

How Does Social Annotation Develop Over Time? Examining Temporal Changes in Discourse Moves During Online Reading Activities

Xiner Liu¹[0009-0004-3796-2251], Yuwei Liang¹[0009-0002-3561-3213], Zhanlan Wei¹[0009-0002-3931-6398], Zhen Xu²[0009-0004-3131-910X], Bodong Chen¹[0000-0003-4616-4353], Ryan S. Baker³[0000-0002-3051-3232], Yunseo Lee⁴[0009-0005-8651-5690], Amanda Barany¹[0000-0003-2239-2271]

¹ University of Pennsylvania, Philadelphia, PA 19104, USA

² Columbia University, New York, NY 10027, USA

³ Adelaide University, Adelaide, SA 5005, Australia

⁴ University of Wisconsin-Madison, Madison, WI 53706, USA

xiner@upenn.edu

Abstract. This study examined how social annotation discussions developed over time during asynchronous online reading activities. Using annotation data from four online courses, we identified and labeled how students engaged with course readings and peers' contributions through a collaborative process involving both human researchers and GPT-5. Statistical and epistemic network analysis results showed that each stage of the annotation discussion demonstrated distinct interactional characteristics. For example, compared to the later stages, earlier discussions involved stronger direct engagement with the reading, during which students more frequently related the text to their prior knowledge and experiences, challenged claims presented in the text, and proposed alternative interpretations. As discussions continued, students became more responsive to one another's ideas and increasingly worked toward collective sensemaking by revisiting earlier comments, reconciling interpretations, and addressing unresolved questions. These findings suggest that participation in social annotation evolves across the course of discussion rather than remaining interactionally uniform over time, and they highlight the importance of considering such temporal dynamics in the design of more effective social annotation environments.

Keywords: Social Annotation, Online Learning, Reading Discussion, Epistemic Networks, Large Language Models, GPT-5

1 Introduction

Social annotation has emerged as an important form of asynchronous collaborative learning in discussion-oriented education environments [27]. It differs from the more familiar genre of online discussion in where the conversation lives. In threaded forums, discussion about a reading is detached from the reading itself and relocated to a separate space organized by reply chains. Social annotation instead binds discussion to the text,

allowing students to attach interpretations, questions, critiques, and reflections to specific passages while creating a visible layer of collaborative interaction around the same text [21]. A growing body of work suggests that this arrangement can support students' engagement with assigned readings [11], their collaborative interpretation of ideas and arguments [6], and their critical discussion of claims raised in the reading [19], by encouraging students to interact not only with the text but also with peers' evolving interpretations of it.

Because annotations accumulate visibly in the discussion environment, the evolving annotation space creates a social-material condition for student participation. Participants engage not only with the reading itself but also with an existing layer of peer interpretation attached to the text [8]. As a result, annotations produced at different time points may constitute substantially different forms of contribution to shaping the collective interpretation of the text. They are made under materially different conditions and produce the material conditions for new contributions.

Yet, existing research has largely looked past this within-activity dynamic. Earlier studies examined annotation discourse either in aggregated form within individual readings or across multiple activities (e.g., [7]), or at the course level over time (e.g., [16]). Comparatively less attention has focused on the feature that most distinguishes social annotation, i.e., the social-material layer that thickens as annotation unfolds within a single reading activity. Such fine-grained temporal analysis is important for activity design. It may reveal how learners participate during the early stages of an assignment, when annotations remain sparse, versus later stages, when the annotation pool becomes rich. These insights could inform strategies for scaffolding the distinct demands of different stages of collaborative reading [15].

To address this gap, the present study examines how students contribute to the annotation space across different phases of a single reading activity. We first identify the discourse moves students demonstrated in social annotation discussions, and then examine how the use of these moves shifted as the discussion progressed. We conducted our study across four graduate-level courses that incorporated a social annotation platform as part of weekly reading assignments. Our research questions ask:

- **RQ1:** How did students respond to readings and peer annotations during social annotation discussions?
- **RQ2:** How did students use different discourse move to advance discussion across different phases of the annotation process?

2 Related Work

Social annotation has emerged as a pedagogical approach to supporting social reading that allows learners to annotate online texts asynchronously and anchor discussions directly to the annotated content [22]. Built on web annotation technologies, social annotation provides learners with simultaneous access to a shared reading environment and creates an interactive layer over online texts [25], affording an online collaborative space in which learners are encouraged to read critically, exchange interpretations, and engage in meaningful knowledge co-construction [14].

A growing body of research has demonstrated the promise of social annotation as a pedagogical strategy. For instance, Brush et al. found that, compared to a traditional discussion board, students using a social annotation system contributed nearly twice as much to online discussions and engaged more contextually with the target reading by anchoring comments directly to specific parts of the text [4]. Kalir et al. similarly reported that undergraduates perceived peer annotations as valuable for enhancing comprehension, validating ideas, and engaging with diverse perspectives [13]. A systematic review synthesizing social annotation activity designs identified social annotation as a tool with unique affordances for fulfilling multiple pedagogical goals in online learning, such as advancing domain knowledge, fostering argumentation and inquiry, improving literacy, and supporting assessment and cross-space learning connections [27].

While the benefits of social annotation have been increasingly documented, enhancing learning experiences within these environments requires a closer examination of learners' behavioral patterns, given their potential relationship with collaborative knowledge construction and students' motivation for continued engagement [10, 26]. Such examination may also help researchers and learning designers better understand how learners engage with social annotation over time, potentially informing the design of more effective tools and features to support engagement. A number of studies have drawn on text mining techniques to model students' social and cognitive behavioral patterns and characterize the overall structure of collaborative discourse. For instance, Li et al. used a meta-clustering approach to identify the optimal number of clusters and then applied K-means clustering to classify undergraduates into distinct annotation and response behavior profiles based on their interactions with course readings across multiple weeks [15]. Their findings revealed meaningful differences in reading engagement, annotation effort, and peer acknowledgment between initiator and responder clusters. Zhu et al. used social network analysis (SNA) to examine role-based interaction patterns across assigned readings and compared SNA measures across groups [28]. The findings showed that *facilitators* occupied the most central interactive positions, *synthesizers* received fewer responses than facilitators, and *summarizers* did not differ significantly from regular participants in levels of social interaction. Liang et al. applied epistemic network analysis to further reveal that, across different stages of the course, facilitators increasingly used questions and prompts to elicit peers' ideas and sustain discussion, whereas summarizers largely mirrored students without roles and showed comparatively limited engagement in collaborative knowledge construction [16]. Extending prior work on behavioral patterns in social annotation, this study traces how students used different interaction strategies to engage with readings and peer contributions at a more granular level within individual reading as they unfold over time.

3 Study Context

The study uses a dataset collected from four seminar-style online classrooms offered at a large research university in the United States. Descriptive statistics of the classroom contexts are presented in Table 1. A total of 55 students participated in these courses, which were all related to the topic of learning and data analytics and taught by the same

instructor. These online classes were designed to support collaborative learning, with the expectation that students would engage in emergent, ongoing social interaction. To facilitate this, the instructor used Hypothesis, a social annotation platform, for asynchronous discussion of the readings. Each week, the instructor posted assigned readings and learning activities. Students were asked to annotate the readings on Hypothesis and discuss their ideas directly in the margins. Although the expectation was one original annotation and two replies per week, students often posted more than the minimum requirement. Students were further encouraged to treat their annotations as a shared knowledge base to support their course projects. No further instructions were provided. The dataset includes a total of 3,052 annotations collected after IRB approval. Each annotation record contains the author ID, timestamp, the ID of the root annotation (if it is a reply), the text of the annotation, and information about the length and location of the highlighted text. All data were de-identified before analysis.

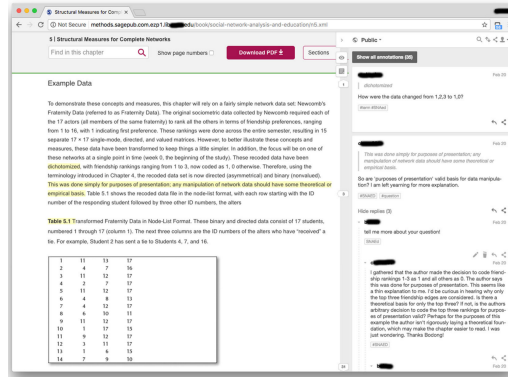


Fig. 1. Interface of Hypothesis. The shared reading appears on the left, while annotations associated with the document are shown on the right (image from Chen [6]).

Table 1. Descriptive statistics of the classroom contexts

Classroom	Students	Readings	Annotation	Thread
Class 1	17	35	1190	421
Class 2	10	18	287	113
Class 3	15	35	811	278
Class 4	13	24	764	250

4 Method

4.1 Data Pre-processing and Codebook Creation

The annotation data were processed to reconstruct the structure of discussion threads and prepare the data for chronological and relational analysis. Annotations were then grouped into threads based on the reply relationships and sorted chronologically.

In addressing RQ1, we developed a codebook of annotation types through an iterative process informed by prior research on discourse analysis and collaborative reasoning (e.g., [18, 20]) and a close examination of the annotation data. To help identify recurring patterns in the data that might not be immediately apparent, the researchers also provided ChatGPT-5 with eight batches of annotation data, each containing approximately 50 annotation threads, sampled evenly across the four courses. We prompted GPT to identify recurring themes and interaction patterns in the data following the procedure developed by Barany et al. [1], with the prompts adapted to describe the context of the present study.

The research team treated the prior literature and GPT-generated output as reference points during codebook development. Based on their own examination and interpretation of the annotation data, the researchers developed an initial set of relationship types that could capture (1) how students engaged with the reading content and (2) how they responded to and built on one another's ideas within their annotation comments. Accordingly, the codebook was designed to apply to both root annotations, where students interacted directly with the highlighted text, and reply annotations, in which students interacted with previous annotation(s) in the discussion thread. The team then iteratively refined the codebook through multiple rounds of review to confirm that the definitions were clear and the examples accurately reflected the intended meaning of each category (see Table 2).

Table 2. Codebook for discourse types in social annotation

Code	Definition	Example
Commending/ Reaffirming	An annotation expresses agreement, appreciation, or praise for a point, claim, or perspective presented in the <i>highlighted text</i> or <i>previous annotation(s)</i> .	<i>"I think that you make a great point by stating that the design process can be iterative..."</i>
Summarizing/ Synthesizing	An annotation condenses or synthesizes points, arguments, explanations, or perspectives presented in the <i>highlighted text</i> or <i>previous annotation(s)</i> into a concise statement.	<i>"...However, I think the issue is that educators struggle to understand the data researchers provide..." in response to a discussion about how the education system made limited use of data.</i>
Clarifying/ Refining	An annotation improves, clarifies, or reframes an idea, argument, or explanation presented in the <i>highlighted text</i> or <i>previous annotation(s)</i> .	<i>"Or very likely 'garbage in, garbage out' without theory." in response to "we are letting data and algorithms to tell us the theory and why/how something works."</i>
Contradicting/ Challenging/ Critiquing	An annotation challenges, opposes, or questions a claim, argument, generalization, or assumption presented in the <i>highlighted text</i> or <i>previous annotation(s)</i> .	<i>"[Company name], transparent? I highly doubt they would want any other editorial company to be able to replicate their work..." in response to a discussion on the transparency of a publishing company.</i>

<i>Branching/ Shifting</i>	An annotation asks a new question or introduces a new theme, perspective, or direction related to the <i>highlighted text</i> or <i>previous annotation(s)</i> .	<i>“This is linked to how we should position analytics in our work. The author implied a more ‘assistive’ role here. Wondering how our views have evolved given the rise of AI in the past few years.”</i>
<i>Question Answering/ Explaining</i>	An annotation directly answers a question or offers an explanation or interpretation of an issue, phenomenon, or idea raised in the <i>highlighted text</i> or <i>previous annotation(s)</i> .	<i>“I think that some internal factors are more obvious and some less to the learner. For instance...” in response to the earlier annotation: “Is it expected that the internal factors are visible to the learners?...”</i>
<i>Connecting/ Relating</i>	An annotation draws connections between the <i>highlighted text</i> or <i>previous annotation(s)</i> and another part of the reading, an existing annotation, a different reading/resource, or personal/real-world experiences.	<i>“...This principle resonates with me because students get mislabeled based solely on a test...”</i>

Note: Constructs are referenced below using their abbreviated names, shown in italics.

4.2 Coding Social Annotation Data

Automated Coding of Reply Annotations. Before applying LLMs to automatically identify discourse relationships within social annotation data, the research team first constructed a human-coded reference dataset for evaluation. Two coders, both with training and experience in qualitative analysis, independently annotated 100 pairs of student annotations across 21 discussion threads randomly selected from the dataset. The threads were randomly selected, with each student appearing at least once, thereby limiting the extent to which the reference dataset reflected the discourse patterns of only a small subset of participants. For each post in a thread, coders assessed how the current annotation related to earlier ones. Coders applied a relationship code as long as the annotation demonstrated evidence of the construct in response to any earlier annotation or discussion context within the thread. This process continued sequentially through the final post in each thread before coders moved to the next thread.

Cohen’s Kappa (κ) was calculated for each construct in the codebook to assess inter-rater reliability (IRR). After the first round of coding, the coders achieved κ values above 0.75 across all relationship types. The sampled dataset was also reviewed to confirm that all constructs appeared in at least 5% of the coded instances. Disagreements were resolved through social moderation and discussion between the coders [9]. The finalized, agreed-upon dataset then served as the human-coded reference standard for evaluating GPT-5’s ability to identify discourse relationships within social annotation discussions.

The prompting design was adapted from Liu et al. [17]; for each construct, GPT-5 was given (1) a definition, (2) representative keywords and/or example(s), (3) previous annotation(s), in sequence, and (4) guidance on how to interpret the social annotation structure. The prompt refinement process involved iterative rounds of testing the model

on the ground truth dataset and revising the instructions or changing the illustrative examples to clarify the boundaries of each relationship type. This cycle continued until the model produced stable results, defined as no decision differing in more than three cases across three repetitions of the same prompt, and coding that closely matched human labels with a κ above 0.70 for all constructs. The refined prompt was then applied to the full dataset.

Manual Coding of Root Annotations. Root annotations were coded separately through a manual process. In many cases, students highlighted only a word or short phrase of the text when adding an annotation. Understanding each root annotation, however, required manually identifying the corresponding passage in the full reading and interpreting the annotation within that context. Given that this contextual matching was already required for manual review, we incorporated the coding into same process rather than adding a separate automated coding step. The two coders labeled how each root annotation related to the reading text based on the annotation relationship codebook using a subset of 100 root posts selected through the same sampling criteria described above. The coders did not achieve $\kappa > 0.70$ for all constructs in the first round. After social moderation to resolve disagreements in the sampled data, another sample of 100 root posts was coded, and κ was recalculated. All constructs met the reliability threshold of 0.7 or higher. At that point, one coder manually coded the remaining root posts.

4.3 Distribution of Discourse Codes Across Reading Annotation Periods

Students completed reading annotation activities asynchronously between two consecutive class meetings. During this period, students could post annotations and replies at any time while engaging with the assigned readings on Hypothesis. Because annotation activity unfolded over multiple days prior to class discussion, we examined whether the occurrence of discourse codes varied across different periods of the annotation assignment (RQ2).

For each reading, we identified the timestamp of the first annotation and calculated the relative day of each subsequent annotation, defined as the number of days between the current annotation and the first annotation associated with that reading. We retained only annotations posted between relative days 0 and 7 (the assignment window) and excluded posts that fell outside the intended pre-class preparation period. To examine temporal variation in annotation activity, annotations were grouped into three temporal categories based on their relative timing within the assignment period. Because the eight-day assignment window could not be divided evenly into three whole-day stages, we used the midpoint of the window to define the *Middle* period (days 3–4) and assigned the three days before and after it to the *Early* (days 0–2) and *Later* (days 5–7) periods, respectively. The final dataset included 2,227 annotations from 47 students. Of these, 928 annotations (41.67%) were posted during the *Early* period, including 542 root posts and 386 replies. A total of 761 annotations (34.17%) were posted during the *Middle* period, including 189 root posts and 572 replies. 538 annotations (24.16%) were

posted during the *Later* period, including 41 root posts and 497 replies. These distributions indicate a gradual decline in annotation activity over the course of the assignment period, alongside a decrease in root posts and an increase in the proportion of replies.

We then examined whether the occurrence of each discourse code differed across these temporal categories. For each discourse code, a 2×3 contingency table was constructed comparing code presence (present vs. absent) across the *Early*, *Middle*, and *Later* annotation periods. Chi-square tests of independence were conducted to determine whether discourse code occurrence was associated with annotation timing.

Because annotation activity was substantially higher during earlier portions of the assignment period, expected frequencies were calculated based on the overall distribution of annotations across the three temporal categories rather than assuming equal annotation activity across time periods. Standardized residuals were also calculated to identify temporal categories in which discourse codes were overrepresented or underrepresented relative to the overall annotation distribution. Positive residuals indicated that a discourse code appeared more frequently than expected within a given temporal category, whereas negative residuals indicated lower-than-expected occurrence (see Section 5.1). To assess the robustness of our findings, we also applied the Benjamini-Yekutieli procedure [2], which controls the false discovery rate under arbitrary dependence among multiple statistical tests. Because all discourse codes were derived from the same set of student annotations, and we could not assume that they satisfied the independence or positive regression dependency conditions under which the Benjamini-Hochberg procedure guarantees such control. We therefore selected this more conservative procedure over the Benjamini-Hochberg procedure.

4.4 Epistemic Network Analysis

To provide a complementary perspective on RQ2, we further applied Epistemic Network Analysis (ENA) to examine how co-occurrences among discourse moves differed across stages. Because the codes were not mutually exclusive, ENA modeled the structure of annotation behavior by identifying which discourse relationship patterns tended to co-occur. We generated three ENA models, one for each stage, and used difference networks to identify which combinations of discourse relationships appeared more strongly at different stages.

Although each reply annotation was coded within its immediate discussion context (e.g., how a reply related to earlier annotations in its thread), the ENA data were segmented at the reading level rather than at the thread level. This decision was driven by the analytical goal of examining the broader discourse ecology surrounding each reading, rather than only the interactional structure within isolated discussion threads. Segmenting by reading allowed annotations from different threads attached to the same text to contribute to the same epistemic network, which made it possible to model how discourse relationship patterns collectively accumulated around a shared reading during each stage. Because annotations were chronologically ordered within each stage, later annotations were still only connected to annotations that had already appeared earlier in the discussion timeline for the same reading. This structure was based on the assump-

tion that students entering the annotation space would encounter comments already visible at the time of their participation, and that those earlier contributions might influence how they constructed subsequent annotations. Consistent with this assumption, we used an infinite stanza window when constructing the epistemic networks. As a result, the network represented how discourse relationship patterns accumulated across the collective annotation activity surrounding a reading. For visual clarity, connections with line weights (lw) below 0.02 or 0.01, depending on the difference model, were excluded from the final visualizations.

5 Results

5.1 Chi-Square Analysis of Discourse Move Types Across Stages

We present the chi-square results and standardized residuals for discourse code occurrence across stages in Table 3. The analysis identified statistically significant differences in the distribution of several discourse move types across the early, mid-, and later stages.

Specifically, the distribution of *Connecting* differed significantly across stages, $\chi^2(2) = 20.32, p < .001$. The standardized residuals showed that *Connecting* appeared more frequently during the *Early* stage (1.51). These results suggest that earlier discussions more often involved establishing connections between ideas and personal contexts. One possible explanation is that fewer peer-generated perspectives were available during the early stage, which may have led students to draw more heavily on their own prior knowledge, personal experiences, and existing understandings when responding to the reading.

The distributions of *Summarizing*, $\chi^2(2) = 9.50, p = .009$, *Question Answering*, $\chi^2(2) = 12.20, p = .002$, and *Clarifying*, $\chi^2(2) = 10.56, p = .005$, also differed significantly across stages. The standardized residuals showed that all three constructs appeared more frequently during the *Later* stage (1.90, 2.00, and 2.29, respectively). These results suggest that later discussions more often involved consolidating ideas across annotations, elaborating on prior comments, and refining earlier understandings. One possible explanation is that students in later stages had more time to process both the reading materials and their own thoughts, which may have led to annotations that were more integrative.

Table 3. Chi-square results and standardized residuals for code occurrence across periods. Constructs are ordered based on the highest residual across the three stages (appear in boldface)

Construct	Early Residual	Middle Residual	Later Residual	Chi-square	p-value
Connecting	1.51	-2.35	0.81	20.32	<0.001*
Contradicting	1.39	-0.70	-1.00	4.61	0.100
Branching	0.54	-0.57	-0.02	1.10	0.578
Commending	-0.93	0.62	0.49	3.44	0.180

Summarizing	-0.83	-0.68	1.90	9.50	0.009*
Question Answering	-0.64	-0.97	2.00	12.20	0.002*
Clarifying	-1.16	-0.64	2.29	10.56	0.005*

* indicates statistical significance after B&Y correction.

5.2 Epistemic Network Visualization of Discourse Types

We next present the difference network comparing the structure of the discourse relationship across the *Early*, *Middle*, and *Later* stages (see Figure 2). In the difference networks, line weight differences (lw_{diff}) stronger in the *Early* stage are shown in red, those stronger in the *Middle* stage are shown in blue, and those stronger in the *Later* stage are shown in purple.

Mann-Whitney tests revealed a significant difference between the *Early* and *Middle* stages along the X-axis ($U = 1063.0$, $p = .04$, $r = -0.27$), but no significant difference between the *Middle* and *Later* stages ($U = 917.5$, $p = .47$, $r = -0.09$). Along the Y-axis, the test revealed a significant difference between the *Early* and *Later* stages ($U = 475.0$, $p = .01$, $r = -0.31$).

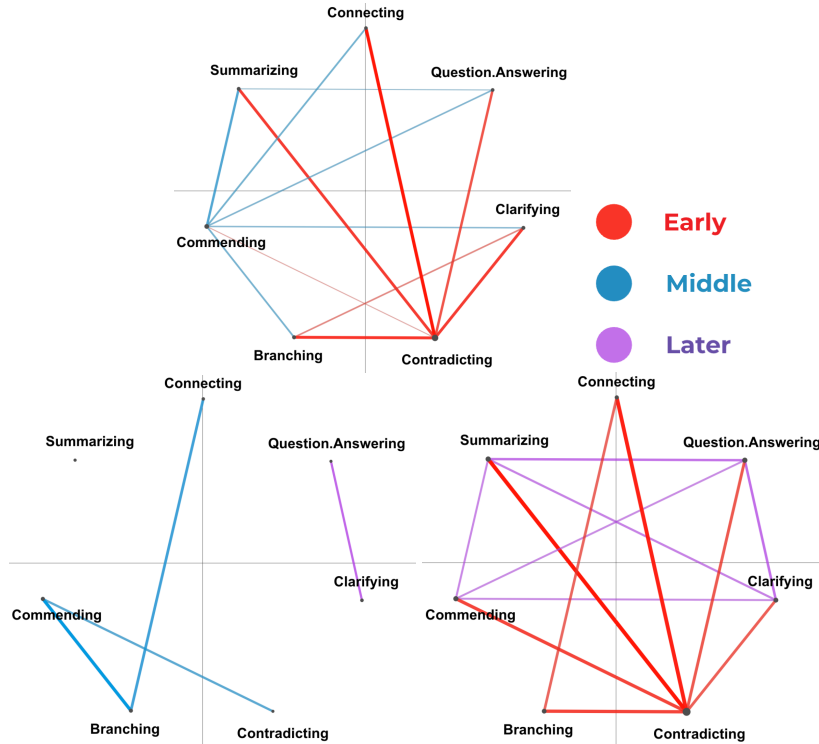


Fig. 2. ENA difference models for *Early* vs. *Middle* stages (top), *Middle* vs. *Later* stages (bottom left), and *Early* vs. *Later* stages (bottom right).

The ENA difference models indicate that the relational structure among discourse codes differed meaningfully across annotation periods. The comparison between the *Early* and *Middle*, as well as between the *Early* and *Later* stages, shows that the *Early* stage had stronger connections involving *Contradicting* (see Table 4). Examination of the annotation data suggests that students during the *Early* stage frequently questioned or challenged claims from the reading itself through root posts (e.g., “*This sounds fascinating, but I’m skeptical...*” and “*I’m skeptical of how much this can actually help...*”), which is expected given that fewer peer responses had accumulated at this point in the discussion process. However, these critical responses rarely appeared as isolated acts of disagreement. Instead, the stronger *Early*-stage edges show that *Contradicting* often co-occurred with other discourse moves like *Connecting* (e.g., “*As a real life example, my school district...*”), *Branching* (e.g., “*This brief portion...makes me ask many more questions...*”), and *Question Answering* (e.g., “*It’s interesting because...*”). In other words, critique during the *Early* stage appeared to function less as rejection and more as a mechanism for opening discussion about the reading and creating opportunities for students to relate to their own experiences. These early exchanges may therefore have helped establish interpretive pathways that later annotations could revisit, refine, or build upon.

This observation aligns with the distribution of the *Contradicting* code across the annotation periods. Of the 570 total instances coded as *Contradicting*, 45% occurred during the *Early* stage, compared to 32% during the *Middle* and 22% during the *Later* stage. Critique became less central as the annotation discussion developed. One possible interpretation is that the object of critique changed as the annotation space became increasingly populated with peer contributions. Early in the annotation period, students primarily interacted with the reading itself, which may have positioned critique as part of the intellectual task of interrogating the text. As peer annotations accumulated, however, disagreement no longer involved only the author’s ideas but also the visible contributions of classmates. Under these conditions, students may have become more selective about expressing contradiction, particularly in a persistent and publicly shared discussion space where responses remained visible to others over time. Another possible interpretation of this pattern is that students’ initial critiques of the reading were gradually clarified and scaffolded through peer annotation activities, which made it easier for later participants to identify common ground and move toward agreement as the discussion progressed.

Compared to the *Early* stage, in which 54% of annotations were coded as *Commending*, 58% of annotations in both the *Middle* and *Later* stages received this code. As also shown in both the ENA plots and the line weight tables, *Commending* became more closely connected with other discourse moves during the *Middle* and *Later* stages. Once more peer contributions had accumulated, annotations frequently opened with acknowledgment or affirmation of a peer’s contribution before advancing a new interpretation or line of discussion (e.g., “*Completely agreed!...*”, “*Great insight...*”, or “*Great point...*”). The stronger presence of explicit validation toward previous participants’ interpretations may suggest that students increasingly treated peer annotations as meaningful interpretive resources. In this sense, meaning-making within the annotation space appeared to become more collectively negotiated through interaction with peers’ contributions emerging from the shared reading.

Meanwhile, we observed that the *Later*-stage epistemic network showed stronger connections between *Summarizing*, *Clarifying*, and *Question Answering* and other discourse moves. In fact, during the *Later* stage, 56% of annotations were labeled as *Summarizing*, compared to 48% in the other two stages. *Clarifying* appeared in 39% of *Later* annotations, compared to 31% in the *Early* stage and 32% in the *Middle* stage. A similar pattern appeared for *Question Answering*, which occurred in 62% of *Later* annotations, compared to 54% in the *Early* stage and 53% in the *Middle* stage.

Close examination of the annotations suggests that students during the *Later* stage often entered discussions that already contained multiple interpretations, unresolved questions, and partially developed lines of reasoning. Under these conditions, summarizing and clarification appeared to function less as simple restatement and more as mechanisms for interpretive coordination. Students frequently attempted to reconcile earlier comments through synthesizing moves (e.g., “*The important thing to note is that...*”, “*My take-way is that...*”) or provided explanatory responses that addressed ambiguities left unresolved in prior exchanges (e.g., “*I had the same confusion too! ... I guess they are talking about...*”). At the *Later* stage, discussion threads often already contained extensive annotation histories and layered interpretations, which appeared to position students more as participants working to organize and stabilize the growing body of collective interpretations surrounding the reading. The stronger interconnectiveness among these discourse moves may therefore suggest that later participation became increasingly oriented toward working through interpretive complexity within the collaborative discussion space.

Table 4. Line weights for prominent connections are shown in the three plots; boldface indicates the highest value within each row

Connection	Early	Middle	Later
Contradicting ↔ Commending	0.21	0.20	0.18
Contradicting ↔ Question Answering	0.20	0.18	0.17
Contradicting ↔ Connecting	0.20	0.17	0.16
Contradicting ↔ Summarizing	0.20	0.17	0.16
Contradicting ↔ Branching	0.17	0.14	0.14
Contradicting ↔ Clarifying	0.15	0.12	0.12
Commending ↔ Connecting	0.25	0.27	0.26
Commending ↔ Question Answering	0.24	0.26	0.26
Commending ↔ Branching	0.25	0.26	0.23
Summarizing ↔ Commending	0.22	0.25	0.24
Summarizing ↔ Question Answering	0.21	0.23	0.23
Commending ↔ Clarifying	0.21	0.22	0.23
Question Answering ↔ Clarifying	0.18	0.18	0.21
Summarizing ↔ Clarifying	0.17	0.18	0.20

6. Discussion and Conclusion

In this study, we explored how social annotation discussions evolved over time within asynchronous online reading activities by examining discourse moves and the relationships among them across different phases of discussion. Using annotation data from four online courses, we developed a codebook based on prior research and themes identified with ChatGPT-5 to capture how students engaged with course readings and peer contributions. We used a combination of manual coding and GPT-5-assisted coding to label discourse moves across the dataset. We then applied chi-square analysis and epistemic network analysis to examine how discourse moves and their relationships differed across the early, middle, and later phases of the weekly annotation window.

We observed that each phase of the annotation discussion demonstrated distinct interactional characteristics. For example, early discussion activity involved strong direct engagement with the reading itself. Students frequently challenged claims and assumptions presented in the text and introduced competing interpretations as they worked through the material. The middle phase became more socially responsive, with students increasingly recognizing, affirming, and extending ideas raised by their peers. Later discussion activity reflected a stronger orientation toward collective sensemaking, as students returned to earlier comments to reconcile interpretations, answer lingering questions, and organize the growing body of discussion surrounding the reading.

These findings suggest that social annotation should not be viewed as a static activity in which all forms of participation occur uniformly across time. Rather, the evolving annotation space creates a unique social-material condition for continuing engagement. As a result, differences in interactional patterns across the discussion timeline should be considered when designing effective asynchronous reading activities. For example, as discussed earlier, direct critique and disagreement appeared to decrease as the annotations accumulated. This pattern is understandable. As more peer contributions accumulated in the later stages, students may have grown increasingly reluctant to contradict peer views, perhaps prioritizing relational harmony over critical engagement. Yet, while supportive interaction may make students more comfortable responding to one another, discussions that become overly focused on agreement may leave fewer opportunities for productive disagreement. In conflict-oriented scenarios, learners are pushed to critically review multiple perspectives, scrutinize peers' reasoning, and construct more substantive arguments in order to reach a consensual conclusion [5, 24]. Such processes play an important role in meaningful collaborative learning and knowledge construction [12]. Therefore, instructors could, for example, incorporate discussion prompts into middle and later phases of the activity that explicitly encourage students to critically examine their peers' interpretations of the text or revisit earlier claims from alternative perspectives.

At the same time, we noticed that early participants primarily engaged directly with the reading before many peer annotations were available. While direct engagement with the assigned text may support independent interpretation of the reading, a meaningful collaborative learning experience also depends on students encountering perspectives they may not have considered on their own, responding to alternative interpretations, and negotiating meaning through interaction with others [3, 23]. Instructors may therefore introduce seeded questions or instructor-posted annotations early in the activity to create opportunities for students to respond to emerging ideas from others. At the same

time, as annotation histories become more visible, students may increasingly orient toward peer interaction rather than the reading itself. Instructors may therefore need to periodically redirect students' attention back to the source material so that discussions remain grounded in the text.

That said, several limitations should still be considered when interpreting the findings of this study. First, the dataset was collected within a relatively similar instructional context, which may limit the generalizability of the findings to other disciplines, educational levels, or annotation environments. Moreover, although the study examined how discourse relationships unfolded over time, the analysis focused primarily on observable annotation interactions rather than students' underlying intentions, motivations, or perceptions during participation. Additional qualitative investigation, such as interviews or reflective accounts, may be needed to better understand how students interpreted and navigated the evolving social dynamics of annotation discussions.

Future research may build on this work by examining more fine-grained interactional dynamics within social annotation discussions. For example, additional studies could investigate what makes some annotation threads more sustained or more diverse in terms of perspectives and participation. Future work may also examine differences between students who initiate annotation threads and those who primarily respond to others, as these participation roles may reflect different forms of engagement or knowledge construction. Ultimately, we hope that continued research on the dynamics of social annotation will provide a stronger foundation for designing more effective instructional interventions and facilitation strategies. In turn, these efforts may create more meaningful opportunities for collaborative knowledge sharing and knowledge construction among students.

References

1. Barany, A., Nasir, N., Porter, C., Zambrano, A. F., ... & Baker, R. S.: ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In *Int. Conf. on Artificial Intelligence in Education*, 134–149 (2024).
2. Benjamini, Y., & Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188 (2001).
3. Bruffee, K. A.: *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. Johns Hopkins University Press, Baltimore, MD (1999).
4. Brush, A. J. B., Barger, D., Grudin, J., Borning, A., & Gupta, A.: Supporting interaction outside of class: Anchored discussions vs. discussion boards. In Stahl, G. (ed.), *Computer support for collaborative learning*, 1st ed., 425–434 (2002).
5. Chan, C. K. K., Burtis, P. J., & Bereiter, C.: Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction* 15(1), 1–40 (1997).
6. Chen, B.: Designing for networked collaborative discourse: An UnLMS approach. *TechTrends* 63(2), 194–201 (2019).
7. Chen, B., Zhu, X., & Shui, H.: Socio-semantic network motifs framework for discourse analysis. In *12th Int. Learning Analytics and Knowledge Conference*, 500–506 (2022).
8. Gao, F., Zhang, T., & Franklin, T.: Designing asynchronous online discussion environments: Recent progress and possible future directions. *British Journal of Educational Technology* 44(3), 469–483 (2013).

9. Herrenkohl, L. R., Cornelius, L.: Investigating elementary students' scientific and historical argumentation. *Journal of the Learning Sciences* 22(3), 413–461 (2013).
10. Huang, X., Wu, H., Liu, X., & Lajoie, S.: Examining the role of peer acknowledgements on social annotations: Unraveling the psychological underpinnings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–9 (2024).
11. Hwang, W. Y., Wang, C. Y., & Sharples, M.: A study of multimedia annotation of web-based materials. *Computers & Education* 48(4), 680–699 (2007).
12. Johnson, D. W., & Johnson, R. T.: An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher* 38(5), 365–379. doi:10.3102/0013189X09339057 (2009).
13. Kalir, J. H.: Social annotation enabling collaboration for open learning. *Distance Education* 41(2), 245–260 (2020).
14. Kear, K., Jones, A., Holden, G., & Curcher, M.: Social technologies for online learning: Theoretical and contextual issues. *Open Learning: The Journal of Open, Distance and e-Learning* 31(1), 42–53 (2016).
15. Li, S., Huang, X., Zhu, G., Du, H., Zhong, T., Hou, C., & Zheng, J.: Exploring behavioural patterns and their relationships with social annotation outcomes. *Journal of Computer Assisted Learning* 40(4), 1389–1399 (2024).
16. Liang, Y., Wei, Z., Liu, X., Zhu, X., Gao, Y., & Chen, B.: Exploring role-based knowledge co-construction in social annotation with epistemic network analysis. In *International Conference on Quantitative Ethnography*, 226–241 (2025).
17. Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R. S.: Assessing the potential and limits of large language models in qualitative coding. In *International Conference on Quantitative Ethnography*, 89–103 (2024).
18. Manca, S., Delfino, M., & Mazzoni, E.: Coding procedures to analyse interaction patterns in educational web forums. *Journal of Computer Assisted Learning* 25(2), 189–200 (2009).
19. Mendenhall, A., & Johnson, T. E.: Fostering the development of critical thinking skills, and reading comprehension of undergraduates using a Web 2.0 tool coupled with a learning system. *Interactive Learning Environments* 18(3), 263–276 (2010).
20. Morales, E., Kalir, J. H., Fleerackers, A., & Alperin, J. P.: Using social annotation to construct knowledge with others: A case study across undergraduate courses. *F1000Research* 11, 235 (2022).
21. Novak, E., Razzouk, R., & Johnson, T. E.: The educational use of social annotation tools in higher education: A literature review. *The Internet & Higher Education* 15(1), 39–49 (2012).
22. Plevinski, J., Weible, J., & DeSchryver, M.: Anchored annotation to support collaborative knowledge construction. Philadelphia, PA: *Int. Society of the Learning Sciences* (2017).
23. Vygotsky, L. S.: *Mind in society: The development of higher psychological processes*. Vol. 86. Harvard University Press (1978).
24. Weinberger, A., Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education* 46(1), 71–95 (2006).
25. W3C Web Annotation Working Group: *W3C web annotation working group* (2016).
26. Zhu, X., & Chen, B.: Understanding idea creation in collaborative discourse through networks: The joint attention-interaction-creation (AIC) framework. *arXiv preprint arXiv:2305.16262* (2023).
27. Zhu, X., Chen, B., ..., & Zhang, R. Z.: Reading and connecting: Using social annotation in online classes. *Information and Learning Sciences* 121(5/6), 261–271 (2020).
28. Zhu, X., Shui, H., & Chen, B.: Beyond reading together: Facilitating knowledge construction through participation roles and social annotation in college classrooms. *The Internet and Higher Education* 59, 100919 (2023).