (0.68 and 0.60 respectively). Annota predictions are roughly comparable to the best student for each given RQ or RQT (a little better in the case of Strengths and Weaknesses and a little worse in the case of Opportunities and Threats). In other words, Annota predictions are succeeding at eliciting the collective expertise within the student crowd to provide a peer-based version of a More Knowledgeable Other for hints about sentence relevancy. They are comparable to sourcing feedback from the best student in the class *without needing to be able to identify them beforehand*. The same overall pattern holds when examining each of the individual RQTs.

5.1.2 Annota hints are much better than majority vote, which is worse than the average student. We note that this does not happen with a naive approach such as a majority vote aggregation. In fact, the majority vote macro-F1 scores are only 0.63 and 0.49, respectively, which is actually worse than the average student (and much worse than Annota predictions). Note that a macro-F1 of 0.49 is essentially the same as random guessing.

5.1.3 Quality gains are due to positive recall and negative precision. These gains can largely be traced to a vastly improved ability to identify relevant sentences (positive recall) and to more accurate 'not relevant' predictions (negative precision), as can be seen in **Figure 3, rows 2-3**. Indeed, Annota predictions have a positive recall score of 0.83 and 0.76 and a negative precision score of 0.97 and 0.96, for the S&W and O&T research questions respectively, outperforming even the best student. This is a good fit for the learning needs of students. It is good to have a better coverage of 'missed passages' to point out things students may not have been attentive to while making sure that 'wrong annotation' hints are more precise to avoid losing credibility when contradicting a students' conscious choice.

In contrast, Majority Vote produces extremely sparse predictions, even producing zero percent recall on three RQTs and only an average of 6.5 relevant sentences per RQT. If one were to use Majority Vote to provide peer-based hints, it would suggest very few missed passages (if ever), and would confuse students by numerous suggestions that almost all of the annotations they made are incorrect.

5.2 Quality can be achieved with reasonable numbers of students and transcripts

We also find that high quality predictions can be achieved without needing very large numbers of students or transcripts. Only ∼15 students and ∼2 transcripts are sufficient, making this approach feasible even in smaller classes. If one is able to leverage previous work (e.g. for the second homework or beyond), high-quality results can be obtained even faster.

5.2.1 Good feedback only requires ∼15 *annotators.* To examine the relationship between the number of annotators who have submitted work and the quality of generated predictions, we simulate the performance for k annotators by

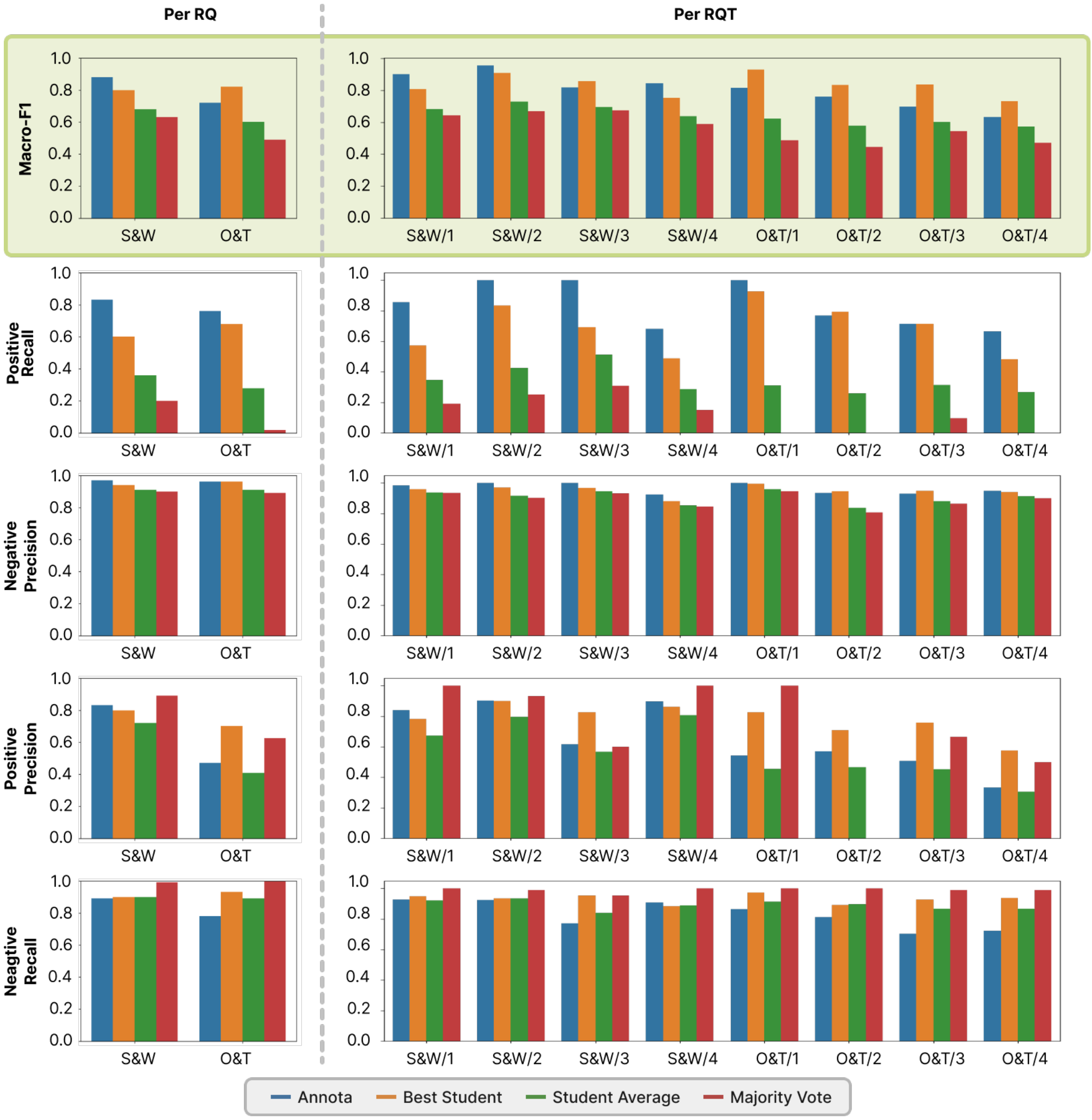


Fig. 3. Comparison of Macro-F1 scores and Positive/Negative Precision/Recall on each of the research questions and RQTs between Annota predictions, the best student, the student average, and majority vote predictions, where the best student is chosen separately for each RQ or RQT. As can be clearly seen, Annota predictions do far better than the student average and majority vote and are comparable to the best student.

sampling 100 sets of k annotators from our real class dataset, applying the Dawid-Skene algorithm to their labels (using their work from the four transcripts in their first homework assignment), and evaluating the performance against our expert labels. We do this separately for each research question. For comparison, we also plot the average student score and the score of the best student out of the sampled k annotators.

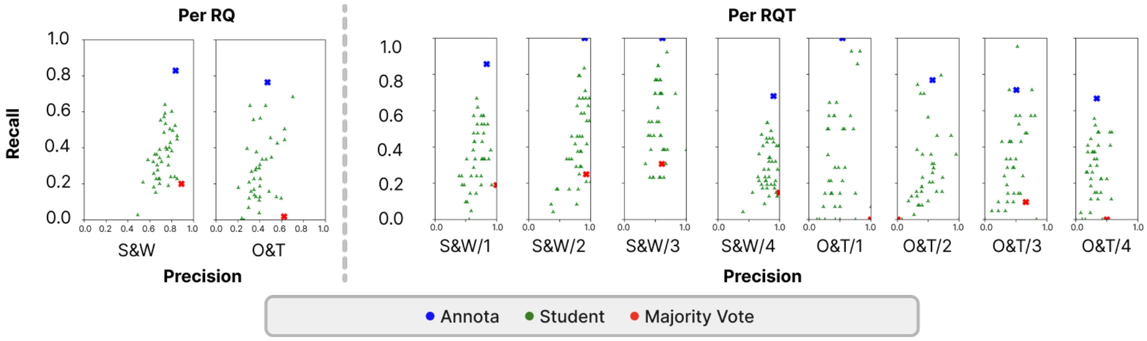


Fig. 4. Comparison of Annota, individual students, and majority vote by Precision-Recall balance. While student annotators are very sparsely distributed and have high variance in quality, Annota robustly outperforms them and majority vote, both on precision and recall.

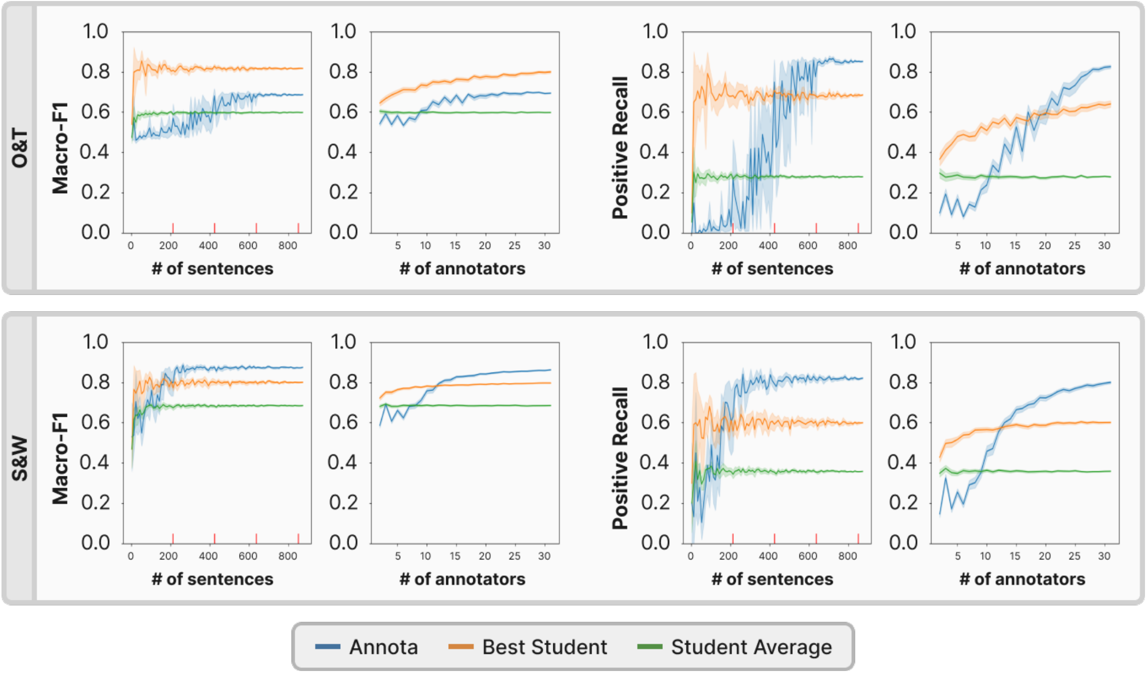


Fig. 5. The above graphs show how quality varies with the number of annotators or sentences. Our graphs varying annotators assume 4 transcripts (882 sentences), which is the number of transcripts in one homework assignment. Our graphs varying sentences use all of the ~40 annotators available for each RQ. The value for k annotators is computed by averaging 100 random samples of k annotators from the set of all students. Similarly, the value for k sentences is computed by averaging 100 random samples of k sentences from the set of all sentences. As can be seen, roughly 15 annotators is enough and roughly 2 transcripts is enough (with the average transcript being roughly 200 sentences). Note that red ticks on sentences vs metrics graphs indicate the passing of the average length of a transcript.

As can be seen in **Figure 5, top-left**, a sharp jump up in the quality (macro-F1) occurs between 8-15 annotators. By 15 annotators, most of the quality gains have been achieved, with quality far exceeding the average student and quality roughly comparable to the best student. The implication is that even small classes can use this technique.

5.2.2 Good feedback only requires assigning ∼2 transcripts. We use a similar approach to examine the relationship between the number of sentences and the quality of generated predictions. We plot using step sizes of 10, i.e. 10 sentences, 20 sentences, etc. For each k, we find the average value from randomly sampling 10 sets of k sentences and applying our algorithm to these labels. We use the entire set of ∼40 annotators per research question. As can be seen, a sharp jump up in the quality (macro-F1) occurs between roughly 200-400 sentences (roughly 1-2 transcripts for 30-minute interviews). By 2 transcripts, most of the quality gains have been achieved, with quality far exceeding the average student and quality roughly comparable to the best student. The implication is that a single homework assignment with students annotating 2 transcripts is enough (we required 4 transcripts in the first assignment).

5.2.3 Leveraging previous work allows good feedback with even fewer annotators. The previous graphs assume that students are annotating for the very first time. But if students are annotating new transcripts for a second homework assignment, the algorithm can leverage the work from their past annotations to enable even faster convergence in its predictions for new transcripts. The reason is because their past work allows the algorithm to more quickly and accurately estimate student confusion matrices.

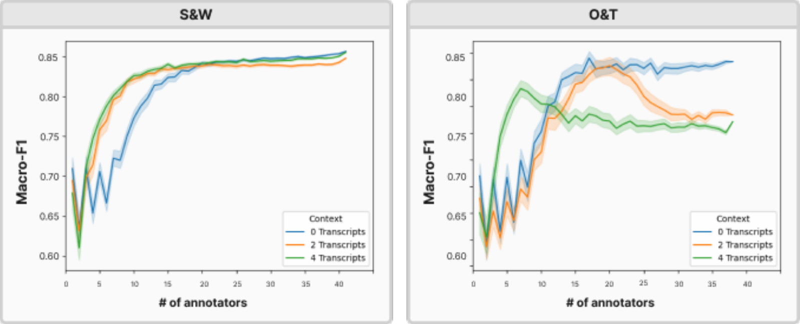


Fig. 6. Leveraging previous work enables one to get high-quality predictions with even fewer annotators.

As can be seen in **Figure 6**, once students have completed 4 transcripts, significant quality gains can already be achieved after just 5 annotators have submitted work on the next transcript.

5.3 Student confusion matrices are very accurate, even after only ∼3 transcripts

Thus far, our focus has been on the accuracy of the produced labels for sentence relevancy. We briefly note that student confusion matrices produced by the Dawid-Skene EM algorithm are also very accurate. Specifically, when computing the average elementwise mean-squared error (MSE) of the computed confusion matrices as compared to the 'ground truth' confusion matrices, we converge to MSE values of less than 0.01 after only ∼3 transcripts (500 - 700 sentences), as shown in **Figure 7**. This lends further evidence that the algorithm is able to identify students more likely to be accurate and appropriately weight their work when computing predictions. These confusion matrices could potentially also be used to highlight work of students who are predicted to be performing better than others.

6 VALUE FOR STUDENT LEARNING

Our analysis revealed three themes regarding the value that peer-based AI hints provided for helping students learn qualitative analysis. Namely, the hints helped students 1) improve their understanding of research questions, 2) more

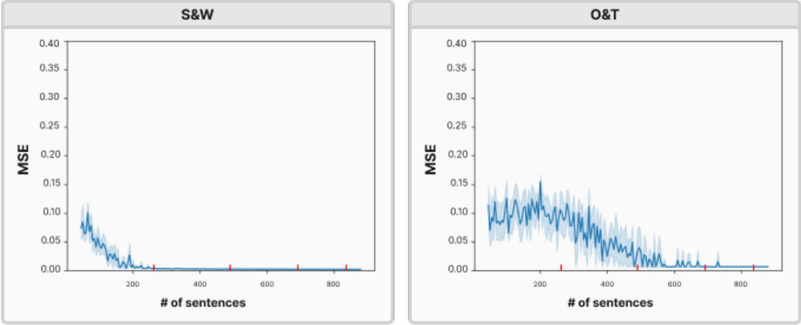


Fig. 7. Inferred student confusion matrices are highly accurate – less than 0.01 average mean-squared error – within only ~3

carefully examine their transcript annotations, and 3) improve their understanding of when they were over-annotating or under-annotating the transcript. In some sense, hints were facilitating the development of a common understanding of research questions, transcript interpretation, and qualitative coding practice that traditionally might be developed in discussions between independent annotators. While students don't always have the ability to engage in such discussions with a more knowledgeable other, hints provide a light form of this.

6.1 Understanding of Research Questions

We found that peer-based AI hints revealed misunderstandings that students had regarding their assigned research questions. For example, Hugo mixed up internal strengths and weaknesses with external opportunities and threats, describing *"that [the hints] double-checked my work"*, helping him identify *"quite a bit of annotations that were more fitting for the other research questions"*. Rafael experienced the same confusion, elaborating:

> *"Yes, I agree with this [hint]... COVID is not a weakness for the [organization], but like, COVID caused the other financial [issues] that caused the weakness." (Rafael)*

In other cases students described the hints giving them a better understanding of the research question intent that had been discussed in class, but was less clear from the question phrasing.

> *"I realized... there's like levels and layers to the values of the organization. And I think I was just a few layers down too much. Yeah. I think I read into it a little bit too much... [the hints were] helping me realize like, what was important what was not, what page everyone is on, what level of values that we're actually looking for made it like a lot more helpful for me" (Isabelle)*

6.2 Reflective Examination of Transcript Annotations

The second way that hints helped students was by prompting deeper reflective examination of their interpretations and annotation decisions, an integral component of experiential learning. For some students, this led to new interpretations of the transcript data and to realizations *"that there are other ways of thinking" (Ella)*. Isabelle described how it *"helped me develop a concrete idea of the company through the perspective of my peers" (Isabelle)* while Justin mentioned that the hints *"show other team members' annotations too which help me better understand the sentences" (Justin)*. Rafael detailed how the hints expanded his view:

> *“The AI hints helped bring to light various perspectives that I might not have considered otherwise... Initially, I thought the main focus was the diversity in the classes they offered. But AI’s insight brought me to understand that the essence of the speech was more about creating a nurturing environment that values each individual’s unique interests and perspectives... The AI pointed out this deeper understanding.” (Rafael)*

In other cases, reflection from hints led to students further solidifying their perspectives. For example, Amber talked about initially coding something as important, then reconsidering her annotation after being prompted by the hints, but ultimately maintaining her perspective:

> *“...it would help me think a little bit more about why I did find that important... [after I thought] about [it] a little more... the second time around... I’m like, well, actually, no, I do find that part important” (Amber)*

Rosa described the ability to see peers’ codes as facilitating the reinforcement of her opinions:

> *“I liked how the hints show me what my peers have also done. I work better in a team environment and it helped me to reinforce my opinions.” (Rosa)*

6.3 Mental Models of Coding Technique

Finally, the hints helped impart mental models of coding practice, leading to realizations of over-annotating or under-annotating the transcripts. Some talked about how seeing all the sentences they missed motivated them to be more detailed or pointed out quotes they had missed:

> *“They were actually helpful because I didn’t expect the AI to generate, like actual like, like helpful hints. But the Annota AI actually generates some pretty useful hints that like, where they find like, quotes that I didn’t even see... It actually got me to, like, redo the rest of the next interviews, like, more in depth, and like, more like detail. So I wouldn’t miss like any more.” (Jackson)*

> *“I was like, oh, I did pick up on the pattern. But there were so many better details or evidence out there.” (Morgan)*

Isabelle talked about the reverse: how she was initially surprised by having many annotations marked as not relevant, but after reviewing them, realized she had over-annotated:

> *“I feel [the hints] helped me like narrow down, the more important, like [codes]... I had a lot [of annotations] that were [marked as] irrelevant. I think [that] was a bit surprising at first. But like, once I went back, I realized like, “oh, this is what everyone thinks is important.” And like, by going by that, I realized that yeah, I did over annotate it.” (Isabelle)*

7 DISCUSSION: DESIGN IMPLICATIONS AND FUTURE DIRECTIONS

Our results show that peer-based AI hints produce predictions comparable to the best students in quality, that these results can be achieved with reasonable numbers of students and transcripts, and that they helped students in ways similar to discussions with a more knowledgeable other (by enhancing their understanding of research questions, facilitating reflection on transcript annotations, and refine their mental models of coding practice). In this section, we discuss broader implications and future directions.

7.1 Crowd-AI collaborative learning and doing-by-learning

Our paper demonstrated two central ideas: (1) that the size of large classes can be an advantage rather than a disadvantage for supporting experiential learning, and (2) that crowdsourcing label aggregation algorithms can be used to harness the collective expertise of student cohorts as peer-based AI hints to support individual learners. Both point to ways in which AI-augmented crowd-based collaborative learning might make it possible for learners to make high-quality real-world contributions as they learn and to obtain the authentic learning experiences often limited to a select few, e.g. in internships requiring participants with prior experience.

We showed one way in which this could take place: through the direct use of crowdsourced label predictions as hints. However, we also see other ways in which this might play out. For example, besides just using the label predictions, one might also use the confusion matrices computed for each student to identify students who need more tailored support from teaching assistants, to suggest high-performing students who might play the role of peer mentors, or to form pairs/groups of students that might have complementary strengths. Researchers have begun to explore the use of AI to support collaborative interpretation in pairs of coders [21]. Peer-based AI hints can be used in a similar way to facilitate discussions at a class-wide level, by prompting students to explain the places where there are differences between their own work and those of their peers.

Prior literature on scaling experiential learning described the importance of making it possible for learners to contribute to real-world work while learning, what they called 'doing-by-learning' [33]. They said, *"Scaling opportunities for learning-by-doing requires finding ways to enable doing-by-learning. If it is possible for learners to make contributions to real-world needs, then people will be incentivized to create opportunities for them to engage. If learners are unable to, then the time it takes to prepare authentic experiences and to mentor learners will continue to be an obstacle... In the whitewater world of automation, many have said that work and learning need to be deeply integrated. Doing-by-learning is a helpful conceptual tool for evaluating our progress towards this goal."* We see our work as contributing to this literature by suggesting a direction for enabling learners to produce high-quality qualitative analysis while learning.

We also see our work as providing a new angle for crowdsourcing algorithms. AI algorithms in crowdsourcing have traditionally centered on improving the quality of aggregated work outcomes for requesters or project clients, but at the cost of reducing work to monotonous micro-tasks that don't provide learning for workers. Our paper shows how crowdsourcing algorithms can be used to augment project-based learning in large classrooms, giving students the scaffolding they need to contribute to high-quality real-world project contributions as they learn.

It may also be possible to explore these benefits in supporting learning in complex crowd work. Instead of taking the traditional crowdsourcing approach of disaggregating a complex task (such as thematic analysis) into smaller micro-tasks (such as determining annotation decisions for individual passages in a transcript) and assigning these micro-tasks to different workers, one might instead assign individual workers to work on the entire complex task, but use the micro-task decomposition to apply crowdsourcing algorithms that can assess and give feedback on their work.

While it may not be practical for a requester to pay a large number of workers to annotate the same transcript, it may be possible for crowd work platforms to, for example, harness the many authentic tasks that exist on their platforms to create uniquely experiential learning tutorials that many workers go through. Then once workers have annotated a few transcripts in the tutorial experience, one can leverage peer-based AI hints for real tasks with a significantly smaller number of annotators (see **Section 5.2.3**). Additionally, given recent research showing that LLMs can simulate some (but not all) of crowdworkers' abilities in crowdsourcing algorithms [52], one can also imagine using multiple AI agents along with an individual worker to form the 'crowd' through which algorithms can be applied. This would facilitate a form of human-AI collaboration that would support humans in learning while working. Finally, if this were combined with government programs for subsidizing apprenticeship learning, one can begin to dream about contexts in which workers can be paid to learn as they transition into new occupations.

7.2 Emotions, perceptions, and presentation of peer-based AI hints for learning

One direction that deserves further discussion and work is to more deeply understand the emotional experience of receiving peer-based AI hints and how to best present them to support learning. While we did find peer-based AI hints to be helpful in many ways, our performance graphs also show that they are imperfect, which means that they can sometimes provide students with 'hints' that are actually incorrect. This could explain some of the negative reactions students described such as being irritated, feeling threatened, and having self-doubt e.g. *"It... irked me sometimes that it would tell me I'm wrong. I know I'm right."* (Morgan), *"I did find it a little threatening... it's in big bold letters: 'Not relevant'."* (Amber), *"I felt kind of like confused. I was questioning myself a lot more."* (Rosa).

We also saw indications of algorithmic aversion [20], with some students perceiving the peer-based AI hints as limited and untrustworthy due to it being an AI that *"doesn't understand human experience"* (Amber) and that *"there's still 20% that only people can understand because we are human"* (Rafael). One student even asked, *"Why is this robot telling me [how] to do my job?"* (Morgan). A separate student understood that the AI hints were peer-based, but still *"[felt] like a lot of the times [she] didn't trust it."* (Rosa) because she thought that her peers were (e.g. *"not... fully reading the statement"*, *"didn't really care about the assignment"*), and not *"doing the same amount of work as [she] was"*.

It would be interesting to explore algorithmic aversion and automation bias in more depth in project-based learning contexts. For example, how much error can students tolerate before trust in the hints as a 'more knowledgeable other' is undermined? Can being transparent with students on precision and recall statistics from the past help them to better calibrate what to expect from the peer-based AI hints (e.g. so that they know that 30% of the hints it presents will be wrong, but that it will help them to identify 80% of the things they missed)? Can using peer-based AI hints to facilitate discussion with other peers alleviate some of the aversion students have to the hints?

7.3 Integrating more traditional approaches to teach qualitative analysis at scale

In this paper, we explored one approach for teaching qualitative analysis at scale through AI algorithms that aggregate student work to provide peer-based AI hints. One important direction is to explore how this approach might blend with more traditional approaches such as: 1) having students work and discuss codes and themes in small groups, or 2) having instructors create their own version for students to learn from. In the first case, students can benefit from more interactive discussions, but are unable to draw from potentially more experienced students outside their group. In the second case, students benefit from expert feedback, but it imposes significant costs on the instructor and may not be as effective for eliciting reflection, e.g. through seeing peers' mistakes or needing to argue for one's own ideas. Compare these with our approach, in which students are able to learn from the entire class through hints that perform comparably to the best student, and all without imposing extra time on the instructor. The downside is that our approach also lacks the interactive discussions of a group or the expert work of an instructor.

We believe that these approaches can be blended together in many ways. For example, instead of responding to peer-based AI hints individually, instructors could ask students to review them in the context of a group discussion. This could also be baked into the design of a platform, e.g. by having AI hints center on a shared group view rather than an individual one, with hints showing the number of people in the class agreeing with either side and facilitating a discussion rather than a simple agree/disagree decision. There could also be a staff/instructor UI that makes it easy for the instructor to comment on the synthesized class-wide work and predictions, either in a live classroom setting or through the platform.

7.4 Subjectivity and learning qualitative analysis

One important issue to discuss is the tension between convergent ML techniques and the subjectivity of qualitative research. As one HCI faculty commented: *"Wouldn't it be good to have variance in coding that helps to identify themes? Isn't this realistically part of how qualitative research is conducted, with iterative sensemaking? Epistemologically I feel this mashup of convergence and qualitative thinking and iteration is a little shaky."* Other research on applying machine learning to qualitative analysis has also discussed these concerns: *"There is also hardly the notion of "ground truth" in qualitative coding, as the ultimate goal is not to build a machine-learnable model of the data, but to discover themes of interest that help answer specific research questions."* [12, 22].

We agree with these sentiments, but we believe that machine learning can still be helpful for qualitative analysis *when used appropriately.* First, it's important to reiterate the limited scope of this study which only centers on helping students determine what is relevant or not relevant for a given (fixed) research question (see Limitations in **Section 4.4**). In a truly inductive thematic analysis process (in which people may be iteratively evolving the research questions themselves), our use of peer-based AI hints could come into play in the later stages of facilitating convergence, e.g. when researchers are engaging in discussions towards consensus or doing IRR calculations.

In either case, it's clear that there are many cases when students are misunderstanding the research question, misinterpreting the transcript data, or operating under misconceptions of coding practice. Our results show that our peer-based AI hints are able to surface these issues to students, and that students benefit from and appreciate it.

But even when there are different valid interpretations, the approach can still be helpful for facilitating discussion. In our deployment, we had students code independently first (to support divergence and variance in coding to identify themes) and only revealed hints to them after they submitted their individual work. Hints are not presented as definitive "ground truth" answers, but as summarizing "what their peers think" for students to evaluate and discuss. And this is truly what they are doing. Unlike approaches where machine learning is used to produce themes directly from the data, peer-based AI hints are surfacing the collective expertise and opinions within the student cohort. They are surfacing areas of consensus or lack of consensus for everyone to consider. In this sense, peer-based AI hints are supporting the same iterative sensemaking that occurs when independent coders in a research team discuss their codes together towards coming to a consensus, not necessarily on 'the ground truth', but on what lens they find most interesting to focus in on. This suggests that an important future direction is to explicitly design how peer-based AI hints might facilitate discussions and collective sensemaking beyond the simple agree/disagree responses and explanations in the current version. This will especially be needed when extending our work to parts of the qualitative analysis process. In these contexts, we think that it can also be valuable to surface the work of peers, but the AI-driven aggregation function could be designed to surface divergent viewpoints rather than convergent ones.

More broadly, it is possible that, in the future, AI systems will be powerful enough to "solve"[4] qualitative analysis or give people the perception that it has solved it. When that happens, it will be tempting for educators to rush it into the classroom to help scale instruction. This could be dangerous if not done well. The result would be the industrialization and standardization of the "human instrument" [35]. There could be an overreliance on the system [16], with students simply agreeing with the interpretation recommended by the system and decreasing the natural diversity of subjective

[4] i.e. reach or surpass human-level performance on the task

interpretations based upon students' unique, lived human experiences (which one could argue *should* be the lenses through which we view the data). Simply put, the danger is AI-induced groupthink [26] that further removes the human element. An appropriate use, on the other hand, might be to use such a system to provide modeling to help students visualize what a high-quality outcome might look like (for a different research question and dataset). Then after students conduct their own analysis, the system could be used to support individuals in improving the clarity, groundedness, or coherence of their themes, and to summarize the unique points of views and insights produced across the classroom.

8 CONCLUSION

In this paper, we introduced a system for generating peer-based hints to help learn qualitative coding at scale through a novel application of the crowdsourcing DS-EM algorithm. We found that this approach was able to achieve high accuracy without needing many students or transcripts, and that when its predictions were presented as hints, they ultimately translated to diverse learning benefits for students. Beyond the value that our particular approach provides by augmenting experiential learning of qualitative analysis, we hope that, more generally, the idea of AI-augmented crowd-based collaborative learning inspires future work in complex crowd work and in unlocking scalable authentic work opportunities for learners across many domains.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful feedback, which helped us refine this paper. We thank Bryan Min, Jessalyn Wang, Daniel Chan, Russell Elliot, Elijah Kirby, Eliza Gonzales, Rina Munakata, Xilian Li, Christopher Toukmaji for work on early portions of the peer-based AI hint implementation or evaluation. We also thank Katherine Isbister for her helpful feedback on early drafts of this paper.

REFERENCES

[1] [n. d.]. ATLAS.ti | The #1 Software for Qualitative Data Analysis. https://atlasti.com

[2] [n. d.]. MAXQDA | All-In-One Qualitative & Mixed Methods Data Analysis Tool. https://www.maxqda.com/

[3] [n. d.]. NVivo. https://lumivero.com/products/nvivo/

[4] 2016. FACT SHEET: President Obama Proposes New "First Job" Funding to Connect Young Americans with Jobs and Skills Training to Start Their Careers. https://obamawhitehouse.archives.gov/the-press-office/2016/02/04/fact-sheet-president-obama-proposes-new-first-job-funding-connect-young

[5] Thomas Bailey, Katherine Hughes, and Tavis Barr. 2000. Achieving scale and quality in school-to-work internships: Findings from two employer surveys. *Educational Evaluation and Policy Analysis* 22, 1 (2000), 41–64.

[6] Lecia Barker. 2009. Student and faculty perceptions of undergraduate research experiences in computing. *ACM Transactions on Computing Education (TOCE)* 9, 1 (2009), 1–28.

[7] Sarah Bawabe, Laura Wilson, Tongyu Zhou, Ezra Marks, and Jeff Huang. 2021. The UX Factor: Using Comparative Peer Review to Evaluate Designs through User Preferences. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–23.

[8] Michael Bernstein. 2023. Teaching HCI Foundations and Frontiers. https://medium.com/@msbernst/teaching-hci-foundations-and-frontiers-76bb22616e0d

[9] John Seely Brown, Allan Collins, and Paul Duguid. 1989. Situated cognition and the culture of learning. *1989* 18, 1 (1989), 32–42.

[10] 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*. 1–13.

[11] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

[12] Nan-chen Chen, Rafal Kocielnik, Margaret Drouhard, Vanessa Peña-Araya, Jina Suh, Keting Cen, Xiangyi Zheng, Cecilia R Aragon, and V Peña-Araya. 2016. Challenges of applying machine learning to qualitative coding. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.

[13] Kabdo Choi, Hyungyu Shin, Meng Xia, and Juho Kim. 2022. AlgoSolve: Supporting subgoal learning in algorithmic problem-solving with learnersourced microtasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[14] Elizabeth F Churchill, Anne Bowser, and Jennifer Preece. 2016. The future of HCI education: a flexible, global, living curriculum. *interactions* 23, 2 (2016), 70–73.

[15] Kevin Crowston, Xiaozhong Liu, and Eileen E Allen. 2010. Machine learning and rule-based automated coding of qualitative data. *proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–2.

[16] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*. 6313.

[17] Sayamindu Dasgupta, William Hale, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Remixing as a pathway to computational thinking. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1438–1449.

[18] Alexander Philip Dawid and Allan 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.

[19] John Dewey. 1986. Experience and education. In *The educational forum*, Vol. 50. Taylor & Francis, 241–252.

[20] 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.

[21] 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 Transactions on Computer-Human Interaction* (2023).

[22] 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*. 1–19.

[23] 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*. 1626–1636.

[24] 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*. 301–304.

[25] Björn Hartmann, Daniel MacDougall, Joel Brandt, and Scott R Klemmer. 2010. What would other programmers do: suggesting solutions to error messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1019–1028.

[26] Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review* 36, 1 (2008), 36.

[27] W Brad Johnson. 2002. The intentional mentor: Strategies and guidelines for the practice of mentoring. *Professional psychology: Research and practice* 33, 1 (2002), 88.

[28] Andreas Kaufmann, Ann Barcomb, and Dirk Riehle. 2020. Supporting Interview Analysis with Autocoding.. In *HICSS*. 1–10.

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

[30] Juho Kim, Philip J Guo, Carrie J Cai, Shang-Wen Li, Krzysztof Z Gajos, and Robert C Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 563–572.

[31] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[32] 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*. 75–84.

[33] 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*. 1–12.

[34] Christophe Lejeune. 2011. From normal business to financial crisis... and back again. An illustration of the benefits of Cassandre for qualitative analysis. In *Forum: Qualitative Sozialforschung*, Vol. 12. Institut fur Klinische Sychologie and Gemeindesychologie, Germany.

[35] Yvonna S Lincoln and Egon G Guba. 1985. *Naturalistic inquiry*. sage.

[36] Chi-Jung Lu and Stuart W Shulman. 2008. Rigor and flexibility in computer-based qualitative research: Introducing the Coding Analysis Toolkit. *International Journal of Multiple Research Approaches* 2, 1 (2008), 105–117.

[37] Megh Marathe and Kentaro Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[38] Jean Piaget, Margaret Cook, et al. 1952. *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

[39] Quizlet. 2023. Quizlet: Learning tools, flashcards, and textbook solutions. https://quizlet.com/. Accessed: 2023-09-05.

[40] 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*. 1–14.

[41] 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*. 1–15.

[42] Johnny Saldaña. 2021. The coding manual for qualitative researchers. *The coding manual for qualitative researchers* (2021), 1–440.

[43] Michelle Salmona, Eli Lieber, and Dan Kaczynski. 2019. *Qualitative and mixed methods data analysis using Dedoose: A practical approach for research across the social sciences*. Sage Publications.

[44] Scratch. 2023. Scratch - Imagine, Program, Share. https://scratch.mit.edu/. Accessed: 2023-08-13.

[45] Rhea Sharma, Atira Nair, Ana Guo, Dustin Palea, and David T Lee. 2022. "It's usually not worth the effort unless you get really lucky": Barriers to Undergraduate Research Experiences from the Perspective of Computing Faculty. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*. 149–163.

[46] Natasha Singer. 2019. Top Universities Join to Push Public Interest Technology. *The New York Times* (March 2019). https://www.nytimes.com/2019/03/11/technology/universities-public-interest-technology.html

[47] David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29.

[48] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes.* Harvard university press.

[49] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing.* 405–416.

[50] 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.* 871–883.

[51] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale.* 379–388.

[52] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. 2023. LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs. *arXiv preprint arXiv:2307.10168* (2023).

[53] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces.* 75–78.

[54] Jasy Liew Suet Yan, Nancy McCracken, and Kevin Crowston. 2014. Semi-automatic content analysis of qualitative data. *IConference 2014 Proceedings* (2014).

Received 9 October 2023