Suggest and Compute Process Performance Indicators from Event Logs

Appendix

Simone Agostinelli[1], Adela del Río Ortega[2], Rocío Goñi Medina[2], Andrea Marrella[1], Manuel Resinas[2], and Jacopo Rossi[1]

[1] Sapienza Universitá di Roma, Rome, italy
{agostinelli,j.rossi,marrella}@diag.uniroma1.it
[2] Universidad de Sevilla, Seville, Spain
{adeladelrio,rgoni,resinas}@us.es

1 Quantitative Evaluation

Tables 1. 2 and 3 show the GPT-4 results for the time category. Specifically, for each run in Table 1 we manually tagged the complete list of PPIs generated for all the 17 activities stored in the log, while for Tables 2 and 3 we manually tagged a randomly selected subset consisting of 30% of the PPIs generated for the 13 and 20 activities stored in the IT incident management and Manuscript review management logs, respectively. Based on the results, we observe that in each run, the number of PPIs that were correctly/incorrectly (A/B) translated from the suggestion stage and then correctly computed consistently exceeded the number of PPIs resulting in an error or empty value (C/D). Furthermore, if we switch our attention on the comparison between columns A and B, it appears that the A values are slightly higher than the B values, except for the Incident Management log, where they are quite similar. This indicates that PPIPilot is able to provide a list of PPIs that are correctly translated from the suggestion stage and then successfully computed. Specularly, C values are lower than D values, except for the Incident Management log, where again they are quite similar, meaning that PPIPilot is able to limit the number of PPIs that cannot be computed due to formatting issues according to the computable PPI definition.

It is worth noticing hallucinations, if present during the suggestion stage, are mitigated in the translation stage. This implies that hallucinated PPIs can fall into categories B, C, or D once they have been translated.

To demonstrate that our approach can be generalized to other LLMs, we repeated the previous experiments with Mistral and LLama. To streamline the tagging process for the PPIs, we combined dimensions A and B, as well as C and D. Consequently, we will now refer to these combined columns as A+B and C+D. Thus, A+B represents the number of PPIs computed with a correct value, while C+D denotes the number of PPIs that resulted in an error or were computed with an empty value.

The experiments in Tables 4, 5, and 6 show a high number of A+B values compared to C+D, indicating that, even in the case of Mistral, the PPIPilot

Table 1: Domestic Declarations (GPT-4)

Runs	A	B	C	D
#1	149	108	5	45
#2	118	105	15	47
#3	131	121	7	47
#4	124	113	16	41
#5	101	117	7	45
#6	161	87	9	42
#7	150	101	8	48
#8	157	95	12	59
#9	121	119	7	52
#10	137	80	20	53
Total	**1349**	**1046**	**106**	**479**

Table 2: IT Incident Management (GPT-4)

Runs	A	B	C	D
#1	24	29	6	15
#2	30	27	3	10
#3	30	25	8	8
#4	33	29	3	10
#5	26	31	6	9
#6	24	34	3	12
#7	31	31	6	7
#8	27	26	5	15
#9	18	27	12	11
#10	30	24	9	8
Total	**273**	**283**	**61**	**105**

Table 3: Manuscript Review (GPT-4)

Runs	A	B	C	D
#1	64	24	6	8
#2	54	41	3	2
#3	49	41	4	5
#4	51	35	7	9
#5	55	38	5	3
#6	62	31	6	2
#7	69	23	6	4
#8	62	27	7	6
#9	73	24	3	2
#10	60	28	6	5
Total	**599**	**312**	**53**	**46**

Table 4: Domestic Declarations (Mistral)

Runs	A + B	C+D
#1	540	119
#2	576	139
#3	520	129
#4	487	93
#5	508	98
#6	516	134
#7	472	107
#8	500	136
#9	620	165
#10	468	112
Total	**5207**	**1232**

Table 5: IT Incident Management (Mistral)

Runs	A + B	C+D
#1	443	44
#2	457	30
#3	434	89
#4	537	80
#5	421	60
#6	416	40
#7	355	80
#8	448	37
#9	395	90
#10	468	44
Total	**4374**	**594**

Table 6: Manuscript Review (Mistral)

Runs	A + B	C+D
#1	584	82
#2	615	176
#3	558	124
#4	661	60
#5	659	100
#6	637	71
#7	600	108
#8	552	110
#9	549	123
#10	540	91
Total	**5955**	**1045**

Table 7: Domestic Declarations (LLaMa)

Runs	A + B	C+D
#1	153	113
#2	164	81
#3	3	4
#4	85	5
#5	175	149
#6	1	4
#7	175	137
#8	121	103
#9	156	119
#10	163	64
Total	**1196**	**779**

Table 8: IT Incident Management (LLaMa)

Runs	A + B	C+D
#1	172	151
#2	119	53
#3	249	124
#4	238	88
#5	147	85
#6	171	72
#7	74	31
#8	63	36
#9	185	87
#10	161	53
Total	**1579**	**780**

Table 9: Manuscript Review (LLaMa)

Runs	A + B	C+D
#1	261	71
#2	25	9
#3	269	55
#4	49	34
#5	180	129
#6	33	9
#7	263	104
#8	186	55
#9	264	89
#10	327	123
Total	**1857**	**678**

approach is able to provide an high number of suggested and computed PPIs while keeping the number of PPIs with errors or empty values relatively low. Similarly, the experiments in Tables 7, 8, and 9 for LLaMa show a greater number of suggested and computed PPIs with respect to the ones resulted in an error or an empty value. However, in this case, the ratio of PPIs with errors or empty values to the total number of suggested and computed PPIs is higher compared to Mistral.

Table 10: Domestic Declarations (GPT-4)

Runs	A	B	C	D
#1	39	178	50	20
#2	49	194	46	19
#3	274	203	42	22
#4	65	218	51	25
#5	58	171	51	22
#6	65	193	51	28
#7	93	214	52	26
#8	80	205	51	23
#9	83	218	63	28
#10	86	195	50	24
Total	**692**	**1989**	**507**	**237**

Table 11: IT Incident Management (GPT-4)

Runs	A	B	C	D
#1	11	51	10	3
#2	16	45	8	3
#3	13	54	8	3
#4	17	49	8	5
#5	21	41	8	7
#6	19	36	9	4
#7	10	57	8	0
#8	19	44	5	2
#9	18	35	7	6
#10	21	36	13	1
Total	**165**	**448**	**84**	**34**

Table 12: Manuscript Review (GPT-4)

Runs	A	B	C	D
#1	21	66	17	2
#2	17	62	19	1
#3	19	72	13	1
#4	26	66	13	4
#5	19	71	8	2
#6	26	66	8	2
#7	27	61	19	1
#8	18	67	14	1
#9	22	71	11	0
#10	21	78	14	1
Total	**216**	**680**	**136**	**15**

Table 10, Table 11 and Table 12 present the results for GPT and the occurrence perspective.

Specifically for each run in Table 10 we manually tagged the full set of PPIs derived from all activities stored in the log. For Tables 11 and 12, we selected 30% of the activities from the log and manually tagged 30% of the PPIs for each selected activity. Based on the results, we observe that in each run, the number of PPIs that were correctly/incorrectly (A/B) translated from the suggestion stage and then accurately computed consistently exceeded the number of PPIs resulting in an error or empty value (C/D). However, it is worth noting that, in contrast to the analysis conducted for the time perspective, the number of incorrectly translated PPIs (B) exceeds the number of correctly translated PPIs (A). This indicates that, for the occurrence category, more PPIs were incorrectly translated from the suggestion stage than were correctly translated.

As already highlighted in the paper, hallucinations if present during the suggestion stage, are mitigated in the translation stage. This implies that hallucinated PPIs can fall into categories B, C, or D.

To demonstrate that our approach can be generalized to other LLMs, we repeated the previous experiment with two additional LLMs: Mistral and LLama. To automate the tagging procedure for the PPIs, we combined dimensions A and B, as well as C and D. Consequently, from now on, we will refer to these columns as A+B and C+D.

The experiments in Table 13, Table 14, and Table 15 show a high number of A+B compared to C+D, indicating that, even in the case of Mistral, the PPIPilot approach provides a robust set of suggested and computed PPIs while keeping the number of PPIs with errors or empty values relatively low. The results we obtained for Mistral related the occurrence category are perfectly aligned with those obtained for the time category counterpart.

Table 13: Domestic Declarations (Mistral)

Runs	**A + B**	**C+D**
#1	444	109
#2	406	77
#3	394	85
#4	441	103
#5	450	88
#6	459	94
#7	378	79
#8	476	112
#9	437	64
#10	425	98
Total	**4310**	**909**

Table 14: IT Incident Management (Mistral)

Runs	**A + B**	**C+D**
#1	293	76
#2	222	76
#3	315	100
#4	289	172
#5	270	80
#6	225	66
#7	324	111
#8	285	57
#9	262	74
#10	297	68
Total	**2782**	**880**

Table 15: Manuscript Review (Mistral)

Runs	**A + B**	**C+D**
#1	558	118
#2	434	89
#3	416	92
#4	643	148
#5	423	102
#6	457	76
#7	418	137
#8	482	108
#9	545	136
#10	448	86
Total	**4824**	**1092**