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

NEURAL INFORMATION PROCESSING SYSTEMS

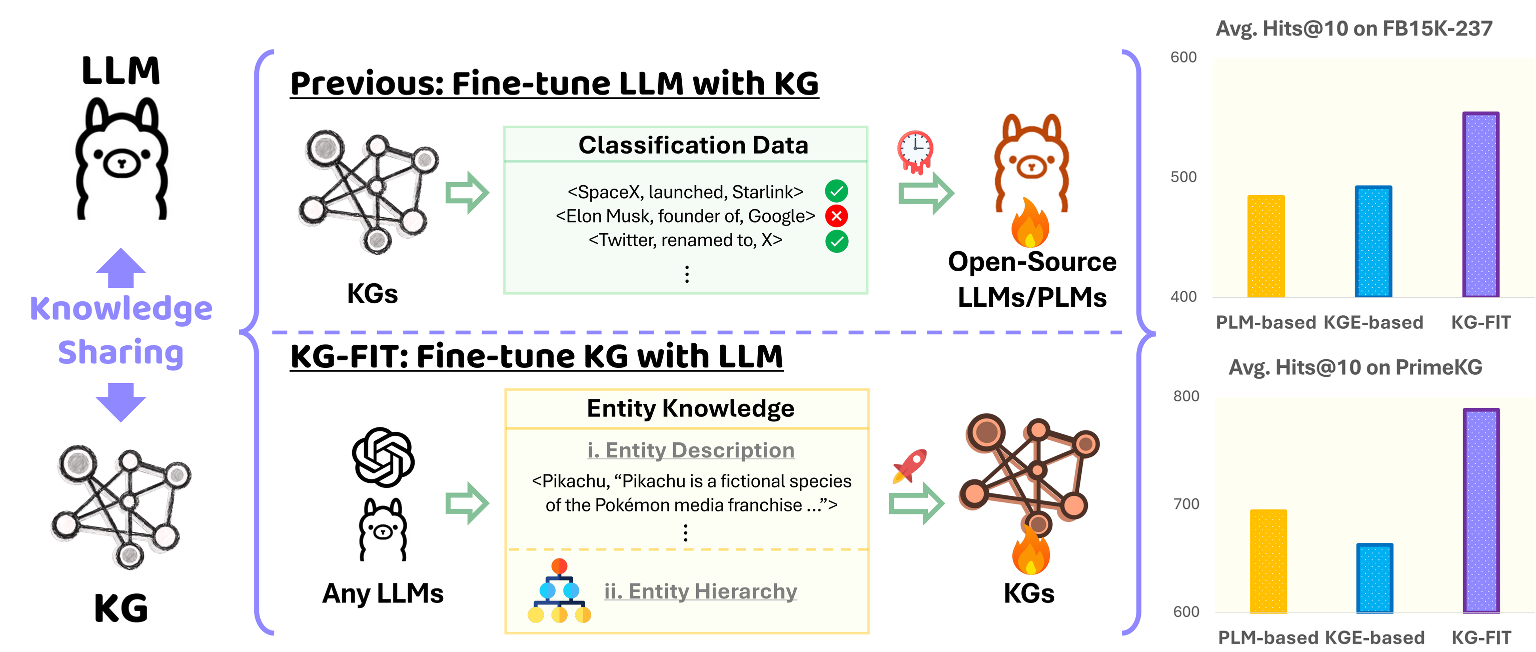
Introduction

Current Challenges:

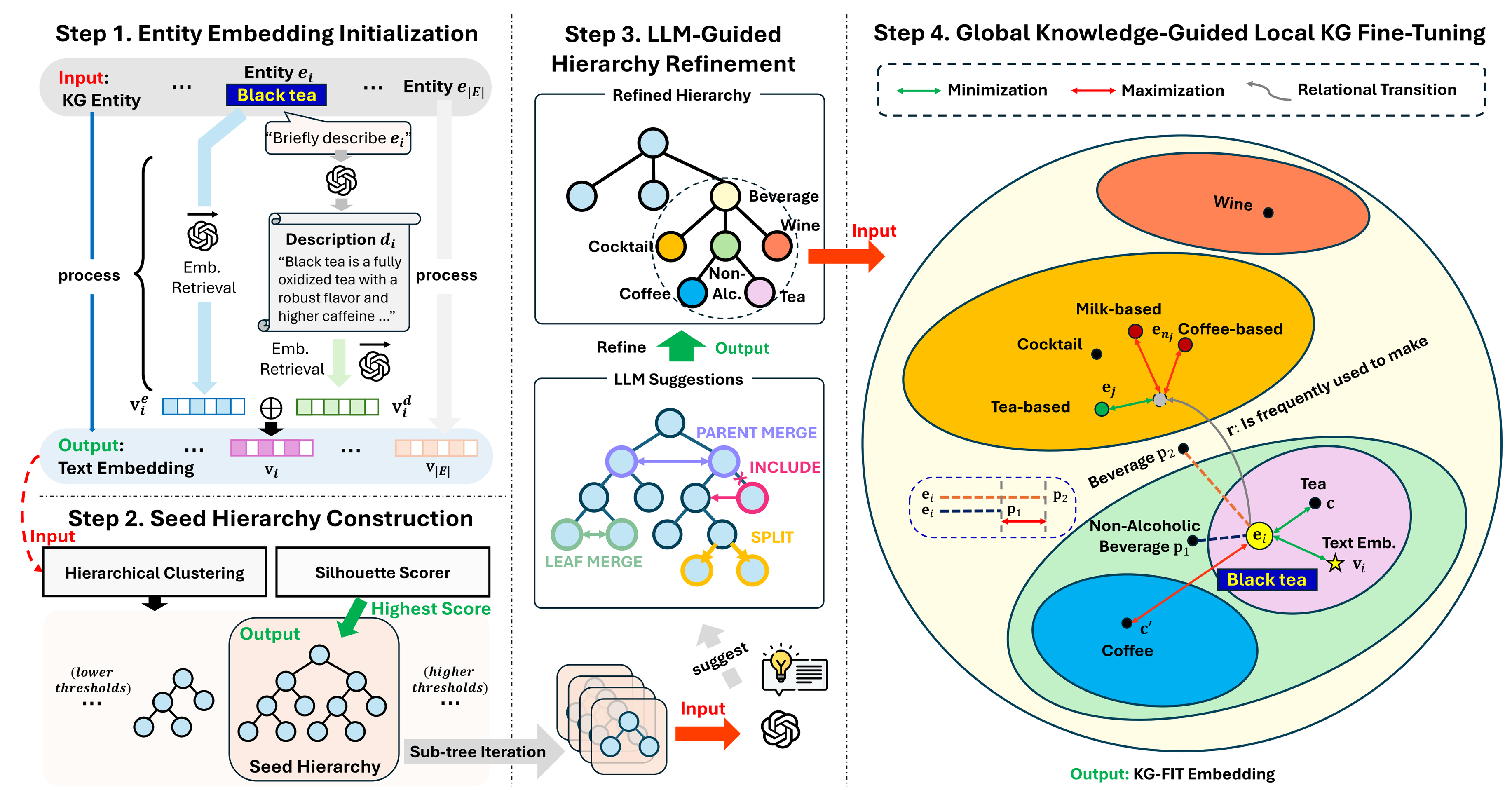
- Knowledge Graph Embeddings are crucial for AI systems but often limited to structure alone
- Existing methods that combine KGs with language models face key limitations:
  - High computational costs during training and inference
  - Limited ability to leverage extensive knowledge in modern LLMs
  - Difficulty keeping up with rapidly evolving generative LLMs

Our Solution: KG-FIT

- A novel framework that directly incorporates knowledge from LLMs into KG embeddings
- Key innovations:
  - LLM-guided hierarchical structure construction
  - Effective integration of global semantics from LLMs with local semantics from KGs
  - No need to fine-tune the language models themselves



Our Method: KG-FIT



KG-FIT framework includes the following steps:

Step 1: Entity Embedding Initialization
Create initial entity embeddings by concatenating:

- Entity name embedding
- Entity description (generated by LLM) embedding

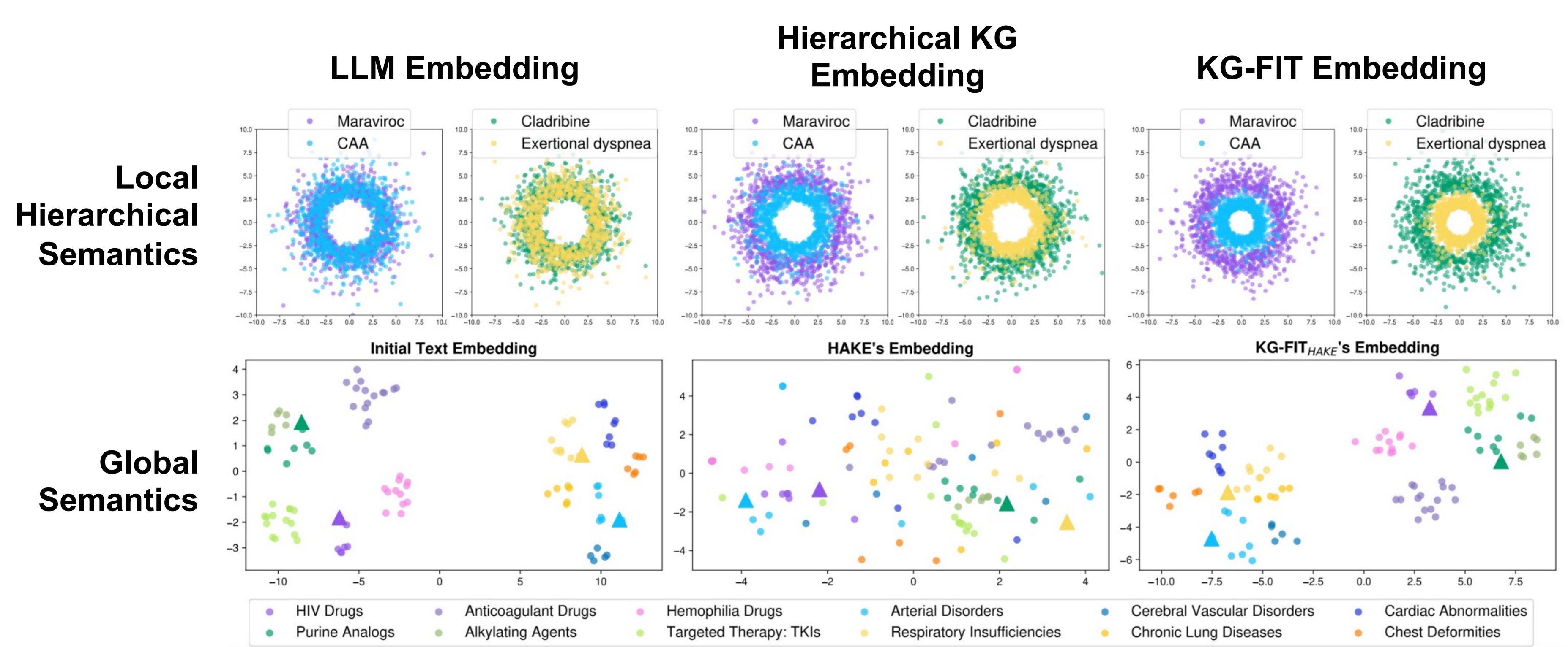
Step 2: Seed Hierarchy Construction
Apply agglomerative clustering to entity embeddings
Select optimal hierarchy using silhouette score

Step 3: LLM-Guided Hierarchy Refinement
Refine the seed hierarchy constructed with LLM's suggestions through an iterative bottom-up tree editing process

Step 4: Knowledge Graph Fine-Tuning

1. Initialize entity and relation embeddings
2. Fine-tune the KG embedding with score functions (defined any base KGE models) and the knowledge in the entity hierarchy.

Embedding Comparison



Our KG-FIT embedding captures both local and global semantics.
In this example, KG-FIT can accurately predict:
Local: (1) Maraviroc has drug effect on coronary artery atherosclerosis (CAA), (2) Cladribine has drug effect of exertional dyspnea
Global: (1) Maraviroc is a type of HIV drugs, (2) Cladribine is a type of Purine Analog.

Performance Comparison

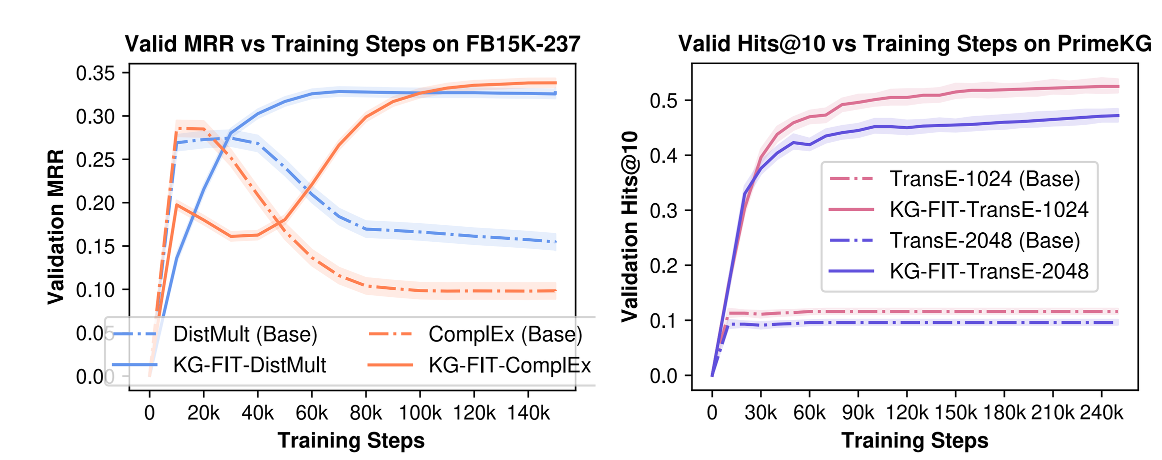
		FB15K-237					YAGO3-10					PrimeKG				
PLM-based Embedding Methods																
Model	PLM	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10	MR	MRR	H@1	H@5	H@10
KG-BERT [22]*	BERT	153	.245	.158	–	.420	–	–	–	–	–	–	–	–	–	–
StAR [23]*	RoBERTa	**117**	.296	.205	–	.482	–	–	–	–	–	–	–	–	–	–
PKGC [28]	RoBERTa	184	.342	.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

			FB15K-237					YAGO3-10					PrimeKG				
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
• • •																	
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**

Key Findings:
(1) KG-FIT significantly outperforms SOTA PLM- and structure-based methods
(2) Performance gain by LHR (LLM-guided hierarchy refinement) is huge

- Suggesting the importance of high-quality hierarchical knowledge of entities

Training Insights



Key Findings:
(1) KG-FIT effectively addresses both overfitting (LHS)and underfitting (RHS) challenges of KGE training
(2) Hierarchical structure provides natural regularization
(3) Integration of LLM knowledge helps reach better convergence points

Paper Code

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

GE HealthCare