https://doi.org/10.51408/issi2025_191

The Effectiveness of Large Language Models in Predicting User Question Preferences on ResearchGate Q&A

Lei Li[1], Yue Hu[2], Hui Peng[3]

[1]leili@bnu.edu.cn,[2]202221260043@mail.bnu.edu.cn,[3]hui_peng@mail.bnu.edu.cn
Department of Information Management, School of government, Beijing Normal University, Beijing (China)

Introduction

In the digital era, academic social platforms such as ResearchGate have become crucial venues for experts, scholars, and researchers across various fields to pose academic questions and receive high-quality answers. Therefore, predicting public preferences for academic questions can help platforms recommend content more accurately, enhance user experience, and assist researchers in understanding current research trends, thereby promoting academic exchange and development. In recent years, methods for predicting user preferences for online-generated content have primarily relied on machine learning algorithms based on feature engineering (Liao et al., 2019), which require high accuracy in feature selection and have limitations in terms of method portability and prediction accuracy. The rapid development of large language models (LLMs) has driven their widespread application across various domains. One of the most promising areas is the use of LLMs for text comprehension and assessment, commonly referred to as "LLMs-as-judges" (Li et al., 2024). This study essentially leverages LLMs to evaluate the popularity of academic questions. The advantages of LLMs in text understanding and assessment provide new possibilities for predicting user preferences for academic questions, offering the potential to reduce excessive reliance on manually crafted features and improve prediction accuracy.

Moreover, unlike general social media platforms, academic Q&A websites place greater emphasis on gaining inspiration and acquiring knowledge of interest through questions. As a result, user question preferences largely depend on the content itself. Leveraging LLMs to predict question popularity by deeply understanding and extracting insights from question content may yield better results. Therefore, this study collects question data from multiple disciplines on ResearchGate Q&A, processes the semantic information in question texts using LLMs, and employs fine-tuning techniques to build a predictive model for user academic question preferences. This approach aims to reveal the potential applications of LLMs in this evaluation task.

Data collection

On the ResearchGate Q&A platform, questioners typically add multiple topic tags to their questions to attract scholars with similar research interests to participate in discussions. This study selected ten specific academic topics from the platform's popular themes, ensuring comprehensive coverage of the five major subject categories in the Web of Science (WOS). Additionally, the broad topics of "learning" and "scientific research" were included to ensure diversity in question types and sufficient data volume, enabling a comprehensive evaluation of LLMs in predicting user academic question preferences.

A Python web crawler was used to collect all questions under these twelve topics, including details such as question titles, descriptions, posting times, view counts, follow counts, answer counts, and recommendation counts. From this dataset, 10,000 questions were randomly sampled for analysis, with the number of questions for each of the 12 topics shown in Table 1.

Table 1. Number of Questions in Each Topic.

Topic	No.	Topic	No.
Molecular Biology	600	Computer Science	111 7
Ethics	404	English	337
Chemistry	170 7	Learning	644
Philosophy	132 2	Journalism	130 8
Economics	336	Social Sciences	411
Psychology	568	Scientific Research	124 6

Model training

Based on mainstream definitions of preferences and features extracted from ResearchGate Q&A, four metrics were selected to measure user preferences: question views, follows, answers, and recommendations. To account for temporal effects, the number of months between the question posting date and data collection date was calculated. If the interval was less than one month, it was recorded as 1 month. Each metric (views, follows, answers, recommendations) was divided by the time interval (in months) and then standardized to derive normalized scores. A pairwise correlation analysis of the four metrics revealed low inter-metric correlations ($r < 0.7$, $p < 0.01$). Consequently, the public preference score for each question was defined as the sum of the time-averaged and standardized values of views, follows, answers, and recommendations. Questions were proportionally divided into "high" and "low" preference tiers based on their aggregated scores, resulting in 5,000 popular questions and 5,000 unpopular questions.

The subsequent step involved constructing the fine-tuning dataset. Each data instance comprised three components: a prompt, an input, and an output. The input consisted of the textual content of each question, which was further divided into two configurations for comparative analysis: (1) question title only, and (2) question title combined with its detailed description. This dual-input approach was designed to evaluate the impact of varying contextual information on prediction performance. The output represented the public preference level for the question, categorized as either "high" or "low." A total of 10,000 questions were randomly sampled and split into training and testing sets at an 8:2 ratio (8,000 for training and 2,000 for testing), with equal representation of both preference categories in each subset to ensure class balance.

Finally, three widely adopted and high-performing base models—GPT-4o-mini (OpenAI), DeepSeek-R1-Distill-Qwen-7B (DeepSeek), and Gemini 1.5 Flash (Google)—were selected for experimentation. The training set was used to fine-tune these models, followed by performance testing to assess their predictive capabilities.

Table 2. Performance Evaluation Results of the Models.

Input	Model name	Popularity Level	Acc (%)	F1 (%)	P (%)	R (%)
Title	GPT-4o-mini	high	70.4	71.4	69.3	73.6
		low		69.4	71.8	67.1
		average		70.4	70.5	70.4
	DeepSeek-R1-Distill-Qwen-7B	high	71.0	70.5	71.7	69.3
		low		71.5	70.3	72.7
		average		71.0	71.0	71.0
	Gemini 1.5 Flash	high	67.8	71.7	63.9	81.7
		low		62.6	74.7	53.9
		average		67.2	69.3	67.8
Title + Description	GPT-4o-mini	high	71.7	72.8	69.9	76.0
		low		70.4	73.8	67.3
		average		71.6	71.9	71.7
	DeepSeek-R1-Distill-Qwen-7B	high	72.7	71.8	74.2	69.6
		low		73.5	71.4	75.8
		average		72.7	72.8	72.7
	Gemini 1.5 Flash	high	72.6	73.6	70.9	76.5
		low		71.4	74.5	68.6
		average		72.5	72.7	72.6

Results

The performance evaluation results of the models are summarized in Table 2. These findings indicate that LLMs exhibit promising potential in predicting user preferences for academic questions, achieving an average prediction accuracy of approximately 70%. Specifically, in the task of predicting preferences based solely on question titles, the DeepSeek-R1-Distill-Qwen-7B model

delivered the best performance, with an accuracy of 71%, while Gemini 1.5 Flash showed comparatively weaker results, achieving 67.8% accuracy. When the input context was expanded from titles only to titles + descriptions, all three models exhibited performance improvements. This confirms that providing richer contextual information enhances LLMs' predictive capabilities. Notably, Gemini 1.5 Flash demonstrated the highest improvement, with a 4.8% increase in accuracy. In contrast, DeepSeek-R1-Distill-Qwen-7B showed a more modest gain of approximately 1% when supplemented with descriptive text. These findings suggest that DeepSeek and GPT-4o-mini may excel at processing concise title-based inputs, where additional detailed information from longer question descriptions contributes marginally to accuracy. Conversely, Gemini 1.5 Flash appears better equipped to leverage complex inputs, effectively integrating both titles and descriptions to refine its predictions.

Conclusions

This study investigates the feasibility and accuracy of LLMs in predicting users' academic information preferences based on textual content from questions on ResearchGate Q&A, an academic social Q&A platform. The findings reveal that, compared to traditional machine learning algorithms reliant on feature engineering, LLMs achieve higher accuracy in predicting user preferences (Li et al., 2015; Li et al., 2020), and providing richer textual information (e.g., question descriptions) positively enhances their predictive performance. Among the tested models, DeepSeek-R1-Distill-Qwen-7B delivered the best results under both input conditions (title-only and title+description), while Gemini 1.5 Flash demonstrated the most significant performance improvement (4.8%) when additional detailed context was introduced. In conclusion, this work validates the preliminary efficacy of LLMs in predicting academic information preferences and provides insights for optimizing LLMs-as-judges in diverse application scenarios. Future research could incorporate external question features, such as the objective attributes of question askers (e.g., expertise, institutional affiliation), to enable more precise question popularity prediction. Furthermore, extending the evaluation of LLMs' predictive capabilities to other social media platforms would strengthen the generalizability of these findings.

References

Liao, D., Xu, J., Li, G., Huang, W., Liu, W., & Li, J. (2019, July). Popularity prediction on online articles with deep fusion of temporal process and content features. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 200-207).

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Li, L., He, D., Jeng, W., Goodwin, S., & Zhang, C. (2015, May). Answer quality characteristics and prediction on an academic Q&A Site: A case study on ResearchGate. In *Proceedings of the 24th international conference on world wide web* (pp. 1453-1458).