Impact of design choices on the quality of multi-modal transformer based embedding for bug localization: Appendix

A.1 Project Description

Table 1: Brief description of the projects included in the dataset

Project Name	**Brief Description**
AspectJ	Programming language
Birt	Data visualization tool
Eclipse UI	User interfcae framework
JDT	IDE plugin framework
SWT	User interface framework
Tomcat	HTTP server
Hadoop	Distributed processing framework
Netbeans	IDE
Storm	Distributed computing system
Kafka	Distributed streaming platform
Logging-log4j2	Logging library
RxJava	Asynchronous event based programming framework
Godot	Game engine
Guava	Java collection library
Libgdx	Game development framework
Lucene-solr	Search library
Groovy	Programming language
Elasticsearch	Distributed search engine
Openjpa	Persistence framework
Beam	Data streaming library
Fresco	Image manipulation library

Jena	Semantic web framework
Aspnetcore	Cross-platform web framework
Spring-boot	Web framework
Maven	Project and dependency management tool
Ignite	Distributed database

A.2 MRR of the embedding models

Table 2: MRR of the embedding models.

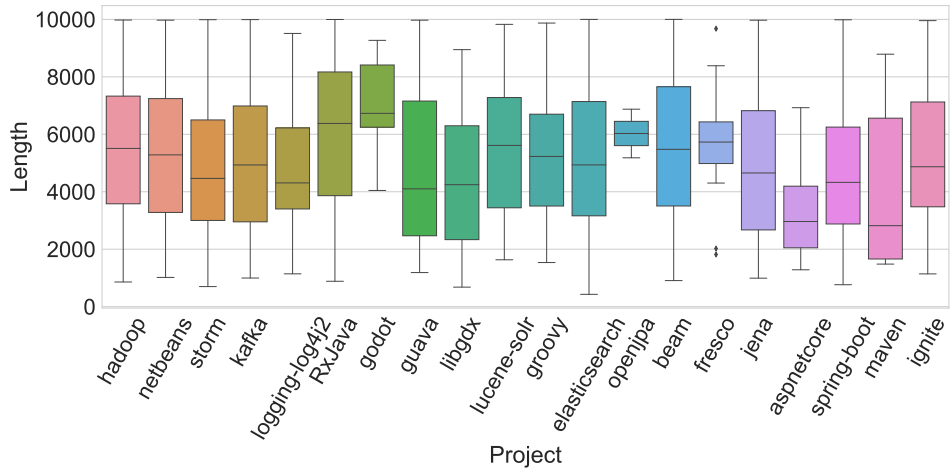
Model Name (RQ3)	Training Strategy (RQ2)	Training Data (RQ1)	MRR for Test Data						
			AspectJ	Birt	Eclipse	JDT	SWT	Tomcat	Overall
Long RoBERTa	Electra	BLDS	0.41	0.37	0.35	0.37	0.31	0.33	0.35
		Bench-BLDS	0.37	0.28	0.40	0.49	0.40	0.40	0.39
	MLM	BLDS	0.28	0.25	0.28	0.33	0.33	0.26	0.29
		Bench-BLDS	0.24	0.25	0.32	0.23	0.35	0.25	0.27
	QA	BLDS	0.16	0.27	0.20	0.13	0.21	0.23	0.20
		Bench-BLDS	0.42	0.43	0.41	0.54	0.40	0.45	0.44
Reformer	Electra	BLDS	0.32	0.29	0.31	0.33	0.37	0.31	0.32
		Bench-BLDS	0.41	0.26	0.35	0.36	0.32	0.28	0.33
	MLM	BLDS	0.13	0.18	0.17	0.15	0.18	0.21	0.17
		Bench-BLDS	0.53	0.34	0.45	0.52	0.42	0.40	0.44
	QA	BLDS	0.16	0.19	0.21	0.21	0.20	0.22	0.20
		Bench-BLDS	0.54	0.46	0.46	0.49	0.41	0.36	0.46
Long CodeBERT	Electra	BLDS	0.45	0.22	0.32	0.33	0.23	0.20	0.31
		Bench-BLDS	0.61	0.31	0.42	0.54	0.45	0.41	0.47
	MLM	BLDS	0.53	0.39	0.46	0.60	0.32	0.36	0.46
		Bench-BLDS	0.55	0.47	0.46	0.51	0.34	0.29	0.46
	QA	BLDS	0.23	0.31	0.34	0.23	0.20	0.22	0.26
		Bench-BLDS	0.27	0.28	0.32	0.19	0.24	0.24	0.26
	Without training	BLDS	0.24	0.23	0.21	0.21	0.22	0.28	0.22
		Bench-BLDS	0.34	0.36	0.34	0.24	0.30	0.28	0.32
CodeBERT	Electra	BLDS	0.13	0.25	0.24	0.14	0.28	0.27	0.21
		Bench-BLDS	0.28	0.24	0.37	0.26	0.31	0.19	0.29
	MLM	BLDS	0.36	0.24	0.27	0.34	0.24	0.26	0.29
		Bench-BLDS	0.24	0.23	0.30	0.29	0.30	0.20	0.27
	QA	BLDS	0.16	0.24	0.34	0.17	0.22	0.23	0.23
		Bench-BLDS	0.26	0.24	0.31	0.31	0.29	0.21	0.28
	Without training	BLDS	0.37	0.25	0.26	0.32	0.24	0.20	0.29
		Bench-BLDS	0.13	0.27	0.30	0.17	0.34	0.25	0.24

A.3 MAP of the embedding models

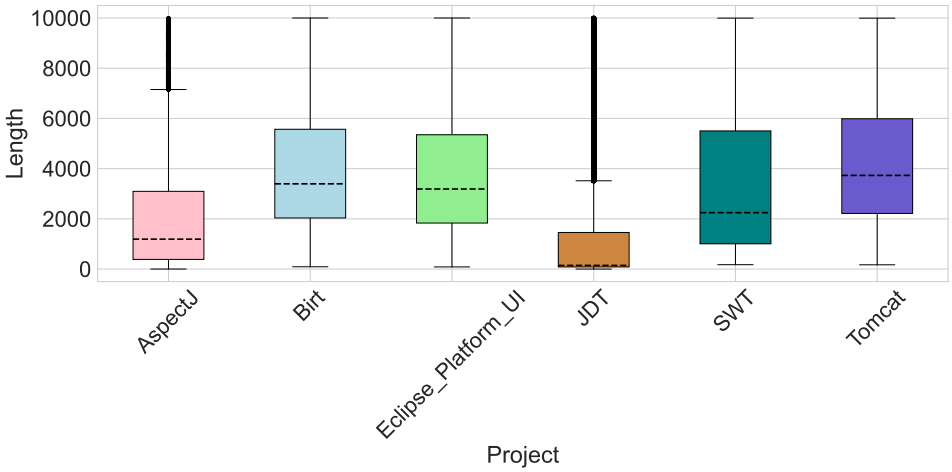
Table 3: MAP of the embedding models.

Model Name (RQ3)	Training Strategy (RQ2)	Training Data (RQ1)	MRR for Test Data						
			AspectJ	Birt	Eclipse	JDT	SWT	Tomcat	Overall
Long RoBERTa	Electra	BLDS	0.33	0.28	0.28	0.27	0.21	0.23	0.26
		Bench-BLDS	0.29	0.21	0.29	0.38	0.31	0.30	0.28
	MLM	BLDS	0.17	0.18	0.18	0.25	0.23	0.16	0.18
		Bench-BLDS	0.15	0.17	0.22	0.13	0.28	0.15	0.17
	QA	BLDS	0.06	0.18	0.12	0.03	0.11	0.15	0.11
		Bench-BLDS	0.34	0.35	0.34	0.44	0.29	0.36	0.33
Reformer	Electra	BLDS	0.23	0.21	0.21	0.24	0.28	0.20	0.22
		Bench-BLDS	0.31	0.17	0.25	0.26	0.26	0.20	0.23
	MLM	BLDS	0.06	0.07	0.11	0.07	0.09	0.12	0.07
		Bench-BLDS	0.45	0.25	0.36	0.42	0.34	0.32	0.35
	QA	BLDS	0.07	0.11	0.12	0.10	0.13	0.12	0.09
		Bench-BLDS	0.46	0.35	0.36	0.39	0.35	0.28	0.36
Long CodeBERT	Electra	BLDS	0.36	0.14	0.22	0.25	0.13	0.12	0.22
		Bench-BLDS	0.51	0.23	0.35	0.43	0.35	0.33	0.36
	MLM	BLDS	0.43	0.30	0.36	0.49	0.25	0.26	0.36
		Bench-BLDS	0.45	0.37	0.38	0.43	0.27	0.22	0.37
	QA	BLDS	0.14	0.22	0.23	0.13	0.13	0.13	0.15
		Bench-BLDS	0.19	0.21	0.26	0.11	0.16	0.14	0.16
	Without training	BLDS	0.14	0.13	0.15	0.12	0.12	0.20	0.13
		Bench-BLDS	0.27	0.27	0.27	0.14	0.23	0.20	0.21
CodeBERT	Electra	BLDS	0.06	0.18	0.17	0.06	0.18	0.18	0.10
		Bench-BLDS	0.21	0.15	0.30	0.16	0.24	0.08	0.19
	MLM	BLDS	0.27	0.15	0.18	0.26	0.16	0.17	0.19
		Bench-BLDS	0.15	0.16	0.22	0.21	0.21	0.09	0.17
	QA	BLDS	0.09	0.15	0.24	0.08	0.13	0.12	0.12
		Bench-BLDS	0.17	0.17	0.21	0.22	0.21	0.12	0.17
	Without training	BLDS	0.28	0.15	0.17	0.23	0.16	0.09	0.18
		Bench-BLDS	0.03	0.17	0.22	0.08	0.24	0.15	0.14

A.4 Length distribution in the dataset



(a) BLDS Dataset



(b) Bench-BLDS Dataset

Figure 1: Length distribution of the source code files in dataset