Agile Effort Estimation: Have We Solved the Problem Yet? Insights From A Second Replication Study

(GPT2SP Replication Report)

Vali Tawosi, Rebecca Moussa, Federica Sarro

{vali.tawosi, rebecca.moussa.18, f.sarro}@ucl.ac.uk

Fu and Tantithamthavorn [1] have recently proposed GPT2SP, a Transformer-based deep learning model for SP estimation of user stories.

They empirically evaluated the performance of GPT2SP on a dataset shared by Choetkiertikul et al. [2] including 16 projects with a total of 23,313 issues. They benchmarked GPT2SP against two baselines (namely the naive Mean and Median estimators) and the method previously proposed by Choetkiertikul et al. [2] (which we will refer to as DL2SP from now on) for both within- and cross-project estimation scenarios, and evaluated the extent to which each components of GPT2SP contribute towards the accuracy of the SP estimates.

Their results show that GPT2SP outperforms DL2SP with a 6%-47% improvement over MAE for the within-project scenario and a 3%-46% improvement for the cross-project scenarios.

However, when we attempted to use the GPT2SP source code made available by Fu and Tantithamthavorn [1] to reproduce their experiments, we found a bug in the computation of the Mean Absolute Error (MAE), which may have inflated the GPT2SP's accuracy reported in their work. Therefore, we had issued a pull request to fix such a bug, which has been accepted and merged into their repository at `https://github.com/awsm-research/gpt2sp/pull/2`.

In this report, we describe the results we achieved by using the fixed version of GPT2SP to replicate the experiments conducted in the original paper for RQ1 and RQ2, Following the original study, we analyse the results considering the Medan Absolute Error (MAE) of the estimation methods over all issues in each project, but we also report the Median Absolute Error (MdAE) and the Standard accuracy (SA) for completeness.

For completeness, we also report the results of three other approaches used as a benchmark to assess and compare GPT2SP's performance, namely DL2SP, Mean, and Median estimators. These results are taken from our previous replication study of DL2SP [3].

RQ1. Does GPT2SP outperform DL2SP and the baselines for the within-project scenario?

Table 1 shows the the results we obtained by running GPT2SP, DL2SP, and the Mean and Median baselines on the dataset used and shared publicly by Choetkiertikul et al. [2] (which is also the dataset used by Fu and Tantithamthavorn in their study [1]).

These results reveal that, for the within-project scenario, GPT2SP outperforms the median baseline and DL2SP in only six cases out of 16. Whereas the results of Wicoxon Ranked-Sum test

($\alpha = 0.05$) on the distribution absolute errors (see Table 2) reveal that the improvement is statistically significant in only three cases against median (two with negligible and one with small effect size), and two cases against DL2SP (both with negligible effect size). If we also consider Bonferroni correction (by setting $\alpha = 0.016$ for $K = 3$ hypothesis) the number of statistically significant cases goes down to two and one for the median baseline and DL2SP, respectively. Compared to the mean baseline, GPT2SP achieves a lower MAE in 14 cases, whereas it slightly underperforms in one case (Mule) and performs as good in one case (Usergrid). However, the difference between the absolute errors produced by the two methods is statistically significant in 12 cases even when considering Bonferroni correction.

Over all 16 cases and among all four estimation methods, the median baseline achieves the lowest MAE in 8 (50%) cases, GPT2SP in 5 (31%) cases, and DL2SP in remaining 3 (19%) cases. Therefore, our replication disproves the results previously obtained by Fu and Tantithamthavorn [1], that is GPT2SP outperforms DL2SP or the median baseline statistically significantly in all cases for the within project scenario.

RQ2. Does GPT2SP outperform DL2SP and the baselines for the cross-project scenario?

Similar to Choetkiertikul et al. [2] and Tawosi et al. [3], Fu and Tantithamthavorn also investigated two cross-project estimation scenarios, namely the within-repository and cross-repository scenarios.

Table 3a shows the results we achieved for GPT2SP, DL2SP and the mean and median baselines for the within-repository scenario. The results of the Wilcoxon test on the distribution of the absolute errors produced by GPT2SP against each of the other methods is provided in the last column of this table.

We can observe that GPT2SP outperforms DL2SP in only two out of eight cases (with marginal statistical significance and a negligible effect size in only one of the two cases), whereas DL2SP outperforms GPT2SP in six cases.

GPT2SP outperforms the Median baseline in only two cases out of eight, with statistical significance but negligible effect size in both cases. GPT2SP achieves a lower MAE than that of the Mean baseline in five cases, a higher one in two, and equal one in the remaining one case out of eight cases, where the difference is significant for three cases but always with a negligible effect size.

When considering the results of cross-repository scenario in Table 3b, we can observe that GPT2SP outperforms DL2SP in six cases out of eight with statistically significant difference in five of them out of which one show a medium effect size, one a small one, and three a negligible one. GPT2SP achieves a lower MAE than that obtained by Median in three cases only, among which one case shows a small effect size and two other cases show a negligible effect size. GPT2SP outperforms the Mean baseline in seven cases with statistically significant difference in six with a large effect size in three, medium one in two and small in one.

Overall, considering both cross-project scenarios, GPT2SP performs better than DL2SP in the cross-repository scenario (the same conclusion achieved by the original study). However, GPT2SP performs poorly when compared against the naïve Median baseline, that is, it achieves statistically significantly better results in less than a third of the cases in both scenarios combined (i.e., five out of 16 cases). Therefore, the replication results does not support the conclusion made in the original study stating that GPT2SP outperforms the baselines statistically significantly.

References

[1] M. Fu and C. Tantithamthavorn, "GPT2SP: A transformer-based agile story point estimation approach," *IEEE Transactions on Software Engineering*, 2022.

[2] M. Choetkiertikul, H. K. Dam, T. Tran, T. Pham, A. Ghose, and T. Menzies, "A deep learning model for estimating story points," *IEEE Transactions on Software Engineering*, vol. 45, no. 7, pp. 637–656, 2019.

[3] V. Tawosi, R. Moussa, and F. Sarro, "Agile effort estimation: Have we solved the problem yet? insights from a replication study," *IEEE Transactions on Software Engineering*, 2022.

Table 1: RQ1. Results obtained by GPT2SP, DL2SP, and the two baselines (Mean and Median) for the Within-Project SP scenario. Best results (among all approaches per project) are highlighted in bold.

Project	Method	MAE	MdAE	SA	Project	Method	MAE	MdAE	SA
Mesos	GPT2SP	1.21	0.98	42.56	Duracloud	GPT2SP	**0.80**	**0.50**	**47.97**
	DL2SP	**1.12**	**0.73**	**46.72**		DL2SP	0.82	0.53	46.97
	Mean	1.41	1.78	33.18		Mean	1.00	1.14	35.17
	Median	1.22	2.00	42.30		Median	0.82	1.00	46.78
Usergrid	GPT2SP	1.19	1.02	22.56	Data Management	GPT2SP	**5.39**	**2.00**	**54.82**
	DL2SP	1.18	**0.80**	23.25		DL2SP	5.86	2.22	50.87
	Mean	1.19	1.23	22.43		Mean	8.66	4.55	27.35
	Median	**1.15**	1.00	**24.97**		Median	6.19	3.00	48.09
Appcelerator Studio	GPT2SP	1.53	0.71	49.88	Moodle	GPT2SP	8.38	8.53	45.76
	DL2SP	1.42	**0.58**	53.49		DL2SP	7.89	**4.93**	48.92
	Mean	1.91	1.52	37.52		Mean	12.63	12.11	18.21
	Median	**1.30**	1.00	**57.41**		Median	**6.59**	6.00	**57.32**
Aptana Studio	GPT2SP	**3.52**	3.10	**37.75**	Mule	GPT2SP	2.61	2.23	28.55
	DL2SP	4.14	**2.52**	26.74		DL2SP	2.59	**1.96**	29.14
	Mean	3.59	3.46	36.53		Mean	2.60	2.22	28.72
	Median	3.61	4.00	36.18		Median	**2.47**	2.00	**32.29**
Titanium	GPT2SP	2.35	1.45	48.78	Mulesoft	GPT2SP	3.70	2.62	23.55
	DL2SP	2.09	**1.34**	54.49		DL2SP	3.67	**2.26**	24.12
	Mean	3.02	1.97	34.29		Mean	3.74	2.80	22.65
	Median	**2.04**	2.00	**55.71**		Median	**3.66**	3.00	**24.32**
Bamboo	GPT2SP	0.77	0.69	54.30	Spring XD	GPT2SP	1.78	1.54	37.09
	DL2SP	0.81	**0.61**	51.78		DL2SP	**1.70**	**1.31**	**39.96**
	Mean	1.22	1.31	27.99		Mean	2.05	2.53	27.59
	Median	**0.75**	1.00	**55.34**		Median	1.71	2.00	39.55
Clover	GPT2SP	3.76	0.97	32.89	Talend Data Quality	GPT2SP	3.65	3.44	29.41
	DL2SP	**3.39**	**0.80**	**39.41**		DL2SP	3.61	**2.92**	30.07
	Mean	4.57	3.06	18.46		Mean	4.56	5.08	11.69
	Median	3.71	2.00	33.77		Median	**3.31**	4.00	**35.85**
Jira Software	GPT2SP	**1.57**	**0.81**	**54.09**	Talend ESB	GPT2SP	**0.86**	**0.55**	**38.81**
	DL2SP	1.70	1.09	50.08		DL2SP	0.90	0.59	36.19
	Mean	2.40	2.15	29.61		Mean	1.04	0.91	26.47
	Median	2.31	2.00	32.40		Median	0.92	1.00	34.86

Table 2: RQ1. Results of the Wilcoxon test (with Vargha-Delaney $\hat{A}_{12}$ effect size in parentheses) comparing the absolute errors of GPT2SP to that of DL2SP and the two baselines (i.e., Mean and Median).

Project	GPT2SP vs		
	Mean	Median	DL2SP
Mesos	0.127 (0.55)	0.955 (0.43)	0.248 (0.53)
Usergrid	<0.001 (0.70)	1.000 (0.43)	0.999 (0.44)
Appcelerator Studio	<0.001 (0.64)	1.000 (0.37)	1.000 (0.43)
Aptana Studio	0.713 (0.48)	0.679 (0.49)	0.044 (0.55)
Titanium	<0.001 (0.70)	0.583 (0.49)	0.516 (0.50)
Bamboo	<0.001 (0.70)	0.021 (0.59)	0.412 (0.51)
Clover	<0.001 (0.76)	0.004 (0.54)	0.001 (0.54)
Jira Software	<0.001 (0.62)	0.988 (0.42)	0.478 (0.50)
Duracloud	<0.001 (0.58)	0.663 (0.49)	0.972 (0.46)
Data Management	<0.001 (0.79)	1.000 (0.31)	1.000 (0.40)
Moodle	0.323 (0.51)	0.678 (0.49)	0.294 (0.52)
Mule	0.497 (0.50)	0.751 (0.48)	0.722 (0.48)
Mulesoft	<0.001 (0.59)	0.962 (0.47)	0.843 (0.48)
Spring XD	<0.001 (0.66)	0.977 (0.45)	0.433 (0.50)
Talend Data Quality	<0.001 (0.61)	0.130 (0.53)	0.325 (0.51)
Talend ESB	<0.001 (0.68)	<0.001 (0.65)	0.217 (0.54)

Table 3: RQ2. Comparing GPT2SP cross-project SP estimation replication results to DL2SP [2], and to the baselines. The results of the Wilcoxon test ($\hat{A}_{12}$ effect size in parentheses) for GPT2SP vs. DL2SP and baselines are shown in the last column. Best results (among all approaches per project) are highlighted in bold.

(a) Within-Repository Training

Source	Target	Method	MAE	MdAE	SA	GPT2SP vs.
MESOS (ME)	USERGRID (UG)	GPT2SP	1.11	0.84	42.49	
		DL2SP	1.16	0.96	39.84	0.049 (0.53)
		Mean	1.02	0.19	46.99	0.998 (0.45)
		Median	0.89	0.00	54.10	1.000 (0.37)
USERGRID (UG)	MESOS (ME)	GPT2SP	1.52	0.92	15.64	
		DL2SP	1.51	1.01	16.18	0.662 (0.50)
		Mean	1.52	0.80	15.57	0.325 (0.50)
		Median	1.50	1.00	16.27	0.692 (0.50)
TISTUD (AS)	APSTUD (AP)	GPT2SP	4.62	3.37	7.99	
		DL2SP	4.37	2.98	12.99	1.000 (0.45)
		Mean	4.27	2.18	15.05	0.992 (0.47)
		Median	4.38	3.00	12.73	1.000 (0.43)
TISTUD (AS)	TIMOB (TI)	GPT2SP	3.28	2.47	22.60	
		DL2SP	3.38	2.39	20.31	0.253 (0.51)
		Mean	3.45	2.82	18.69	<0.001 (0.56)
		Median	3.17	2.00	25.24	1.000 (0.44)
APSTUD (AP)	TISTUD (AS)	GPT2SP	2.86	2.45	45.15	
		DL2SP	2.70	2.07	48.25	1.000 (0.47)
		Mean	3.38	3.24	35.20	<0.001 (0.58)
		Median	3.17	3.00	39.30	<0.001 (0.55)
APSTUD (AP)	TIMOB (TI)	GPT2SP	4.09	3.52	30.38	
		DL2SP	3.51	2.53	40.34	1.000 (0.39)
		Mean	4.36	4.24	25.78	<0.001 (0.56)
		Median	4.19	4.00	28.67	0.004 (0.52)
MULE (MU)	MULESTUDIO (MS)	GPT2SP	3.48	2.29	21.32	
		DL2SP	3.64	2.04	17.61	0.206 (0.51)
		Mean	3.34	2.71	24.42	0.943 (0.48)
		Median	3.26	3.00	26.26	0.999 (0.45)
MULESTUDIO (MS)	MULE (MU)	GPT2SP	3.03	2.65	30.27	
		DL2SP	2.77	2.47	36.28	0.985 (0.47)
		Mean	3.05	1.77	29.83	0.078 (0.52)
		Median	2.60	3.00	40.24	1.000 (0.43)

(b) Cross-Repository Training

Source	Target	Method	MAE	MdAE	SA	GPT2SP vs.
TISTUD (AS)	USERGRID (UG)	GPT2SP	2.18	2.03	37.67	
		DL2SP	3.47	3.50	0.98	<0.001 (0.76)
		Mean	3.08	2.82	12.02	<0.001 (0.73)
		Median	2.30	2.00	34.40	<0.001 (0.56)
TISTUD (AS)	MESOS (ME)	GPT2SP	2.65	2.75	28.83	
		DL2SP	3.18	3.20	14.70	<0.001 (0.60)
		Mean	3.28	2.82	11.95	<0.001 (0.64)
		Median	2.58	3.00	30.85	0.994 (0.47)
MDL (MD)	APSTUD (AP)	GPT2SP	4.37	3.07	68.12	
		DL2SP	5.03	3.77	63.29	<0.001 (0.55)
		Mean	9.84	8.95	28.21	<0.001 (0.86)
		Median	3.97	3.00	71.05	0.989 (0.47)
MDL (MD)	TIMOB (TI)	GPT2SP	3.60	2.53	73.90	
		DL2SP	3.34	1.96	75.82	1.000 (0.45)
		Mean	11.19	11.95	18.91	<0.001 (0.91)
		Median	4.19	4.00	69.63	<0.001 (0.60)
MDL (MD)	TISTUD (AS)	GPT2SP	2.16	1.96	83.92	
		DL2SP	2.64	1.66	80.35	<0.001 (0.57)
		Mean	11.45	11.95	14.90	<0.001 (0.98)
		Median	3.17	3.00	76.44	<0.001 (0.61)
DM	TIMOB (TI)	GPT2SP	3.58	2.31	61.99	
		DL2SP	3.81	2.65	59.52	<0.001 (0.53)
		Mean	5.61	5.03	40.35	<0.001 (0.75)
		Median	3.46	1.00	63.22	0.999 (0.47)
USERGRID (UG)	MULESTUDIO (MS)	GPT2SP	4.02	2.16	4.57	
		DL2SP	3.95	2.11	6.18	0.865 (0.48)
		Mean	4.04	2.20	4.00	0.145 (0.52)
		Median	3.91	2.00	7.16	0.996 (0.46)
MESOS (ME)	MULE (MU)	GPT2SP	3.13	2.24	9.57	
		DL2SP	3.20	2.31	7.37	0.347 (0.51)
		Mean	2.89	1.81	16.56	0.999 (0.46)
		Median	2.92	2.00	15.65	0.999 (0.46)

5