How User Learning Behavior Affects Answer Quality in Online Q&A Communities?

Yinzhi Luo[1,a], Feng Hu[1,b*], Jianmei Xiao[2,c]

[1] Beijing University of Technology, Beijing 100124, China
[2] University of Sheffield, Sheffield, The United Kingdom
{[a]luoyinzhi01@emails.*[b]jut.edu.cn,[c]jianmei472@gmail.com}

Abstract. The quality of answers in online Q&A communities is essential for their sustainable development. Existing research has mainly focused on factors such as answer features, user characteristics, and community mechanisms, but the impact of user learning behavior on answer quality remains underexplored. This study, based on data from the Zhihu platform, applies social cognitive theory and the SBERT model to calculate the semantic similarity between answers. By quantifying user learning behavior, the study examines how it affects answer quality. The results show a negative correlation between user learning behavior and answer quality, primarily due to content homogeneity and a lack of innovation. Information entropy moderates this relationship, with higher entropy mitigating the negative effects of learning behavior and improving answer quality. This study introduces a new definition of user learning behavior and offers a quantitative method, providing theoretical support and practical recommendations for improving the answer recommendation system and user experience on online Q&A platforms.

Keywords: Online Q&A Community, User Learning Behavior, Answer Quality, Social Cognitive Theory, SBERT Model

1 Introduction

With the rapid development of internet technology, online Q&A communities have become key platforms for acquiring knowledge and solving problems [1]. These platforms integrate individual wisdom with collective intelligence, providing significant societal value. However, as the user base grows, uneven answer quality has become a prominent issue. This stems from variations in users' knowledge, communication skills, and the platform's incentive and recommendation mechanisms. The proliferation of low-quality answers negatively impacts user experience, knowledge dissemination, and the platform's reputation. Therefore, exploring the factors influencing answer quality is essential for improving content and optimizing user experience.

Existing research on answer quality primarily focuses on three aspects: 1) the impact of answer characteristics such as content, relevance, and timeliness; 2) the role of user characteristics like social status, reputation, and intrinsic motivation [2]; and 3) the influence of community-level feedback and incentive mechanisms [3]. Some studies also

M. Zhang et al. (eds.), *Proceedings of the 2025 5th International Conference on Informatization Economic Development and Management (IEDM 2025)*, Advances in Economics, Business and Management Research 330,
https://doi.org/10.2991/978-94-6463-724-3_21

examine how user learning behavior affects answer quality. To better understand this aspect, we turn to social cognitive theory, which provides a framework for understanding how individuals learn from their environment, behavior, and cognition.

Social cognitive theory, proposed by Bandura, emphasizes that individuals learn through the interaction of environment, behavior, and cognition [4]. It suggests that learning occurs via direct and observational learning. In online Q&A communities, users' answers and feedback are clearly presented on both the question page and personal page, providing an intuitive learning environment. Users can adjust their response strategies based on past experiences of receiving high likes (direct learning), or analyze others' responses to identify characteristics of high-quality answers (observational learning). Feedback indicators like likes, comments, and favorites guide users' learning behaviors. For example, Jin et al. (2022) measured direct learning by the number of answers posted by users in the past week, representing the knowledge and skills users accumulate by personally answering questions; and defined observational learning using the number of answers from users' followers in the past week, representing learning through observing others [5]. Similarly, Shi et al. (2021) measured direct learning by the average length of high-vote answers from the user's past contributions, reflecting skills gained through personal success, and observational learning by the average length of high-vote answers from other users on the same question, indicating learning through others' successful answers. These indicators explore how past contributions and community feedback influence users' subsequent contributions [6].

While previous studies have analyzed behavioral data to measure learning behavior, they have not fully explored how users learn from existing answers for the same question or optimized the measurement of learning behavior. This study redefines learning behavior as the process by which users learn from the content characteristics of existing answers to the same question while answering. By calculating the semantic similarity between answers, the study directly measures users' learning behavior and explores its impact on answer quality. This provides a new perspective for improving answer quality in online Q&A communities.

2 Review and Hypotheses

2.1 The Impact of User Learning Behavior on Answer Quality

Previous studies have explored the impact of user learning behavior on answer quality. For instance, Jin et al. (2022) integrated the Ability-Motivation-Opportunity theory with social learning theory, finding that users enhance their abilities and contribute higher-quality knowledge through both direct and indirect experiences [5]. Shi et al. (2021), based on social cognitive theory, measured the impact of direct and observational learning on knowledge contribution, finding both had a significant positive effect on the information provided in subsequent answers [6]. Krishna et al. (2024) observed that for novice users, continuous answering improves answer quality, but excessive contribution leads to a focus on quantity over quality [7].

These studies suggest that user learning behavior positively impacts answer quality. However, this study redefines it as the process of learning from existing answers when responding to questions. While learning helps users quickly develop strategies, imitative learning may lead to answer convergence, causing reader fatigue and diminishing answer recognition, potentially negatively affecting answer quality. Therefore, the study proposes the following hypothesis:

H1: User learning behavior toward the content features of existing answers for the same question will negatively impact the quality of their own answers.

2.2 The Moderating Role of Information Entropy

Information entropy, a core concept in information theory, quantifies uncertainty and the amount of information [8]. In text analysis, it measures the diversity and unpredictability of the text. High entropy indicates an unpredictable distribution of information, while low entropy suggests a more predictable one. Studies show that higher entropy increases text attractiveness. For instance, Fresneda & Gefen (2019) found Amazon reviews with new information had higher entropy and greater value [9], and Sun et al. (2024) found higher entropy in health consultation responses provided more useful information [10]. In online Q&A communities, information entropy is a key indicator of answer quality. Repetitive information leads to low entropy and value, while new insights or a change in presentation increases entropy and enhances answer quality [11]. Thus, information entropy may moderate the relationship between learning behavior and answer quality, mitigating the negative effects of answer convergence. Based on this, the study proposes the following hypothesis:

H2: Information entropy positively moderates the relationship between learning behavior and answer quality, meaning that higher entropy reduces the negative impact of learning behavior on answer quality.

3 Model and Variables

3.1 Data Collection and Variable Definition

Data Collection. The data for this study were sourced from Zhihu, a leading online Q&A community in China. According to data from Eastmoney (2024), as of the end of 2023, Zhihu had 105 million monthly active users, 71.3 million content creators, and over 775 million pieces of content [12]. We focus on the "learning English" topic, excluding time-related factors, with data collected from February 2011 to November 2023. The dataset includes 2,327 questions and 234,540 answers from 151,028 users. To ensure data quality, we excluded non-text answers, blanks, links, and answers lacking user interface elements. We also removed time-sensitive or controversial questions and limited answers per question to between 10 and 1,000. To control for the impact of answer length on semantic similarity, we restricted answer length to between 10 and 7,051 characters. The final dataset includes 1,813 questions, 141,455 answers, and 93,252 users.

Variable Definition. *Dependent Variable.* Answer quality. The number of upvotes reflects the community's recognition of the answer, and we use upvotes as a proxy for answer quality. To account for time differences, we calculate the ratio of upvotes to the time difference between the posting time and the data collection time. We set the 85th percentile of this ratio (0.0131) as the threshold for quality assessment. Answers with a ratio above this threshold are considered high-quality, while those below it are considered low-quality. *Independent Variable.* We measure user learning behavior by calculating the semantic similarity between answers using the para-phrase-multilingual-MiniLM-L12-v2 model from the SBERT framework [13]. For long texts exceeding 512 tokens, we split them into segment of no more than 512 tokens. We compute the embedding vector for each segment and average them to obtain the final embedding representing of the answer. The semantic similarity between answers is calculated using cosine similarity, with the average similarity between each answer and existing answers for the same question as the measure. Additionally, answer length and the number of images are included as supplementary indicators of user effort. *Moderating Variable.* Answer Information Entropy. The calculation formula is as follows:

$$\text{entropy}(a_j) = -\sum_{i=1}^{k} P(x_i) \log P\,(x_i) \tag{1}$$

Where k is the number of distinct words in the answer, and $P(x_i)$ is $\frac{\text{frequency of a word } i}{\text{total number of words } n \text{ in answer}}$

As shown in Table 1, the table summarizes the model variables and their descriptions.

Table 1. Variable list and measurement

Variable type	Variable	Description
DV	Answer quality	Binary Variable, 1 for high-quality answers, 0 for low-quality answers
IV	User learning behavior	Sort answers by time for the same question. Calculate the average semantic similarity between an answer and existing answers
	Answer length	Length of the answer in characters
	Answer image	Number of images included in the answer
Moderators	Information entropy	Information entropy value of the answer
Controls	User expertise	Total number of upvotes received by the user / (Data collection time - user's first answer time)
	User status	Number of followers the user has / (Data collection time - user's first answer time)
	Answer time position	For the same question, map the answer's posting time order to a range between 0 and 1
	Question popularity	Pageviews of the question / (Data collection time - question posting time)
	Question Type	Binary Variable, 1 for recommendation-type questions, 0 for fact-type questions

The descriptive statistics of the variables are shown in Table 2.

Table 2. Descriptive statistics of variables (N = 141,455)

Variable	Min	Max	Mean	S.D	Skewness
User learning behavior	-0.12	0.97	0.49	0.15	-0.08
Answer length	11.00	7049.00	614.32	1003.31	2.79
Answer image	0.00	179.00	1.19	3.32	8.44
Information entropy	0.00	10.61	5.54	1.97	-0.07
User status	0.00	4453.58	8.44	134.30	28.48
User expertise	0.00	3874.81	12.54	98.07	20.50
Question popularity	1.07	69375.74	1668.89	5346.27	9.83

3.2 Empirical Model

Since the dependent variable "Answer quality" is a binary response variable, this study uses the Generalized Linear Mixed-Effects Model (GLMM) for analysis. Given the nested structure of the data, with answers nested within questions and users, random effects for question ID and user ID are included to control for potential biases. To address significant skewness in some explanatory variables, natural logarithmic transformations are applied to improve the robustness of the coefficient estimates. The research model is as follows:

$$\begin{aligned}\mathrm{logit}(P\,(Is\_good\_answer_{ij}=1)) = \ & \beta_0+\beta_1\cdot \mathrm{User\_learning\_behavior}_{ij}\\ & +\beta_2\cdot \mathrm{Answer\_length}_{ij}\\ & +\beta_3\cdot \mathrm{Answer\_image}_{ij}\\ & +\beta_4\cdot \mathrm{Answer\_information\_entropy}_{ij}\\ & +\beta_5\cdot \mathrm{User\_learning\_behavior}_{ij}\cdot \mathrm{Answer\_information\_entropy}_{ij}\\ & +\beta_6\cdot \mathrm{User\_social\_status}_{j}\\ & +\beta_7\cdot \mathrm{User\_expertise}_{j}\\ & +\beta_8\cdot \mathrm{Answer\_time\_position}_{i}\\ & +\beta_9\cdot \mathrm{Question\_type}_{i}\\ & +\beta_{10}\cdot \mathrm{Question\_popularity}_{i}\\ & +\gamma_i+\gamma_j\end{aligned}$$

Where:

- β_i are fixed effect coefficients;
- $\gamma_i \sim N(0, \sigma^2_{\mathrm{question}})$ represents the random intercept at the question level;
- $\gamma_j \sim N(0, \sigma^2_{\mathrm{user}})$ represents the random intercept at the user level;

3.3 Model Results

As shown in Table 3, in examining the impact of learning behavior on answer quality, both Model 1 and Model 2 show that the coefficient for learning behavior is negative at the 0.001 significance level, supporting Hypothesis H1. This suggests that increased learning behavior reduces the likelihood of an answer being recognized as high-quality, likely due to content repetition and lack of innovation, which leads to reader fatigue. Model 2 introduces an interaction term between learning behavior and answer information entropy, which has a significant positive effect on answer quality, supporting Hypothesis H2. High information entropy mitigates the negative impact of learning behavior, allowing answers with similar content to maintain quality through novelty in expression or thinking. Additionally, both models show that the number of images and answer length positively affect answer quality, consistent with findings by Song et al. (2024) [14], indicating that detailed answers with visual elements are more likely to be considered high-quality.

Table 3. Empirical results on answer quality

Dependent Variable	Answer quality			
	Model 1		Model 2	
	Estimate	Z value	Estimate	Z value
Intercept	-1.59***	-15.04	-1.61***	-15.11
User learning behavior	-0.36***	-18.69	-0.39***	-19.80
Answer length	1.11***	14.54	1.01***	13.10
Answer image	0.35***	27.08	0.34***	26.12
Information entropy	0.64***	8.20	0.78***	9.63
User status	0.22***	7.93	0.83***	31.74
User expertise	0.84***	31.85	0.22***	8.03
Answer time position	-1.73***	-34.25	-1.76***	-34.62
Question popularity	0.36***	10.55	0.35***	10.33
Question type	-0.64***	-5.95	-0.67***	-6.18
User learning behavior * Information entropy			0.10***	6.42
Random effect				
$\sigma^2_{\text{question}}$	1.313		1.340	
σ^2_{user}	2.088		2.057	

Note: ***p < 0.001, **p < 0.01, *p < 0.05.

3.4 Robustness Test

To test robustness, the study adjusts the high-quality answer threshold to the 75th and 80th percentiles and re-estimates the model. As shown in Table 4, the negative effect of learning behavior on answer quality remains significant across all thresholds, confirming its stability. The interaction between learning behavior and information entropy consistently shows a positive effect, validating the robustness of this interaction.

Table 4. Robustness test results on answer quality

Dependent Variable	Answer quality			
	Model 70		Model 80	
	Estimate	Z value	Estimate	Z value
Intercept	-0.27**	-2.68	-0.86***	-8.44
User learning behavior	-0.32***	-20.10	-0.35***	-20.29

Answer length	0.84***	13.53	0.89***	12.93
Answer image	0.30***	27.07	0.32***	27.00
Information entropy	0.79***	12.25	0.83***	11.66
User status	0.31***	12.90	0.27***	10.83
User expertise	0.74***	34.36	0.77***	33.01
Answer time position	-1.73***	-42.20	-1.75***	-38.90
Question popularity	0.25***	7.76	0.30***	9.05
Question type	-0.53***	-5.12	-0.61***	-5.80
User learning behavior * Information entropy	0.07***	5.43	0.09***	6.36
Random effect				
$\sigma^2_{\text{question}}$	1.352		1.317	
σ^2_{user}	1.346		1.644	

Note: ***p < 0.001, **p < 0.01, *p < 0.05.

4 Discussion and Future Directions

4.1 Theoretical Contributions

This study, based on data from Zhihu, applies social cognitive theory and data analysis to explore how user learning behavior impacts answer quality. Unlike previous studies suggesting that learning improves answer quality, our findings reveal a negative correlation, showing that learning behavior can lead to content homogenization and reduced innovation, thereby decreasing quality.

We propose a quantification method using the SBERT model to directly measure user learning behavior through semantic similarity. Additionally, we introduce information entropy as a moderating factor, showing that diverse content can mitigate the negative effects of learning behavior. High information entropy, even with similar content, allows for innovative expression, enhancing answer quality.

4.2 Practical Implications

The study shows that user learning behavior negatively affects answer quality, mainly due to content homogenization. To mitigate this, platforms should optimize recommendation algorithms to avoid overly similar answers, prioritizing those with higher information entropy. These answers, being more diverse and innovative, reduce user fatigue and improve quality. Additionally, platforms could introduce a "Unique Perspective" label to recognize and promote answers with original insights, encouraging more valuable contributions.

4.3 Future Research Directions

Although this study has explored the impact of user learning behavior on answer quality, several areas require further investigation. Future research could focus on improving behavioral analysis methods by using larger models or reinforcement learning to enhance the accuracy of learning behavior analysis in Q&A communities. Additionally, with the rise of AI-Generated Content (AIGC), future studies should explore how it interacts with user learning behavior and its long-term effects on answer quality. This would help optimize the synergy between AI- and user-generated content, ultimately improving community contributions. Moreover, this study focuses on Zhihu, and it is unclear whether the findings apply to other Q&A platforms or are specific to Zhihu' cultural and platform characteristics. Future research should explore the generalizability of these findings across different platforms and cultural contexts. Additionally, future studies should address challenges in measuring information entropy, such as potential cultural biases in entropy perception, to offer a more balanced view.

5 Conclusion

To enhance answer quality in online Q&A communities, this study investigates user learning behaviors through the lens of social cognitive theory, constructs a model of factors influencing answer quality. The SBERT model is employed to calculate semantic similarity between answers under the same question, thereby quantifying user learning behaviors. To account for random effects at both the question and user levels, a generalized linear mixed-effects model (GLMM) is applied for empirical analysis.

The results demonstrate a significant negative correlation between user learning behaviors and answer quality, primarily attributable to content homogeneity and a lack of innovation, which diminish community recognition of answer quality. Information entropy plays a moderating role in this relationship, with higher entropy mitigating the negative impact of learning behaviors. Specifically, introducing novel ideas or employing diverse expressions in answers enhances quality.

Based on these findings, the study recommends that platforms optimize answer recommendation mechanisms by introducing a "unique perspective" label and prioritizing high-quality answers to elevate overall content standards. This research not only proposes a novel operational definition and quantitative measurement method for user learning behaviors but also provides theoretical support and practical recommendations for improving answer recommendation systems and user experiences on online Q&A platforms.

Acknowledgements

This research is supported by Beijing Natural Science Foundation ("The Research on the Individual Learning Behavior in Online Knowledge Sharing Community and Its Effect on the Performance of Knowledge Sharing based on the Dynamic Perspective" (9242001)).

References

1. LE L T, SHAH C. Retrieving people: Identifying potential answerers in Community Question-Answering [J]. Journal of the Association for Information Science and Technology, 2018, 69(10): 1246-58.
2. WU J, XU Z, HUANG Q, Cai J. What motivates knowledge sharing? Evaluating the quality of answer contribution in online Q&A communities [J]. Data and Information Management, 2024.
3. WANG N, LIU Y, XIAO S. Which feedback matters? The role of expressions and valence in continuous high-quality knowledge contribution in the online Q&A community [J]. Decision Support Systems, 2022, 156: 113750.
4. BANDURA A. Social foundations of thought and action [M]. Englewood Cliffs, NJ: Prentice-Hall, 1986.
5. JIN Y, TAN Y, HUANG J. Managing contributor performance in knowledge-sharing communities: A dynamic perspective [J]. Production and Operations Management, 2022, 31(11): 3945-62.
6. SHI C, HU P, FAN W, Qiu L. How learning effects influence knowledge contribution in online Q&A community? A social cognitive perspective [J]. Decision Support Systems, 2021, 149: 113610.
7. KRISHNA V, SHRESTHA Y R, VON KROGH G. “The double-edged sword of inflated help”: Unravelling the motivation crowding in community question-answering platforms [J]. PLOS ONE, 2024, 19(3): e0297627.
8. SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27(3): 379-423.
9. FRESNEDA J E, GEFEN D. A semantic measure of online review helpfulness and the importance of message entropy [J]. Decision Support Systems, 2019, 125: 113117.
10. SUN W, GUO X, KWOK R C-W. Emotion or information? Differences and relationships between two types of professional support in online health communities [J]. Information & Management, 2024, 61(3): 103944.
11. PATIL S, LEE K. Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors [J]. Social Network Analysis and Mining, 2015, 6(1): 5.
12. Bai, Yang. Zhihu 2023 revenue of 4.199 billion yuan, net loss reduced by 44.9%. Eastmoney. Retrieved from http://www.eastmoney.com/a/202403263024801060.html, last accessed March 26, 2024.
13. ZHOU L, TAO J, ZHANG D. Does Fake News in Different Languages Tell the Same Story? An Analysis of Multi-level Thematic and Emotional Characteristics of News about COVID-19 [J]. Information Systems Frontiers, 2023, 25(2): 493-512.
14. SONG Y, XU X, DUTTA K, Li Z. Improving answer quality using image-text coherence on social Q&A sites [J]. Decision Support Systems, 2024, 180: 114191.