

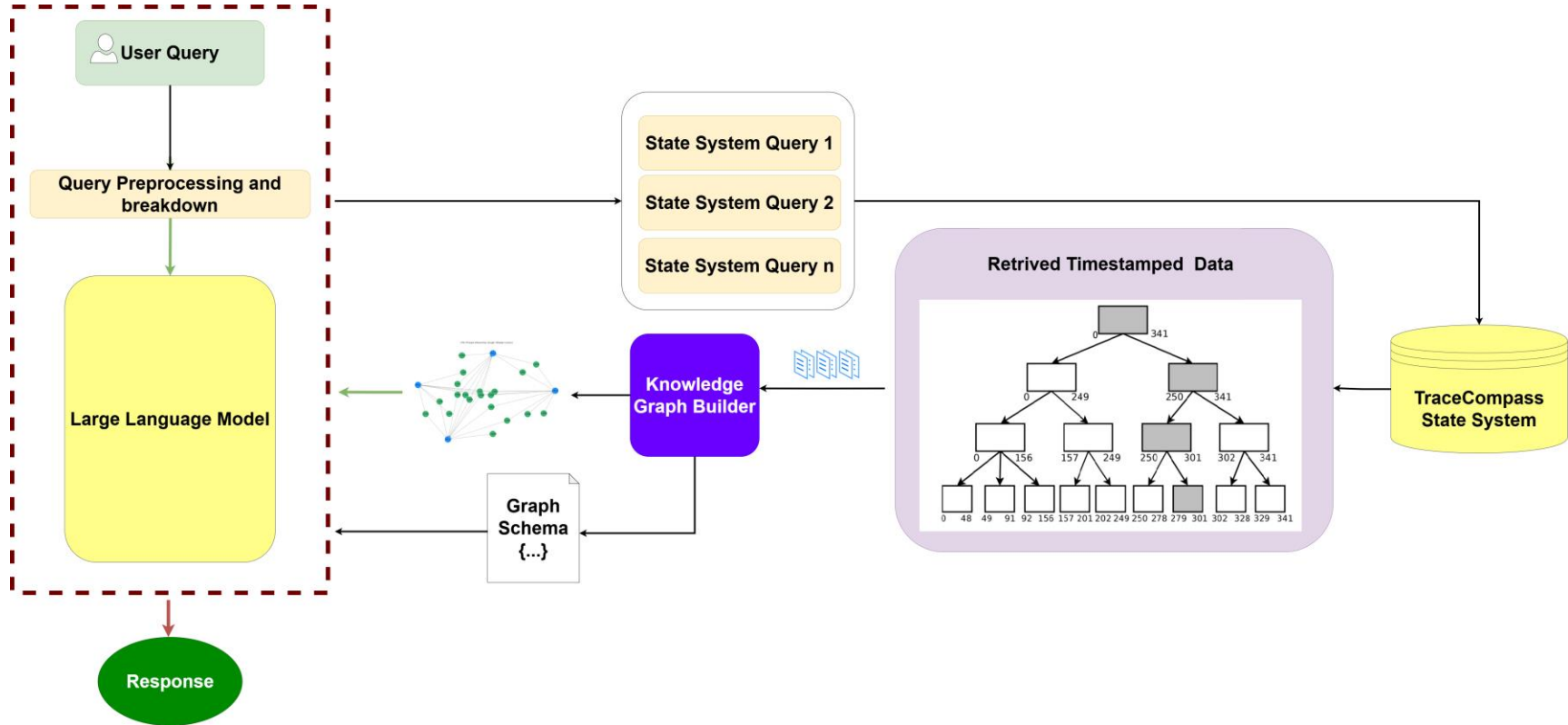
Dec 5



TAAF: A Trace Abstraction and Analysis Framework Synergizing Knowledge Graphs and LLMs

Alireza Ezaz

Methodology Architecture



Research Questions

- **RQ1:** To what extent does the incorporation of a Knowledge Graph improve the accuracy of the TAAF when answering trace-related queries?
- **RQ2:** Does providing the LLM with the graph schema (node types and features) enhance the accuracy of its responses?
- **RQ3:** How does the accuracy and quality of TAAF responses vary across different LLM models (GPT-4 .1 nano (small), GPT-4o, o4-mini (Reasoning))?
- **RQ4:** What is the effect of time interval length on the performance of TAAF's answers?
- **RQ5:** How does TAAF's accuracy vary across different query types (e.g., multiple-choice, true/false, explanatory), and graph structures (single-hub vs. multi-hub)?

Approach

- We designed **100** unique questions.
- Each question is asked **3 times** under the raw-data condition, yielding **300** data points.
- Each question is also asked three times under the KG-powered condition, adding another **300** data points.
- Therefore, each experiment produces **600** data points in total.
- To date, we've run **10** experiments, resulting in **6,000** data points overall.
- The main Metric we use is **Accuracy:**

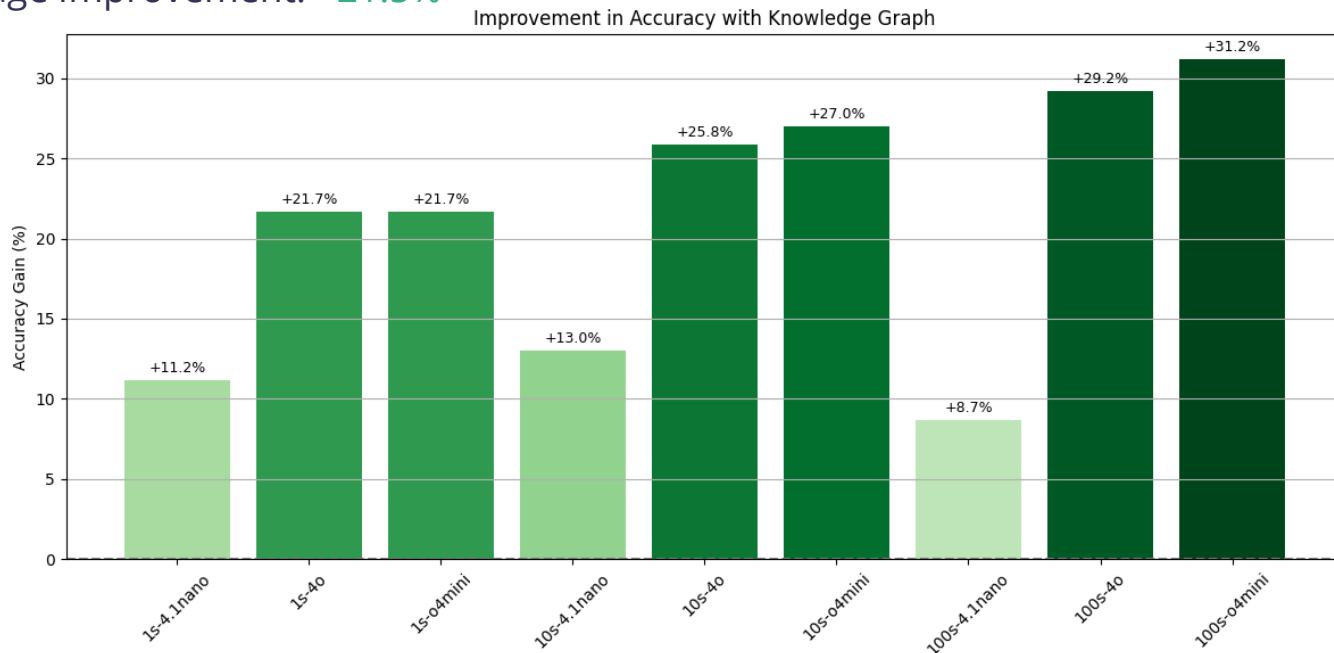
$$\frac{(Number\ of\ 0s \times 0) + (Number\ of\ 0.5s \times 0.5) + (Number\ of\ 1s \times 1)}{300} \times 100$$

RQ1: To what extent does the incorporation of a Knowledge Graph improve the accuracy of the TAAF when answering trace-related queries?

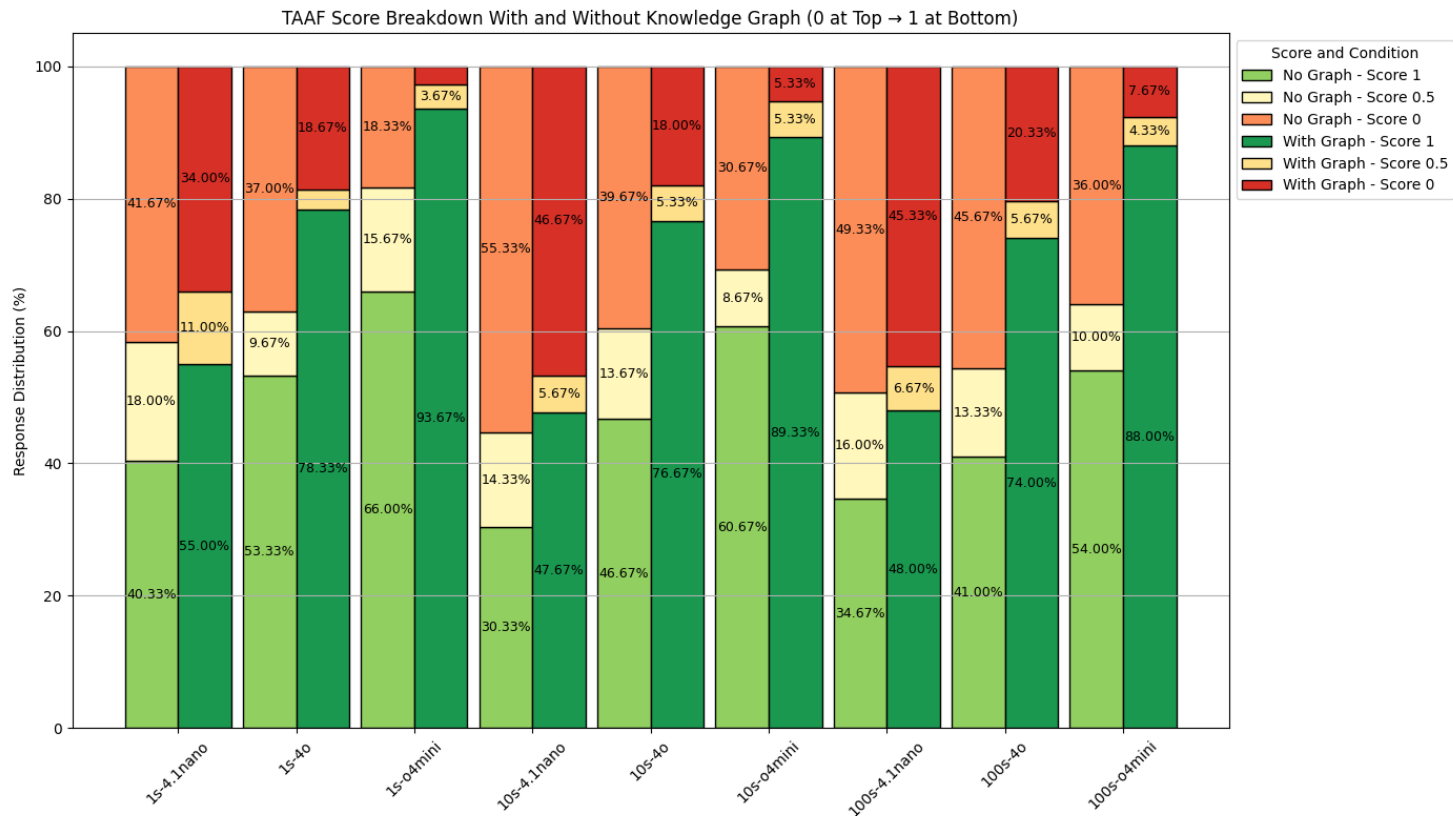
Min Improvement: +8.7%

Max Improvement: + 31.2%

Average Improvement: +21.5%



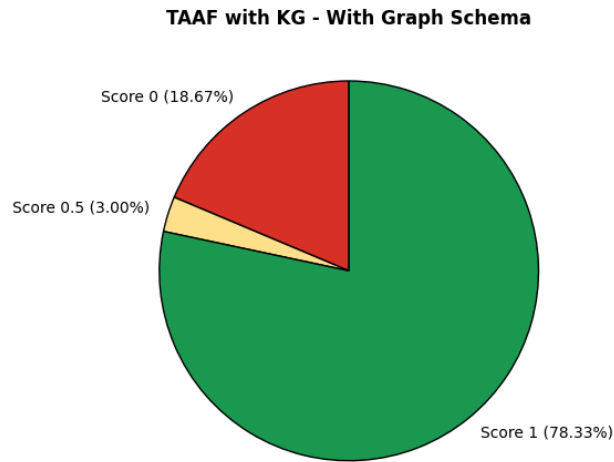
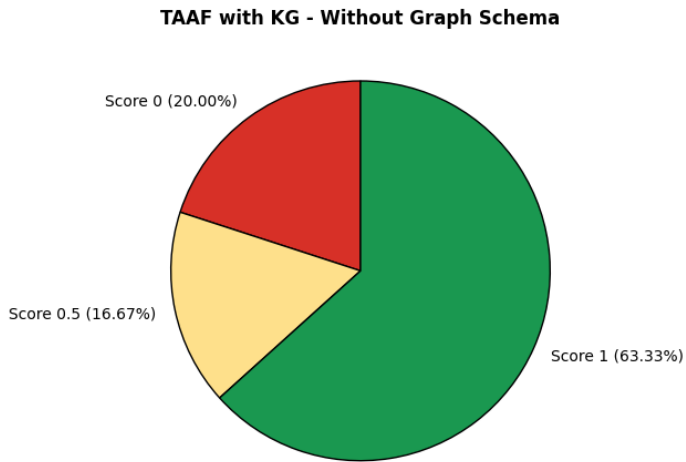
RQ1: To what extent does the incorporation of a Knowledge Graph improve the accuracy of the TAAF when answering trace-related queries?



RQ2: Does providing the LLM with the graph schema (node types and features) enhance the accuracy of its responses?

Experiment was done on 1s of data from the mid timestamp with 4o model
+21.67% Improvement

Effect of Graph Schema on Accuracy (Improvement: +21.67%)

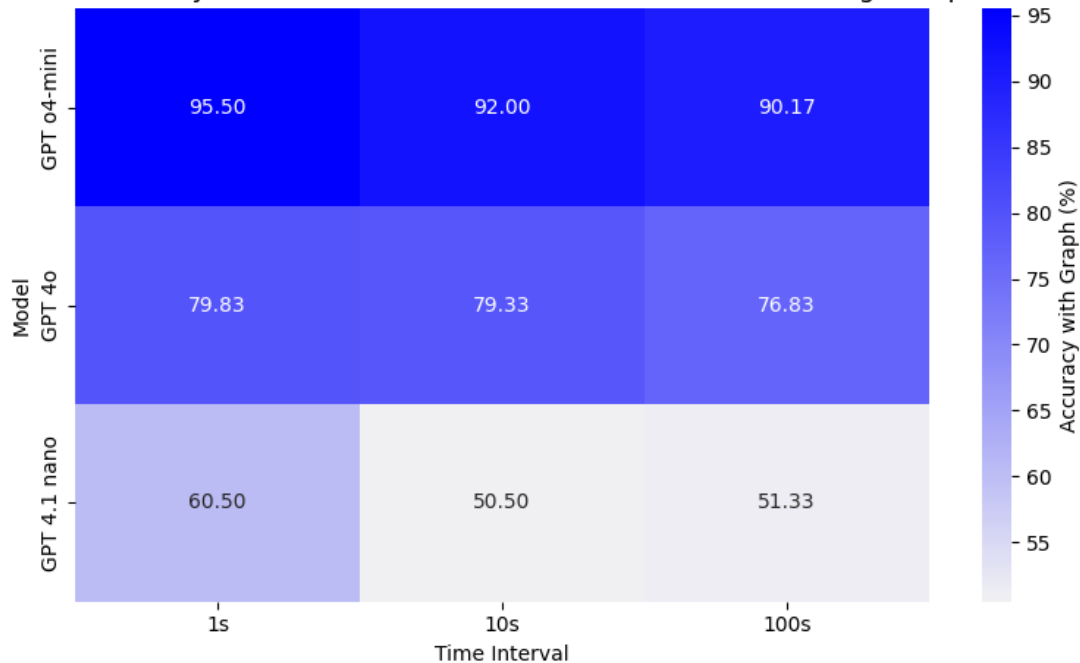


RQ3: How does the accuracy and quality of TAAF responses vary across different LLM models (GPT-4 .1 nano (small), GPT-4o, o4-mini (Reasoning))?

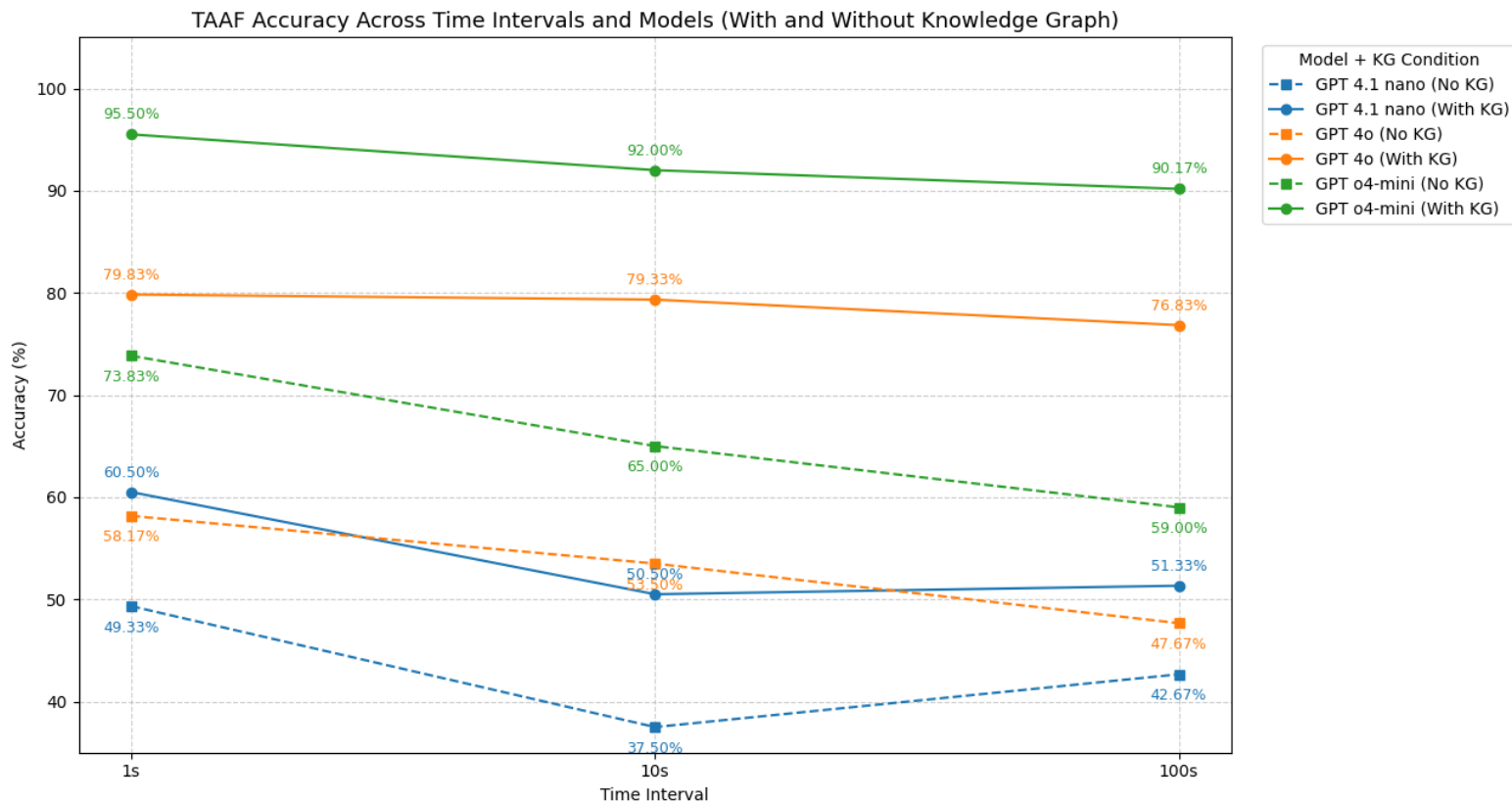
Min Accuracy: 50.50

Max Accuracy: 95.50

TAAF Accuracy Across Models and Time Intervals (With Knowledge Graph)



RQ4: What is the effect of time interval length on the performance of TAAF's answers?

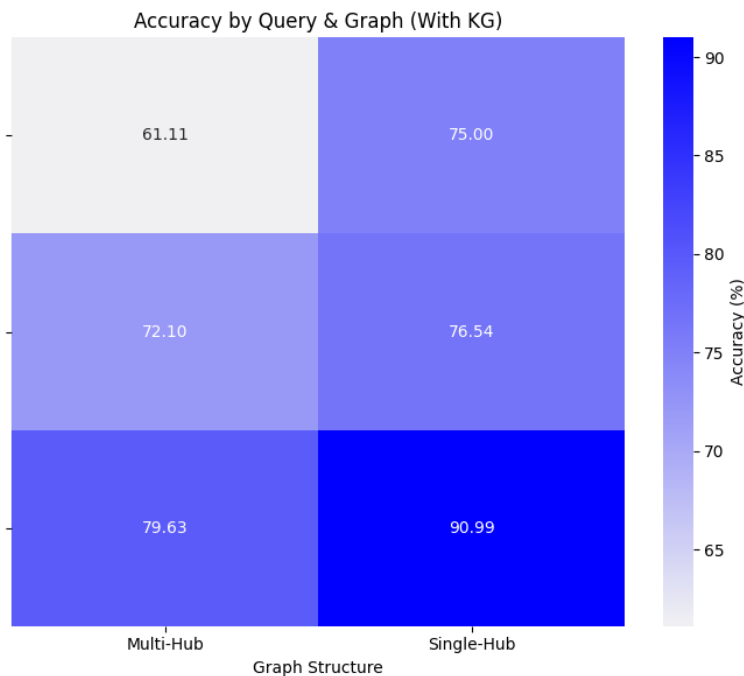
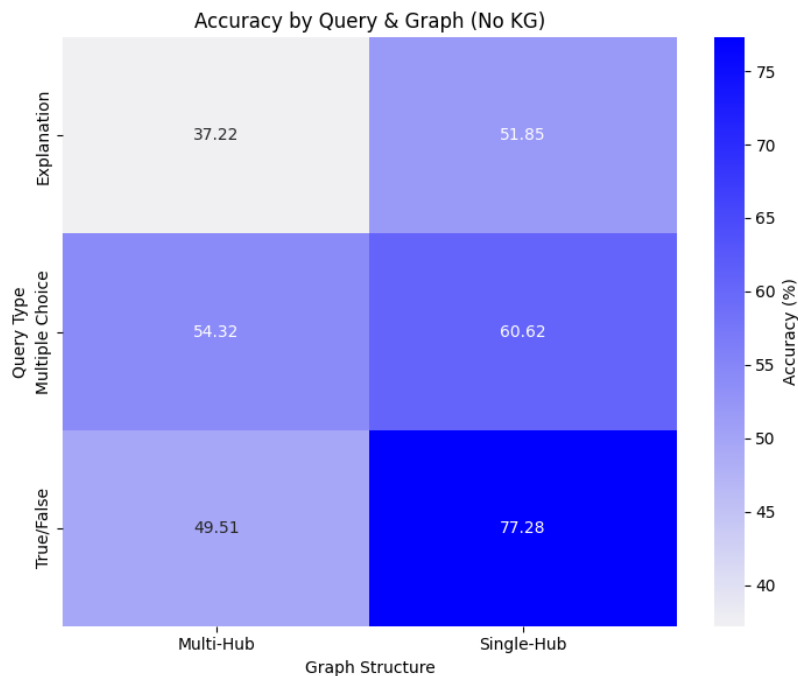


RQ5: How does TAAF's accuracy vary across different query types (e.g., multiple-choice, true/false, explanatory), and graph structures (single-hub vs. multi-hub)?

Note: Aggregated results across all models

TAAF Best Accuracy: True/False Single-Hub 90.99%

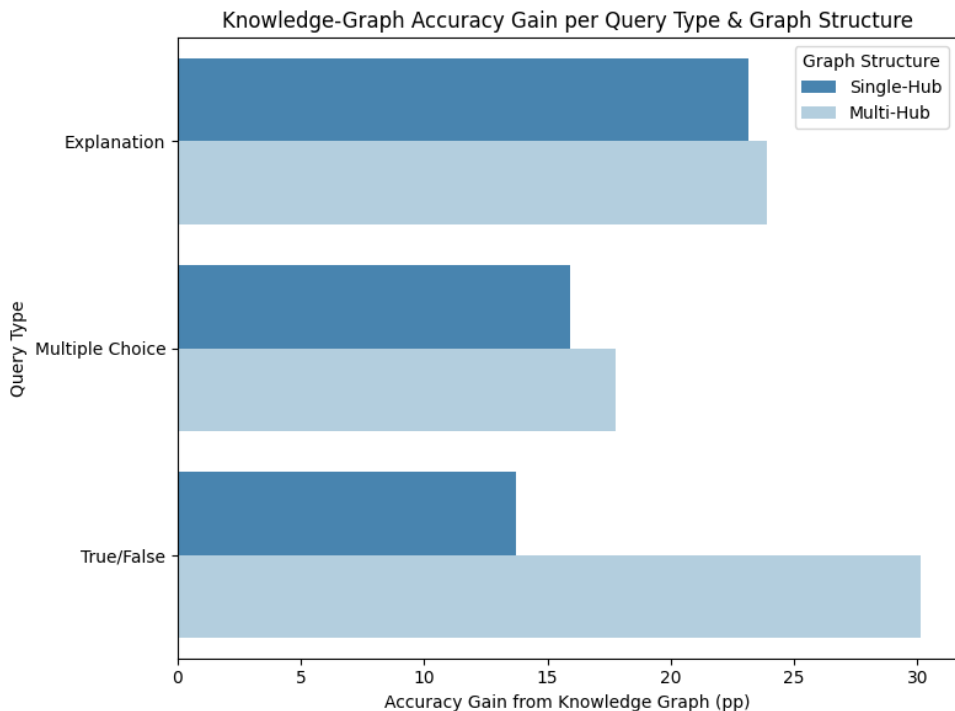
TAAF Worst Accuracy: Explanation Multi-Hub 37.22%



RQ5: How does TAAF's accuracy vary across different query types (e.g., multiple-choice, true/false, explanatory), and graph structures (single-hub vs. multi-hub)?

Note: Aggregated results across all models

TAAF Best Accuracy Gain: True/False Multi-Hub 30%



RQ5: How does TAAF's accuracy vary across different query types (e.g., multiple-choice, true/false, explanatory), and graph structures (single-hub vs. multi-hub)?

Table 4.1: GPT 4.1 nano — 1 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single-Hub	45.00	13.33	41.67	48.33	33.33	8.33	58.33	62.50
	Multi-Hub	61.67	20.00	18.33	28.33	46.67	20.00	33.33	43.33
Multiple Choice	Single-Hub	35.56	17.78	46.67	55.56	11.11	5.56	83.33	86.11
	Multi-Hub	40.00	11.11	48.89	54.44	26.67	4.44	68.89	71.11
True/False	Single-Hub	20.00	8.89	71.11	75.56	10.00	2.22	87.78	88.89
	Multi-Hub	41.11	21.11	37.78	48.33	16.67	12.22	71.11	77.22

RQ5:How does TAAF's accuracy vary across different query types (e.g., multiple-choice, true/false, explanatory), and graph structures (single-hub vs. multi-hub)?

Table 4.2: GPT 4.1 nano — 10 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	46.67	16.47	36.67	45.00	43.33	5.00	54.67	54.17
	Multi Hub	56.67	21.47	21.67	32.50	46.67	13.33	40.00	46.67
Multiple Choice	Single Hub	50.00	13.33	36.67	43.33	29.17	5.83	65.00	67.92
	Multi Hub	55.67	13.80	31.33	37.83	37.14	4.29	58.57	60.71
True/False	Single Hub	27.78	9.44	62.78	67.50	11.11	4.44	84.44	86.97
	Multi Hub	55.56	16.47	27.78	36.11	29.17	9.17	61.67	66.25

Table 4.3: GPT 4.1 nano — 100 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	49.03	15.97	34.96	42.38	45.10	7.84	47.06	50.00
	Multi Hub	59.15	21.13	19.72	30.29	49.30	19.72	30.99	40.45
Multiple Choice	Single Hub	42.40	19.81	36.78	46.70	28.61	6.42	68.97	70.18
	Multi Hub	48.50	13.30	38.00	44.75	33.58	6.72	59.70	62.96
True/False	Single Hub	35.63	14.38	50.00	57.19	14.13	6.38	79.50	82.69
	Multi Hub	55.10	17.45	27.25	36.97	26.13	10.92	62.95	67.41

Table 4.4: GPT 4o — 1 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	38.00	8.00	62.00	66.00	18.00	2.67	79.33	80.67
	Multi Hub	46.00	12.00	42.00	48.00	29.33	9.33	61.33	66.00
Multiple Choice	Single Hub	26.00	6.00	68.00	71.00	7.00	1.33	96.67	92.33
	Multi Hub	33.33	8.33	58.33	62.50	13.33	3.33	83.33	85.00
True/False	Single Hub	13.33	5.00	81.67	84.17	4.00	0.67	96.33	95.67
	Multi Hub	32.00	14.47	53.33	60.67	12.67	3.33	84.00	85.57

Table 4.5: GPT 4o — 10 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	34.48	11.49	54.02	59.76	33.33	3.45	63.22	64.94
	Multi Hub	50.00	12.50	37.50	43.75	38.46	11.54	50.00	55.77
Multiple Choice	Single Hub	34.55	12.73	52.73	58.09	13.64	2.73	83.64	85.00
	Multi Hub	46.43	11.90	41.67	47.62	22.32	3.57	74.11	75.89
True/False	Single Hub	18.27	6.73	75.00	78.47	6.25	2.50	98.25	92.50
	Multi Hub	41.67	16.47	41.67	50.00	17.44	10.47	72.09	76.32

Table 4.6: GPT 4o — 100 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	39.39	12.12	48.48	54.55	29.54	8.41	62.95	66.26
	Multi Hub	53.17	15.52	29.31	36.67	35.29	14.74	49.47	56.84
Multiple Choice	Single Hub	38.22	14.71	46.08	53.44	11.36	3.52	84.31	86.27
	Multi Hub	50.88	14.04	35.09	42.11	23.81	5.95	70.24	72.22
True/False	Single Hub	26.92	9.62	63.46	68.27	7.21	2.88	89.90	91.35
	Multi Hub	47.83	17.39	34.78	43.48	15.85	6.71	77.44	80.80

Table 4.7: GPT o4-mini — 1 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	17.78	15.56	66.67	74.44	4.44	5.00	90.56	93.86
	Multi Hub	35.24	22.86	41.90	53.33	8.57	14.29	77.14	84.29
Multiple Choice	Single Hub	17.02	8.54	74.47	78.72	3.19	1.06	95.74	96.27
	Multi Hub	21.21	14.14	64.65	71.72	5.41	2.70	91.89	93.24
True/False	Single Hub	8.33	7.22	84.44	88.96	2.22	1.11	96.67	97.22
	Multi Hub	21.75	13.75	64.50	71.88	4.37	2.46	93.17	94.35

Table 4.8: GPT o4-mini — 10 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	9.52	13.10	77.38	83.33	5.95	5.95	88.10	91.87
	Multi Hub	27.40	19.86	52.74	42.87	9.59	13.70	76.71	83.56
Multiple Choice	Single Hub	13.68	10.26	76.06	81.19	3.08	3.08	93.85	95.28
	Multi Hub	18.10	11.43	70.48	76.19	5.24	2.86	91.90	93.33
True/False	Single Hub	3.33	5.00	91.67	94.17	0.58	0.58	98.89	99.17
	Multi Hub	14.29	12.14	73.57	79.94	1.78	3.56	94.67	96.44

Table 4.9: GPT o4-mini — 100 s interval

Query Type	Graph Structure	Without KG				With KG			
		0%	0.5%	1%	Acc %	0%	0.5%	1%	Acc %
Explanation	Single Hub	12.50	12.50	75.00	81.25	6.15	6.15	87.69	90.76
	Multi Hub	25.64	17.85	56.41	55.38	7.69	12.82	79.49	85.90
Multiple Choice	Single Hub	17.59	11.69	71.81	78.43	4.98	2.48	92.56	94.80
	Multi Hub	22.22	11.11	66.67	72.22	5.88	2.94	91.18	92.65
True/False	Single Hub	6.58	7.89	85.53	89.47	1.54	1.92	96.54	97.50
	Multi Hub	18.25	13.14	68.61	75.18	2.55	3.63	93.83	95.64

Future Research Questions

- **RQ6:** To what extent does the choice of temporal location (early vs. late in the trace) affect system performance and reasoning accuracy?
- **RQ7:** How does the temperature (sampling randomness) parameter affect the accuracy and consistency of LLM responses within TAAF?
- **RQ8:** How does introducing an AI agent (e.g., multi-turn dialogue, clarification loops) improve the interpretability and correctness of TAAF responses?
- **RQ9:** How does TAAF's performance compare to using only raw event data with LLMs, without leveraging the state system or knowledge graph?



Thanks!

Do you have any questions?



sezaz@brocku.ca



www.linkedin.com/in/s-alireza-ezaz

