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

This Appendix reports the complete list of results obtained from the quantitative evaluation, since in the paper for the sake of space we have shown only the computed aggregated average values.

Tables 1, 2 and 3 show the GPT-4 results for the time category. Specifically, for the Domestic Declaration log, we manually tagged each PPI in the complete list generated by each run of PPIPilot for all 17 activities in the log, by assigning one value among A, B, C, or D. For IT incident management and Manuscript review management logs, we manually tagged a randomly selected subset comprising 30% of the PPIs generated by each run of PPIPilot for the 13 and 20 activities, 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 exceeds 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. This suggests that, in most cases, the PPI textual descriptions (output of the Suggestion Stage) align with the computable PPI definitions (output of the Translation Stage). Specularly, C values are lower than D values, except for the Manuscript Review 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.

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 dimensions B, C, or D once they have been translated into PPI computable definitions.

Instead, Tables 10, 11 and 12 present the results of GPT-4 for the occurrence perspective which were omitted from the main paper due to space constraints. 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. This means

Table 1: Domestic Declarations (GPT-4) - Time

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) - Time

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) - Time

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) - Time

Runs	A + B	C + D
#1	540	103
#2	576	122
#3	520	112
#4	487	76
#5	508	81
#6	516	117
#7	472	92
#8	500	120
#9	620	148
#10	468	112
Total	**5207**	**1083**

Table 5: IT Incident Management (Mistral) - Time

Runs	A + B	C + D
#1	443	31
#2	457	17
#3	434	76
#4	537	67
#5	421	48
#6	416	27
#7	355	69
#8	448	24
#9	395	77
#10	468	31
Total	**4374**	**467**

Table 6: Manuscript Review (Mistral) - Time

Runs	A + B	C + D
#1	584	63
#2	615	157
#3	558	106
#4	661	41
#5	659	82
#6	637	52
#7	600	89
#8	552	92
#9	549	103
#10	540	72
Total	**5955**	**857**

Table 7: Domestic Declarations (LLaMa) - Time

Runs	A + B	C + D
#1	231	142
#2	164	70
#3	93	88
#4	210	80
#5	185	170
#6	82	100
#7	175	124
#8	121	92
#9	156	107
#10	251	135
Total	**1668**	**1108**

Table 8: IT Incident Management (LLaMa) - Time

Runs	A + B	C + D
#1	172	140
#2	119	48
#3	249	114
#4	238	79
#5	147	76
#6	171	64
#7	154	89
#8	146	92
#9	185	76
#10	161	46
Total	**1742**	**824**

Table 9: Manuscript Review (LLaMa) - Time

Runs	A + B	C + D
#1	261	56
#2	175	49
#3	269	43
#4	196	114
#5	180	118
#6	267	56
#7	287	96
#8	398	142
#9	264	76
#10	327	110
Total	**2624**	**860**

that GPT-4 suggestions more often result in textual PPI descriptions that do not correspond one-to-one with computable PPI definitions in PPINAT [1].

To demonstrate that our approach is also feasible with other LLMs, we repeated the previous experimentation 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

Table 10: Domestic Declarations (GPT-4) - Occ.

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: Incident Management (GPT-4) - Occ.

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) - Occ.

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 13: Domestic Declarations (Mistral) - Occ.

Runs	A + B	C + D
#1	444	93
#2	406	62
#3	394	71
#4	441	86
#5	450	71
#6	459	78
#7	378	65
#8	476	95
#9	437	49
#10	425	82
Total	**4310**	**752**

Table 14: Incident Management (Mistral) - Occ.

Runs	A + B	C + D
#1	293	64
#2	222	65
#3	315	87
#4	289	159
#5	270	68
#6	225	55
#7	324	98
#8	285	45
#9	262	61
#10	297	55
Total	**2782**	**757**

Table 15: Manuscript Review (Mistral) - Occ.

Runs	A + B	C + D
#1	558	98
#2	434	71
#3	416	74
#4	643	128
#5	423	85
#6	457	58
#7	418	119
#8	482	89
#9	545	116
#10	448	68
Total	**4824**	**906**

Table 16: Domestic Declarations (LLaMa - Occ.)

Runs	A + B	C + D
#1	116	278
#2	126	230
#3	72	182
#4	126	208
#5	70	217
#6	86	191
#7	116	203
#8	55	160
#9	104	166
#10	108	189
Total	**979**	**2024**

Table 17: Incident Management (LLaMa) - Occ.

Runs	A + B	C + D
#1	148	73
#2	185	94
#3	135	95
#4	102	229
#5	163	131
#6	193	106
#7	191	66
#8	81	62
#9	92	147
#10	190	117
Total	**1480**	**1110**

Table 18: Manuscript Review (LLaMa) - Occ.

Runs	A + B	C + D
#1	191	199
#2	198	134
#3	93	84
#4	107	243
#5	106	179
#6	104	191
#7	122	263
#8	126	190
#9	92	286
#10	151	293
Total	**1290**	**2062**

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 approach is able to provide an high number of suggested and computed PPIs while

Table 19: yes descr. yes goal - Domestic Declarations (GPT-4) - Time

Runs	A + B	C + D
#1	257	50
#2	223	62
#3	252	54
#4	237	57
#5	218	52
#6	248	51
#7	251	56
#8	252	71
#9	240	59
#10	217	73
Total	**2395**	**585**

Table 20: yes descr. yes goal - Incident Management (GPT-4) - Time

Runs	A + B	C + D
#1	161	66
#2	173	38
#3	173	54
#4	182	49
#5	181	31
#6	155	46
#7	180	47
#8	166	54
#9	153	55
#10	163	45
Total	**1687**	**485**

Table 21: yes descr. yes goal - Manuscript Review (GPT-4) - Time

Runs	A + B	C + D
#1	286	27
#2	288	26
#3	270	34
#4	273	49
#5	284	28
#6	284	30
#7	292	29
#8	286	30
#9	300	20
#10	279	28
Total	**2842**	**301**

Table 22: no descr. yes goal - Domestic Declarations (GPT-4) - Time

Runs	A + B	C + D
#1	216	67
#2	235	57
#3	246	65
#4	225	67
#5	220	61
#6	222	60
#7	242	58
#8	227	64
#9	233	65
#10	222	62
Total	**2288**	**626**

Table 23: no descr. yes goal - Incident Management (GPT-4) - Time

Runs	A + B	C + D
#1	185	46
#2	173	33
#3	148	59
#4	140	72
#5	148	63
#6	164	55
#7	181	38
#8	163	42
#9	152	57
#10	173	41
Total	**1627**	**506**

Table 24: no descr. yes goal - Manuscript Review (GPT-4) - Time

Runs	A + B	C + D
#1	275	41
#2	275	42
#3	282	35
#4	310	26
#5	285	32
#6	286	36
#7	268	47
#8	289	30
#9	263	40
#10	290	41
Total	**2823**	**370**

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 PPIs computed with a value (0.46) is three times higher than that for Mistral (0.15). We emphasize that the purpose of this experimentation is not to claim that Mistral outperforms GPT-4 or LLaMa. Instead, our goal is to demonstrate the feasibility of the PPIPilot approach in suggesting and computing PPIs across state-of-the-art LLMs. The specific reasons why Mistral performs better are beyond the scope of this experimentation.

The experiments in Tables 13, 14, and 15 highlight the results we obtained for Mistral related the occurrence category and are perfectly aligned with those obtained for the time category counterpart. However, this is not the case for the experiments related to LLaMa, as shown in Tables 16 and 18, where the number of A + B is lower than C + D, except in Table 17, where the values are quite similar. This discrepancy arises because LLaMa struggles to accurately interpret

Table 25: yes descr. no goal - Domestic Declarations (GPT-4) - Time

Runs	A + B	C + D
#1	237	59
#2	224	65
#3	217	84
#4	225	75
#5	239	57
#6	219	76
#7	242	63
#8	238	66
#9	234	54
#10	211	76
Total	**2286**	**675**

Table 26: yes descr. no goal - Incident Management (GPT-4) - Time

Runs	A + B	C + D
#1	178	63
#2	157	64
#3	169	55
#4	170	53
#5	164	48
#6	170	53
#7	150	66
#8	165	58
#9	164	55
#10	163	66
Total	**1650**	**581**

Table 27: yes descr. no goal - Manuscript Review (GPT-4) - Time

Runs	A + B	C + D
#1	277	47
#2	286	32
#3	283	28
#4	285	30
#5	262	33
#6	285	22
#7	284	29
#8	277	40
#9	285	33
#10	288	30
Total	**2812**	**324**

Table 28: no descr. no goal - Domestic Declarations (GPT-4) - Time

Runs	A + B	C + D
#1	204	84
#2	211	72
#3	200	90
#4	217	68
#5	216	66
#6	208	75
#7	220	68
#8	222	63
#9	208	70
#10	214	56
Total	**2120**	**712**

Table 29: no descr. no goal - Incident Management (GPT-4) - Time

Runs	A + B	C + D
#1	152	80
#2	151	66
#3	173	49
#4	170	50
#5	169	80
#6	151	78
#7	166	52
#8	170	62
#9	167	78
#10	148	88
Total	**1617**	**683**

Table 30: no descr. no goal - Manuscript Review (GPT-4) - Time

Runs	A + B	C + D
#1	287	37
#2	267	34
#3	280	28
#4	279	35
#5	263	45
#6	279	35
#7	289	27
#8	294	28
#9	290	30
#10	277	44
Total	**2805**	**343**

the translation of textual PPI description (as output of the PPIs Suggestion Stage) according to the PPI Definition model in PPINAT [1] (as dictated by the PPIs Translation Stage).

What the previous experiments have in common is that all the elements of the approach depicted in the main paper have been always fixed. However, what happens if we remove the process description from the prompt used as input for the PPIs Suggestion Stage? What if we remove the organizational goal from the prompt used as input for the same stage? And what if both are removed? These questions are addressed below. This experimentation was conducted using GPT-4 as the employed LLM, focusing on both the time perspective and the occurrence perspective.

What the previous experiments have in common is that all the elements of the approach depicted in the main paper have been always fixed. However, in the next set of experiments, we aim to explore the impact of removing certain elements from the prompt used as input for the PPIs Suggestion Stage. Specifically, we investigate the impact of removing the process description, the organizational

Table 31: yes descr. yes goal - Domestic Declarations (GPT-4) - Occ.

Runs	A + B	C + D
#1	217	70
#2	243	65
#3	477	64
#4	283	76
#5	229	73
#6	258	79
#7	307	78
#8	205	74
#9	301	91
#10	281	74
Total	**2681**	**744**

Table 32: yes descr. yes goal - Incident Management (GPT-4) - Occ.

Runs	A + B	C + D
#1	189	45
#2	190	28
#3	205	33
#4	210	38
#5	200	39
#6	173	38
#7	196	35
#8	184	35
#9	190	33
#10	181	38
Total	**1918**	**362**

Table 33: yes descr. yes goal - Manuscript Review (GPT-4) - Occ.

Runs	A + B	C + D
#1	288	40
#2	266	40
#3	277	49
#4	287	47
#5	292	29
#6	278	27
#7	273	51
#8	267	35
#9	288	34
#10	305	46
Total	**2821**	**398**

Table 34: no descr. yes goal - Domestic Declarations (GPT-4) - Occ.

Runs	A + B	C + D
#1	216	72
#2	214	76
#3	225	89
#4	233	74
#5	213	87
#6	232	71
#7	222	79
#8	225	82
#9	210	71
#10	215	74
Total	**2205**	**775**

Table 35: no descr. yes goal - Incident Management (GPT-4) - Occ.

Runs	A + B	C + D
#1	182	37
#2	183	32
#3	187	48
#4	192	36
#5	188	46
#6	203	25
#7	180	38
#8	180	41
#9	188	42
#10	187	39
Total	**1870**	**384**

Table 36: no descr. yes goal - Manuscript Review (GPT-4) - Occ.

Runs	A + B	C + D
#1	289	53
#2	285	60
#3	273	59
#4	267	43
#5	247	55
#6	281	55
#7	286	64
#8	297	70
#9	296	53
#10	275	77
Total	**2796**	**589**

goal, or both from the prompt. This experimentation was conducted using GPT-4 as the employed LLM, with a focus on the time perspective.

From the time perspective, the presence of both a description and a goal achieves the highest scores for A + B and the lowest scores for C + D across all logs, indicating that the inclusion of both elements significantly improves the quality of the results (cf. Tables 19, 20, 21). The absence of either a description (cf. Tables 22, 23, 24) or a goal (cf. Tables 25, 26, 27) slightly reduces the scores for A + B while minimally affecting the C + D values. The combined absence of both description and goal (cf. Tables 28, 29, 30) results in the lowest scores for A + B across all logs. In summary, while the individual contribution of either a description or a goal is positive, their combination is essential to maximize scores, highlighting the importance of context for LLMs like GPT-4.

Tables 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42 present the results for the occurrence perspective. These findings are perfectly aligned with those from the time perspective, offering no new insights compared to the previous analysis.

Table 37: yes descr. no goal - Domestic Declarations (GPT-4) - Occ.

Runs	A + B	C + D
#1	202	96
#2	219	90
#3	205	88
#4	222	76
#5	242	93
#6	207	90
#7	250	75
#8	233	83
#9	242	99
#10	248	86
Total	**2270**	**876**

Table 38: yes descr. no goal - Incident Management (GPT-4) - Occ.

Runs	A + B	C + D
#1	195	59
#2	166	58
#3	173	48
#4	185	61
#5	167	48
#6	172	49
#7	184	55
#8	164	51
#9	184	52
#10	91	21
Total	**1681**	**502**

Table 39: yes descr. no goal - Manuscript Review (GPT-4) - Occ.

Runs	A + B	C + D
#1	275	63
#2	291	61
#3	281	54
#4	289	62
#5	268	56
#6	284	63
#7	298	64
#8	273	65
#9	259	75
#10	266	49
Total	**2784**	**612**

Table 40: no descr. no goal - Domestic Declarations (GPT-4) - Occ.

Runs	A + B	C + D
#1	223	80
#2	234	74
#3	227	86
#4	133	56
#5	238	73
#6	249	88
#7	232	77
#8	227	84
#9	219	81
#10	245	83
Total	**2227**	**782**

Table 41: no descr. no goal - Incident Management (GPT-4) - Occ.

Runs	A + B	C + D
#1	172	42
#2	164	47
#3	165	51
#4	164	48
#5	161	58
#6	159	53
#7	163	47
#8	174	47
#9	121	41
#10	165	56
Total	**1608**	**490**

Table 42: no descr. no goal - Manuscript Review (GPT-4) - Occ.

Runs	A + B	C + D
#1	264	57
#2	288	56
#3	284	57
#4	293	58
#5	271	58
#6	273	53
#7	294	62
#8	250	63
#9	288	50
#10	258	57
Total	**2763**	**571**

References

1. Resinas, M., del Río-Ortega, A., van der Aa, H.: From text to performance measurement: Automatically computing process performance using textual descriptions and event logs. In: Business Process Management. pp. 266–283. Springer Nature Switzerland, Cham (2023)