

Analyzing Stack Overflow Question Types and Answer Characteristics: Implications for Designing and Benchmarking Coding Assistant LLMs

Omid Mohammadi Kia*, Mahmood Neshati

*- Corresponding author: Faculty Member / Researcher in Computer Science and Engineering
Shahid Beheshti Faculty of Computer Science and Engineering
Shahid Beheshti University, Tehran, Iran

2- Faculty Member / Researcher in Computer Science and Engineering
Shahid Beheshti Faculty of Computer Science and Engineering
Shahid Beheshti University, Tehran, Iran

Abstract

Stack Overflow is a central source of programming knowledge and a key training and evaluation resource for large language model (LLM)-based coding assistants. Because developer questions differ in intent, effective answers vary in structure and content. We analyze 50,000 high-engagement Stack Overflow questions from 2020–2024, categorizing them into three developer intents—conceptual, debugging, and best practice—using manual annotation and a fine-tuned RoBERTa classifier. For each category, we examine answer characteristics including length, code usage, images, and external references. Our findings show consistent patterns: debugging questions receive short, code-focused answers; conceptual questions favor longer, reference-supported explanations; and best-practice questions combine code with visual or explanatory elements. These structural differences are strongly associated with answer acceptance and perceived usefulness. We discuss how incorporating question-type awareness can improve adaptive LLM prompting, response formatting, and stratified benchmarking beyond code correctness, and we release an annotated dataset to support future research on context-aware coding assistants.

Research method

The study used a hybrid methodology combining manual annotation and machine learning-based classification. Data were collected from the official Stack Overflow Data Dump, including 50,000 high-quality questions with accepted answers and sufficient engagement. Questions were categorized into conceptual, debugging, and best-practice classes. A subset of 5,000 questions was manually annotated by three software engineering experts to create a gold-standard dataset. A RoBERTa-based classifier was then fine-tuned using the annotated data and achieved an F1-score of 89% for automatic question classification. Answer characteristics such as:

- answer length,
- code snippet presence,
- image inclusion,
- external references
- and code complexity

were automatically extracted using NLP and HTML parsing tools including spaCy, BeautifulSoup, and NLTK. Statistical analyses including Chi-square tests, ANOVA, and logistic regression were applied to evaluate relationships between question types and answer effectiveness.

conclusion

This study analyzes 50,000 high-engagement Stack Overflow posts (2020–2024) to examine how answer characteristics vary across three question types: conceptual, debugging, and best practice. By systematically evaluating answer length, code usage, images, and references, we show that effective responses align closely with developer intent—debugging questions favor concise, code-focused answers; conceptual questions require longer, reference-supported explanations; and best-practice questions combine multiple modalities.

Using a hybrid annotation strategy and a RoBERTa-based classifier (F1 = 0.89), we enable scalable and reliable question categorization. Our findings reveal statistically significant relationships between question type, answer format, and acceptance outcomes, and we release an annotated dataset to support further research.

These results inform the design and evaluation of LLM-based coding assistants by motivating intent-aware prompting, adaptive response formatting, and stratified benchmarks that extend beyond code correctness. Future work includes real-time intent detection, cross-platform validation, and user-centered evaluation of LLM-generated responses.

The purpose of the research

This research investigates how different types of Stack Overflow questions influence the structure and characteristics of effective answers. The study focuses on three major developer intents: conceptual questions, debugging questions, and best-practice questions. By analyzing 50,000 high-engagement Stack Overflow posts from 2020–2024, the research aims to identify patterns in answer length, code usage, images, and external references. The findings are intended to support the design of more adaptive and context-aware LLM-based coding assistants and to improve benchmarking methods beyond simple code correctness evaluation.

Practical or simulation results

The experimental results revealed strong relationships between developer intent and effective answer structure:

Debugging questions received short and highly code-focused answers.

- 91% contained code snippets
- Average answer length: 112 words
- Highest acceptance rate: 75%

Conceptual questions required longer explanatory responses with external references.

- Average length: 512 words
- 82% contained references
- Lower code usage (28%)

Best-practice questions combined multiple modalities including code, diagrams, and references.

- 72% included code snippets
- 56% included images
- Average length: 285 words

Statistical analysis confirmed significant associations between question type and answer characteristics ($p < 0.001$). Logistic regression showed that code inclusion significantly increased answer acceptance for debugging questions, while references improved acceptance for conceptual questions. The study demonstrates that adaptive, intent-aware response generation can substantially improve the effectiveness of LLM-based coding assistants and provides a foundation for future benchmarking frameworks that consider contextual and structural response quality.

References

[1] S. Mustafa and W. Zhang, “Why Do I Share? Participants’ Personality Traits and Online Participation,” *Int J Hum Comput Interact*, vol. 40, no. 14, pp. 3763–3781, 2024, doi: 10.1080/10447318.2023.2201551.

[2] C. Fu, X. Yue, B. Shen, S. Yu, and Y. Min, “Patterns of interest change in stack overflow,” *Scientific Reports* 2022 12:1, vol. 12, no. 1, pp. 1–10, Jul. 2022, doi: 10.1038/s41598-022-15724-3.

[3] S. Mustafa, W. Zhang, and M. M. Naveed, “How to mend the dormant user in Q&A communities? A social cognitive theory-based study of consistent geeks of StackOverflow,” *Behaviour & Information Technology*, vol. 43, no. 10, pp. 2024–2043, Jul. 2024, doi: 10.1080/0144929X.2023.2237604.

[4] M. Chen et al., “Evaluating large language models trained on code,” *arxiv.org* M Chen, J Tworek, H Jun, Q Yuan, HPDO Pinto, J Kaplan, H Edwards, Y Burda, N JosepharXiv preprint arXiv:2107.03374, 2021•arxiv.org, Accessed: Nov. 01, 2025. [Online]. Available: <https://arxiv.org/abs/2107.03374>

[5] A. Barua, S. Thomas, A. H.-E. software engineering, and undefined 2014, “What are developers talking about? an analysis of topics and trends in stack overflow,” *SpringerA Barua, SW Thomas, AE HassanEmpirical software engineering*, 2014•Springer, Accessed: Nov. 01, 2025. [Online]. Available: <https://link.springer.com/article/10.1007/s10664-012-9231-y>