Appendix: Automatic Commit Message Generation: A Critical Review and Directions for Future Work

Yuxia Zhang, Zhiqing Qiu, Klaas-Jan Stol, Wenhui Zhu, Jiaxin Zhu, Yingchen Tian, and Hui Liu

Appendix A
RQ2: Statistical analysis of the impact of small commit size

The following four tables show the results of Mann–Whitney U test and the effect size between the three approaches' performance in their original testing set and other testing sets with larger diffs. In each table, the first part indicates the corresponding model is trained on 'New Benchmark' in RQ1, i.e., diff lengths of training commits are limited to (0, 100] (and max. 200 for FIRA); the models in the second part are trained on commits whose diff length is between (0, 632] (and max. 512 for CCRep).

Appendix B
Developer Survey

Current learning or retrieving-based approaches can generate a short sentence (less than 30 characters) as commit message for a given code change, i.e., it is the 'subject' of a commit message. Our research team is eager to hear your voice on the usefulness of generating subjects for commits. Your response will greatly impact the future research direction of this task.

The survey consists of 9 questions and should take less than 5 minutes to finish. We would much appreciate it if you could please answer them (note all the answers will be kept anonymous and only for research purposes).

We may ask further questions if we encounter any problems while processing your response.

Q1. How many years have you participated in open source software?

a) Less than three months
b) More than three months but less than one year
c) More than one year but less than three years
d) More than three years

Q2. What is the hardest work when writing commit messages?

a) Writing the subject of a commit message
b) Writing the body of a commit message
c) Other:

Q3. Please explain your choice for Q2: _____

Q4. What is the most time-consuming part when writing commit messages?

a) Writing the subject of a commit message
b) Writing the body of a commit message
c) Other:

Q5. Please explain your choice for Q4. _____

Q6. What do you usually write in the subject of a commit message?

a) Describe what was changed
b) Explain why made this change
c) Other:

Q7. If a tool can automatically generate a short subject for your code change, please rate its usefulness (ranging from 0 to 4, and a higher score indicates higher usefulness).

Q8. Please explain your choice for Q7: _____

Q9. Please tell us your further advice on commit message generation.

Appendix C
List of papers containing Bot Accounts

The following papers were used to search for bot accounts: [1–26]

References

[1] B. Baudry, Z. Chen, K. Etemadi, H. Fu, D. Ginelli, S. Kommrusch, M. Martinez, M. Monperrus, J. Ron, H. Ye *et al.*, "A software-repair robot based on continual learning," *IEEE Software*, vol. 38, no. 4, pp. 28–35, 2021.

- *Yuxia Zhang, ZhiQing Qiu, Wenhui Zhu, and Hui Liu are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. E-mail: yuxiazh@bit.edu.cn*
- *Klaas-Jan Stol is with Lero, the Science Foundation Ireland Research Centre for Software and the School of Computer Science and IT, University College Cork, Ireland.*
- *Jiaxin Zhu is with the Institute of Software, Chinese Academy of Sciences, Beijing, China.*
- *Yingchen Tian is with Tmall Technology Co., Zhejiang, China.*

TABLE 1: Significance and Effect size of NNGen's testing set with diff length [0,100] compared to other five testing sets

	[100-200]		[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.12	<0.001	0.11	<0.001	0.11	<0.001	0.10	<0.001	0.10
ROUGE-L	<0.001	0.32	<0.001	0.34	<0.001	0.36	<0.001	0.37	<0.001	0.37
METEOR	<0.001	0.33	<0.001	0.37	<0.001	0.36	<0.001	0.36	<0.001	0.36

[1] The length of the training set diff is (0, 100].

	[100-200]		[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.07	<0.001	0.07	<0.001	0.09	<0.001	0.07	<0.001	0.11
ROUGE-L	<0.001	0.2	<0.001	0.22	<0.001	0.24	<0.001	0.22	<0.001	0.27
METEOR	<0.001	0.21	<0.001	0.22	<0.001	0.24	<0.001	0.22	<0.001	0.28

[1] The length of the training set diff is (0, 632].

TABLE 2: Significance and Effect size of CoRec's testing set with diff length [0,100] compared to other five testing sets

	[100-200]		[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.12	<0.001	0.13	<0.001	0.14	<0.001	0.14	<0.001	0.15
ROUGE-L	<0.001	0.31	<0.001	0.34	<0.001	0.35	<0.001	0.36	<0.001	0.36
METEOR	<0.001	0.31	<0.001	0.35	<0.001	0.36	<0.001	0.37	<0.001	0.38

[1] The length of the training set diff is (0, 100].

	[100-200]		[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.07	<0.001	0.07	<0.001	0.09	<0.001	0.07	<0.001	0.11
ROUGE-L	<0.001	0.2	<0.001	0.22	<0.001	0.24	<0.001	0.22	<0.001	0.27
METEOR	<0.001	0.21	<0.001	0.22	<0.001	0.24	<0.001	0.22	<0.001	0.28

[1] The length of the training set diff is (0, 632].

TABLE 3: Significance and Effect size of FIRA's testing set with diff length [0,200] compared to other five testing sets

	[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.08	<0.001	0.10	<0.001	0.10	<0.001	0.10
ROUGE-L	<0.001	0.07	<0.001	0.09	<0.001	0.08	<0.001	0.09
METEOR	<0.001	0.07	<0.001	0.09	<0.001	0.08	<0.001	0.09

[1] The length of the training set diff is (0, 200].

	[200-300]		[300-400]		[400-500]		[500-632]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.07	<0.001	0.10	<0.001	0.10	<0.001	0.09
ROUGE-L	<0.001	0.07	<0.001	0.09	<0.001	0.07	<0.001	0.08
METEOR	<0.001	0.06	<0.001	0.08	<0.001	0.06	<0.001	0.07

[1] The length of the training set diff is (0, 632].

[2] A. Carvalho, W. Luz, D. Marcilio, R. Bonifácio, G. Pinto, and E. D. Canedo, "C-3pr: a bot for fixing static analysis violations via pull requests," in *2020 IEEE 27th International conference on software analysis, evolution and reengineering*. IEEE, 2020, pp. 161–171.

[3] M. Cernat, A. N. Staicu, and A. Stefanescu, "Towards automated testing of rpa implementations," in *11th ACM SIGSOFT International Workshop on Automating TEST Case Design, Selection, and Evaluation*, 2020, pp. 21–24.

[4] B. da Silva, C. Hebert, A. Rawka, and S. Sereesathien, "Robin: a voice controlled virtual teammate for software developers and teams," in *2020 IEEE International Conference on Software Maintenance and Evolution*. IEEE, 2020, pp. 789–791.

TABLE 4: Significance and Effect size of CCRep's testing set with diff length [0,100] compared to other five testing sets

	[100-200]		[200-300]		[300-400]		[400-512]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.55	<0.001	0.62	<0.001	0.62	<0.001	0.63
ROUGE-L	<0.001	0.55	<0.001	0.63	<0.001	0.63	<0.001	0.65
METEOR	<0.001	0.57	<0.001	0.65	<0.001	0.65	<0.001	0.67

[1] The length of the training set diff is (0, 100].

	[100-200]		[200-300]		[300-400]		[400-512]	
	p-value	effect	p-value	effect	p-value	effect	p-value	effect
BLUE	<0.001	0.21	<0.001	0.26	<0.001	0.32	<0.001	0.33
ROUGE-L	<0.001	0.22	<0.001	0.25	<0.001	0.31	<0.001	0.32
METEOR	<0.001	0.22	<0.001	0.25	<0.001	0.30	<0.001	0.33

[1] The length of the training set diff is (0, 512].

[5] H. Ed-Douibi, G. Daniel, and J. Cabot, "Openapi bot: A chatbot to help you understand rest apis," in *Web Engineering: 20th International Conference, ICWE 2020, Helsinki, Finland, June 9–12, 2020, Proceedings*. Springer, 2020, pp. 538–542.

[6] L. Erlenhov, F. G. de Oliveira Neto, M. Chukaleski, and S. Daknache, "Challenges and guidelines on designing test cases for test bots," in *IEEE/ACM 42nd International Conference on Software Engineering workshops*, 2020, pp. 41–45.

[7] L. Erlenhov, F. G. D. O. Neto, and P. Leitner, "An empirical study of bots in software development: Characteristics and challenges from a practitioner's perspective," in *28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 445–455.

[8] C. Lebeuf, A. Zagalsky, M. Foucault, and M.-A. Storey, "Defining and classifying software bots: A faceted taxonomy," in *2019 IEEE/ACM 1st international workshop on bots in software engineering*. IEEE, 2019, pp. 1–6.

[9] C. R. Lebeuf, "A taxonomy of software bots: towards a deeper understanding of software bot characteristics," Ph.D. dissertation, University of Victoria, 2018.

[10] C. Matthies, F. Dobrigkeit, and G. Hesse, "An additional set of (automated) eyes: chatbots for agile retrospectives," in *2019 IEEE/ACM 1st international workshop on bots in software engineering*. IEEE, 2019, pp. 34–37.

[11] M.-A. Storey and A. Zagalsky, "Disrupting developer productivity one bot at a time," in *2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 2016, pp. 928–931.

[12] Y. Tian, F. Thung, A. Sharma, and D. Lo, "Apibot: question answering bot for api documentation," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering*. IEEE, 2017, pp. 153–158.

[13] S. Urli, Z. Yu, L. Seinturier, and M. Monperrus, "How to design a program repair bot? insights from the repairnator project," in *40th International Conference on Software Engineering: Software Engineering in Practice*, 2018, pp. 95–104.

[14] R. van Tonder and C. Le Goues, "Towards s/engineer/bot: Principles for program repair bots," in *2019 IEEE/ACM 1st international workshop on bots in software engineering*. IEEE, 2019, pp. 43–47.

[15] M. Wessel, I. Steinmacher, I. Wiese, and M. A. Gerosa, "Should i stale or should i close? an analysis of a bot that closes abandoned issues and pull requests," in *2019 IEEE/ACM 1st international workshop on bots in software engineering (BotSE)*. IEEE, 2019, pp. 38–42.

[16] M. Wyrich and J. Bogner, "Towards an autonomous bot for automatic source code refactoring," in *IEEE/ACM 1st international workshop on bots in software engineering (BotSE)*. IEEE, 2019, pp. 24–28.

[17] B. Xu, Z. Xing, X. Xia, and D. Lo, "Answerbot: Automated generation of answer summary to developers' technical questions," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017, pp. 706–716.

[18] M. Jain, R. Kota, P. Kumar, and S. N. Patel, "Convey: Exploring the use of a context view for chatbots," in *2018 chi conference on human factors in computing systems*, 2018, pp. 1–6.

[19] C. Lebeuf, M.-A. Storey, and A. Zagalsky, "Software bots," *IEEE Software*, vol. 35, no. 1, pp. 18–23, 2017.

[20] S. Pérez-Soler, E. Guerra, J. de Lara, and F. Jurado, "The rise of the (modelling) bots: Towards assisted modelling via social networks," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017, pp. 723–728.

[21] F. A. Valério, T. G. Guimarães, R. O. Prates, and H. Candello, "Here's what i can do: Chatbots' strategies to convey their features to users," in *XVI Brazilian Symposium on Human Factors in Computing Systems*, 2017, pp. 1–10.

[22] X. Wu, A. Gao, Y. Zhang, T. Wang, and Y. Tang, "A preliminary study of bots usage in open source community," in *13th Asia-Pacific Symposium on Internetware*, 2022, pp. 175–180.

[23] M. Golzadeh, A. Decan, D. Legay, and T. Mens, "A ground-truth dataset and classification model for detecting bots in github issue and pr comments," *Journal of Systems and Software*, vol. 175, p. 110911, 2021.

[24] M. Wyrich, R. Ghit, T. Haller, and C. Müller, "Bots don't mind waiting, do they? comparing the interaction with automatically and manually created pull requests," in *IEEE/ACM Third International Workshop on Bots in Software Engineering*. IEEE, 2021, pp. 6–10.

[25] T. Dey, S. Mousavi, E. Ponce, T. Fry, B. Vasilescu, A. Filippova, and A. Mockus, "Detecting and characterizing bots that commit code," in *17th International Conference on Mining Software Repositories*. ACM, 2020, pp. 209–219.

[26] M. Wessel, B. M. De Souza, I. Steinmacher, I. S. Wiese, I. Polato, A. P. Chaves, and M. A. Gerosa, "The power of bots: Characterizing and understanding bots in oss projects," *ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–19, 2018.