Appendix C

Test results and failure classification workflow

Table 4: Performance comparison of various LLMs and prompt engineering techniques.

Techniques	Success rate				Average score				Weighted average score			
	GPT-4o	GPT-4	GPT-3.5	Total	GPT-4o	GPT-4	GPT-3.5	Total	GPT-4o	GPT-4	GPT-3.5	Total
Zero-shot	0.68	0.39	0.06	0.37	40.65	40.38	36.17	40.37	27.64	15.75	2.17	14.80
Baseline prompt	0.50	0.20	0.06	0.25	37.96	35.50	34.00	37.00	18.98	7.10	2.04	9.37
Baseline detailed prompt	0.86	0.58	0.06	0.48	42.21	42.07	38.33	42.15	36.30	24.40	2.30	20.23
Role-based	0.67	0.41	0.04	0.37	40.91	41.17	37.50	40.88	27.41	16.88	1.50	15.26
Role-based prompt	0.48	0.12	0.02	0.21	38.33	37.00	35.00	37.97	18.40	4.44	0.70	7.85
Role-based detailed prompt	0.86	0.70	0.06	**0.54**	42.35	41.89	38.33	42.00	36.42	29.32	2.30	**22.68**
Role-based best practices	0.70	0.41	0.01	0.37	41.01	41.66	41.00	41.25	28.71	17.08	0.41	15.40
Role-based best practices prompt	0.60	0.10	0.00	0.23	38.80	37.80	0.00	38.66	23.28	3.78	0.00	9.02
Role-based best practices detailed prompt	0.80	**0.72**	0.02	0.51	42.68	42.19	41.00	42.43	34.14	**30.38**	0.82	21.78
Zero-shot CoT	0.71	0.46	0.11	0.43	40.63	40.39	34.82	40.05	28.85	18.58	3.83	17.09
Human prompt	0.68	0.30	0.18	0.39	39.06	37.13	34.22	37.81	26.56	11.14	6.16	14.62
Human detailed prompt	0.74	0.62	0.04	0.47	42.08	41.97	37.50	41.90	31.14	26.02	1.50	19.55
Zero-shot CoT APE	0.64	0.44	0.17	0.42	40.80	39.89	34.00	39.55	26.11	17.55	5.78	16.48
APE prompt	0.60	0.36	**0.20**	0.39	39.27	37.50	32.30	37.52	23.56	13.50	**6.46**	14.51
APE detailed prompt	0.68	0.52	0.14	0.45	42.15	41.54	36.43	41.31	28.66	21.60	5.10	18.45
Meta prompting	0.72	0.43	0.04	0.40	41.33	40.56	36.00	40.87	29.76	17.44	1.44	16.21
Meta prompt	0.56	0.16	0.06	0.26	39.71	39.38	36.00	39.36	22.24	6.30	2.16	10.23
Meta detailed prompt	**0.88**	0.70	0.02	0.53	42.36	40.83	36.00	41.61	**37.28**	28.58	0.72	22.19
Meta meta prompting	0.18	0.09	0.02	0.10	52.83	50.33	42.50	51.34	9.51	4.53	0.85	4.96
Meta meta prompt	0.08	0.02	0.04	0.05	**58.00**	43.00	42.50	51.43	4.64	0.86	1.70	2.40
Meta meta detailed prompt	0.28	0.16	0.00	0.15	51.36	51.25	0.00	51.32	14.38	8.20	0.00	7.53
ToT-style prompt 1	0.66	0.31	0.00	0.32	40.61	39.23	0.00	40.16	26.80	12.16	0.00	12.99
Prompt 1	0.60	0.26	0.00	0.29	38.20	35.31	0.00	37.33	22.92	9.18	0.00	10.70
Detailed prompt 1	0.72	0.36	0.00	0.36	42.61	42.06	0.00	42.43	30.68	15.14	0.00	15.27
ToT-style prompt 2	0.59	0.11	0.01	0.24	40.54	40.73	36.00	40.51	23.92	4.48	0.36	9.59
Prompt 2	0.42	0.06	0.00	0.16	37.86	36.67	0.00	37.71	15.90	2.20	0.00	6.03
Detailed prompt 2	0.76	0.16	0.02	0.31	42.03	42.25	36.00	41.94	31.94	6.76	0.72	13.14
ToT-style prompt 3	0.56	0.02	0.00	0.19	40.21	39.50	0.00	40.19	22.52	0.79	0.00	7.77
Prompt 3	0.52	0.02	0.00	0.18	38.19	35.00	0.00	38.07	19.86	0.70	0.00	6.85
Detailed prompt 3	0.60	0.02	0.00	0.21	41.97	44.00	0.00	42.03	25.18	0.88	0.00	8.69
Total												
Total detailed prompts	0.72	0.45	0.03	0.40	42.64	42.13	37.20	42.31	30.61	19.13	1.12	16.95
Total non-detailed prompts	0.50	0.16	0.06	0.24	38.96	37.00	34.32	38.16	19.63	5.92	1.92	9.16
Total Zero-shot	0.65	0.40	0.03	0.37	40.86	41.08	36.63	40.84	26.48	16.57	0.98	15.16
Total Zero-shot CoT	0.68	0.45	0.14	**0.42**	40.71	40.14	34.32	39.80	27.48	18.07	4.80	16.78
Total meta prompting	0.45	0.26	0.03	0.25	43.63	42.25	38.17	42.93	19.64	10.99	1.15	10.59
Total ToT-style prompts	0.60	0.15	0.00	0.25	40.46	39.61	36.00	40.28	24.41	5.81	0.12	10.11
Total	0.60	0.31	0.04	0.32	41.12	40.79	35.33	40.76	24.58	12.52	1.52	13.06

Table 5: Failures overview of various LLMs and prompt engineering techniques.

Techniques	Compliance failure		Syntax errors		Deployment failures		Cluster state failures						Total failures
	Not wrapped	Wrong names	Invalid YAML	Not Kubeconform	Duplicate resources	Other errors	PVC storage class	Missing resources	Missing key	MySQL failure	WordPress not ready	Other errors	
GPT-4o	2	12	9	42	4	3	12	3	0	7	280	15	389
Zero-Shot	0	0	0	0	0	0	1	0	0	0	31	0	32
Role-Based	0	0	0	1	1	2	0	0	0	1	58	0	63
Zero-Shot CoT	2	0	4	2	0	0	3	0	0	1	50	3	65
Meta prompting	0	0	0	1	0	0	0	0	0	2	25	0	28
Meta meta prompting	0	6	2	35	0	1	3	2	0	3	22	8	82
ToT-style prompts	0	6	3	3	3	0	5	1	0	0	94	4	119
GPT-4	10	61	22	53	39	7	2	66	4	9	418	2	693
Zero-Shot	0	0	0	5	0	2	0	5	1	0	48	0	61
Role-Based	0	0	1	2	0	1	0	8	2	0	104	0	118
Zero-Shot CoT	0	3	2	6	0	1	0	3	0	0	95	0	110
Meta prompting	0	0	0	2	0	1	0	1	0	0	53	0	57
Meta meta prompting	0	6	0	13	0	0	2	16	0	9	45	0	91
ToT-style prompts	10	52	19	25	39	2	0	33	1	0	73	2	256
GPT-3.5	4	120	22	59	51	53	11	86	9	5	526	11	957
Zero-Shot	0	3	1	1	2	10	1	12	0	0	67	0	97
Role-Based	0	0	3	0	1	20	7	23	1	0	140	0	195
Zero-Shot CoT	0	3	5	3	0	6	0	7	4	0	144	0	172
Meta prompting	0	1	0	2	0	6	2	2	0	1	82	0	96
Meta meta prompting	0	24	1	36	0	2	0	4	0	2	25	4	98
ToT-style prompts	4	89	12	17	48	9	1	38	4	2	68	7	299
Total													
Zero-Shot	0	3	1	6	2	12	2	17	1	0	146	0	190
Role-Based	0	0	4	3	2	23	7	31	3	1	302	0	376
Zero-Shot CoT	2	6	11	11	0	7	3	10	4	1	289	3	347
Meta prompting	0	1	0	5	0	7	2	3	0	3	160	0	181
Meta meta prompting	0	36	3	84	0	3	5	22	0	14	92	12	271
ToT-style prompts	14	147	34	45	90	11	6	72	5	2	235	13	674
Total detailed prompts	12	93	25	78	34	25	8	101	13	11	486	13	899
Total not detailed prompts	4	100	28	76	60	38	17	54	0	10	738	15	1140
Total	16	193	53	154	94	63	25	155	13	21	1224	28	2039

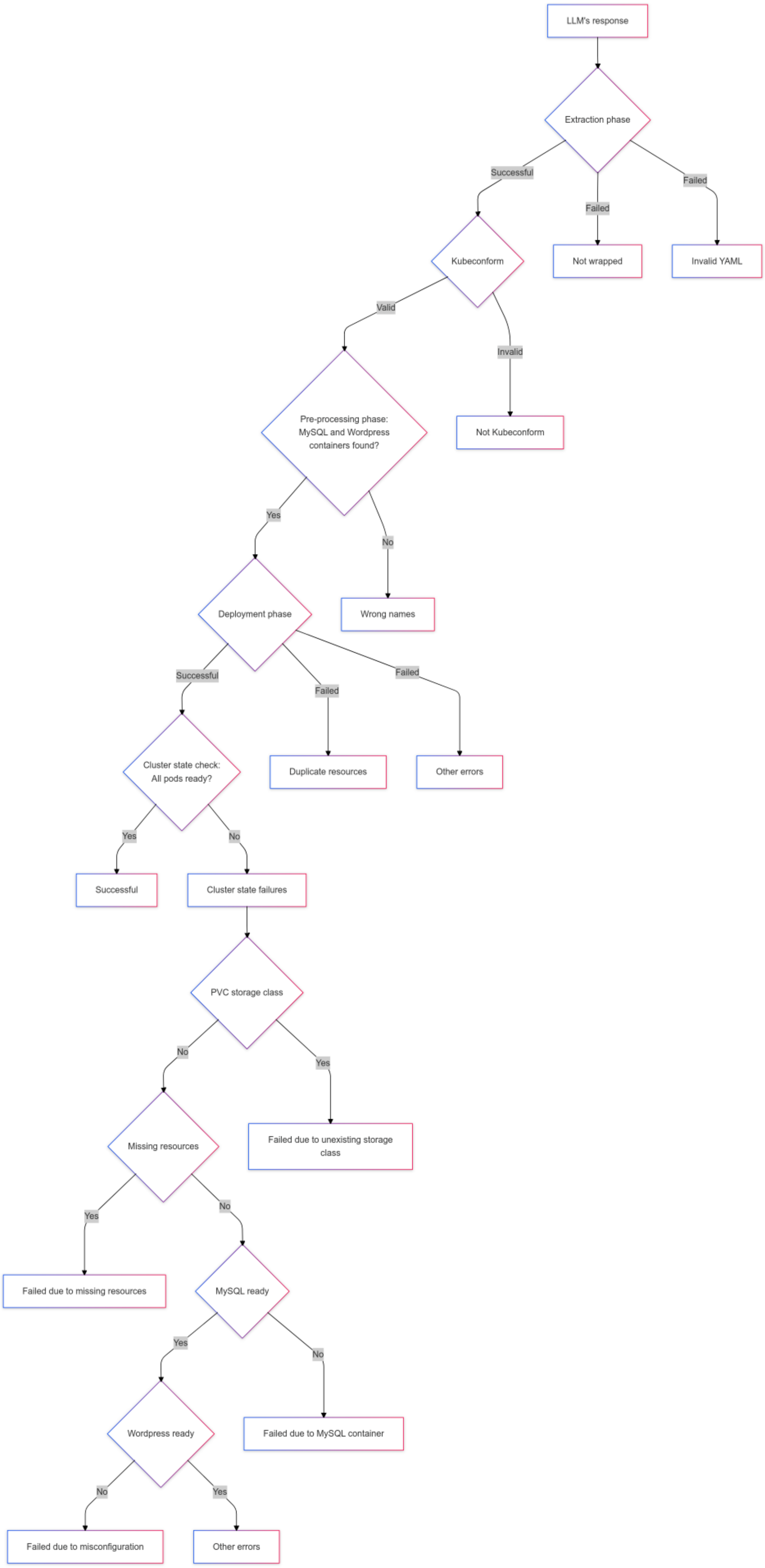

Figure 7: Flow diagram of the failures categorization framework.