KG-FIT: Knowledge Graph Fine-Tuning Upon Open-World Knowledge

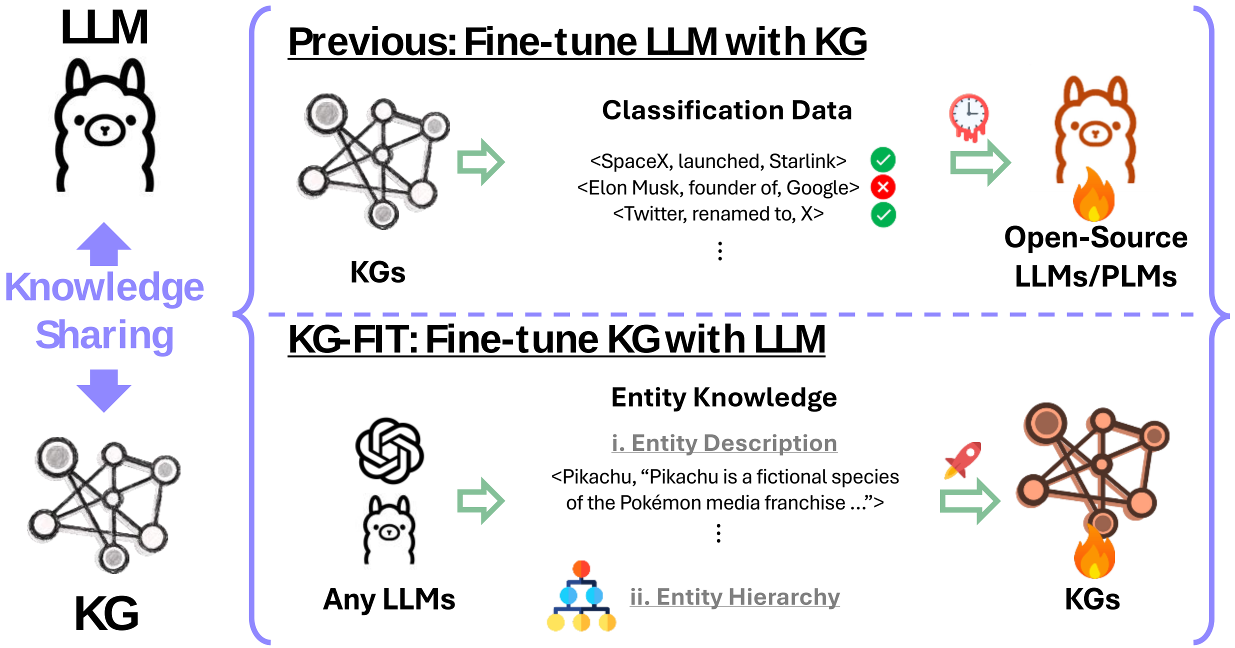
Pengcheng Jiang, Lang Cao, Cao Xiao*, Parminder Bhatia*, Jimeng Sun, and Jiawei Han
University of Illinois at Urbana Champaign GE HealthCare*

Patrick (Pengcheng) Jiang
CS Ph.D. at UIUC

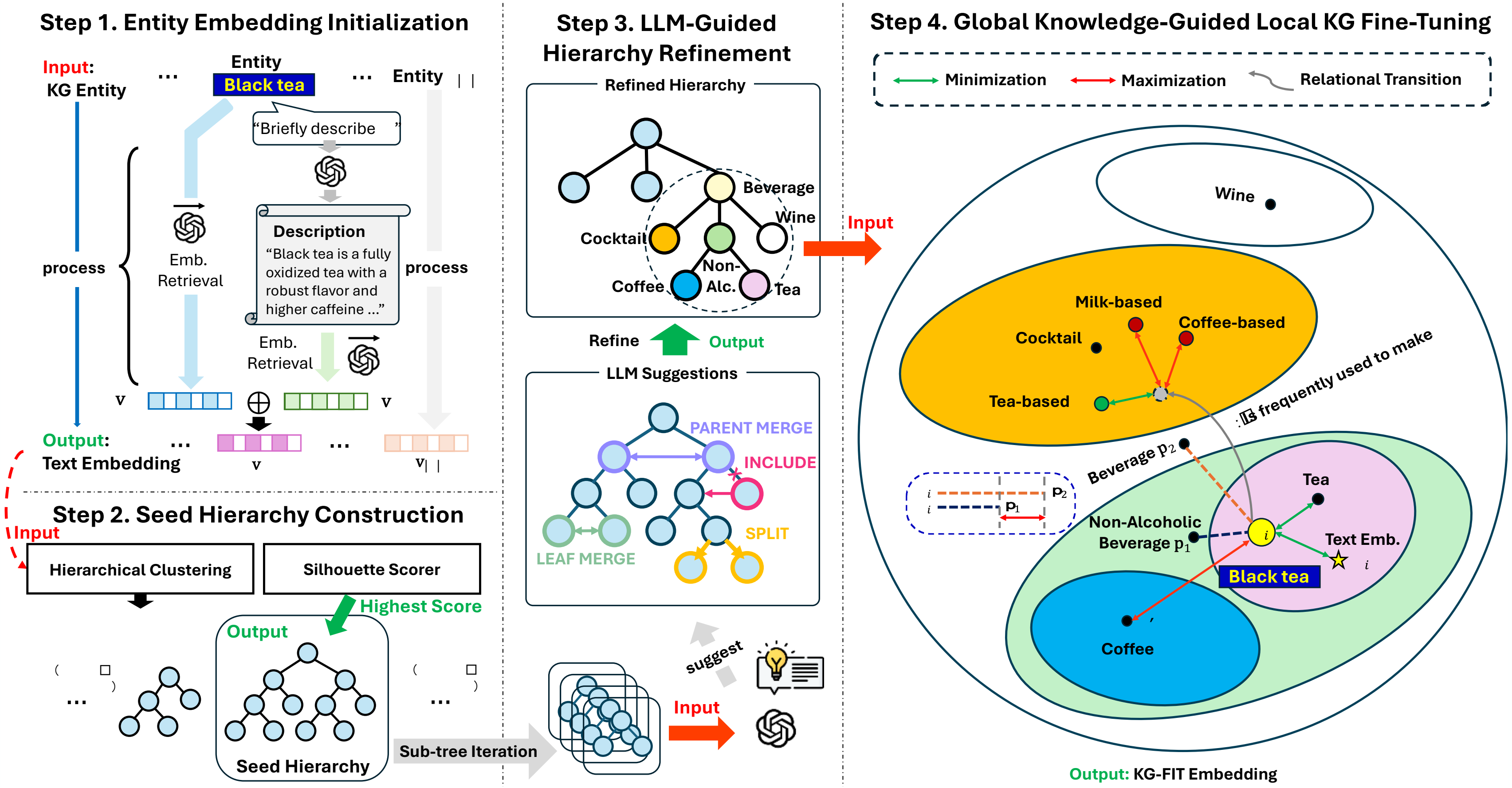
Motivation & Problem

Challenges:
- Structure-based KGE methods are limited to graph structure
- PLM-based methods are computationally expensive
- Need to leverage LLM knowledge efficiently

Our Solution:
- Fine-tune KG with LLM instead of fine-tuning LLM with KG
- Combine advantages of both approaches



KG-FIT Framework



Two-Stage Approach:

1. LLM-Guided Hierarchy Construction
   - Entity description generation
   - Seed hierarchy construction
   - LLM-guided refinement

2. Knowledge Graph Fine-Tuning
   - Hierarchical clustering constraint
   - Semantic anchoring
   - Link prediction objective

Datasets

- **Datasets**

Table 1: **Datasets statistics.** #Ent./#Rel: number of entities/relations. #Train/#Valid/#Test: number of triples contained in the training/validation/testing set.

Dataset	#Ent.	#Rel.	#Train	#Valid	#Test
FB15k-237	14,541	237	272,115	17,535	20,466
YAGO3-10	123,182	37	1,079,040	5,000	5,000
PrimeKG	10,344	11	100,000	3,000	3,000

FB15K-237:
- A subset of Freebase, a large collaborative knowledge base focusing on common knowledge.

YAGO3-10:
- A subset of YAGO, a large knowledge base derived from multiple sources including Wikipedia, WordNet, and GeoNames.

PrimeKG:
- A biomedical KG integrates 20 biomedical resources, detailing 17,080 diseases through 4,050,249 relationships. In this study, we extract a subset of PrimeKG, which contains 106,000 triples.

- **Metrics**

Mean Rank (MR):
- Measures the average rank of true entities.

Mean Reciprocal Rank (MRR):
- Averages the reciprocal ranks of true entities.

Hits@N:
- Measures the proportion of true entities in the top N predictions.

Main Results

(1) KG-FIT consistently and significantly outperforms state-of-the-art PLM-based and structure-based methods across all datasets and metrics.

(2) With LLM-guided hierarchy refinement, KG-FIT achieves huge performance gains compared to the base models and KG-FIT with seed hierarchy.

(3) KG-FIT is more effective for smaller KGs, e.g., more performance gains on PrimeKG (~ 0.1 million triples) than YAGO3-10 (~1 million triples).

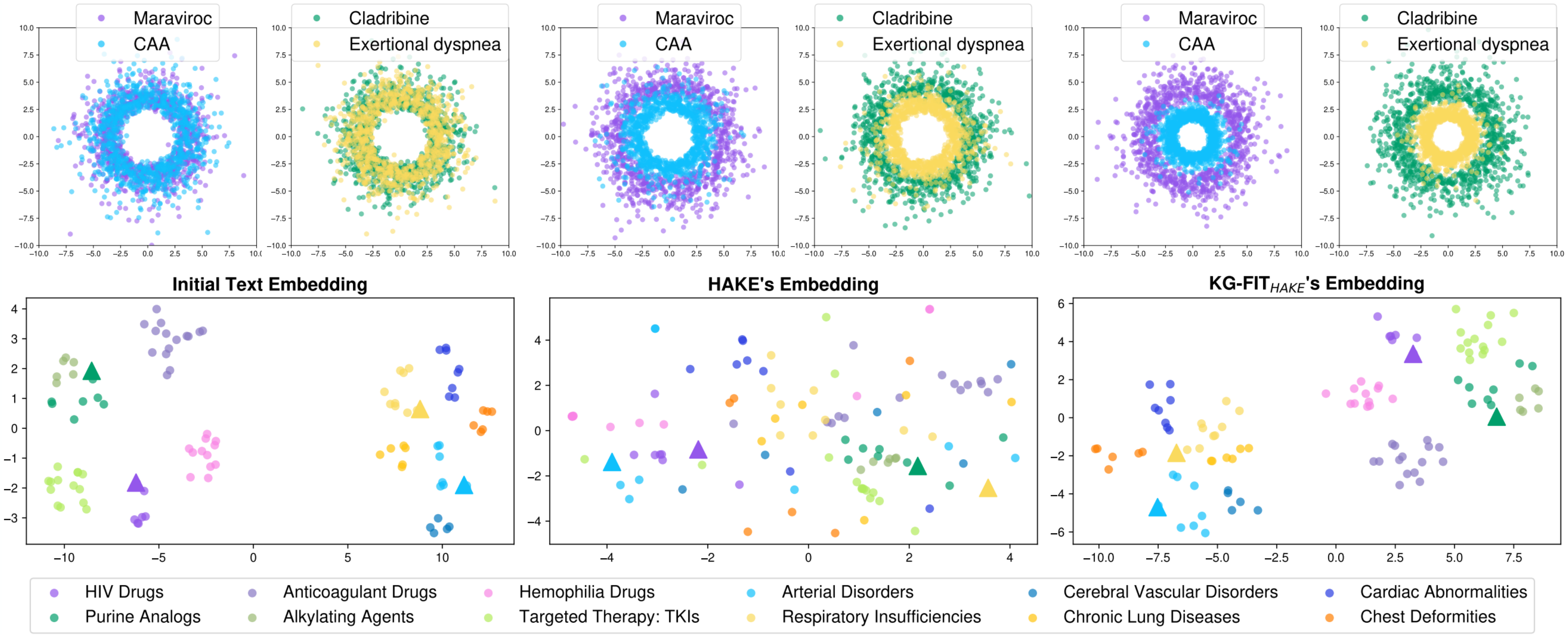
PLM-based Embedding Methods

		FB15K-237					YAGO3-10					PrimeKG				
Model	PLM	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10
KG-BERT [22]*	BERT	153	.245	.158	–	.420	–	–	–	–	–	–	–	–	–	–
StAR [23]*	RoBERTa	**117**	.296	.205	–	.482	–	–	–	–	–	–	–	–	–	–
PKGC [28]	RoBERTa	184	.342	.236	.441	.525	1225	.501	.426	.596	.660	219	.485	.391	.565	.625
C-LMKE [26]*	BERT	141	.306	.218	–	.484	–	–	–	–	–	–	–	–	–	–
KGT5 [25]*	T5	–	.276	.210	–	.414	–	.426	.368	–	.528	–	–	–	–	–
KG-S2S [24]*	T5	–	.336	.257	–	.498	–	–	–	–	–	–	–	–	–	–
SimKGC [27]	BERT	–	.336	.249	–	.511	–	–	–	–	–	168	.527	.524	.679	.742
CSProm-KG [32]	BERT	–	.358	.269	–	.538	1145	.488	.451	.624	.675	157	.540	.492	.652	.745
LLM Emb. (zero-shot)	TE-3-S	2044	.023	.002	.035	.068	22741	.009	.000	.016	.024	5581	.000	.000	.000	.000
	TE-3-L	1818	.030	.004	.048	.085	18780	.015	.000	.019	.032	4297	.001	.000	.000	.000

Structure-based Embedding Methods

Model	Frame	$\mathcal{H}$	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10
TransE	Base [14]	—	233	.287	.192	.389	.478	1250	.500	.398	.626	.685	182	.048	.000	.043	.124
	KG-FIT	Seed	142	.345	.242	.457	.547	952	.520	.429	.638	.700	80	.298	.000	.315	.516
		LHR	122	**.362**	.264	.478	.568	**529**	.544	.463	.650	.705	69	.334	.000	.342	.536
DisMult	Base [15]	—	283	.260	.163	.349	.437	5501	.451	.365	.553	.615	174	.577	.475	.699	.782
	KG-FIT	Seed	184	.316	.198	.415	.512	963	.486	.413	.591	.673	107	.589	.495	.715	.799
		LHR	154	.331	.226	.433	.529	861	.527	.441	.636	.682	78	.617	.526	.747	.813
ComplEx	Base [16]	—	347	.252	.161	.344	.439	6681	.463	.384	.560	.612	202	.614	.522	.728	.789
	KG-FIT	Seed	201	.325	.223	.436	.523	997	.491	.422	.603	.669	94	.638	.548	.767	.823
		LHR	151	.344	.247	.458	.551	842	.544	.460	.646	.697	82	**.651**	**.566**	**.772**	**.835**
ConvE	Base [17]	—	341	.312	.224	.401	.508	1105	.529	.451	.619	.673	144	.516	.456	.645	.760
	KG-FIT	Seed	181	.318	.237	.411	.521	912	.535	.455	.628	.685	93	.627	.534	.757	.812
		LHR	177	.318	.241	.415	.525	885	.541	.461	.647	.695	72	.648	.547	.767	.824
TuckER	Base [18]	—	363	.320	.230	.417	.505	1110	.529	.454	.633	.690	171	.543	.442	.663	.737
	KG-FIT	Seed	175	.330	.241	.433	.521	874	.538	.458	.651	.703	77	.640	.542	.770	.805
		LHR	144	.349	.255	.448	.543	838	.545	**.466**	**.654**	.708	62	.648	.550	**.779**	.820
pRotatE	Base[19]	—	188	.310	.205	.399	.502	974	.477	.385	.573	.655	118	.491	.399	.593	.681
	KG-FIT	Seed	160	.355	.257	.461	.558	910	.525	.436	.622	.693	75	.635	.538	.745	.809
		LHR	**119**	**.371**	**.277**	**.483**	**.572**	829	**.550**	.464	.648	**.710**	69	**.649**	**.574**	**.779**	**.833**
RotatE	Base [19]	—	190	.333	.241	.428	.528	1620	.495	.402	.550	.670	57	.539	.447	.646	.727
	KG-FIT	Seed	141	.354	.261	.464	.555	790	.529	.440	.643	.708	**46**	.622	.517	.740	.805
		LHR	**120**	**.369**	**.274**	**.488**	**.570**	**744**	**.563**	**.475**	**.658**	**.722**	**34**	.645	.532	.758	.817
HAKE	Base [20]	—	184	.344	.247	.435	.538	1220	530	.431	.634	.681	95	.595	.515	.708	.760
	KG-FIT	Seed	162	.358	.268	.470	.563	854	.541	.455	.647	.703	82	.638	.540	.747	.808
		LHR	137	**.362**	**.275**	**.485**	**.572**	**810**	**.568**	**.474**	**.662**	**.718**	**42**	**.682**	**.605**	**.785**	**.835**

Visualization



Conclusion

We introduced KG-FIT, a novel framework that enhances knowledge graph (KG) embeddings by integrating open-world entity knowledge from Large Language Models (LLMs).

- KG-FIT effectively combines the knowledge from LLM and KG to preserve both global and local semantics, achieving state-of-the-art link prediction performance on benchmark datasets.

- It shows significant improvements in accuracy compared to the base models. Notably, KG-FIT can seamlessly integrate knowledge from any LLM, enabling it to evolve with ongoing advancements in language models.

- Future work will explore using the KG-FIT embedding for precise knowledge retrieval, which can set a strong foundation for retrieval augmented generation (RAG) by LLMs.

ILLINOIS

Thank you!

Patrick Jiang