Appendix

Anonymous submission

1 Detail of Factors

All factors we studied are listed in Table 4.

2 Details of Experiments

2.1 Datasets

- GraphextQA: A retrieval-independent dataset designed to study LLMs for KG understanding and utilization. The test set has 2890 samples, each sample consisting of a question and a high-quality retrieved sub-KG, and using WikiData (Vrandečić and Krötzsch 2014) as the KG source. Because this dataset excludes the effect of retrieval, most of our experiments were performed on it.
- CWQ and WQSP: Two widely used datasets that use FreeBase (Bollacker et al. 2008) as the KG source. Their test sets have 3531 and 1639 samples, respectively. Since they do not provide retrieved sub-KGs, we use the retrieval results of LUO et al. (2024).

The statistics of the datasets are shown in Table 1.

Dataset	sample_num (test set)	KG	nodes_num (avg±std)	edges_num (avg±std)	answer_cover_rate
WQSP	1628	FreeBase	18.74±23.81	30.84±70.59	0.93
CWQ	3531	FreeBase	23.92±28.55	45.02±76.33	0.78
GraphextQA	2890	WikiData	4.73±1.64	4.42±1.87	0.98

Table 1: Dataset statistics

2.2 Hyperparameters

All hyperparameters are listed in Table 2.

2.3 Evaluation Metrics

We use Hits@1 to evaluate the correctness of the prediction, which is defined as:

$$His@1 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i \in y_i^*) \tag{1}$$

where N is the number of samples, y_i is the model generated answer and $y*_i$ is the correct answers at sample i. Generally, it's the same as Recall.

	Parameters	Values
General	seed	42
Generation	temperature	0
	top-p	0.25
	max_new_tokens	128
	max_seq_len_to_capture	2048
	max_model_len	2048
	max_num_seqs	2048
	decoding strategy	greedy
	precision	float16
	GPU	single GeForce RTX 3090 24GB
Fine-Tuning (LoRA)	learning_rate	5.00E-05
	micro_batch_size	2
	batch_size	64
	max_epochs	10
	early_stop_patience	1
	max_seq_length	1024
	warmup_proportion	0.1
	weight_decay	0.02
	lora_r	64
	lora_alpha	64*4
	precision	bfloat16
	GPU	single NVIDIA A800 40GB

Table 2: Hyperparameters

We use Rank-biased Overlap (RBO) (Webber, Moffat, and Zobel 2010) to measure the consistency of different factors across settings to indicate generalizability, which is defined as:

$$I_d = S_{1:d} \cap T_{1:d} \tag{2}$$

$$A_d = \frac{|I_d|}{d} = \frac{|S_{1:d} \cap T_{1:d}|}{d} \tag{3}$$

$$RBO(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d \tag{4}$$

where S and T are two infinite rankings, $1:d$ means the set of the elements from position 1 to position d in the list. A_d is the overlap of lists S and T to depth d. The parameter p determines how steep the decline in weights is: the smaller p, the more top-weighted the metric is.

2.4 Implementation Details

Feature analysis methods

- Permutation Importance: Randomize the values of each feature and then monitor how much the model performance decreases, if a larger decrease is obtained it means that the feature is more important.
- feature importance in Random Forests: Feature importance scores within the random forest model, show the contribution of each feature to the final prediction.
- feature importance in XGBoost: Feature importance scores within the XGBoost, show the contribution of each feature to the final prediction.
- Pearson Correlation: Measures the linear relationship (linear correlation) between each feature and the target variable, the higher the correlation the more important the feature.
- Recursive Feature Elimination: Recursively remove features and see how they affect the model performance, features that lead to greater degradation after removal are more important.
- PCA: Perform a principal component analysis on the features and look at the explained variance ratio for each principal component.
- ANOVA: Analysis of Variance (ANOVA) tests whether there is a significant difference between multiple sample means. It is based on the concept of variance and determines whether there is a significant difference between groups by comparing within-sample (intra-group) and between-sample (inter-group) variation.

We use the Python library sklearn[1] to implement the above methods.

Graph features importance We use the matrix consisting of all the features as input, and the model's Hits@1 on each sample as the prediction target. Then, we use the above seven feature analysis methods to calculate the importance of each feature. For node-level and edge-level graph features, since they have multiple values on each sample, we use their average, maximum, and minimum values. In practice, we also use text features such as number of questions, number of answers, length of input text, etc., in addition to graph features.

After that, for each graphical feature, there are seven importance scores from seven feature analysis methods. We rank features based on their mean of the seven importance scores, with higher rankings indicating that the feature has a greater impact on the final generation of the model.

Details of interpretability analysis We use t-SNE to search patterns in the information flows of the LLM (Figure 5 in our paper). We follow previous work (Ferrando and Voita 2024), recording the importance values of all the sub-edges corresponding to individual attention heads, as well as FFN blocks. To use t-SNE, we need to represent the information flow corresponding to each token (position) as a vector, each vector corresponding to the *pos*-th position is defined as:

$$(\sum_j e_{pos,j}^{1,1}, \sum_j e_{pos,j}^{1,2}, ..., \sum_j e_{pos,j}^{L,H}, e_{pos}^{ffn_1}, ..., , e_{pos}^{ffn_L})$$

[1]https://scikit-learn.org/stable/index.html

where $e_{pos,j}$ is the importance value in the information flow of the model, i.e., the weight of the edge from *pos*-th token to *j*-th token, L is the number of layers and H is the number of attention heads.

Similarly, for each token, we list the importance value of each attention head and FFN in each layer (Figure 6 in paper).

3 More Results

3.1 Graph Features

Top-30 graph features are listed in Table 3.

Graph Features	avg	std	range
node_avg_degree_centrality	42.20	13.79	69.06
edge_avg_edge_betweenness_centrality	45.78	11.36	69.78
node_avg_average_neighbor_degree	46.40	11.16	64.41
node_avg_information_centrality	48.50	11.64	69.07
node_avg_current_flow_betweenness_centrality	49.65	14.95	75.57
global_reaching_centrality	49.77	11.50	74.08
node_avg_closeness_centrality	50.57	13.15	70.71
node_avg_katz_centrality	52.29	13.51	64.21
node_avg_eccentricity	52.33	16.08	90.92
non_randomness	52.76	11.75	73.68
node_avg_communicability_betweenness_centrality	54.03	13.49	73.51
node_avg_harmonic_centrality	54.98	10.75	67.11
edge_avg_preferential_attachment	55.42	16.47	83.53
degree_assortativity_coefficient	58.35	12.73	84.27
edge_avg_edge_current_flow_betweenness_centrality	58.98	10.75	76.54
node_avg_laplacian_centrality	59.16	12.88	75.54
s_metric	60.48	13.14	77.43
node_avg_eigenvector_centrality	61.32	14.28	79.15
edge_avg_jaccard_coefficient	62.30	10.70	85.58
kemeny_constant	65.21	11.84	81.59
node_avg_betweenness_centrality	65.57	8.75	73.25
node_avg_subgraph_centrality	67.02	15.76	86.38
average_node_connectivity	68.95	12.09	88.05
edge_avg_common_neighbor_centrality	69.66	13.49	85.37
edge_avg_edge_load_centrality	69.67	14.97	79.60
edge_avg_adamic_adar_index	70.44	10.55	85.30
is_tree	71.46	10.08	80.49
number_attracting_components	71.68	12.59	87.41
bridges_num	72.39	10.95	82.91
diameter	72.68	14.70	105.44

Table 3: Top-30 graph features. Those starting with node or edge indicate that they are node-level or edge-level features, while others are graph-level features.

3.2 Formats

As shown in Table 5, we find that the path-based methods may discard some triples that are not on the path, which is why the path-based formats are worse than flat triples. In addition, the directed path-based formats are the worst because there are fewer reachable paths in the directed graph, leading to a more severe loss of information. This suggests that when sub-KGs are of high quality, we should keep their original information intact. However, when there is more noise in the sub-KG (e.g., on the CWQ and WQSP datasets), the path-based formats achieve the opposite performance, outperforming the flat triples because of their effective removal of irrelevant information. This illustrates the importance of retrieval in KG+RAG, where the problem of knowledge noise can impair generation performance.

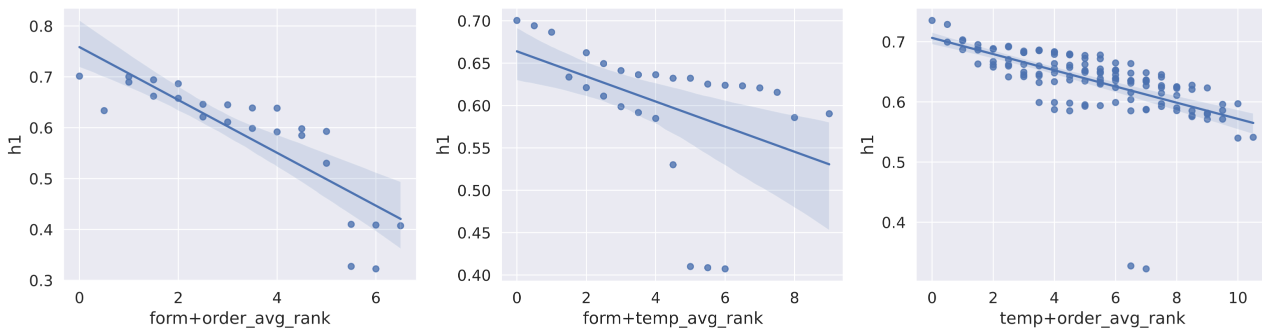

Figure 1: The average rankings of different factor combinations are positively correlated with the generation performance, indicating the combinability between different factors.

3.3 Orders

In Table 6, We find that different orders perform on multiple models with similar means and variances, and there is no significantly better order. Except for the two methods based on directed graphs, which perform significantly worse because of their loss of information. The order is model-dependent, with each model having its own preferred order and changing little with the influence of the dataset or the few-shot.

3.4 Templates

Figure 2: The performance of different size models across templates, based on the GraphextQA test set. As the number of parameters increases, the mean of the model's performance on different templates improves while the variance decreases, suggesting that large models are less sensitive to templates.

As model parameters (capabilities) increase, models become less sensitive to templates (see Figure 2 due to larger models having a denser information flow, using information from all tokens and not relying on separators (templates) only (see Figure 3).

3.5 Usability

We find the factors in the linearization phase are combinable (see Figure 1). We try combinations between formats+orders, formats+templates, and orders+templates and find that their rankings are positively correlated with the model generation performance. This means using two optimal methods simultaneously usually remains optimal.

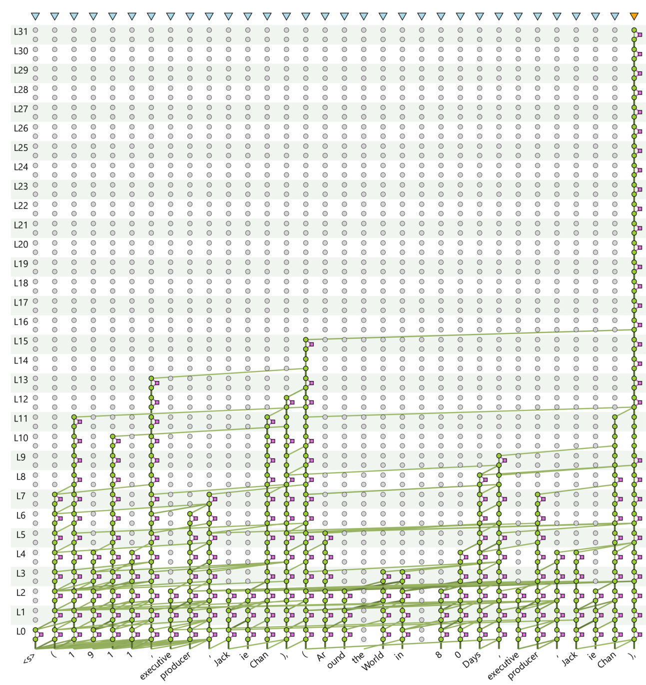

Figure 3: Information flow of llama-7b.

<table>
<tr><th>Phases</th><th>Categories</th><th>Factors</th></tr>
<tr><td rowspan="3">Graph Transformation Phase</td><td>node-level</td><td>degree_centrality, eigenvector_centrality, katz_centrality, closeness_centrality,
information_centrality, betweenness_centrality, current_flow_betweenness_centrality,
communicability_betweenness_centrality, load_centrality, subgraph_centrality,
harmonic_centrality, percolation_centrality, second_order_centrality, laplacian_centrality,
average_neighbor_degree, number_of_cliques, triangles, clustering, core_number,
eccentricity, pagerank, hits, constraint, effective_size, closeness_vitality</td></tr>
<tr><td>edge-level</td><td>edge_betweenness_centrality, edge_current_flow_betweenness_centrality,
edge_load_centrality, resource_allocation_index, jaccard_coefficient,
adamic_adar_index, preferential_attachment, common_neighbor_centrality</td></tr>
<tr><td>graph-level</td><td>estrada_index, global_reaching_centrality, node_connectivity,
maximum_independent_set_size, large_clique_size, average_clustering,
diameter, treewidth, min_weighted_vertex_cover_size, minimum_cut,
degree_assortativity_coefficient, asteroidal_triple_num, bridges_num,
clique_num, transitivity, number_connected_components,
number_strongly_connected_components, number_weakly_connected_components,
number_attracting_components, number_bridge_components, average_node_connectivity,
node_connectivity, edge_connectivity, minimum_edge_cut_num, minimum_node_cut_num,
min_edge_cover_num, simple_cycles_num, girth, kemeny_constant, radius, periphery_num,
dominating_set_num, local_efficiency, global_efficiency, min_cost_flow_cost, flow_hierarchy,
number_of_isolates, max_maximal_matching, non_randomness, overall_reciprocity, s_metric,
wiener_index, is_eulerian, is_planar, is_regular, is_tournament, is_tree, is_triad</td></tr>
<tr><td rowspan="3">Linearization Phase</td><td>Formats</td><td>flat triples
KG-to-Text: MVP, KG-to-Text: QA,
GDL: GML, GDL: DOT,
Path: Dijkstra short path (undirected), Path: Dijkstra short path (directed),
Path: Bellman short path (undirected), Path: Bellman short path (directed),
Path: simple path (undirected), Path: simple path (directed),
flat triples + global node, flat triples + reverse edges</td></tr>
<tr><td>Orders</td><td>dictionary,
sim:Q descend, sim:Q ascend,
sim:A descend, sim:A ascend,
sim:Q descend (BGE), sim:Q ascend (BGE),
sim:A descend (BGE), sim:A ascend (BGE),
travel:BFS (directed), travel:DFS (directed),
travel:BFS (undirected), travel:DFS (undirected)</td></tr>
<tr><td>Templates</td><td>(h, r, t),
(h,r,t),
h r t
(h,r,t);
<h,r,t>,
[h,r,t],
(h;r;t),
(h->r->t),
h |r |t [sep]
[triple]h,r,t[/triple][sep]
<triple>h,r,t</triple><sep>
[head]h[relation]r[tail]t[sep]
[/head]h[/relation]r[/tail]t[/sep]
<head>h<relation>r<tail>t<sep></td></tr>
<tr><td rowspan="3">Generalizability</td><td>Models</td><td>Llama-2-7B-chat, Llama-2-7B, Llama-3-8B, Llama-3-8B-inst,
Mistral-7B-v0.1, Mistral-7B-inst,
Phi-3-mini-128k-inst,
OPT-125M, OPT-350M, OPT-1.3B, OPT-2.7B, OPT-6.7B,
Llama-2-13B-chat,
ChatGPT,
GPT-4o</td></tr>
<tr><td>Datasets</td><td>GraphextQA
CWQ
WQSP</td></tr>
<tr><td>Tricks</td><td>0-shot
2-shot
LoRa</td></tr>
</table>

Table 4: All studied factors

	Methods	Formats													Statistics	
		flat triples + reverse edges	flat triples + global node	flat triples	Path: simple (undirected)	Path: Dijkstra (undirected)	Path: Bellman (undirected)	KG-to-Text: MVP	GDL: DOT	KG-to-Text: QA	GDL: GML	Path: Bellman (directed)	Path: simple (directed)	Path: Dijkstra (directed)	avg	std
Models	GPT-4o	88.93	87.54	87.54	80.28	80.28	79.58	70.93	87.89	79.58	87.89	50.17	49.13	49.48	75.33	15.52
	ChatGPT	77.51	78.55	76.12	76.82	78.20	76.82	65.74	79.58	53.98	75.43	41.52	40.83	41.87	66.38	15.83
	Llama-3-8B	88.20	84.01	77.40	72.25	73.94	72.11	65.81	51.42	61.07	58.55	38.58	38.55	38.37	63.10	17.17
	Llama-3-8B-inst	70.66	73.39	74.12	67.27	67.44	67.44	61.18	74.08	69.20	67.89	40.69	40.52	40.97	62.68	12.99
	Mistral-7B-inst	70.03	67.92	69.76	69.24	69.76	69.52	60.42	66.12	53.46	70.83	40.14	40.17	40.28	60.59	12.58
	Phi-3-mini-inst	69.69	72.35	73.25	68.62	69.10	69.69	63.04	70.76	44.95	61.00	37.20	36.99	36.78	59.49	14.75
	Llama-2-7B-chat	62.11	59.86	63.36	68.65	70.03	69.41	59.17	61.11	53.01	58.48	41.00	40.87	40.73	57.52	10.61
	Llama-2-7B	68.24	72.73	72.25	64.53	63.70	64.15	66.44	19.13	16.23	0.38	28.20	28.03	27.99	45.54	25.74
	Llama-2-13B-chat	83.60	76.30	74.05	63.01	55.92	55.71	45.16	9.13	39.45	0.14	25.12	26.16	25.12	44.53	26.37
	OPT-6.7B	52.80	41.31	54.78	61.42	60.24	61.45	58.06	31.94	14.84	28.17	25.19	25.19	25.05	41.57	17.09
	Mistral-7B-v0.1	69.48	68.48	53.63	61.18	60.59	60.59	54.95	7.72	3.49	0.66	31.25	31.49	31.52	41.16	25.02
	OPT-2.7B	42.35	47.51	36.68	34.39	34.91	34.36	27.65	22.73	22.98	17.99	14.81	14.78	14.43	28.12	11.09
	OPT-1.3B	44.39	33.67	35.81	16.92	16.82	16.09	14.29	13.98	1.70	9.45	9.38	9.45	9.27	17.79	12.44
	OPT-125M	9.97	6.23	17.82	18.51	19.27	18.89	22.25	11.73	14.91	0.00	9.31	9.27	9.34	12.89	6.35
	OPT-350M	4.98	1.52	3.63	12.66	11.52	12.28	22.46	20.76	28.17	0.14	7.06	7.02	6.85	10.70	8.52
	avg	60.20	58.09	58.01	55.72	55.45	55.21	50.50	41.87	37.13	35.80	29.31	29.23	29.20	45.83	12.32
	std	25.50	26.76	24.27	22.93	23.14	22.97	19.10	29.05	24.45	33.37	13.78	13.59	13.78	22.51	6.05
Datasets	GraphextQA	62.11	59.86	63.36	68.65	70.03	69.41	59.17	61.11	53.01	58.48	41.00	40.87	40.73	57.52	10.61
	WQSP	29.48	33.35	37.04	49.14	52.70	53.01	31.45	36.30	29.98	21.13	52.58	54.12	52.95	41.02	11.66
	CWQ	16.99	18.83	20.90	41.63	53.36	53.50	18.66	21.52	16.57	15.77	53.44	53.38	53.07	33.66	17.43
	avg	36.20	37.35	40.43	53.14	58.70	58.64	36.43	39.64	33.18	31.79	49.01	49.45	48.92	44.07	13.24
	std	23.30	20.80	21.43	13.95	9.82	9.33	20.71	20.00	18.43	23.26	6.95	7.45	7.09	12.22	3.67
Tricks	lora	98.65	98.17	98.69	91.66	91.73	91.52	92.42	98.37	98.27	97.75	75.22	74.74	74.98	90.94	9.55
	2-shot	76.47	71.80	78.51	77.92	79.58	79.58	71.52	75.71	73.70	74.50	35.92	36.33	36.06	66.74	17.66
	0-shot	62.11	59.86	63.36	68.65	70.03	69.41	59.17	61.11	53.01	58.48	41.00	40.87	40.73	57.52	10.61
	avg	79.08	76.61	80.18	79.41	80.45	80.17	74.37	78.40	74.99	76.91	50.72	50.65	50.59	71.73	12.61
	std	18.41	19.60	17.72	11.58	10.87	11.07	16.81	18.78	22.66	19.75	21.38	20.99	21.25	17.26	4.41

Table 5: All results of formats

	Methods	Orders													Statistics	
		sim:Q ascend	sim:Q descend	sim:A descend	dictionary	sim:Q ascend (BGE)	sim:A descend (BGE)	sim:Q descend (BGE)	sim:A ascend (BGE)	sim:A ascend	travel: DFS (undirected)	travel: BFS (undirected)	travel: DFS (directed)	travel BFS (directed)	avg	std
Models	GPT-4o	88.58	87.89	87.20	87.54	87.54	87.20	87.54	89.62	88.58	88.93	88.93	46.02	44.98	81.58	16.03
	Llama-3-8B	77.02	77.65	77.51	77.40	77.23	77.96	78.20	77.99	76.85	76.57	76.61	34.36	34.53	70.76	16.13
	ChatGPT	77.16	78.20	74.05	76.12	76.47	74.05	76.12	79.24	79.93	76.82	78.20	37.02	35.99	70.72	15.28
	Llama-3-8B-inst	72.18	74.22	68.55	74.12	71.14	68.58	74.15	77.13	77.82	72.56	73.39	36.02	36.30	67.40	14.13
	Phi-3-mini-inst	69.72	76.92	68.30	73.25	70.17	69.90	75.92	77.37	77.09	73.81	73.94	28.51	28.17	66.39	17.16
	Llama-2-7B	74.91	70.66	78.41	72.25	75.16	76.96	71.83	66.92	65.71	67.79	68.24	23.29	23.18	64.25	18.62
	Mistral-7B-inst	70.00	70.07	68.96	69.76	69.38	68.55	71.04	72.11	71.35	70.83	70.66	30.62	30.14	64.11	15.00
	Llama-2-7B-chat	65.78	64.50	70.14	63.36	66.19	68.96	64.60	59.79	59.27	63.84	63.88	32.73	32.25	59.64	12.42
	Llama-2-13B-chat	77.47	72.77	62.53	74.05	62.08	62.66	62.46	59.86	58.96	61.28	60.10	21.63	21.25	58.24	17.38
	Mistral-7B-v0.1	54.53	54.43	54.98	53.63	54.19	54.71	54.64	53.43	52.73	56.61	54.98	20.35	20.80	49.23	12.75
	OPT-6.7B	59.34	53.08	59.31	54.78	59.03	56.75	52.84	52.25	50.28	46.19	45.81	13.01	13.77	47.42	15.74
	OPT-1.3B	36.16	36.96	39.38	35.81	36.33	38.34	37.06	33.98	34.88	24.57	25.74	4.78	4.67	29.90	12.02
	OPT-2.7B	39.24	35.02	40.07	36.68	39.31	37.06	34.91	35.22	33.46	24.22	23.98	4.08	4.22	29.81	12.46
	OPT-125B	17.02	15.95	17.30	17.82	16.85	17.85	15.61	15.81	15.85	15.29	14.88	4.46	4.57	14.56	4.56
	OPT-350B	3.88	3.63	4.57	3.63	3.94	3.84	3.49	3.70	3.29	4.01	3.84	1.49	1.35	3.44	0.95
	avg	58.87	58.13	58.08	58.01	57.67	57.56	57.36	56.96	56.40	54.89	54.88	22.56	22.41	51.83	13.08
	std	24.37	24.79	23.30	24.27	23.69	23.53	24.48	24.98	24.96	25.93	26.06	14.17	13.91	22.96	4.04
Datasets	GraphextQA	65.78	64.50	70.14	63.36	66.19	68.96	64.60	59.79	59.27	63.84	63.88	32.73	32.25	59.64	12.42
	WQSP	35.20	36.49	42.14	37.04	32.56	43.06	37.53	31.88	32.00	36.86	37.41	36.18	36.61	36.53	3.37
	CWQ	20.22	21.50	35.17	20.90	19.31	36.56	21.58	15.41	16.20	19.09	20.16	20.11	20.59	22.06	6.40
	avg	40.40	40.83	49.15	40.43	39.35	49.53	41.24	35.69	35.83	39.93	40.48	29.67	29.82	39.41	7.40
	std	23.22	21.83	18.51	21.43	24.17	17.14	21.75	22.44	21.79	22.53	22.02	8.46	8.28	18.95	4.61
Tricks	lora	98.24	98.41	97.72	98.69	98.41	98.24	97.51	97.85	97.85	97.85	98.06	78.93	78.37	95.09	7.30
	2-shot	81.04	77.30	83.84	78.51	80.62	83.11	78.55	74.39	75.43	75.29	76.12	32.60	32.46	71.48	17.54
	0-shot	65.78	64.50	70.14	63.36	66.19	68.96	64.60	59.79	59.27	63.84	63.88	32.73	32.25	59.64	12.42
	avg	81.68	80.07	83.90	80.18	81.74	83.44	80.22	77.35	77.52	79.00	79.35	48.09	47.69	75.40	12.42
	std	16.24	17.12	13.79	17.72	16.14	14.64	16.52	19.20	19.38	17.31	17.32	26.71	26.57	18.05	5.12

Table 6: All results of orders

	Methods	Templates														Statistics	
		<triple>h,r,t </triple><sep>	[triple]h,r,t [/triple][sep]	<h,r,t>,	[h,r,t],	<head>h <relation>r <tail>t<sep>	[/head]h [/relation]r [/tail]t[/sep]	[head]h [relation]r [tail]t[sep]	(h, r, t),	h |r |t [sep]	h r t	(h,r,t);	(h,r,t),	(h;r;t),	(h->r->t),	avg	std
	GPT-4o	88.93	87.89	87.89	88.93	85.81	86.85	85.12	87.54	88.58	85.47	88.58	87.89	88.93	87.89	87.59	1.30
	Llama-3-8B	79.79	82.32	76.71	77.44	80.93	81.11	80.00	77.40	76.68	80.35	78.37	78.34	78.13	77.54	78.94	1.79
	ChatGPT	76.47	75.43	77.16	77.85	75.09	71.97	72.66	76.12	74.74	77.16	76.47	75.78	75.43	76.82	75.65	1.67
	Llama-3-8B-inst	72.94	72.49	74.01	73.77	72.46	71.66	70.83	74.12	71.94	74.12	74.05	73.77	74.39	75.12	73.26	1.22
	Llama-2-7B	75.67	73.46	74.15	71.97	71.07	78.24	71.59	72.25	72.01	74.50	72.60	71.14	70.31	68.37	72.67	2.44
	Phi-3-mini-inst	74.57	74.43	74.60	73.18	68.37	65.36	65.29	73.25	73.15	70.90	73.15	71.49	71.80	70.14	71.41	3.13
	Mistral-7B-inst	67.44	68.58	69.72	70.24	65.92	66.12	65.22	69.76	67.30	66.40	70.62	70.73	69.62	68.65	68.31	1.89
	Llama-2-13B-chat	68.20	53.94	72.63	62.66	65.64	81.76	64.64	74.05	52.11	64.39	60.80	54.33	55.92	51.73	63.06	9.04
Models	Llama-2-7B-chat	64.95	63.22	63.22	64.12	61.56	66.23	62.32	63.36	63.63	62.39	63.63	62.08	62.53	58.58	62.99	1.75
	Mistral-7B-v0.1	59.48	63.56	66.68	60.66	62.80	56.57	63.84	53.63	58.37	40.35	57.34	57.92	55.92	63.36	58.61	6.41
	OPT-6.7B	38.79	68.03	41.49	53.39	27.72	72.01	32.84	54.78	49.13	57.85	42.04	47.40	46.12	50.76	48.74	12.27
	OPT-1.3B	82.87	61.00	77.89	40.80	75.64	42.80	46.40	35.81	45.09	23.91	34.53	33.81	33.67	28.93	47.37	19.23
	OPT-2.7B	56.19	62.21	26.23	42.28	46.16	69.00	53.81	36.68	50.42	36.37	44.12	41.11	38.06	36.47	45.65	11.57
	OPT-125M	63.88	30.80	68.82	78.72	75.78	2.98	55.64	17.82	6.85	7.68	13.01	11.38	12.91	6.57	32.35	29.21
	OPT-350M	14.33	46.71	9.58	12.60	6.51	22.46	3.56	3.63	19.03	40.31	3.53	3.22	1.80	2.46	13.55	14.30
	avg	65.63	65.61	64.05	63.24	62.76	62.34	59.58	58.01	57.94	57.48	56.86	56.03	55.70	54.89	60.01	3.80
	std	18.68	14.20	21.46	19.39	21.10	23.10	20.18	24.27	22.00	22.73	24.73	24.67	24.91	25.78	21.94	3.13
	GraphextQA	64.95	63.22	63.22	64.12	61.56	66.23	62.32	63.36	63.63	62.39	63.63	62.08	62.53	58.58	62.99	1.75
	WQSP	36.36	38.02	37.65	38.21	35.38	34.77	34.77	37.04	35.01	37.78	36.86	36.79	36.98	37.96	36.68	1.24
Datasets	CWQ	19.77	19.80	21.13	21.27	19.37	16.94	19.37	20.90	20.67	21.35	21.13	21.24	21.01	22.37	20.45	1.33
	avg	40.36	40.35	40.67	41.20	38.77	39.31	38.82	40.43	39.77	40.51	40.54	40.04	40.17	39.64	40.04	0.70
	std	22.85	21.80	21.21	21.58	21.30	24.96	21.76	21.43	21.87	20.65	21.49	20.61	20.94	18.16	21.47	1.45
	lora	98.24	98.24	98.20	98.10	97.99	98.20	97.92	98.69	97.92	98.34	98.55	98.24	98.27	98.03	98.21	0.22
	2-shot	78.30	82.08	76.33	76.64	73.22	69.45	78.51	78.51	76.26	75.05	77.30	75.64	75.54	71.97	76.06	3.10
Tricks	0-shot	64.95	63.22	63.22	64.12	61.56	66.23	62.32	63.36	63.63	62.39	63.63	62.08	62.53	58.58	62.99	1.75
	avg	80.50	81.18	79.25	79.62	77.59	77.96	79.58	80.18	79.27	78.59	79.83	78.65	78.78	76.19	79.08	1.69
	std	16.75	17.53	17.67	17.18	18.61	17.60	17.83	17.72	17.34	18.24	17.59	18.27	18.09	20.06	17.81	1.44

Table 7: All results of templates

References

Bollacker, K., Evans, C., Paritosh, P.; et al. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 1247–1250.

Ferrando, J.; and Voita, E. 2024. Information Flow Routes: Automatically Interpreting Language Models at Scale. arXiv:2403.00824.

LUO, L., Li, Y.-F., Haf, R.; and Pan, S. 2024. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *ICLR*.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Webber, W., Moffat, A.; and Zobel, J. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4): 1–38.