Knowledge Graph

Topics

Knowledge Graph Construction with LLM

- [HI-AI@KDD, 2024] Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema
- [IEEE, BIBM, 2024] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Appplication of Knowledge Graph

[AAAI, 2025] K-ON: Stacking Knowledge On the Head Layer of LLM

Topics

Knowledge Graph Construction with LLM

- [HI-AI@KDD, 2024] Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema
- [IEEE, BIBM, 2024] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Appplication of Knowledge Graph

■ [AAAI, 2025] K-ON: Stacking Knowledge On the Head Layer of LLM

Intro

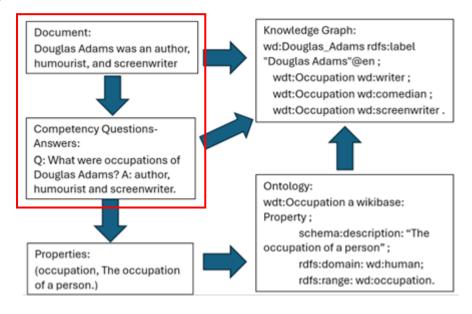
- Knowledge Graph(KG) power tasks like <u>search</u> and <u>QA</u>, but manual curation is costly and hard to scale.
- Directly using LLMs yields inconsistent or redundant facts due to the absence of a unified schema(Ontology), and alignment with existing Knowledge Base(KB) is difficult, especially for proprietary documents beyond pretraining.

Goal

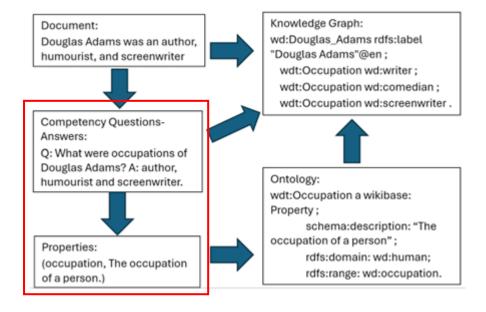
 Ontology-grounded LLM pipeline (grounded in Wikidata) that constructs high-quality, interpretable, and interoperable KGs with minimal human intervention.

- 1. Competency Question¹⁾ (CQ) and Answer Generation using LLM
- 2. Relation Extraction from CQs and Ontology Matching
- 3. Ontology Formatting
- 4. KG Construction

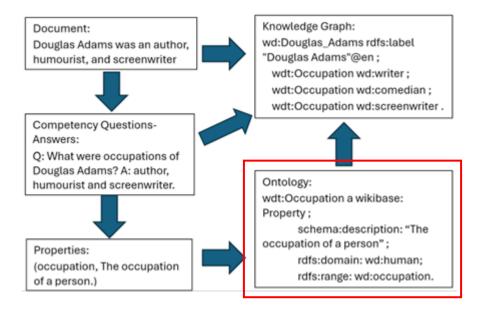
- 1. Competency Question¹⁾ (CQ) and Answer Generation using LLM
- 2. Relation Extraction from CQs and Ontology Matching
- 3. Ontology Formatting
- 4. KG Construction



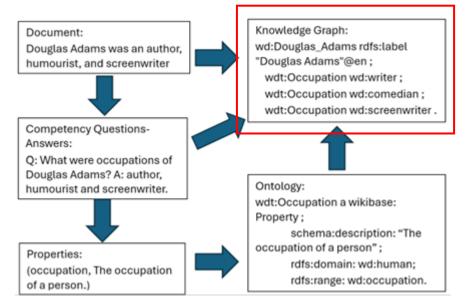
- 1. Competency Question¹⁾ (CQ) and Answer Generation using LLM
- 2. Relation Extraction from CQs and Ontology Matching
- 3. Ontology Formatting
- 4. KG Construction



- 1. Competency Question¹⁾ (CQ) and Answer Generation using LLM
- 2. Relation Extraction from CQs and Ontology Matching
- 3. Ontology Formatting
- 4. KG Construction



- 1. Competency Question¹⁾ (CQ) and Answer Generation using LLM
- 2. Relation Extraction from CQs and Ontology Matching
- 3. Ontology Formatting
- 4. KG Construction



1. Competency Question (CQ) and Answer Generation using LLM

- Provide the LLM with instructions, examples, and a template, and have it generate CQs for the given document
- Input the CQs generated by the LLM back into the LLM to generate answers for each CQ.

Prompt.1. CQ generation

Write competency questions based on the abstract level concepts in the document. Write questions that can be answered using the document only.

Write up to 3 questions per document.

Below are the examples and follow the same format when generating competency questions:

####

Document: Douglas Noel Adams (11 March 1952 - 11 May 2001) was an English author, humourist, and screenwriter, best known for The Hitchhiker's Guide to the Galaxy (HHGTTG). Originally a 1978 BBC radio comedy, The Hitchhiker's Guide to the Galaxy developed into a "trilogy" of five books that sold more than 15 million copies in his lifetime. It was further developed into a television series, several stage plays, comics, a video game, and a 2005 feature film. Adams's contribution to UK radio is commemorated in The Radio Academy's Hall of Fame.

Prompt.1. CQ generation

```
####
Questions:
CQ1. What is the date of birth of Douglas Noel Adams?
CQ2. What is the date of death of Douglas Noel Adams?
CQ3. What is the occupation of Douglas Noel Adams?
CQ4. What is the country of citizenship of Douglas Noel Adams?
CQ5. What is the most notable work of Douglas Noel Adams?
CQ6. What is the original medium of The Hitchhiker's Guide to the Galaxy?
CQ7. In what year was The Hitchhiker's Guide to the Galaxy originally broadcast?
CQ8. How many books are in The Hitchhiker's Guide to the Galaxy "trilogy"?
CQ9. What other media adaptations were created based on The Hitchhiker's Guide to the Galaxy?
```

Prompt.1. CQ generation

```
####
Document:
{document to be processed}
####
Questions:
```

Prompt.2. CQ answering

```
Use the provided document to answer user query. If you don't know the answer, just say that you don't know, don't try to make up an answer.

Passage: {doc}
Query: {query}
```

2. Relation Extraction from CQs and Ontology Matching

- A candidate property list is pre-populated from Wikidata²⁾ Ontology.
- The LLM extracts properties from CQs and provides brief descriptions, including their domain and range.
- These property descriptions are converted into sentence embeddings³⁾, and a vector similarty search retrieves the top-1 (closest) candidate for each property.
- Then an LLM vets each pair to decide semantic similarity as a final dedupulication; if they are
 'similar', replace with the Wikidata property, otherwise the newly proposed property is retained.

³⁾ bge-small-en (sentence embedding model)

Prompt.3. Relation extraction

You are an assistant in building a knowledge graph. Analyze the following competency questions and identify all relationships and concepts concepts mentioned in the question.

Extract relation first, then describe the usage of each relation based on your understanding given the context of competency questions.

Afterwards, extract all relation-related concepts.

You should only extract properties between entities and literals, not entities themselves, or classes of entities. Therefore, not all CQs contain valid properties.

If you don't know the answer, just say that you don't know, don 't try to make up an answer.

Merge all relations into one list and all concepts into one list.

Do not reply using a complete sentence, and only give the answer in the following format.

Below are the examples and follow the same format to extract the relations:

Prompt.3. Relation extraction

####

Document: Douglas Noel Adams (11 March 1952 - 11 May 2001) was an English author, humourist, and screenwriter, best known for The Hitchhiker's Guide to the Galaxy (HHGTTG). Originally a 1978 BBC radio comedy, The Hitchhiker's Guide to the Galaxy developed into a "trilogy" of five books that sold more than 15 million copies in his lifetime. It was further developed into a television series, several stage plays, comics, a video game, and a 2005 feature film. Adams's contribution to UK radio is commemorated in The Radio Academy's Hall of Fame.

Prompt.3. Relation extraction

Hitchhiker's Guide to the Galaxy?

```
####
Questions:
CQ1. What is the date of birth of Douglas Noel Adams?
CQ2. What is the date of death of Douglas Noel Adams?
CQ3. What is the occupation of Douglas Noel Adams?
CQ4. What is the country of citizenship of Douglas Noel Adams?
CQ5. What is the most notable work of Douglas Noel Adams?
CQ6. What is the original medium of The Hitchhiker's Guide to the Galaxy?
CQ7. In what year was The Hitchhiker's Guide to the Galaxy originally broadcast?
CQ8. How many books are in The Hitchhiker's Guide to the Galaxy "trilogy"?
```

CQ9. What other media adaptations were created based on The

```
####
Relations:
(date of birth, The date on which the subject was born.)
(date of death, The date on which the subject died.)
(occupation, The occupation of a person.)
(country of citizenship, The country of which the subject is a citizen.)
(notable work, The most notable work of a person.)
(genre, The genre or type of work.)
(publication date, The date or period when a work was first published or released.)
(has part, Indicates that the subject has a certain part, component, or element.)
(series, Indicates that the subject is part of a series, such as a book series, film series, or television series.)
```

Prompt.3. Relation extraction

```
####
Document:
{document to be processed}
####
Questions:
{CQs}
####
Relations:
```

Prompt.4. Ontology matching

```
Decide if the two properties are semantically similar in an ontology.

You should say yes if you decide that these propties are similar, or if they are inverse properties.

Answer in "yes" or "no" only.

Property 1: {p1}

Property 2: {p2}
```

3. Ontology Formatting

Use LLM to generate an OWL ontology based on the matched and newly created properties.

4. KG Construction

- Use LLM to construct a KG using the CQs and answers, grounded in the previously generated ontology.
- For each (CQ, answer) pair, LLM extracts relevant entities and maps them to the ontology using the defined properties.
- The final output is a set of RDF triples that forms the complete KG.

Prompt.5. Ontology formatting

```
Use the relations (properties) and their usage comments to
   build an ontology in RDF format.
If you don't know the answer, just say that you don't know, don
   't try to make up an answer.
Don't provide anything other than an ontology in RDF format.
Infer and summarize classes for domain and range of the
   relations across the concepts provided, and add these
   classes to relations only if required for clousre of
   relations.
For each relation, add relevant ontology entry for it.
Add rdfs:comment based on the usage comments.
Use wdt: namespace for all relations discovered. Use entities
  under these prefixes if necessary:
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: < http://www.w3.org/2002/07/owl#>.
@prefix wikibase: <http://wikiba.se/ontology#> .
@prefix schema: <http://schema.org/> .
@prefix wd: <http://www.wikidata.org/entity/> .
@prefix wdt: <http://www.wikidata.org/prop/direct/> .
Use turtle syntax.
```

Prompt.5. Ontology formatting

```
Below is an example:
####
Relations:
(results, results: results of a competition such as sports or
   elections)
####
Ontology:
wdt: Results a wikibase: Property;
    schema: description "results of a competition such as sports
        or elections";
    rdfs:label "results";
    rdfs:domain wd:referendum, wd:competition, wd:party
       conference, wd: sporting event;
    rdfs:range wd:electoral result, wd:voting result, wd:sport
       result, wd:race result.
####
Relations:
{relation}
####
Ontology:
```

Prompt.6. KG generation

```
Your task is to construct a knowledge graph based on the
   provided ontology.
Focus on understanding relationships from the question answer
   pair and document,
and extract related entities, then mapping them to the ontology
    using the properties defined in the ontology.
Do not include new properties other than those in ontology.
   Only use those properties in the ontology.
Output in turtle format following the ontology provided.
You should only include knowledge in question answer pairs and
   the document.
Do not make up answers.
Use this ontology based on Wikidata as the starting point:
{ont}
Below is an example:
```

Prompt.6. KG generation

####

Document:

Douglas Noel Adams (11 March 1952 - 11 May 2001) was an English author, humourist, and screenwriter, best known for The Hitchhiker's Guide to the Galaxy (HHGTTG). Originally a 1978 BBC radio comedy, The Hitchhiker's Guide to the Galaxy developed into a "trilogy" of five books that sold more than 15 million copies in his lifetime. It was further developed into a television series, several stage plays, comics, a video game, and a 2005 feature film. Adams's contribution to UK radio is commemorated in The Radio Academy's Hall of Fame.

Prompt.6. KG generation

Question answer pairs: Q: What is Douglas Adams an instance of? A: Douglas Adams is an instance of human. Q: What is Douglas Adams' sex or gender? A: Douglas Adams' sex or gender is male. Q: Where was Douglas Adams born? A: Douglas Adams was born in Cambridge. Q: Where did Douglas Adams die? A: Douglas Adams died in Santa Barbara, California. O: When was Douglas Adams born? A: Douglas Adams was born on 1952-03-11. Q: On what date did Douglas Adams die? A: Douglas Adams died on 2001-05-11. Q: What occupation did Douglas Adams have? A: Douglas Adams was a writer, comedian, and dramatist.

Prompt.6. KG generation

```
####
Ontology:
@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#.
@prefix rdfs: http://www.w3.org/2000/01/rdf-schema#.
@prefix wdt: http://www.wikidata.org/prop/direct/.
@prefix wd: http://www.wikidata.org/entity/.
@prefix xsd: http://www.w3.org/2001/XMLSchema# .
wd: Douglas Adams rdfs: label "Douglas Adams"@en;
    wdt:InstanceOf wd:human ;
    wdt:SexOrGender wd:male :
    wdt: PlaceOfBirth wd: Cambridge;
    wdt:PlaceOfDeath wd:Santa_Barbara_California;
    wdt: DateOfBirth "1952-03-11" ^ xsd: date ;
    wdt: DateOfDeath "2001-05-11" ^ ^ xsd: date :
    wdt: Occupation wd: writer ;
    wdt: Occupation wd: comedian ;
    wdt: Occupation wd: dramatist ;
    wdt: LanguagesSpokenWrittenOrSigned wd: English;
    wdt:EducatedAt wd:St_Johns_College_Cambridge;
    wdt: EducatedAt wd: Brentwood School Essex ;
    wdt: AlumniOf wd: St Johns College;
    wdt: NotableWork wd: The_Hitchhikers_Guide_to_the_Galaxy ;
    wdt: NotableWork wd: Dirk Gentlys Holistic Detective Agency;
```

```
####
Document:
{doc}

####
Questions and Answer pairs:
{qa}

####
Knowledge Graph:
```

Experiment

- Dataset:
 - Wiki-NRE: 1,000 test samples, 45 relation types.
 - SciERC: 974 test samples, 7 relation types.
 - WebNLG: 1,165 test samples, 159 relation types.
- Evaluation Metric:
 - Partial F1-score for KG triples based on WebNLG+2020

Method	Wiki-NRE	SciERC	WebNLG
Non-LLM Baseline	0.484	0.532	0.767
LLM Baseline	0.647	0.07	0.728
Proposed (Mistral)	0.66/0.60	0.73/0.58	0.74/0.68
Proposed (GPT-4o)	0.71 /N/A	0.77/N/A	0.76/N/A

Topics

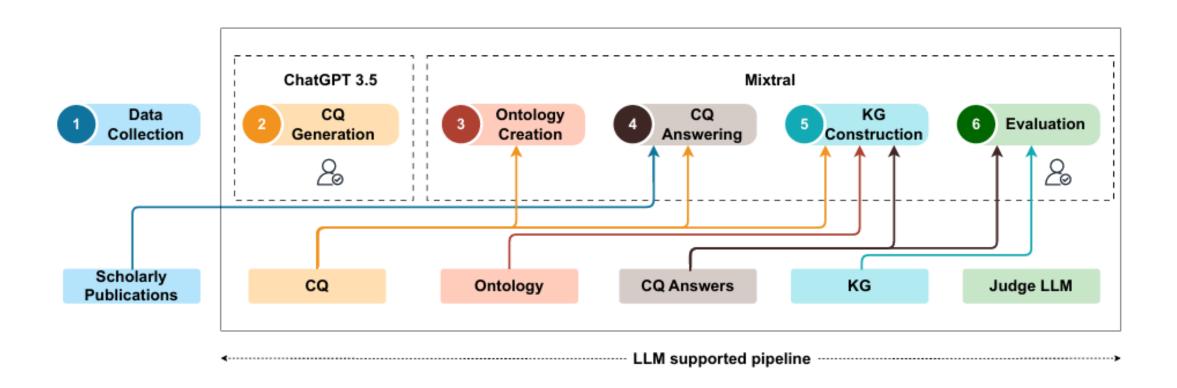
Knowledge Graph Construction with LLM

- [HI-AI@KDD, 2024] Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema
- [IEEE, BIBM, 2024] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Appplication of Knowledge Graph

■ [AAAI, 2025] K-ON: Stacking Knowledge On the Head Layer of LLM

From human experts to machines: An LLM supported approach to ontology and knowledge graph construction



From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Evaluation

- 1) Generated CQ Answers
 - The LLM evaluated the alignment of generated answers with human-generated ground truth on a 0–10 scale.
 - Scores were classified as Right (≥6), Wrong (<3), and Partial (3–5).
- 2) Automatically Extracted KG Concepts
 - The LLM verified whether KG concepts (triple) appeared in the generated CQ answers.
 - Only five scholarly articles were manually annotated by experts to test the feasibility of the pipeline.

From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Result

Publication / Combination of	Prompt	Prompt	Prompt	Prompt
prompt and CQ answers	v1 and	v1 and	v2 and	v2 and
	CQ an-	CQ an-	CQ an-	CQ an-
	swer v1	swer v2	swer v1	swer v2
Klein et al. 2015 [12]	24.32	X	X	9.42
Khalighifar et al. 2021 [11]	85.71	x	x	68.06
Choe et al. 2021 5	73.21	64.41	59.26	64.29
Mahmood et al. 2016 [16]	91.53	81.48	82.76	77.55
Younis et al. 2020 [25]	66.67	x	67.51	61.67

Topics

Knowledge Graph Construction with LLM

- [HI-AI@KDD, 2024] Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema
- [IEEE, BIBM, 2024] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction

Appplication of Knowledge Graph

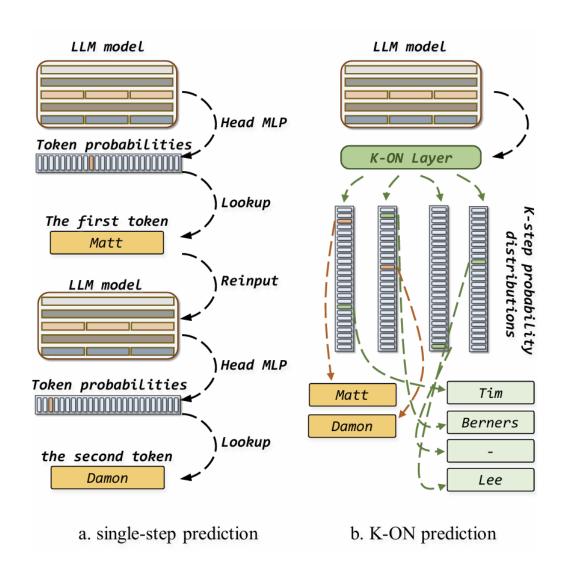
[AAAI, 2025] K-ON: Stacking Knowledge On the Head Layer of LLM

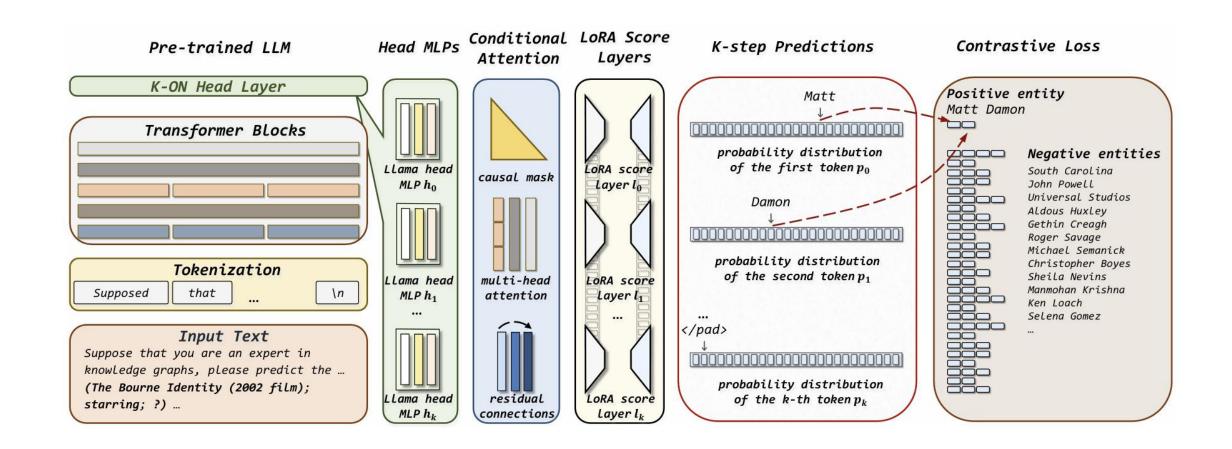
Problem background

- Granularity mismatch:
 - LLMs operate at the token level while KGs are organized around entities spanning multiple tokens to represent a single entity. This mismatch hinders direct entity prediction by LLMs.

Goal

• Propose a lightweight head-centric fine-tuning that adds K prediction heads to perform next-K-token one-shot entity prediction for KGs, trained with entity-level contrastive learning and HTT to align distributions.





0. Problem Setup

- Let $G = \{E, R, T\}$, where E is the set of entities, R the set of relations
- Goal:
 - Given an incomplete triple $(e_1, r_1, ?)$ or $(?, r_1, e_2)$, the task is to predict the missing entity

1. Tokenization & Encoding

Given an input sentence, tokenize it:

$$t_{0:N-1} = \{t_n \in V\}_{n=0}^{N-1}$$

Encode the token sequence using a transformer M:

$$h_{0:N-1}^{(m)} = M(t_{0:N-1})$$

Let $h_0^{(m)}$ denote the hidden representation at the final position of the query.

The original LLM head *H* produces the output distribution:

$$p_n = H(h_n^{(m)}) \in \mathbb{R}^{|V|}$$

2. K-ON: Creating K 'step representations' with head MLPs

• Project the shared hidden representation $h_0^{(m)}$ through K MLP heads:

$$h_{0:K-1}^{(h)} = \{L_k^{(h)}(\sigma(W_k^{(h)}h_0^{(m)}))\}_{k=0}^{K-1}$$

Each $h_k^{(h)}$ represents the intermediate feature for predicting the k-th token of the entity.

3. Conditional Attention (small Transformer) to restore sequential dependencies

• Define a causal maks $M \in \mathbb{R}^{K \times K}$:

$$M_{ij} = egin{cases} 1 & i \geq j \ 0 & i < j \end{cases}$$

Feed the sequence $h_{0:k}^{(h)}$ into a small transformer M_s with causal masking, then apply residual connection:

$$h_k^{(a)} = M_s(h_{0:k}^{(h)}, M) + h_0^{(m)}$$

Now $h_k^{(a)}$ serves as the final prediction vector for the k-to token.

4. Vocabulary Distribution via LoRA Score Layer

Apply LoRA to the score layer for each step:

$$W_k^{(S)} = W^{(S)} + A_k B_k, \quad p_k = W_k^{(S)} h_k^{(a)} \in \mathbb{R}^{|V|}$$

• Each p_k is the predicted distribution for the k-th token.

5. Label Entity as a Token Sequence of Length K

• Tokenize the entity label I_e and normalize to fixed length K using padding/truncation:

$$t_{0:K-1}(e) = P(\tau(l_e), K)$$

Ex) for K = 4:
"Matt Damon" \rightarrow [Matt, Damon, <pad>, <pad>]

6. Stack Token Distributions and Gather Entity Token Probabilities

Stack the token distribution:

$$P = egin{pmatrix} p_0 \ p_1 \ dots \ p_{K-1} \end{pmatrix}$$

• Extract the probabilities corresponding to entity token positions:

$$p(e) = [P_{0,t_0(e)}, P_{1,t_1(e)}, \dots, P_{K-1,t_{K-1}(e)}]$$

7. Aggregate Token-Level Scores into Entity Score

Apply weighted sum over token positions with learned weights:

$$p_e = \sum_{k=0}^{K-1} lpha_k p_k(e)$$

• Here, p_e represents the aggregated score for entity e.

8. Loss

Entity-Level Contrastive Loss (NCE)

$$L_{ ext{NCE}}(e) = -\log p_e + rac{1}{|N|} \sum_{e_i \in N} \log p_{e_j}$$

where e: correct entity and a sampled set of negatives $N \subset E$

- 2) Head Trajectory Tuning
 - (1) Supervised Fine-Tuning

$$\mathcal{L}_{\text{sft}}(e) = \sum_{k=0}^{K-1} \left(-\log p_k^e + \frac{1}{\mathcal{V}} \sum_{e_j \in \mathcal{V}} \log p_k^{e_j} \right)$$

(1) Token Distribution Tuning

$$\mathcal{L}_{\text{tdt}}(e) = \sum_{k=0}^{K-1} D_{\text{KL}}(\mathbf{p}_k^{\text{e, k-on}}, \mathbf{p}_k^{\text{e, llm}})$$

Results (MRR, Hits@1/3/10)

Model	Modality	DB15K			MKGW				
		MRR↑	Hits@1↑	Hits@3↑	Hits@10↑	MRR↑	Hits@1↑	Hits@3↑	Hits@10↑
TransE (Bordes et al. 2013)	S	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23
DistMult (Yang et al. 2015)	S	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86
RotatE (Sun et al. 2019)	S	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73
IKRL (Xie et al. 2017)	S+I	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07
TransAE (Wang et al. 2019)	S+I	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72
KG-Bert (Yao, Mao, and Luo 2019)	S+T	23.94	11.98	31.05	46.54	28.68	21.12	32.57	43.46
MMKRL (Lu et al. 2022)	S+T+I	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69
OTKGE (Cao et al. 2022)	S+T+I	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88
MMRNS (Xu et al. 2022)	S+T+I	32.68	23.01	37.86	51.01	<u>35.03</u>	28.59	37.49	<u>47.47</u>
KGLM (Youn and Tagkopoulos 2022)	S+T	28.47	17.66	36.02	48.89	34.12	27.01	36.87	.46.62
QEB (Wang et al. 2023)	S+T+I	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32
VISTA (Lee et al. 2023)	S+T+I	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61
MANS (Zhang, Chen, and Zhang 2023)	S+T+I	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78
FLT-LM (Lin et al. 2023)	S+T	33.45	24.56	37.67	50.12	32.75	25.89	32.87	44.56
AdaMF (Zhang et al. 2024c)	S+T+I	32.51	21.31	<u>39.67</u>	<u>51.68</u>	34.27	27.21	<u>37.86</u>	47.21
KG-Llama-7b (Yao et al. 2023)	S+T	-	13.46	-	-	-	20.20	-	-
GPT 3.5 Turbo (Zhu et al. 2023)	S+T	-	21.71		_		22.66		_
K-ON	S+T	38.10	30.13	42.77	53.59	36.64	30.05	38.72	48.26