Mining how curiosity spreads in student-led online discussions

Farhan Ali
National Institute of Education,
Nanyang Technological University
farhan.ali@nie.edu.sg

ABSTRACT

Curiosity is central to learning and human development, yet it is typically studied as an individual psychological state. Using a social cognitive lens, this study examines whether and how curious behaviors are transmitted socially in student-led online discussions. We first constructed and validated a new large-scale dataset from a student-driven educational forum on Reddit (full dataset: ~55,000 users, ~39,000 posts; ~667,000 comments). Curiosity was operationalized as question-posing using a 2×2 question type taxonomy involving open/closed and possibility thinking dimensions. Generalized linear models revealed clear social transmission effects: question-posing in comments generally increased questioning in subsequent replies, regardless of question type. Parallel mediation analyses involving two categories of discussion features - cognitive and socio-affective - uncovered pathways of transmission of curiosity involving social endorsement, cognitive alignment, and selective mimicry. These findings advance the field by demonstrating that curiosity is socially transmitted, highlighting its contingent spread, and providing insights into how educators and platform developers can intervene and design features to steer discussion trajectories. Our research further contributes descriptively via a new taxonomy of question-posing and the dissemination of a new online discussion corpus useful for future research.

Keywords

adolescents, curiosity, online discussions, Singapore, Reddit, question-posing

1. INTRODUCTION & BACKGROUND

Curiosity is central to learning yet often treated as an internal state. There is a need to broaden our understanding in order to better foster curious learners. Drawing on social cognitive theory, we examine how curiosity, expressed through question-posing, spreads in peer online discussions where there are rich intersections of formal learning, informal learning, and community inquiry. Using large-scale data from an education-oriented Reddit community, data mining, and quantitative modeling, we investigate curiosity's transmission patterns and mechanisms in authentic student digital environments.

1.1 Social cognitive theory

Social learning theory conceptualizes learning as inherently social, shaped by observing others' behaviors and their consequences. Social cognitive theory further emphasizes cognitive and motivational processes that determine whether observed behaviors are adopted [3, 4]. A central process is the transmission, or diffusion, of thoughts and actions across individuals. The proposed mechanisms include imitation, modeling, and vicarious reinforcement, whereby behaviors perceived as rewarded are more likely to be replicated. Importantly, transmission is thought to be selective: individuals evaluate relevance and value before adopting the observed behaviors. In education, parallel theoretical traditions such as social constructivism highlight knowledge as co-constructed through dialogue and collaboration [20].

1.2 Curiosity

Curiosity - the desire for new information and knowledge [16] - is a foundational driver of learning and development and predicts important outcomes such as academic achievement, well-being, and longevity [22, 24, 27]. Yet contemporary schooling is often argued to dampen curiosity [7], with cross-cultural evidence indeed documenting systematic declines from elementary to secondary years [11]. Understanding how curiosity can be triggered and fostered is, therefore, an important educational challenge.

Most research conceptualizes curiosity as an intrapersonal state. However, curiosity is also expressed behaviorally in social contexts, most visibly through question-posing, a behavior widely studied in curiosity research in education [7, 13]. A question is not merely a private attempt to resolve an information gap; it signals to others what is worth asking, how to ask, and whether such behavior is socially valued. In this sense, curiosity becomes embedded within a social learning ecology, with potential to spread or dissipate depending on peer responses and contextual cues.

While question-posing is often used as an indicator of curiosity, questioning may not all reflect intrinsically motivated curiosity. Drawing on a goal-systems perspective, it is useful to distinguish between curiosity as a motive (ends) and curious behaviors that serve diverse goals (means) [23] such as performance-enhancement or tension-reduction. This distinction is especially relevant in large-scale online data, where inferring internal motivation is difficult without additional contextual or self-report measures. Accordingly, we follow framing by Szumowska & Kruglanski [23] to operationalize curious behaviors (as opposed to curious motivations) with question-posing. Importantly, in our context, questions are voluntary rather than teacher-prompted and non-evaluative with no immediate benefit, making them a meaningful indicator of momentary curious behaviors in authentic peer learning interactions.

Emerging evidence suggests that curiosity may indeed be socially transmissible from human to human in classroom settings [2], and even from robot to human in experimental settings [9]. However, the extent, mechanisms, and boundary conditions of curiosity transmission, particularly in authentic, peer-driven digital environments where many adolescents now regularly interact, remain

Farhan Ali. Mining how curiosity spreads in student-led online discussions. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 500–507. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040018

insufficiently understood. An important question is the nature of the transmission. Transmission may be broad, where exposure increases questioning generally, or narrow, where specific question types are preferentially replicated. Distinguishing these forms clarifies whether curious behaviors spread as generalized activation or subtype-specific processes, informing strategies for interventions or nudging on online platforms or by instructors.

1.3 Educational context: Reddit online discussions

Digital communities such as Reddit provide a valuable setting for examining curiosity in authentic peer interaction. Reddit hosts sustained, community-driven discussions spanning personal, academic, and social topics, leaving behind rich textual traces suitable for large-scale educational research. Prior studies have used Reddit data to illuminate important educational issues, including impact of COVID-19 [26] and the emergence of ChatGPT [18] . The platform's affordances also align well with a social learning perspective: for instance, upvotes function as social reinforcement, threaded replies make interaction paths explicit, and norms such as quoting or referencing can signal uptake and engagement.

Within this Reddit context, curiosity can be operationalized through question-posing, a visible indicator that researchers have emphasized as being important [7, 13]. Building on prior educational research and recent frameworks distinguishing open/closed and possibility-oriented questions[1], we propose crossing these dimensions to derive four nuanced question types, enabling automated analysis of how curiosity is expressed and potentially transmitted in adolescent online discussions.

To understand how curiosity spreads, discussion patterns must be analyzed. Prior work integrating social learning and community-of-inquiry perspectives identified three interdependent features [10]: cognitive presence (collective meaning-making through discourse), social presence (interpersonal and affective expressions), and teaching presence (directed facilitation toward learning goals). In adolescent-driven communities with minimal adult moderation, cognitive and social presence are especially central. This framework highlights how epistemic contributions (e.g., explanations, meaning) and socio-affective signals (e.g., endorsement, emotional tone) may shape peer transmission processes.

Guided by this perspective, we identified an online community using three criteria: predominantly student-driven participation with little to no adult or bot or spam activity; high activity and diverse threads to support robust linguistic analyses; and a focus on schooling-related experiences, not general lifestyle or pop culture chatter. We identified r/SGExams as meeting these criteria. SGExams is Singapore's largest youth-oriented education forum, where students discuss academic subjects, examinations, study strategies, transitions, and peer issues. Its shared national context enhances the likelihood that expressions of curiosity are recognized and taken up by peers, making it well-suited for examining social transmission in authentic digital learning environments.

1.4 Current study

The current study investigated the extent of social transmission of curiosity in online educational discussions. To the extent that social transmission exists, we further asked about the nature of its transmission in terms of specificity and features that mediate these transmission effects. To this end, we analyzed our newly curated, large-scale dataset from an education-related Reddit community, SGExams. Curiosity was operationalized as question-posing using a new 2×2 taxonomy, crossing open/closed and possibility/non-possibility question forms. We analyzed transmission using generalized linear models and quantified mediating pathways using Monte Carlo-based counterfactual contrasts with ten cognitive and socio-affective discussion features modeled.

2. METHOD

2.1 Dataset construction and validation

Institutional Review Board (IRB) approval for retrieving and analyzing publicly-available deidentified data with a waiver of consent was obtained prior to the start of research. We built a custom data collection pipeline using Python Reddit API (PRAW v7) to retrieve posts and comments from SGExams. The dataset was compiled through an iterative, query-based procedure. This iteration was necessary as there are constraints in returns from a single API query. An initial word set (e.g., "exam," "teacher," common function words) was used to generate automated queries and returns. Unique terms were then incorporated to expand subsequent queries and returns. Retrieval continued until saturation, defined as no new returned posts or comments from additional queries (as identified by post/comment ID). The final dataset comprised 38,920 posts and 666,978 comments from January 2019 to October 2023, including metadata (timestamps, identifiers, flairs, etc.). For this paper, we applied two-step preprocessing based on post flairs (topic labels assigned to a discussion thread). Flairs with fewer than 4,000 comments or activity concentrated within a few months were excluded, retaining 12 flairs grouped into five broader topic categories. We also reconstructed the discussion hierarchy by pairing parent posts/comments with their child replies. Note that "parent-child" here refers to the discussion hierarchy for ease of communication, and not familial relationship. Because users generally do not reply to themselves, these parent-child links largely reflect between-user interactions.

To assess completeness, we conducted two validation procedures. First, we restricted automated queries to a one-month window and compared retrieved counts with manual tallies of posts and comments directly on SGExams subreddit website. Second, we cross-validated our dataset with a third-party aggregator of Reddit statistics (https://subredditstats.com/). To verify that contributors were adolescent students, we manually reviewed 200 highly active users (≥10 posts/comments). For each, up to 50 posts and comments were examined for explicit age disclosures (e.g., "I'm 16") and contextual indicators (e.g., references to secondary school levels, examinations, or post-secondary pathways). These checks assessed the dataset's coverage and demographic relevance to the target population.

2.2 Curiosity measures

Curiosity was operationalized as question-posing, aligned with past work on indicators of curiosity [7, 13]. Because questions typically start with question words, we used a rule-based classification at the sentence level. Table 1 presents the full categorization of question types based on the rules and their implied intent and stance. Each sentence was assigned to one and only one category based on the rules. Individual comments were often composed of multiple sentences and thus could contain more than one type of question. We performed a validation of our automated classification by manually labelling 1,807 randomly selected sentences.

2.3 Discussion mediators

We identified and engineered 10 literature-informed mediators capturing shared meaning-making and affective exchange in Reddit discussions. Cognitive mediators indexed epistemic alignment

between parent and child comments (e.g., lexical overlap, conceptual similarity, content uptake), whereas socio-affective mediators reflected social valuation, social referencing, and emotional alignment. Together, they operationalized potential mechanisms for the transmission of curiosity. The mediators were engineered using natural language processing, unsupervised learning as well as a hybrid lexical-semantic mining process that we developed. First, we started with a small dictionary of lexicons (e.g., "you're right,") and embedded them along with the full corpus' n-grams (up to 5) using a pretrained sentence transformer (all-MiniLM-L6-v2). An iterative algorithm identified semantically close and frequent n-grams, followed by manual screening for a final lexical list. This hybrid lexical-semantic mining thus incorporates both lexical, semantic information, and contextual information. This approach contrasts with more standard methods, such as LIWC which typically lacks incorporation of context and is not grounded in actual use frequencies. This method was applied to two mediators of agreement and social referencing. This automated mining was then validated against 1,000 human-labeled sentences. See Table 2 for a detailed exposition of the mediators, measurements, and sources.

Table 1. Question types: description, rule, and examples

Type	Implied intent/stance	Sentence starters	Examples
O-NP: Open, Non-possibility questions	Factual clarification; descriptive or explanatory intent; assumption of a knowable answer	('who', 'when', 'what', 'where', 'how', 'why') AND (NOT ('can', 'may', 'might', 'could', 'would', 'if', 'will'))	"what is the DAE process like" "where is my monohybrid dihybrid epistasis"
O-P: Open, possibility questions	Exploratory reasoning; hypothetical thinking; acknowledgment of multiple possible outcomes	('who', 'when', 'what', 'where', 'how', 'why') AND ('can', 'may', 'might', 'could', 'would', 'if', 'will')	"how might we interpret the data here" "how can i explain and link this to evaluation"
C-NP: Closed, non-possibility questions	Verification, or confirmation; primary expectation of a yes/no answer; settled knowledge assumption	'must', 'should', 'do', 'is', 'are', 'does', 'did', 'have', 'has'	"Is the form for your course only?" "should i go back to sec 2 algebra"
C-P: Closed, possibility questions	Feasibility checking; bounded outcomes	'can', 'may', 'might', 'could', 'would', 'will'	"could it be such that the seats are all full and some are left out" "can it explain your action"

Note. 'if' was not included as sentence starter in the possibility-closed category as we found that sentences starting with 'if' were almost all non-questions.

Table 2. Cognitive and socio-affective discussion mediators to assess pathways for transmission effect

	Description	Measurement	Source
Cognitive			
Target lexicon mimicry	Extent of target lexicon similarity, capturing specific surface-level interactional alignment	Proportion of parent target lexicons in children's target lexicons. Target lexicons refer to the list of words used to classify question types (Table 1)	Combined from parent and child comments
General mimicry	Extent of overall lexicon similarity, indicating general surface-level interactional alignment	Jaccard similarity of content words (i.e., excluding stop words)	Combined from parent and child comments
Semantic similarity	Level of conceptual alignment beyond specific word use, indicating shared meaning and focus	Cosine distance between parent and child comments in the embedding space of a pre-trained sentence transformer model (all-MiniLM-L6-v2, 384 dimensions)	Combined from parent and child comments
Local referencing	Degree of uptake of specific content, reflecting focused engagement	Counts of '>' in children. '>' is the community markdown norm for replying to specific statements/sentences made by the parent comment	Child comments
Agreement	Extent of public concurrence or endorsement, signaling public alignment	Counts of any of the following lexicons as mined by our hybrid lexical-semantic approach: 'me too', 'sameeee','i agree','you're right', 'ur right', 'you right', 'you make sense', 'u make sense'	Child comments
Socio-affective			
Popularity	Level of community endorsement, reflecting a comment's perceived importance and value	Number of upvotes; similar to likes on other social media platforms	Parent comments
Social referencing	Extent of relational references, signaling interpersonal links	Counts of any of the following lexicons as mined by our hybrid lexical-semantic approach: 'you say that', 'u mentioned', 'u say', 'what you say', 'you mentioned', 'you explain', 'youve mentioned', 'saying you', 'you mention'	Child comments
Sentiment - Parent	Affective valence, capturing overall emotional tone	Sentiment analysis using a transformer model fine-tuned for social media sentiment classification into three classes [17]: positive, negative, neutral	Parent comments

	Description	Measurement	Source
Sentiment - Child	Affective valence, capturing overall emotional tone	Same as Sentiment - Parent but for child comments	Child comments
Sentiment mimicry	Degree of affective valence matching, indicating emotional alignment	Whether the parent and child sentiments match	Combined from parent and child comments

2.4 Quantitative modeling

We first fitted generalized linear models (GLMs), specifically negative binomial regressions. Alternatives such as linear models were not appropriate given the zero-bounded, count data with substantially larger variance than means. Each GLM model predicted one child question type count from all four parent question types. Thus, four separate GLM models were run as there were four question type outcomes. Wald tests compared regression coefficients to assess narrow versus broad transmission. To strengthen inference, we conducted 1,000 randomization tests within a discussion thread that preserved marginal count distributions while breaking parent-child links. This allowed us to evaluate whether observed effects exceeded chance while taking into account discussion threads and topics. We also fitted GLMs for non-examination topics to probe boundary conditions. A total of 40 Wald tests were performed. We used Šidák multiple-testing correction, a less conservative procedure compared to Bonferron correction. As a robustness check, we also implemented logistic regression (binarizing outcomes). GLMs were implemented in MATLAB (2025b).

To examine transmission mechanisms, we applied Monte Carlo g-computation, a flexible and rigorous approach to mediation originating from causal inference framework [12, 21]. It estimates how an outcome would change under hypothetical (counterfactual) manipulations of predictors and mediators. Intuitively, the method uses fitted models to simulate what would happen if the parent comment had different levels of a predictor, while allowing mediators to respond accordingly.

For each parent predictor-child outcome pair, we fitted a negative binomial model for the outcome and separate models for ten candidate mediators (Table 2), all included simultaneously in a parallel mediation framework. Let A denote the parent predictor, M the vector of mediators, and Y the child outcome. Using 1,000 Monte Carlo draws from the estimated parameter distributions, we simulated counterfactual expectations under contrasting predictor levels a_1 and a_0.

Effects were decomposed into indirect and direct components using the potential outcomes framework. The indirect effect (IE) is defined as:

$$IE = \mathbb{E}[Y_{a,M(a_1)}] - \mathbb{E}[Y_{a,M(a_0)}],$$

which captures the change in the outcome due to changes in mediators induced by the predictor, holding the predictor fixed. The direct effect (DE) is defined as:

$$DE = \mathbb{E}[Y_{a_1,M(a_0)}] - \mathbb{E}[Y_{a_0,M(a_0)}],$$

which captures the change in the outcome due to the predictor, holding mediators fixed at levels corresponding to a_0. The total effect satisfies:

$$TE = IE + DE = \mathbb{E}[Y_{a_1}] - \mathbb{E}[Y_{a_0}].$$

Mediator-specific indirect effects were evaluated using 95% confidence intervals, with the proportion mediated computed as $IE/(IE + DE)$. Given the observational nature of the data, these effects are interpreted descriptively rather than causally as key assumptions such as temporal ordering and sequential ignorability (e.g., no unmeasured confounders) are not fully satisfied. Analyses were conducted in R (MASS and related packages).

3. RESULTS

3.1 SGExams: Descriptive statistics and validation

Table 3 presents comment distributions across consolidated SGExams topics. Original post flairs were grouped based on shared themes, educational stages, and temporal trends into five categories: Exam Preparation, Exam Outcomes, Pre-university, University, and Others. Exam Preparation includes study strategies, content questions, and challenges surrounding high-stakes national examinations held annually. Exam Outcomes centers on exam results and admissions. Pre-university covers broader secondary and post-secondary experiences beyond exams (e.g., relationships, school activities, internships). University focuses on college admissions and courses. The heterogeneous "Others" category (e.g., jobs, scholarships, rants) was excluded from further analysis. Our primary analyses focused on Exam Preparation, which constituted the majority of activity and was theoretically most conducive to transmission effects due to its high-stakes context; results for the other topics are reported in Appendix Table 1.

We conducted extensive validation before analysis. Retrieval completeness was assessed in two ways: comparison with manual counts over a one-month window (99.8% retrieval) and cross-checking with third-party statistics from https://subredditstats.com/ (97.8%). Although third-party methodologies are not fully transparent, converging evidence suggests near-complete retrieval, lending confidence to complete conversational hierarchy.

Demographic validation of 200 randomly sampled users showed that 174 provided age-indicative statements; of these, 93.7% appeared to be adolescents broadly defined as students between secondary school and college. Finally, automated question classification was validated against 1,804 human-labeled sentences (κ = 0.97 for double-coded subset), yielding a weighted F1 of 0.90. Together, these extensive validation procedures support the dataset's completeness and validity. The code and dataset are publicly available at our institutional data repository (*https://doi.org/10.25340/R4/SYASFR*). We have removed Reddit usernames as the dataset is made up of discussions by minors in order to afford some level of privacy. Linked data is available upon request.

3.2 Social transmission of curiosity

Four GLMs were estimated, each predicting the count of a specific child question type from the four parent question types. As hypothesized, we uncovered strong and reliable transmission effects of curiosity as measured by question-posing. Results for Exam Preparation consolidated topic specifically are presented in Table 4 (N = 136,010 filtered parent-child pair). A parent question type strongly predicted the same question type in the child comments. For example, a one unit increase in parent open, non-possibility question type was associated with a 205% increase in expected count of child open, non-possibility question type. Evidence for narrow transmission where the same parent question type

Table 3. Consolidated discussion topics, original post flairs, and their distribution

Consolidated topics	Original Flair Title	Count (% of Total)
Exam Preparation (50.3%)	Exam Megathread	119608 (20.9%)
	O Levels	108271 (19.0%)
	A Levels	47347 (8.3%)
	N Levels	12070 (2.1%)
Exam Outcomes (4.4%)	Results Megathread	24966 (4.4%)
Pre-University (17.1%)	Junior Colleges	48726 (8.5%)
	JC vs Poly	3878 (0.7%)
	Polytechnic	45126 (7.9%)
University (17.0%)	University	97016 (17.0%)
Others (11.2%)	Rant	54005 (9.5%)
	Jobs	4493 (0.8%)
	Scholarships	5585 (1.0%)

Note. Data shown is post-filtering (see Methods). "O Levels" and "N levels" are Singapore's national exam for secondary school while "A levels" is the national exam for high school. "Junior Colleges" or "JC" are senior high schools, while "Polytechnic" or "Poly" post-secondary vocational institutions.

preferentially predicted its own type in child comments was tested using pairwise Wald test. Each GLM, superscript [a] in Table 4 indicates whether the bolded parent question type similar to child question type predicted the child question type significantly better than the other parent question types. For example, parent O-NP predicted child O-NP significantly more strongly than other parent question types, as indicated by superscripts [a] for O-P, C-NP, C-P. A narrow transmission effect will produce superscripts for all other parent types. Narrow transmission was observed for open type questions, both non-possibility and possibility (though with one non-significant result) types. However, closed question types still statistically significantly predicted open question types, though with smaller effect sizes. Our randomization procedure identified 5.2% of GLMs coefficients that were statistically significant, in line with a Type 1 error of 5%, suggesting that our results cannot be explained simply by chance or thread clustering. In the Appendix, we performed the same GLM analyses for the remaining three consolidated topics that do not directly relate to the Exam Preparation topic. We found transmission effects though not as strongly and statistically reliably as Exam Preparation. Interestingly, Exam Outcome discussions were more likely to exhibit suppression effects, the opposite of transmission (Appendix Table 1), indicating that the nature and goals of discussions play important roles in social transmission. In summary, curiosity as measured by question-posing in student discussions strongly exhibits social transmission in high-stakes exam-oriented student discussions, though with more limited evidence for specificity in its transmission.

Table 4. GLM results predicting the outcome variable of child question type from parent question type

	Beta	SE	z	Corrected p	Relative ratio
O-NP (outcome variable)					
Intercept	-2.51	0.01	-176.95	<0.01	0.08
O-NP	1.11	0.04	31.13	<0.01	3.05
O-P[a]	0.19	0.15	1.31	0.99	1.21
C-NP[a]	0.36	0.04	9.61	<0.01	1.43
C-P[a]	0.27	0.06	4.28	<0.01	1.32
O-P (outcome variable)					
Intercept	-5.43	0.05	-119.77	<0.01	0.00
O-NP[a]	0.57	0.09	6.23	<0.01	1.77
O-P	1.22	0.29	4.21	<0.01	3.39
C-NP[a]	0.16	0.11	1.42	0.99	1.17
C-P	0.46	0.17	2.65	0.26	1.58
C-NP (outcome variable)					
Intercept	-2.64	0.01	-188.12	<0.01	0.07
O-NP[a]	0.84	0.03	24.36	<0.01	2.31
O-P	0.46	0.14	3.33	<0.01	1.58
C-NP	0.45	0.04	12.82	<0.01	1.57
C-P	0.35	0.06	5.65	<0.01	1.41
C-P (outcome variable)					
Intercept	-3.36	0.02	-179.49	<0.01	0.03
O-NP	0.57	0.04	13.12	<0.01	1.77
O-P[a]	-0.12	0.22	-0.54	0.99	0.89
C-NP[a]	0.36	0.05	7.69	<0.01	1.43
C-P	0.67	0.08	8.90	<0.01	1.95

Note. Four GLMs were used to predict each dependent variable of child question type from the four parent question types (+ intercept). For each GLM, Beta: GLM regression coefficient estimate (log scale). SE: standard error of the coefficient. z: Wald test statistic (Beta/SE). p: p-value for the Wald test. Relative ratio: exponentiated coefficient, exp(Beta), indicating multiplicative change in expected count of outcome variable for a one-unit increase in the predictor. The number minus 1 times 100% indicates the percentage change. For example, a unit increase in the count of parent O-NP was linked to a 205% increase in the expected count of child O-NP.
For each GLM, superscript [a] indicates whether the bolded parent question type predicts the same question type in the child comment significantly better than the other parent question types (based on pairwise Wald test). For example, in the GLM to predict O-NP, parent O-NP predicted child O-NP significantly more strongly than other parent question types, as indicated by superscript [a] for O-P, C-NP, C-P. A narrow transmission effect will produce superscripts for all other parent types. Such an effect was observed only for parent O-NP and O-P question types.
O-NP: open, non-possibility questions; O-P: open, possibility questions; C-NP: closed, non-possibility questions; C-P: closed, possibility questions. * Corrected p-values: *p < 0.05; ** p < 0.01; *** p < 0.001

3.3 Mediating pathways

To examine pathways mediating transmission, we applied a Monte Carlo simulation-based g-computation approach. Ten discussion features, five each in cognitive and socio-affective categories were tested as potential parallel mediators. Table 5 presents the results. Numbers are the percentage of total effects of the predictor on the outcome mediated by the respective mediator, i.e., indirect effects. The vast majority of indirect effects were statistically significant, suggesting reliable mediations that were, however, generally small on the order of ~5-15%. Bolded estimates represent the top 3 mediators for each of the question types. Popularity, as measured by upvotes, was frequently amongst the top mediators. This result indicates that the popularity of parent comment functions as a marker of social endorsement, increasing the likelihood that subsequent child replies reproduce curiosity. Agreement, as measured by public concurrence and endorsement, also played an important mediating role for all question types, suggesting follow-up comments were more likely to display curiosity when students were more aligned in their views and opinions.

Interestingly, target lexicon mimicry also significantly mediated the social transmission of curiosity, indicating that adolescent students are prone to question word re-use in their curiosity expressions. However, this question word re-use contrasted with negative mediation by general mimicry, which measures overall lexical re-use. This pattern of results highlights the selective epistemic uptake, and not merely broad language imitation. We also observed that the percentage of indirect effects tended to be quite similar for the different question types. Indeed, formal quantification revealed very high correlation in indirect effects estimates for all question types (Spearman's r = 0.79). This provides a plausible explanation for the general finding of broad curiosity transmission:

similar mediating mechanisms appear to operate across different question types, suggesting shared cognitive and socio-affective pathways rather than type-specific processes.

Table 5. Proportion of total effects of the parent question type on child question type mediated by 10 conversational features.

	O-NP ↓ Media-tor ↓ O-NP	O-P ↓ Media-tor ↓ O-P	C-NP ↓ Media-tor ↓ C-NP	C-P ↓ Media-tor ↓ C-P
Mediator: Cognitive				
Target lexicon mimicry	**7.27***	2.54	**5.64***	**12.24***
General mimicry	**-7.44***	-2.34	-1.50	**-6.06***
Semantic similarity	0.19*	-0.27	1.06*	-0.21
Local referencing	3.07*	2.40*	4.04*	2.98*
Agreement	6.05*	**6.05***	**6.97***	**6.00***
Mediator: Socio-affective				
Popularity	**15.20***	**11.32***	**10.05***	3.28*
Social referencing	0.79*	-1.96	0.27	0.92*
Sentiment - Parent	0.35*	**-3.25***	0.21*	1.09*
Sentiment - Child	1.64*	2.12	-1.38*	1.06*
Sentiment mimicry	-0.26	0.86	0.51	0.26

Note. Numbers are estimated percentages of the total effects of predictor on outcome mediated by the respective mediator. Bolded percentages are the top 3 mediators for each predictor-outcome relationship. * indicates statistically significant indirect effects ($p < 0.05$).
O-NP: open, non-possibility questions; O-P: open, possibility questions; C-NP: closed, non-possibility questions; C-P: closed, possibility questions.

4. DISCUSSION

Our study analyzed a new dataset of online educational discussions led by adolescent students. Leveraging a newly created large-scale and topic-rich nature of the dataset, we mined for intricate signals of curiosity transmission and its potential pathways using advanced regression models followed by more flexible Monte Carlo g-computation to further quantify mediation. We found strong social transmission of curiosity: parent question-posing robustly predicted child question-posing, in agreement with past studies suggesting social transmission of curiosity [2, 9]. There was limited evidence of narrow transmission, though broader cross-type effects emerged as strongly. Social transmission was especially strongest in exam discussions. Monte Carlo g-computation identified statistically reliable though modest mediating pathways. We found that popularity (upvotes) and agreement consistently mediated social transmission.

The finding that comment popularity plays one of the most important roles in socially transmitting curiosity is supported by previous findings. Participants indicated greater curiosity about the answers to scientific questions when the questions were presented with experimentally manipulated high vs. a low number of up-votes [6], highlighting the importance of social cues in promoting curiosity. Our study further extends this finding by demonstrating that popularity not only heightens individual curiosity toward content, but also shapes subsequent behavioral expression of curiosity in peer interaction. Specifically, highly upvoted parent questions were more likely to elicit follow-up questions from others, indicating that social endorsement functions as a diffusion mechanism of question-posing. Cognitive alignment, as operationalized by our agreement conversational feature, is also an important mediator of social transmission. Students were more likely to be curious when they agreed with the stance or perspective expressed in the prior comment. One possible explanation is that cognitive alignment creates a psychologically safe and supportive space in which extending the discussion through further questions becomes more likely, thereby facilitating the social propagation of curiosity.

We also found a dissociation between question word mimicry and general lexical mimicry, in which the former increases, but the latter decreases, curiosity propagation. This pattern can be interpreted through a social cognitive lens. Social cognitive theory posits that observational learning is selective: individuals attend to, encode, and reproduce behaviors perceived as meaningful, functional, or socially valued, not merely any behavior. The positive mediation by target lexicon mimicry suggests that adolescents selectively model epistemically relevant features (e.g., interrogative structures signaling curiosity), rather than broadly imitating surface language features. An experimental study lends support to this idea: children and adults re-use or novelly re-combine previously exposed questions, especially when they are informative, likely in the service of efficient inquiry and effective learning [15]. Our work also links to prior work on linguistic style matching in discourse analysis [8, 19], though our results provide more nuances, such as what linguistic aspects are being matched by adolescent students.

Our study extends past online discussion analyses in educational data mining/learning analytics. These studies have fruitfully applied computational linguistics methods with a focus on better understanding and predicting course participation and performance [e.g., 25]. We extend these studies to focus more on social transmission, particularly its mediating mechanisms, applying it to curiosity expressions by way of questioning.

Our study makes further contributions to research by introducing a scalable, multidimensional taxonomy of question-posing and applying it to a large, naturalistic dataset. Existing taxonomies often dichotomize questions into lower- versus higher-order types or rely on domain-specific distinctions (e.g., causal vs. descriptive in science), which can be overly reductive or limited in generalizability or scalability. By crossing openness (open vs. closed) with possibility thinking based on a previous framework [1], our extension captures variation in information-seeking intent while remaining adaptable across domains. Moreover, its reliance on identifiable linguistic markers enables highly reliable automated classification at scale. Coupled with a large corpus of peer-driven discussions that we have curated and continue to build on, this approach advances methodological tools for studying many aspects of educational discussions, including curiosity, as they unfold dynamically in authentic digital environments. While discussions in the context of formal settings (group work in a classroom or online course discussions) are relatively well studied, the dynamics of informal, peer-driven learning communities remain underexplored.

Our study has implications for educators as discussion facilitators and for designers of social learning platforms, though with important caveats. For educators, the findings suggest that curiosity can be socially amplified through visible endorsement and constructive alignment. Publicly recognizing thoughtful questions and modeling how to extend peers' inquiries can help cultivate inquiry-oriented norms [2]. However, overemphasis on popularity could privilege confident or dominant students; highlighting alignment might limit epistemic diversity. Educators, therefore, need to balance public reinforcement with inclusive facilitation strategies that invite diverse voices and stances. For platform designers, features such as upvotes, threading, and content highlighting can shape epistemic behavior by signaling what is valued. Yet, simple popularity metrics may also distort attention toward socially appealing rather than substantively rich questioning. Designing

systems that surface high-quality questions, such as via rotating visibility or algorithmic weighting, may better support equitable and sustained discussions.

We acknowledge limitations in the study. First, while our dataset is grounded in general educational discussions (unlike other general adolescent-based Reddit communities which can be dominated by pop culture and general lifestyle), it is nonetheless limited to one particular national educational context. Second, our modeling of mediation based on g-computation which originates from a causal inference framework. However, failing to meet strict causal assumptions such as sequential ignobility (e.g., no unmeasured confounding) and temporal ordering limits strong causal statements. For instance, due to the limitations of the Reddit API, upvote counts were returned as a single total with no temporal information. Another limitation is that the data has a hierarchical and networked structure. As such, observations are not fully independent, and standard errors from our GLMs may be underestimated. While our analyses partially controlled for discussion context by analyzing consolidated topic categories separately and performing within-thread randomization tests, we did not explicitly model all forms of clustering such as users and higher-order sequences (e.g., subheads within threads) operating at different scales [5, 14].

Considering these limitations, future studies can be extended in several ways. One interesting question is whether social networks play a role. While Reddit discussions are largely anonymous, repeated interactions create recognizable usernames and emergent micro-communities. An important next step is to examine whether curiosity transmission is influenced by network position or prior interaction history between users. Another promising direction is to investigate temporal dynamics: e.g., does repeated engagement with curious peers increase an individual's baseline propensity to ask questions over time, suggesting longer-term socialization effects beyond immediate momentary exchanges? We believe addressing these questions can illuminate fine-grained processes of how students learn digitally socially beyond the classroom.

5. CONCLUSION

This study demonstrates that curiosity, operationalized as volitional question-posing, is not merely an individual disposition but a socially transmitted behavior in peer-driven online discussions. Using large-scale data and computational modeling, we show that questioning propagates across interactions, with both broad and selective patterns shaped by cognitive alignment and social endorsement. By framing curiosity at the behavioral level, our findings highlight how learning-related inquiry emerges and spreads in authentic digital collaborative environments, offering insights for designing educational contexts that cultivate sustained, self- and peer-driven questioning.

6. ACKNOWLEDGMENTS

None.

7. APPENDIX

Appendix Table 1. GLM results predicting the outcome variable of child question type from parent question type for each of the three remaining consolidated topics

	Exam outcome	Pre-University	University
O-NP (outcome variable)			
Intersect	0.07**	0.08**	0.09**
O-NP	2.01**	1.56**	1.50**
O-P	2.68	1.80*	1.66*
C-NP	0.66	1.29**	1.21**
C-P	0.49	1.31**	1.25**
O-P (outcome variable)			
Intercept	<0.01	<0.01**	<0.01**
O-NP	1.85	1.31*	1.47*
O-P	<0.01	6.67**	3.51*
C-NP	0.77	1.28*	1.19
C-P	<0.01	1.00	1.04
C-NP (outcome variable)			
Intercept	0.15**	0.10**	0.13**
O-NP	1.80*	1.25**	1.14**
O-P	3.38	1.34	1.65*
C-NP	0.48**	1.49**	1.35**
C-P	0.41**	1.39**	1.08
C-P (outcome variable)			
Intercept	0.06**	0.04**	0.07**
O-NP	1.98**	1.26**	1.15**
O-P	0.94	1.24	1.32
C-NP	0.64*	1.34**	1.13**
C-P	0.78	1.54**	1.36**
N (parent-child pairs)	11,134	50,053	53,356

Note. Statistics are relative ratios (regression coefficients exponentiated), which are multiplicative change in expected count of outcome variable for a one-unit increase in the predictor. The number minus 1 times 100% indicates the percentage change. Results can be compared to Table 4 in the main text.
O-NP: open, non-possibility questions; O-P: open, possibility questions; C-NP: closed, non-possibility questions; C-P: closed, possibility questions.
Corrected p-values: *p < 0.05; ** p < 0.01; *** p < 0.001

8. REFERENCES

[1] Acar, S., Berthiaume, K., and Johnson, R., 2023. What kind of questions do creative people ask? *Journal of Creativity 33*, 3, 100062. DOI= http://dx.doi.org/10.1016/j.yjoc.2023.100062.

[2] Ali, F., Wang, Y., Wang, S.J.-W., and Zhu, G., 2025. Triggers of curiosity in social constructivist classroom discourse. *npj Science of Learning 10*, 1, 33. DOI= http://dx.doi.org/10.1038/s41539-025-00330-5.

[3] Bandura, A., 1986. *Social foundations of thought and action: A social cognitive theory*.

[4] Bandura, A. and Walters, R.H., 1977. *Social learning theory*. Prentice Hall.

[5] Chen, B. and Poquet, O., 2023. Uncovering socio-temporal dynamics in online discussions: An event-based approach. *Australasian Journal of Educational Technology 39*, 6, 1-16. DOI= http://dx.doi.org/10.14742/ajet.8618.

[6] Dubey, R., Mehta, H., and Lombrozo, T., 2021. Curiosity is contagious: A social influence intervention to induce curiosity. *Cognitive Science 45*, 2, e12937. DOI= http://dx.doi.org/10.1111/cogs.12937.

[7] Engel, S., 2011. Children's need to know: Curiosity in schools. *Harvard Educational Review 81*, 4, 625-645. DOI= http://dx.doi.org/10.17763/haer.81.4.h054131316473115.

[8] Gonzales, A.L., Hancock, J.T., and Pennebaker, J.W., 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research 37*, 1, 3-19. DOI= http://dx.doi.org/10.1177/009365020935146.
[9] Gordon, G., Breazeal, C., and Engel, S., 2015. Can children catch curiosity from a social robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 91-98. DOI= http://dx.doi.org/https://doi.org/10.1145/2696454.269646.
[10] Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., and Paulin, D., 2018. Learning in the wild: coding for learning and practice on Reddit. *Learning, Media and Technology 43*, 3, 219-235. DOI= http://dx.doi.org/10.1080/17439884.2018.1498356.
[11] Huang, H., Tang, X., and Salmela-Aro, K., 2024. Facilitating youth's curiosity in learning: Needs-based ecological examinations. *Journal of Youth and Adolescence 53*, 3, 595-608. DOI= http://dx.doi.org/10.1007/s10964-023-01936-x.
[12] Imai, K., Keele, L., and Yamamoto, T., 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*, 1, 51-71. DOI= http://dx.doi.org/10.1214/10-STS321.
[13] Jirout, J.J., Zumbrunn, S., Evans, N.S., and Vitiello, V.E., 2022. Development and testing of the curiosity in classrooms framework and coding protocol. *Frontiers in Psychology 13*, 875161. DOI= http://dx.doi.org/10.3389/fpsyg.2022.875161.
[14] Knight, S., Wise, A.F., and Chen, B., 2017. Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics 4*, 3, 7–17. DOI= http://dx.doi.org/10.18608/jla.2017.43.2.
[15] Liquin, E.G., Rhodes, M., and Gureckis, T.M., 2025. Seeking new information with old questions: Children and adults reuse and recombine concepts from prior questions. *Open Mind 9*, 885-925. DOI= http://dx.doi.org/10.1162/opmi.a.12.
[16] Litman, J., 2019. Curiosity: Nature, dimensionality, and determinants. In *The Cambridge handbook of motivation and learning*, K.A. Renninger and S.E. Hidi Eds. Cambridge University Press, 418-442.
[17] Loureiro, D., Barbieri, F., Neves, L., Anke, L.E., and Camacho-Collados, J., 2022. TimeLMs: diachronic language models from Twitter. *arXiv*, 2202.03829. DOI= http://dx.doi.org/10.48550/arXiv.2202.03829.
[18] Na, H., Staudt Willet, K.B., Shi, H., Hur, J., He, D., and Kim, C., 2024. Initial discussions of ChatGPT in education-related subreddits. *Journal of Research on Technology in Education 57*, 5, 953-971. DOI= http://dx.doi.org/10.1080/15391523.2024.2338091.
[19] Niederhoffer, K.G. and Pennebaker, J.W., 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology 21*, 4, 337-360. DOI= http://dx.doi.org/10.1177/02619270223795.
[20] Palincsar, A.S., 2005. Social constructivist perspectives on teaching and learning. *An introduction to Vygotsky 2*, 285-314.
[21] Snowden, J.M., Rose, S., and Mortimer, K.M., 2011. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology 173*, 7, 731-738. DOI= http://dx.doi.org/10.1093/aje/kwq472.
[22] Swan, G.E. and Carmelli, D., 1996. Curiosity and mortality in aging adults: A 5-year follow-up of the Western Collaborative Group Study. *Psychology and Aging 11*, 3, 449. DOI= http://dx.doi.org/10.1037/0882-7974.11.3.449.
[23] Szumowska, E. and Kruglanski, A.W., 2020. Curiosity as end and means. *Current Opinion in Behavioral Sciences 35*, 35-39. DOI= http://dx.doi.org/10.1016/j.cobeha.2020.06.008.
[24] von Stumm, S., Hell, B., and Chamorro-Premuzic, T., 2011. The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science 6*, 6, 574-588. DOI= http://dx.doi.org/10.1177/17456916114212.
[25] Wen, M., Yang, D., and Rosé, C., 2014. Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the International AAAI Conference on Web and Social Media*, 525-534.
[26] Yan, T. and Liu, F., 2022. COVID-19 sentiment analysis using college subreddit data. *PloS One 17*, 11, e0275862. DOI= http://dx.doi.org/10.1371/journal.pone.0275862.
[27] Zainal, N.H. and Newman, M.G., 2022. Curiosity helps: Growth in need for cognition bidirectionally predicts future reduction in anxiety and depression symptoms across 10 years. *Journal of Affective Disorders 296*, 642-652. DOI= http://dx.doi.org/10.1016/j.jad.2021.10.001.