Detecting Missing Information of Questions in Q&A Chatrooms

***Abstract*—In developer chatrooms, questions lacking essential information often result in longer response times and lower response rates, posing a significant challenge for newcomers who are unfamiliar with the topics or interaction patterns. To address this issue and facilitate the integration of newcomers into the community, we have identified two key types of information necessary for well-formed questions based on Stack Overflow guidelines: Problem Statement (PS) and Expect to Do (ETD). We manually analyzed 2,000 chatroom questions from eight active domains (channels) and discovered that developers use 66 distinct *lexico-syntactic patterns* to describe ETD and PS, highlighting the potential for automated detection of such content. Based on these findings, we designed and evaluated four automated approaches to detect missing ETD and/or PS in chatroom questions, including two baselines, pre-trained models, and Chain-of-Thought (CoT) prompting on Large Language Models (LLMs). In their optimal configurations, *GPT-4o*, BERT, and ELECTRA outperform other approaches, being the only models to achieve an F1 score exceeding 80% in detecting missing ETD and PS. Analysis of the outputs generated by *GPT-4o* demonstrates that LLMs have the potential to automatically annotate *lexico-syntactic patterns* similar to those produced manually through open coding.**

***Index Terms*—Chatroom, Lexico-syntactic Pattern, Pre-trained, LLM, CoT Prompt**

APPENDIX

A. Performance of pre-trained models on reduced validation set

In addition to the 10-fold cross-validation (10CV) experiments on 1,000 chatroom questions with pre-trained models presented in the main paper, we conducted another set of experiments to evaluate the performance of these models when reducing the size of the *validation set* from 1,000 to 200, decreasing by 200 each time. Given the limited size of the reduced *validation set*, we employed a 5-fold cross-validation (5CV) strategy, allocating 80%, 10%, and 10% of the data for training, parameter tuning, and testing per fold, respectively.

According to Figure 1, F1 score dramatically decreased when the *validation set* size was reduced to 400 for detecting missing ETD and 600 for detecting missing PS. However, the F1 score remained almost unchanged for detecting missing PS when the *validation set* size reduced from 1000 to 600, showing only a 10%-15% drop in F1 score between the best and worst cases. This indicates a certain level of robustness to the reduced *validation set* size for detecting missing PS. In contrast, the F1 score for detecting missing ETD exhibited a linear decrease before the dramatic drop at a *validation set* size of 400, with a 20%-40% decline in F1 score between the best and worst cases.

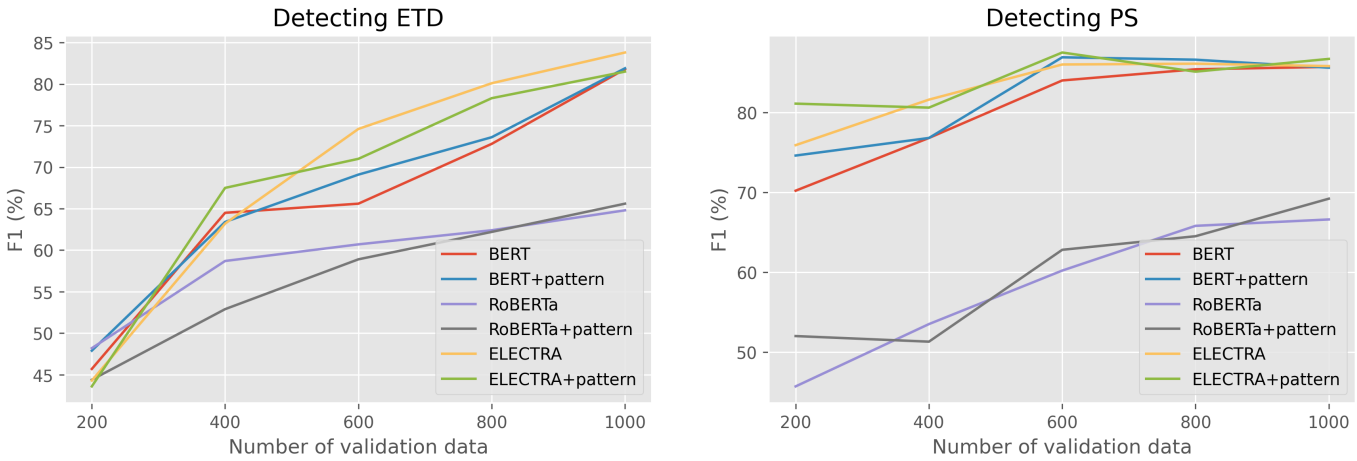

Fig. 1. Distribution of detecting missing ETD and PS for pre-trained models when reducing the size of *validation set*