Models or Code? Evaluating the Quality of LLM-Generated Specifications: A Case Study in Optimization at Kinaxis

NEGIN AYOUGHI, DAVID DEWAR, SHIVA NEJATI AND MEHRDAD SABETZADEH

B STATISTICAL TEST RESULTS FOR RQ1–RQ4 (COMPLEMENT TO SECTION 5.7 RESULTS)

Table 1. Statistical tests for **RQ1** comparing **AMPL-based variants** and **Python-based variants** based on (a) execution success rate (Success) and zero relative error rates (#Zero/#Exec) using the Z-test; and (b) based on relative error using the Mann-Whitney test. Blue cells 🟦 indicate significant improvements of AMPL-based over Python-based variants. No significant improvements are observed in the opposite direction.

(a) Comparing the variants based ib the Success and #Zero/#Exec metrics

Variant			**Public**				**Industry**			
LLM	**Structuring**	**Refinement**	Success		#Zero/#Exec		Success		#Zero/#Exec	
			Z	***p-value***	**Z**	***p-value***	**Z**	***p-value***	**Z**	***p-value***
Gemini 1.5-Flash	Unstructured	One-off	-5.31	1.00	**3.15**	**0.00**	-2.34	0.99	0.83	0.20
		Refinement	**3.02**	**0.00**	-1.64	0.95	-1.10	0.86	0.04	0.48
	Structured	One-off	-0.32	0.63	-0.67	0.75	-2.61	1.00	0.78	0.22
		Refinement	**6.64**	**0.00**	**2.12**	**0.02**	0.00	0.50	0.00	0.50
GPT-4o	Unstructured	One-off	-8.56	1.00	**2.05**	**0.02**	0.00	0.50	-0.94	0.83
		Refinement	-3.72	1.00	**2.30**	**0.01**	0.27	0.40	0.47	0.32
	Structured	One-off	-2.76	1.00	-3.60	1.00	-4.39	1.00	0.63	0.27
		Refinement	**6.24**	**0.00**	-2.74	1.00	-0.55	0.71	**2.82**	**0.00**
Gemini 2.5-Pro	Unstructured	One-off	-4.82	1.00	0.69	0.25	-1.39	0.92	-0.09	0.53
		Refinement	-1.06	0.86	-3.49	1.00	-0.47	0.68	-0.19	0.57
	Structured	One-off	-2.62	1.00	0.32	0.38	-2.09	0.98	1.65	0.05
		Refinement	**2.20**	**0.01**	**2.68**	**0.00**	0.00	0.50	-0.34	0.63
o4-mini	Unstructured	One-off	-4.33	1.00	**3.73**	**0.00**	-1.13	0.87	-0.69	0.76
		Refinement	0.26	0.40	1.15	0.13	0.31	0.38	-0.35	0.64
	Structured	One-off	**3.71**	**0.00**	-1.35	0.91	-0.55	0.71	0.51	0.30
		Refinement	**7.77**	**0.00**	-4.77	1.00	0.31	0.38	0.07	0.47

(b) Comparing the variants based on the relative error metric

Variant			**Public Dataset**		**Industry Dataset**	
LLM	**Structuring**	**Refinement**	***p-value***	**$\hat{A}_{12}$**	***p-value***	**$\hat{A}_{12}$**
Gemini 1.5-Flash	Unstructured	One-off	0.23	0.47	0.53	0.50
		Refinement	1.00	0.58	0.93	0.57
	Structured	One-off	0.99	0.58	0.86	0.54
		Refinement	**0.00**	**0.41(S)**	0.56	0.51
GPT-4o	Unstructured	One-off	0.13	0.46	0.57	0.51
		Refinement	**0.04**	**0.45(N)**	0.32	0.45
	Structured	One-off	1.00	0.60	0.50	0.48
		Refinement	0.98	0.56	**0.01**	**0.32(M)**
Gemini 2.5-Pro	Unstructured	One-off	0.23	0.48	0.84	0.56
		Refinement	1.00	0.57	0.51	0.50
	Structured	One-off	0.11	0.47	1.00	0.50
		Refinement	**0.00**	**0.42(S)**	0.72	0.53
o4-mini	Unstructured	One-off	**0.01**	**0.45(N)**	0.68	0.53
		Refinement	0.17	0.48	0.48	0.50
	Structured	One-off	0.82	0.52	0.17	0.47
		Refinement	1.00	0.59	0.49	0.50

Table 2. Statistical tests for **RQ2** comparing **structured variants** and **unstructured variants** based on (a) execution success rate (Success) and zero relative error rates (#Zero/#Exec) using the Z-test; and (b) based on relative error using the Mann-Whitney test. Blue cells 🟦 indicate significant improvements of structured over unstructured variants. No significant improvements are observed in the opposite direction.

(a) Comparing the variants based on the Success and #Zero/#Exec metrics

Variant			Public				Industry			
LLM	Language	Refinement	Success		#Zero/#Exec		Success		#Zero/#Exec	
			Z	*p-value*	Z	*p-value*	Z	*p-value*	Z	*p-value*
Gemini 1.5-Flash	AMPL	One-off	**2.63**	**0.00**	-2.98	1.00	0.00	0.50	0.00	0.50
		Refinement	0.96	0.17	**6.06**	**0.00**	**1.69**	**0.04**	-0.21	0.58
	Python	One-off	-2.39	0.99	0.75	0.22	0.29	0.39	0.07	0.47
		Refinement	-2.79	1.00	**1.88**	**0.03**	0.61	0.27	-0.18	0.57
GPT-4o	AMPL	One-off	**2.44**	**0.00**	-2.65	1.00	-1.39	0.92	0.21	0.41
		Refinement	**5.73**	**0.00**	-3.58	1.00	0.00	0.50	**1.91**	**0.03**
	Python	One-off	-3.53	1.00	**3.02**	**0.00**	**3.16**	**0.00**	-1.50	0.93
		Refinement	-4.25	1.00	1.47	0.07	0.81	0.21	-0.49	0.69
Gemini 2.5-Pro	AMPL	One-off	0.67	0.25	-3.65	1.00	-1.06	0.85	**1.83**	**0.03**
		Refinement	0.31	0.38	**3.13**	**0.00**	0.38	0.35	0.42	0.34
	Python	One-off	-1.62	0.95	-3.54	1.00	-0.33	0.63	0.23	0.41
		Refinement	-2.92	1.00	-3.04	1.00	-0.09	0.53	0.58	0.28
o4-mini	AMPL	One-off	**1.92**	**0.03**	-2.15	0.98	0.00	0.50	1.44	0.07
		Refinement	**2.15**	**0.02**	-2.95	1.00	0.00	0.50	0.00	0.50
	Python	One-off	-6.07	1.00	**2.89**	**0.00**	-0.58	0.72	0.37	0.36
		Refinement	-5.80	1.00	**3.14**	**0.00**	0.00	0.50	0.41	0.34

(b) Comparing the variants based on the relative error metric

Variant			Public Dataset		Industry Dataset	
LLM	Language	Refinement	*p-value*	$\hat{A}_{12}$	*p-value*	$\hat{A}_{12}$
Gemini 1.5-Flash	AMPL	One-off	1.00	0.54	1.00	0.50
		Refinement	**0.00**	**0.34(M)**	**0.04**	**0.41(S)**
	Python	One-off	0.49	0.50	0.34	0.47
		Refinement	0.11	0.47	0.58	0.51
GPT-4o	AMPL	One-off	0.95	0.56	1.00	0.50
		Refinement	1.00	0.57	0.07	0.39
	Python	One-off	**0.01**	**0.44(S)**	0.79	0.57
		Refinement	0.17	0.47	0.78	0.57
Gemini 2.5-Pro	AMPL	One-off	1.00	0.58	0.18	0.46
		Refinement	**0.00**	**0.42(S)**	0.55	0.51
	Python	One-off	1.00	0.59	0.48	0.50
		Refinement	1.00	0.58	0.32	0.47
o4-mini	AMPL	One-off	0.99	0.55	0.54	0.50
		Refinement	1.00	0.56	0.44	0.49
	Python	One-off	0.07	0.46	0.85	0.53
		Refinement	**0.03**	**0.46(N)**	0.52	0.50

Table 3. Statistical tests for **RQ3** comparing **refinement variants** and **one-of variants** based on (a) execution success rate (Success %) and zero relative error rates (#Zero/#Exec) using the Z-test; and (b) based on relative error using the Mann-Whitney test. Blue cells 🟦 indicate significant improvements of refinement over one-of variants. No significant improvements are observed in the opposite direction.

(a) Comparing the variants based on the Success % and #Zero/#Exec metrics

Variant			**Public**				**Industry**			
LLM	**Language**	**Structuring**	Success %		#Zero/#Exec		Success %		#Zero/#Exec	
			Z	***p-value***	**Z**	***p-value***	**Z**	***p-value***	**Z**	***p-value***
Gemini 1.5-Flash	AMPL	Unstructured	**10.68**	**0.00**	-4.91	1.00	1.55	0.06	-1.00	0.84
		Structured	**9.10**	**0.00**	**3.49**	**0.00**	**3.16**	**0.00**	-1.19	0.88
	Python	Unstructured	**2.71**	**0.00**	-0.31	0.62	0.29	0.39	-0.30	0.62
		Structured	**2.30**	**0.01**	0.69	0.24	0.61	0.27	-0.56	0.71
GPT-4o	AMPL	Unstructured	**7.13**	**0.00**	-0.63	0.73	**1.81**	**0.03**	-0.11	0.58
		Structured	**10.15**	**0.00**	-0.77	0.78	**3.13**	**0.00**	1.13	0.13
	Python	Unstructured	**2.23**	**0.01**	-0.60	0.73	1.50	0.06	-1.58	0.94
		Structured	1.48	0.07	-2.13	0.98	-0.89	0.81	-0.72	0.76
Gemini 2.5-Pro	AMPL	Unstructured	**5.13**	**0.00**	-3.74	1.00	0.79	0.21	-0.69	0.75
		Structured	**4.80**	**0.00**	**3.04**	**0.00**	**2.19**	**0.01**	-2.04	0.98
	Python	Unstructured	1.43	0.08	0.25	0.40	-0.15	0.56	-0.66	0.75
		Structured	0.08	0.47	0.75	0.22	0.09	0.46	-0.30	0.62
o4-mini	AMPL	Unstructured	**5.45**	**0.00**	-0.29	0.62	1.43	0.08	0.37	0.36
		Structured	**5.53**	**0.00**	-0.88	0.81	1.43	0.08	-1.16	0.88
	Python	Unstructured	0.92	0.18	**2.53**	**0.01**	0.00	0.50	-0.41	0.66
		Structured	1.15	0.12	**2.43**	**0.01**	0.58	0.28	-0.37	0.64

(b) Comparing the variants based on the relative error metric

Variant			**Public Dataset**		**Industry Dataset**	
LLM	**Language**	**Structuring**	***p-value***	$\hat{A}_{12}$	***p-value***	$\hat{A}_{12}$
Gemini 1.5-Flash	AMPL	Unstructured	1.00	0.64	0.86	0.58
		Structured	**0.00**	**0.41(S)**	0.82	0.56
	Python	Unstructured	0.68	0.51	0.69	0.53
		Structured	0.26	0.48	0.63	0.53
GPT-4o	AMPL	Unstructured	0.72	0.52	0.40	0.47
		Structured	0.78	0.52	0.16	0.41
	Python	Unstructured	0.78	0.52	0.88	0.61
		Structured	0.99	0.56	0.79	0.56
Gemini 2.5-Pro	AMPL	Unstructured	1.00	0.58	0.72	0.54
		Structured	**0.00**	**0.43(S)**	0.98	0.38
	Python	Unstructured	0.38	0.49	0.77	0.54
		Structured	0.21	0.48	0.62	0.52
o4-mini	AMPL	Unstructured	0.65	0.51	0.38	0.48
		Structured	0.78	0.52	0.88	0.56
	Python	Unstructured	0.07	0.47	0.64	0.52
		Structured	**0.01**	**0.45(N)**	0.67	0.52

Table 4. Statistical tests for **RQ4** compare EXEOS variants when used with **reasoning-based** LLMs, i.e., Gemini 2.5-Pro and o4-mini, versus when used with **instruction-following** LLMs, i.e., Gemini 1.5-Flash and GPT-4o, based on (a) execution success rate (Success) and zero relative error rates (#Zero/#Exec) using the Z-test, and (b) relative error using the Mann-Whitney test. Blue cells 🟦 indicate significant improvements in results obtained with reasoning-based LLMs over those obtained with instruction-following LLMs. No significant improvements are observed in the opposite direction.

(a) Comparing the variants based on the Success and #Zero/#Exec metrics

Variant			Public				Industry			
Language	**Structuring**	**Refinement**	Success %		#Zero/#Exec		Success %		#Zero/#Exec	
			Z	***p-value***	**Z**	***p-value***	**Z**	***p-value***	**Z**	***p-value***
AMPL	Unstructured	One-off	**12.78**	**0.00**	1.34	0.09	**2.30**	**0.01**	-0.02	0.51
		Refinement	**8.02**	**0.00**	**4.07**	**0.00**	**1.99**	**0.02**	0.65	0.26
	Structured	One-off	**11.23**	**0.00**	**2.51**	**0.01**	**2.58**	**0.00**	1.66	0.05
		Refinement	**4.10**	**0.00**	**2.38**	**0.01**	1.07	0.14	-0.27	0.60
Python	Unstructured	One-off	**9.79**	**0.00**	**3.09**	**0.00**	**2.59**	**0.00**	0.18	0.43
		Refinement	**8.03**	**0.00**	**5.84**	**0.00**	1.65	0.05	1.00	0.16
	Structured	One-off	**8.04**	**0.00**	-0.41	0.66	-0.42	0.66	**1.92**	**0.03**
		Refinement	**11.35**	**0.00**	-0.03	0.51	0.43	0.33	**2.28**	**0.01**

(b) Comparing the variants based on the relative error metric

Variant			Public Dataset		Industry Dataset	
Language	**Structuring**	**Refinement**	***p-value***	$\hat{A}_{12}$	***p-value***	$\hat{A}_{12}$
AMPL	Unstructured	One-off	**0.04**	**0.47(N)**	0.40	0.49
		Refinement	**0.00**	**0.42(S)**	0.38	0.48
	Structured	One-off	**0.01**	**0.45(N)**	0.05	0.44
		Refinement	**0.01**	**0.47(N)**	0.60	0.51
Python	Unstructured	One-off	**0.00**	**0.45(N)**	0.41	0.49
		Refinement	**0.00**	**0.42(S)**	0.19	0.46
	Structured	One-off	0.76	0.51	**0.03**	**0.42(S)**
		Refinement	0.62	0.51	**0.01**	**0.39(M)**

Table 5. Results comparing the baseline with the best-performing AMPL and Python variants of EXEOS using the Public dataset. Blue cells 🟦 indicate significant improvements of EXEOS over the baseline. No significant improvements are observed in the opposite direction.

(a) Executability and correctness results for the baseline on the Public dataset

Metric	**Gemini 1.5 Flash**	**GPT-4o**	**Gemini 2.5 Pro**	**o1-mini**
#Exec (Succ.%)	171 (57%)	194 (65%)	268 (89%)	209 (70%)
Mean (RelErr)	1.45	4.05	1.30	1.55
Med (RelErr)	0	0	0	0
Std (RelErr)	7.01	40.70	5.95	4.02
#Zero (RelErr)	98	127	158	155

(b) Statistical tests comparing Ampl4 and Python4 and the baseline

LLM	Ampl4 vs. Baseline						Python4 vs. Baseline					
	Success		#Zero/#Exec		RelErr		Success		#Zero/#Exec		RelErr	
	p-val	Z	*p-val*	Z	*p-val*	$\hat{A}_{12}$	*p-val*	Z	*p-val*	Z	*p-val*	$\hat{A}_{12}$
Gemini 1.5 Flash	**0.00**	**4.88**	**0.00**	**3.53**	**0.00**	**0.37(M)**	0.97	-1.84	0.11	1.25	0.17	0.48
GPT-4o	**0.00**	**7.27**	0.98	-2.11	0.96	0.55	0.14	1.10	0.29	0.55	0.24	0.48
Gemini 2.5 Pro	**0.00**	**8.12**	**0.00**	**2.82**	**0.00**	**0.39(M)**	**0.00**	**6.19**	0.37	0.32	0.14	0.47
o4-mini	**0.00**	**9.49**	0.99	-2.27	0.94	0.54	**0.02**	**1.99**	**0.01**	**2.28**	**0.00**	**0.43(S)**