Statistical Test Results for RQ1–RQ4 (COMPLEMENTS RESULTS in SECTION 5)

NEGIN AYOUGHI, DAVID DEWAR, SHIVA NEJATI AND MEHRDAD SABETZADEH

Table 1. Statistical tests for **RQ1** comparing **AMPL-based variants** and **Python-based variants** based on (a) execution success rate (Success) using the Z-test; and (b) based on relative-error using the Mann-Whitney U test. Blue cells indicate significant improvements of AMPL-based over Python-based variants. Green cells indicate significant improvements of Python-based over AMPL-based variants.

(a) Comparing the variants based on the Success metric

Variant			Public		Industry	
LLM	**Structuring**	**Refinement**	**Z**	***p-value***	**Z**	***p-value***
Gemini 1.5-Flash	Unstructured	One-off	-5.31	0.00	-2.34	0.02
		Refinement	3.02	0.00	-1.10	0.28
	Structured	One-off	-0.32	0.74	-2.61	0.00
		Refinement	6.64	0.00	0.00	1.00
GPT-4o	Unstructured	One-off	-8.56	0.00	0.00	1.00
		Refinement	-3.72	0.00	0.27	0.80
	Structured	One-off	-2.76	0.00	-4.39	0.00
		Refinement	6.24	0.00	-0.55	0.58
Gemini 2.5-Pro	Unstructured	One-off	-4.82	0.00	-1.39	0.16
		Refinement	-1.06	0.28	-0.47	0.64
	Structured	One-off	-2.62	0.00	-2.09	0.04
		Refinement	2.20	0.02	0.00	1.00
o4-mini	Unstructured	One-off	-4.33	0.00	-1.13	0.26
		Refinement	0.26	0.80	0.31	0.76
	Structured	One-off	3.71	0.00	-0.55	0.58
		Refinement	7.77	0.00	0.31	0.76

(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) using the Z-test; and (b) based on relative-error using the Mann-Whitney U test. Blue cells indicate significant improvements of structured over unstructured variants. Green cells indicate significant improvements of unstructured over structured variants. All reported p-values are rounded to two decimal places.

(a) Comparing the variants based on the Success metric

Variant			Public		Industry	
LLM	**Language**	**Refinement**	**Z**	***p-value***	**Z**	***p-value***
Gemini 1.5-Flash	AMPL	One-off	2.63	0.00	0.00	1.00
		Refinement	0.96	0.34	1.69	0.08
	Python	One-off	-2.39	0.02	0.29	0.78
		Refinement	-2.79	0.00	0.61	0.54
GPT-4o	AMPL	One-off	2.44	0.00	-1.39	0.16
		Refinement	5.73	0.00	0.00	1.00
	Python	One-off	-3.53	0.00	3.16	0.00
		Refinement	-4.25	0.00	0.81	0.42
Gemini 2.5-Pro	AMPL	One-off	0.67	0.50	-1.06	0.30
		Refinement	0.31	0.76	0.38	0.70
	Python	One-off	-1.62	0.10	-0.33	0.74
		Refinement	-2.92	0.00	-0.09	0.94
o4-mini	AMPL	One-off	1.92	0.06	0.00	1.00
		Refinement	2.15	0.04	0.00	1.00
	Python	One-off	-6.07	0.00	-0.58	0.56
		Refinement	-5.80	0.00	0.00	1.00

(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-off variants** based on (a) execution success rate (Success) using the Z-test; and (b) based on relative-error using the Mann-Whitney U test. Blue cells indicate significant improvements of refinement over one-off variants. No significant difference where one-off variants outperform refinement variants. All reported p-values are rounded to two decimal places.

(a) Comparing the variants based on the Success metric

Variant			Public		Industry	
LLM	**Language**	**Structuring**	**Z**	***p-value***	**Z**	***p-value***
Gemini 1.5-Flash	AMPL	Unstructured	10.68	0.00	1.55	0.12
		Structured	9.10	0.00	3.16	0.00
	Python	Unstructured	2.71	0.00	0.29	0.78
		Structured	2.30	0.02	0.61	0.54
GPT-4o	AMPL	Unstructured	7.13	0.00	1.81	0.06
		Structured	10.15	0.00	3.13	0.00
	Python	Unstructured	2.23	0.02	1.50	0.12
		Structured	1.48	0.14	-0.89	0.38
Gemini 2.5-Pro	AMPL	Unstructured	5.13	0.00	0.79	0.42
		Structured	4.80	0.00	2.19	0.02
	Python	Unstructured	1.43	0.16	-0.15	0.88
		Structured	0.08	0.94	0.09	0.92
o4-mini	AMPL	Unstructured	5.45	0.00	1.43	0.16
		Structured	5.53	0.00	1.43	0.16
	Python	Unstructured	0.92	0.36	0.00	1.00
		Structured	1.15	0.24	0.58	0.56

(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** comparing EXEOS variants when used with **reasoning** 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) using the Z-test, and (b) relative-error using the Mann-Whitney U test. Blue cells indicate significant improvements in results obtained with reasoning LLMs over those obtained with instruction-following LLMs. No significant difference where instruction-following LLMs outperforms reasoning. All reported p-values are rounded to two decimal places.

(a) Comparing the variants based on the Success metric

Variant			Public		Industry	
Language	**Structuring**	**Refinement**	**Z**	***p-value***	**Z**	***p-value***
AMPL	Unstructured	One-off	12.78	0.00	2.30	0.02
		Refinement	8.02	0.00	1.99	0.04
	Structured	One-off	11.23	0.00	2.58	0.00
		Refinement	4.10	0.00	1.07	0.28
Python	Unstructured	One-off	9.79	0.00	2.59	0.00
		Refinement	8.03	0.00	1.65	0.10
	Structured	One-off	8.04	0.00	-0.42	0.68
		Refinement	11.35	0.00	2.28	0.02

(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)**