Technical Appendix

Anonymous submission

Data filtering

We apply the following filtering rules to improve the quality of the training corpus:

- Exclude samples where the query is not in English.
- Exclude samples where the query is less than 3 tokens or more than 256 tokens.
- Exclude samples where the function name contains "test".
- Exclude constructor samples.
- Exclude samples where the function body is fewer than 3 lines.

The statistical results of the pre-training data are shown in the Table 1.

Table 1: Statistics of training data.

Language	#Functions
Python	714,352
Java	718,902
Go	715,664
Php	707,424
JavaScript	707,778
Ruby	64,157
Total	3,628,277