

CS639 Deep Learning for NLP

LLMs and Knowledge Graphs

Junjie Hu



Slides adapted from Graham, Zhengbao
and the survey from Pan et al.

<https://junjiehu.github.io/cs639-spring26>

Goal for Today

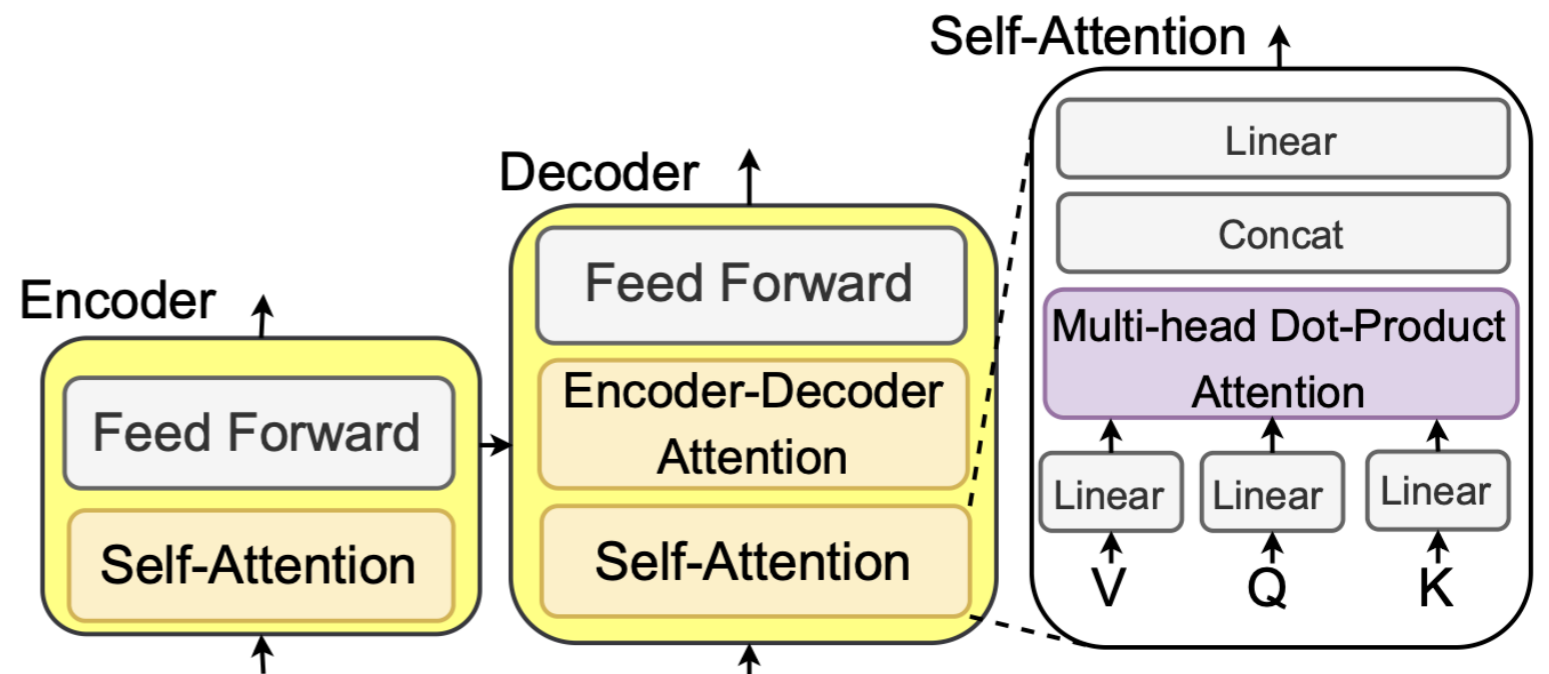
- Types of Knowledge Graph (KG)
- Integration of LLMs and KG
 - **KG-enhanced LLMs**
 - **LLM-augmented KGs**
 - **Synergized LLMs + KGs**
- Future Research Directions

Recap: Transformer LLMs

- **An encoder-decoder Transformer** (a.k.a. seq2seq model) consists of
 - A Transformer **encoder**: summarize input contexts
 - Another Transformer **decoder**: attend to input contexts, and iteratively generate words one by one
- **A Transformer layer**: consists of a multi-head self-attention layer, and two feedforward layers.
- **Autoregressive LMs**: generates the next word given the prefix context

$$P(X) = \prod_{i=1}^I P(x_i | x_1, \dots, x_{i-1})$$

Next Word Context

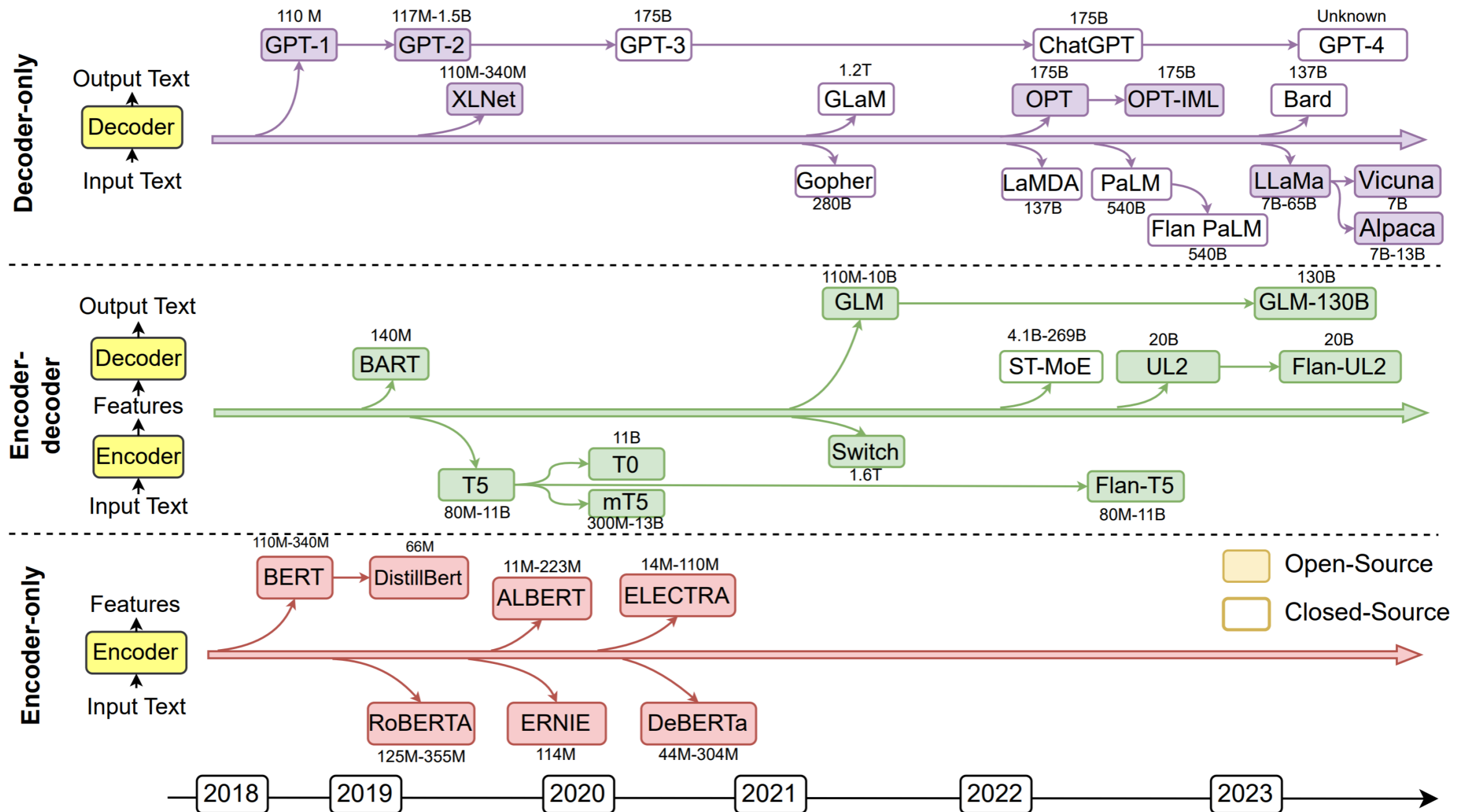


(See LM lecture #5)

(See Transformer lecture #7)

Large Language Models (LLMs)

- Overparameterized NNs pretrained on massive texts by self-supervised objectives



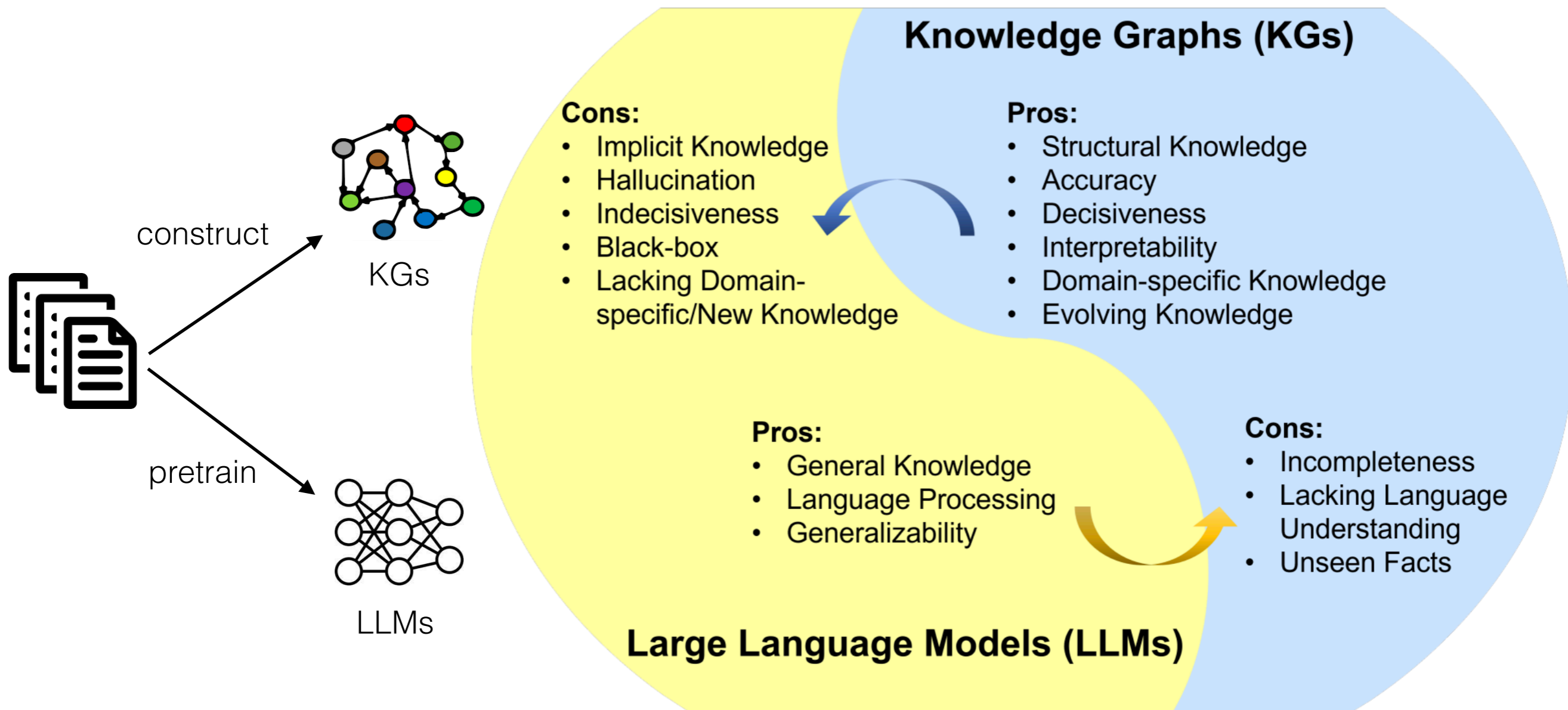
(See Pretraining lecture #8)

Knowledge Graph (a.k.a. Knowledge Base)

- Structured databases of knowledge usually containing
 - Entities (nodes in a graph)
 - Relations (edges between nodes)
- KG Constructions:
 - Manual curation by human experts
 - Information extraction by ML models or heuristics
 - Hybrid human-in-the-loop approach

LLM vs. KG

- Both LLMs and KGs store knowledge in a parametric and non-parametric way



LLMs + KGs

- Research questions:
 - How can we **learn to create/expand knowledge graphs** with large neural networks?
 - How can we **learn from the information in knowledge graphs** to improve LLMs?
 - How can we **use structured knowledge to answer questions?**

Types of Knowledge Graphs

Four Major Types of KGs

• Encyclopedic KGs

- Extracted from *diverse sources* (human experts, databases, encyclopedias)
- Examples: WikiData, Freebase, DBpedia, YAGO, NELL

• Commonsense KGs

- *Daily concepts* (e.g., objects and events) and *tacit knowledge*.
- Examples: ConceptNet, ATOMIC, ASER, TransOMCS, CausalBank

• Domain-specific KGs

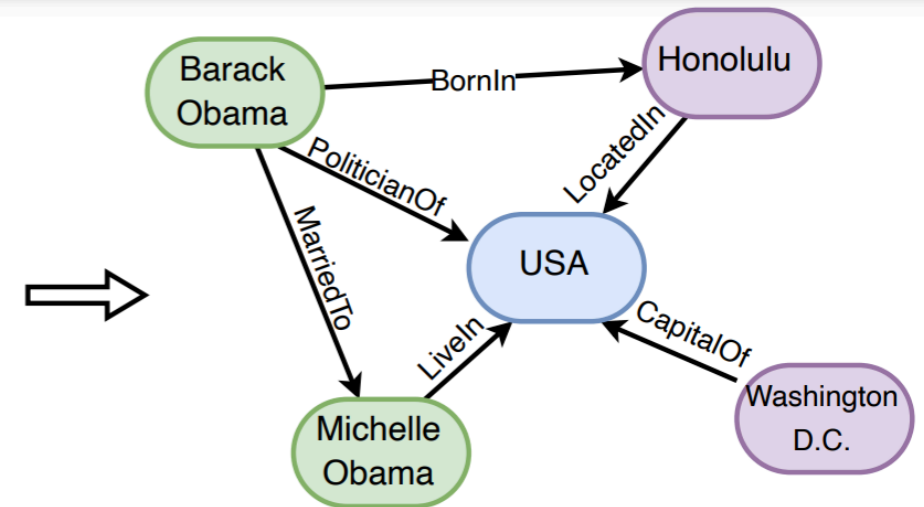
- Smaller but more accurate/reliable in a specific domain: medical, biology, and finance.
- Example: UMLS (medical)

• Multimodal KGs

- Multimodal facts (images, sounds, videos)
- Examples: IMGpedia, MMKG,

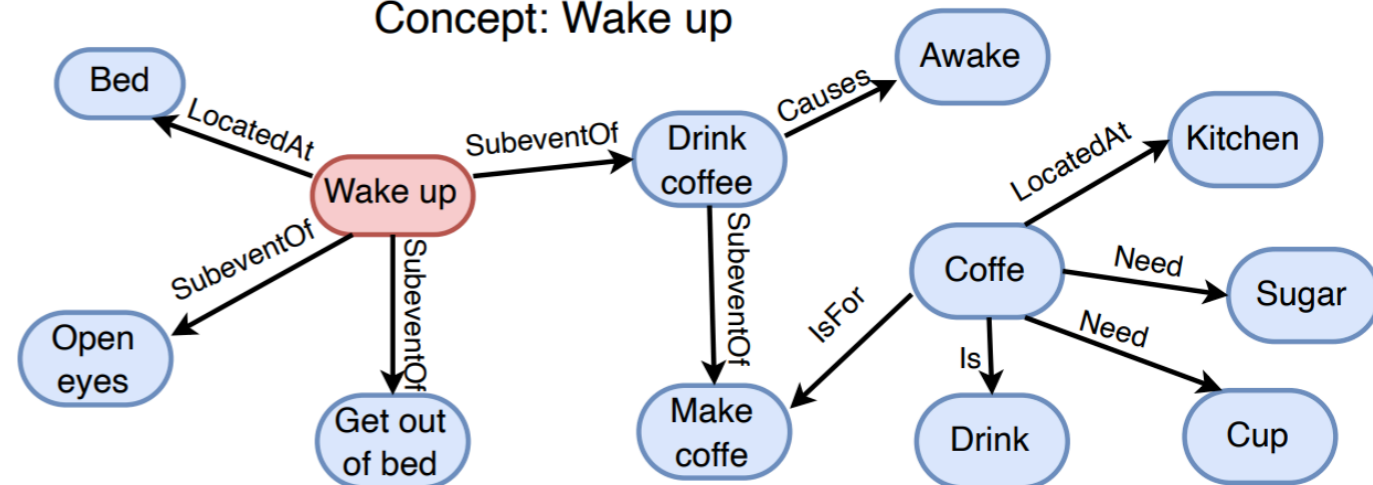
Encyclopedic Knowledge Graphs

Wikipedia



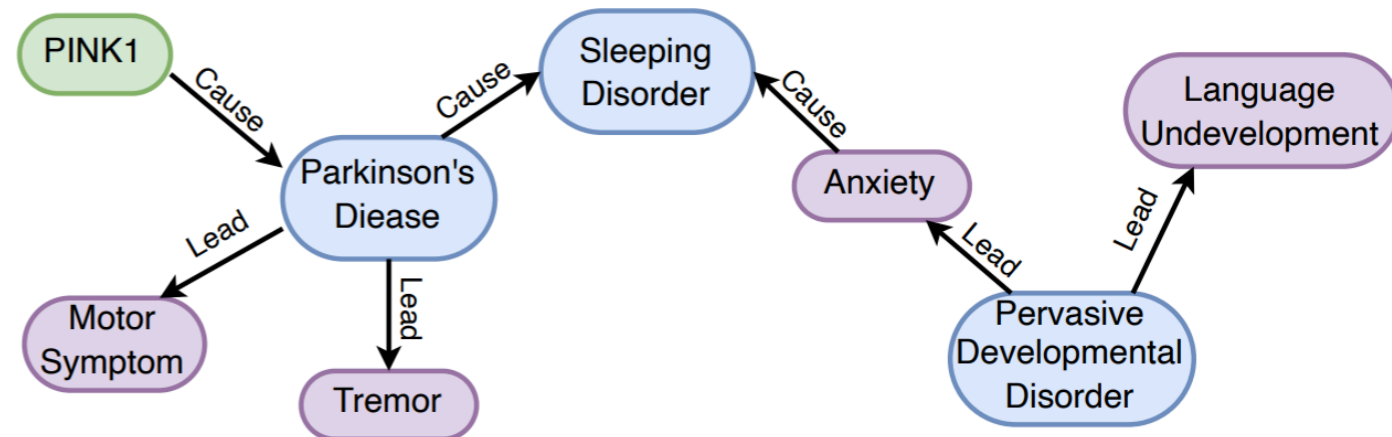
Commonsense Knowledge Graphs

Concept: Wake up

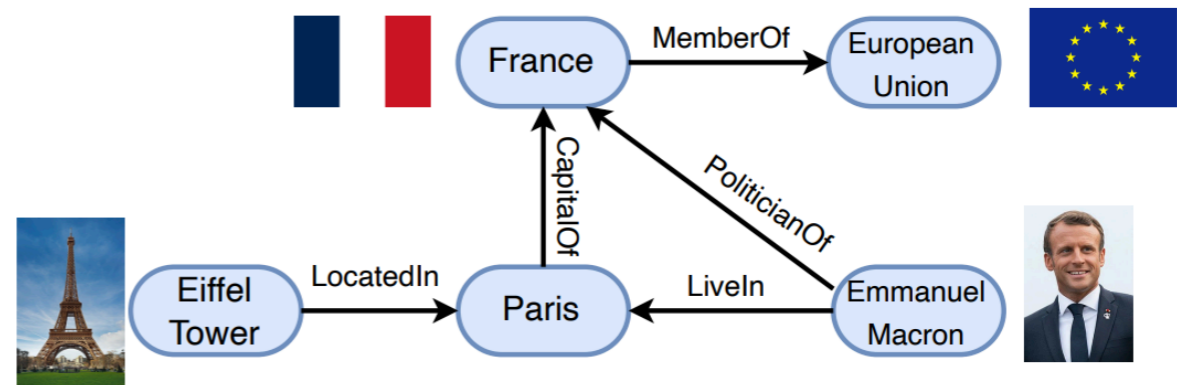


Domain-specific Knowledge Graphs

Medical Knowledge Graph

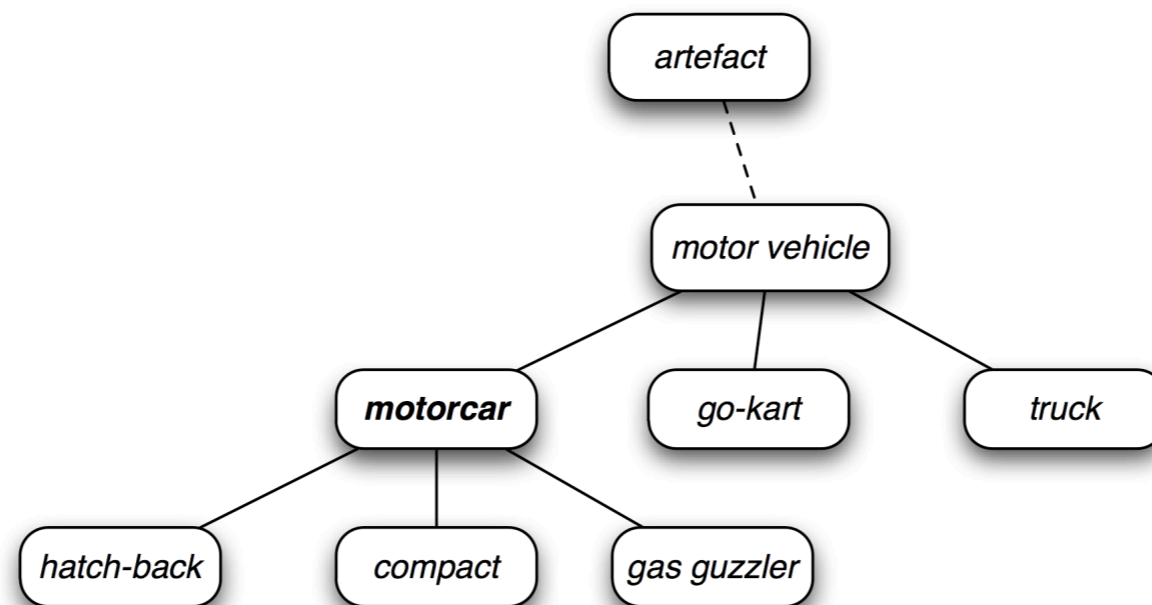


Multi-modal Knowledge Graphs



WordNet (Miller 1995)

- WordNet is a large database of words including parts of speech, semantic relations



- Nouns: is-a relation (hatch-back/car), part-of (wheel/car), type/instance distinction
- Verb relations: ordered by specificity (communicate -> talk -> whisper)
- Adjective relations: antonymy (wet/dry)

DBPedia (Auer et al. 2007)

- Extraction of structured data from Wikipedia

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Carnegie Mellon University (**Carnegie Mellon** or **CMU** /ˈkɑːrniɡi ˈmɛlən/ or /kɑːrˈneɪɡi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by [Andrew Carnegie](#) as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the [Mellon Institute of Industrial Research](#) to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (5 km) from [Downtown Pittsburgh](#). Carnegie Mellon has seven colleges and independent schools: the [College of Engineering](#), [College of Fine Arts](#), [Dietrich College of Humanities and Social Sciences](#), [Mellon College of Science](#), [Tepper School of Business](#), [H. John Heinz III College of Information Systems and Public Policy](#), and the [School of Computer Science](#). The university also has campuses in [Qatar](#) and [Silicon Valley](#), with degree-granting programs in six continents.

Carnegie Mellon is ranked 25th in the United States and 77th in the world by *U.S. News & World Report*.^[9] It is home to the world's first degree-granting Robotics and Drama programs,^[10] as well as one of the first Computer Science departments.^[11] The university was ranked 89th for R&D in 2015 having spent \$242 million.^[12]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 20 Nobel Prize Laureates,^[13] 12 Turing Award winners, 22 Members of the American Academy of Arts & Sciences,^[14] 19 Fellows of the American Association for the Advancement of Science, 72 Members of the [National Academies](#), 114 Emmy Award winners, 44 Tony Award laureates, and 7 Academy Award winners.^[15]

Structured data

Coordinates: 40.443322°N 79.943583°W﻿ / ﻿40.443322°N 79.943583°W﻿ / 40.443322; -79.943583

Carnegie Mellon University



Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) ^[1] Carnegie Mellon University (1988–present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie

- [owl:Thing](#)
- [dul:Agent](#)
- [dul:SocialPerson](#)
- [wikidata:Q24229398](#)
- [wikidata:Q3918](#)
- [wikidata:Q43229](#)
- [dbo:Agent](#)
- [dbo:EducationalInstitution](#)
- [dbo:Organisation](#)
- [dbo:University](#)
- [geo:SpatialThing](#)
- [schema:CollegeOrUniversity](#)
- [schema:EducationalOrganization](#)
- [schema:Organization](#)
- [umbel-rc:Business](#)
- [umbel-rc:EducationalOrganization](#)
- [umbel-rc:Organization](#)
- [umbel-rc:University](#)

WikiData (Bollacker et al. 2008)

- *Curated* database of entities, linked, and extremely large scale, multilingual

The screenshot shows the WikiData page for Richard Feynman. The page includes a header with the name "Richard Feynman" and a dropdown menu. Below the header are links for "Discuss 'Richard Feynman'" and "Hide Empty Fields". A small image of Feynman is shown on the left. The main content area lists various properties of Feynman, such as "Types", "Also known as", "Gender", "Date of Birth", "Place of Birth", "Country Of Nationality", "Profession", "Religion", "Parents", "Children", and "Siblings". A tooltip is visible over the "Siblings" list, showing a list of names and a detailed description for Joan Feynman. The right sidebar contains sections for "Page History", "Web Link(s)", "Employment history", "Education", "Quotations", and "Books Written".

Richard Feynman

Discuss "Richard Feynman" Hide Empty Fields

Types: Person (People), Author (Publishing), Physicist (Science), Deceased Person (People), Film writer (Film), Influence Node (mikelove's types), Person Or Being In Fiction (Fictional Universes), Book Subject (Publishing)

Also known as: Richard Phillips Feynman

Gender: Male

Date of Birth: May 11, 1918

Place of Birth: Far Rockaway, Queens

Country Of Nationality: United States

Profession: Physicist, Scientist

Religion: Atheism

Parents: double-click to add

Children: Michelle Louise Feynman, Carl Feynman

Siblings:

- Sibling
- Joan Fey
- Joan Feynman** (Person)
- Richard Feynman ... (Richard Phillips Feynman) (Person, Author, Physicist, Deceased Person, Film writer)
- Ana Gasteyer (Person, Film actor, TV Actor, Theater Actor)
- Gervase of Tilbury (Person)
- Alec Baldwin ... (Alexander Rae Baldwin) (Person, Film actor, Film director, Film producer, TV personality)
- Ernest Thesiger (Person, Film actor, Deceased Person)
- Mean Girls (Film)
- Riverside Drive (Landscape project)
- Portrait of Jennie (Film)
- Television Personalities ... (The Television Personalities) (Band)

Page History
Created by Melaweb Oct 22, 2006
Last edited by robert Oct 29, 2007

Web Link(s)
double-click to add

Employment history
Cornell University
California Institute of Technology
Thinking Machines

Education
Princeton University • 1942 • Ph.D.
Massachusetts Institute of Technology • 1939 • Bachelor's degree

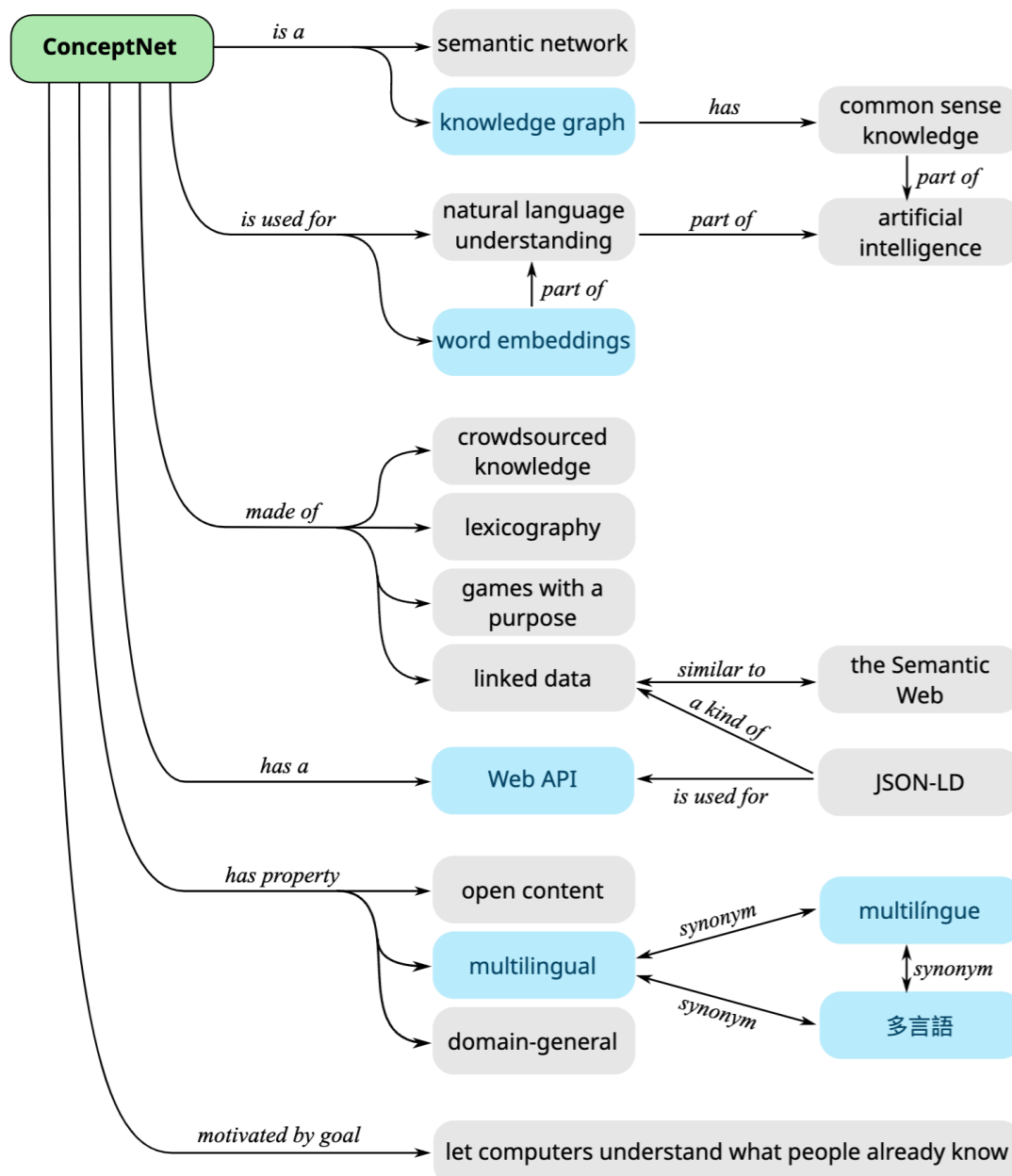
Quotations
I like sex: sure, it may give some results, but that's not why we do it.
I do not create, I do not understand.

Books Written
What Do You Care What Other People Think?
The Pleasure of Finding Things Out
The Feynman Lectures on Physics
Surely You're Joking, Mr. Feynman!

Joan Feynman
Person
Joan Feynman (b. 31 March 1928) is an astrophysicist who made original studies of the interactions between the solar wind and the Earth's magnetosphere. While working at the NASA Ames Research Centre in 1971, Feynman showed that coronal mass injections could be identified by the presence of helium in...

ConceptNet

- An open, multilingual KG originated at MIT Media Lab in 1999
- It has since grown to include knowledge from other crowdsourced resources, expert-created resources, and games with a purpose.



en knowledge

An English term in ConceptNet 5.8

Sources: Open Mind Common Sense contributors, CC-CEDICT 2017-10, DBpedia 2015, JMDict 1.07, Verboesity players, German Wiktionary, English Wiktionary, French Wiktionary, and Open Multilingual WordNet
View this term in the API

Documentation
FAQ
Chat
Blog

Synonyms

sh znanje →
en cognition (n, wn) →
ja ナレッジ (n) →
ja ノリッジ (n) →
ja ノレッジ (n) →
ja 人智 (n) →
ja 人知 (n) →
ja 学 (n) →
ja 学力 (n) →
ja 学殖 (n) →
ja 弁え (n) →
ja 心得 (n) →
ja 承知 (n) →
ja 智識 (n) →
ja 知得 (n) →
ja 知見 (n) →
ja 知識 (n) →
ja 認識 (a) →
ja 辨え (n) →
zh 知识 →

More »

Related terms

en learn →
ceb walay earth (a) →
en key stage (n) →
en know (v) →
en driver →
en london →
en route →
sh znanje (n) →
en science →
en test →
ady ШІЭНЫГЪЭ (n) →
ang andgiet (n) →
ang andgiete (n) →
ang cann (n) →
ang cunnung (n) →
ang cybpu (n) →
ang gefræge (n) →
ang gewitt (n) →
ang wittig (a) →
ar تعالم (v) →

More »

Causes of knowledge

en learning →
en remembering →
en listening to the radio →
en memorising →
en study →
en studying →
en analysing something →
en discovering the truth →
en expressing information →
en finding information →
en going to school →
en knowing how the stock market performed →
en looking through a telescope →
en reading →
en reading a magazine →
en studying for a subject →
en surfing the web →
en thinking →

Things that require knowledge

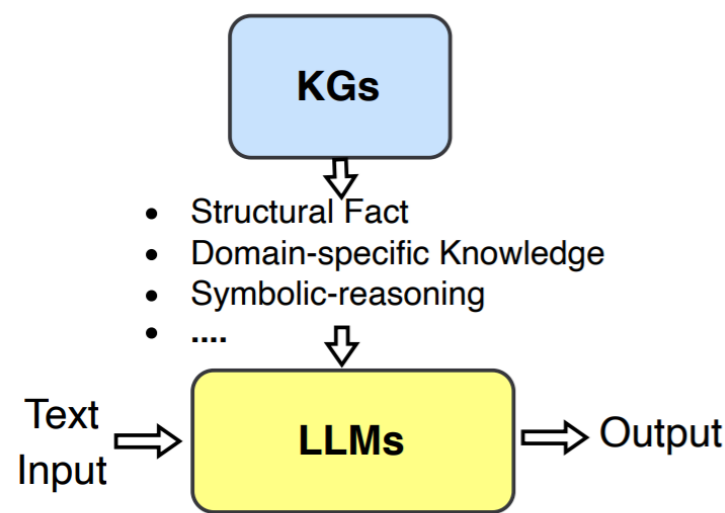
en answering questions →
en checking vital signs →
en testing the car →
en weeding the garden →
en debating politics →
en designing software →
en discovering the truth →
en doing a crossword puzzle →
en expressing information →
en fix a computer →
en giving a clue →
en handling proposals →
en having an examination →
en helping someone →
en passing your university exams →
en preparing for a vote →
en program →
en putting on the stand →

<https://conceptnet.io/c/en/knowledge>

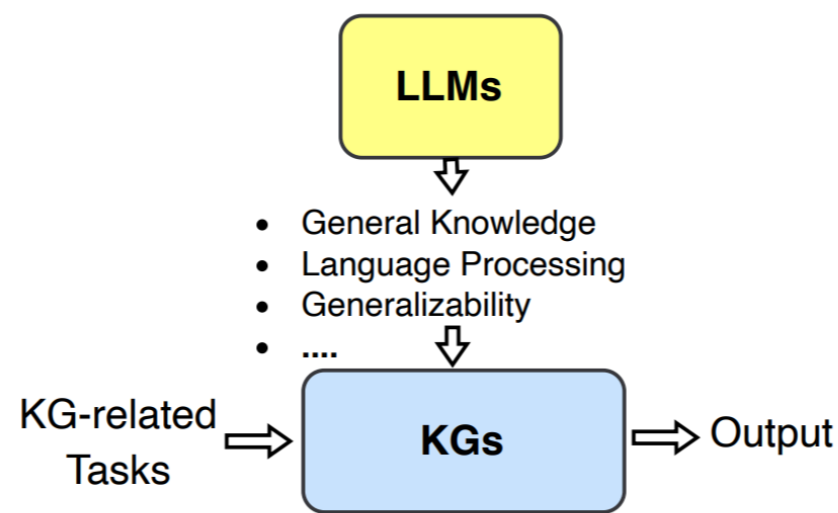
Unifying LLMs and KGs

General Roadmap of Unifying KGs and LLMs

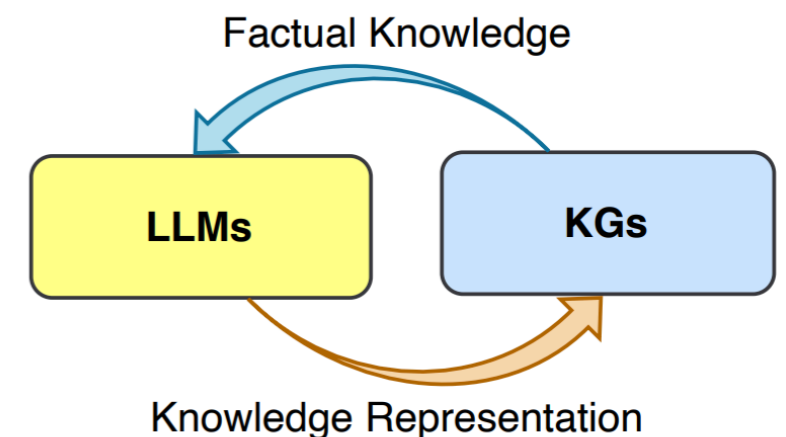
- Three general frameworks in the NLP/Data Mining communities
 - **KG-enhanced LLMs**: incorporate KGs into LLM pretraining/inference
 - **LLM-augmented KGs**: use LLMs to augment the incomplete KGs (improve KG representations, KG construction)
 - **Synergized LLMs + KGs**: mutually improve each other from all perspectives (including data, model, optimization and applications)



a. KG-enhanced LLMs

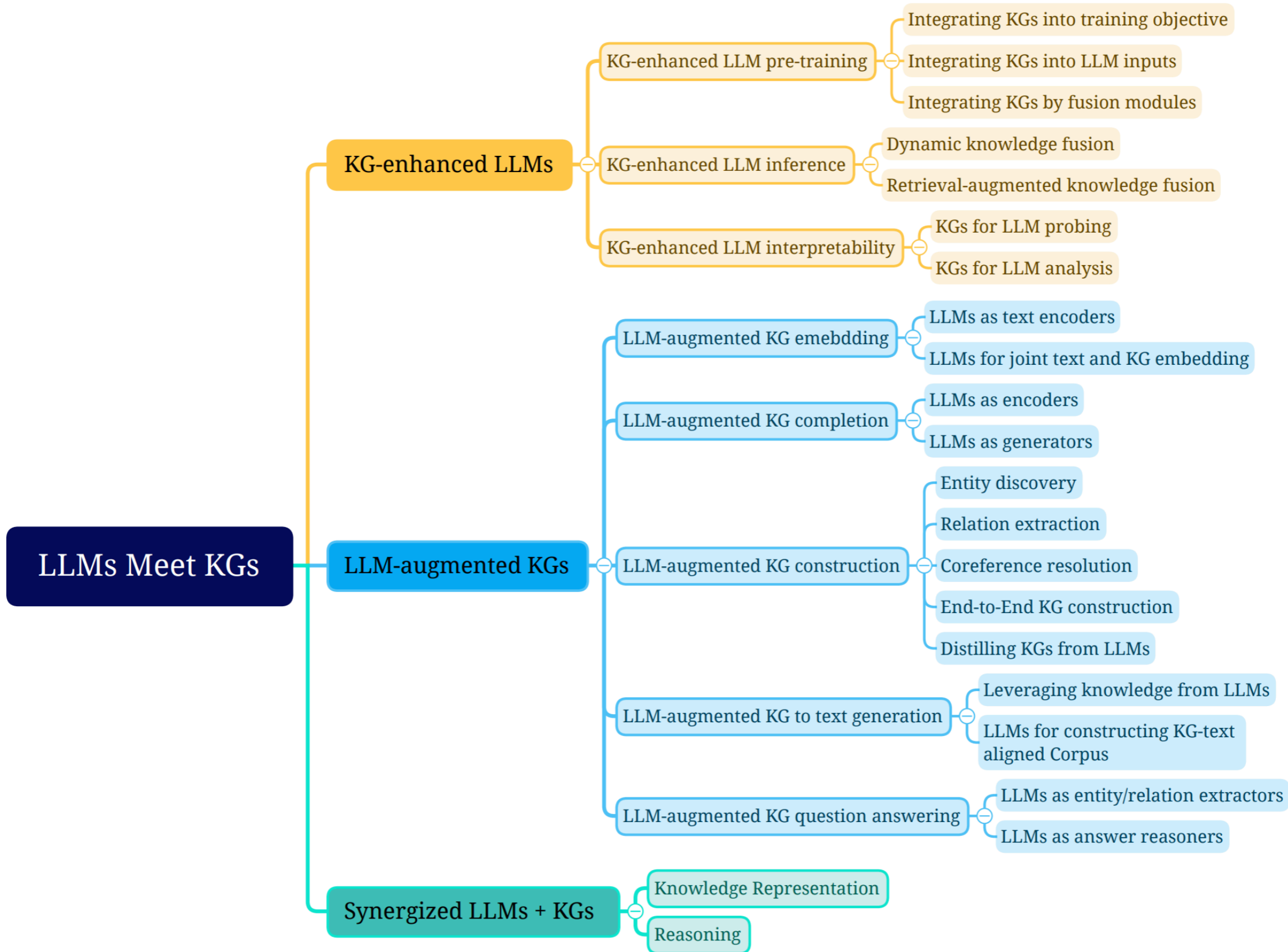


b. LLM-augmented KGs

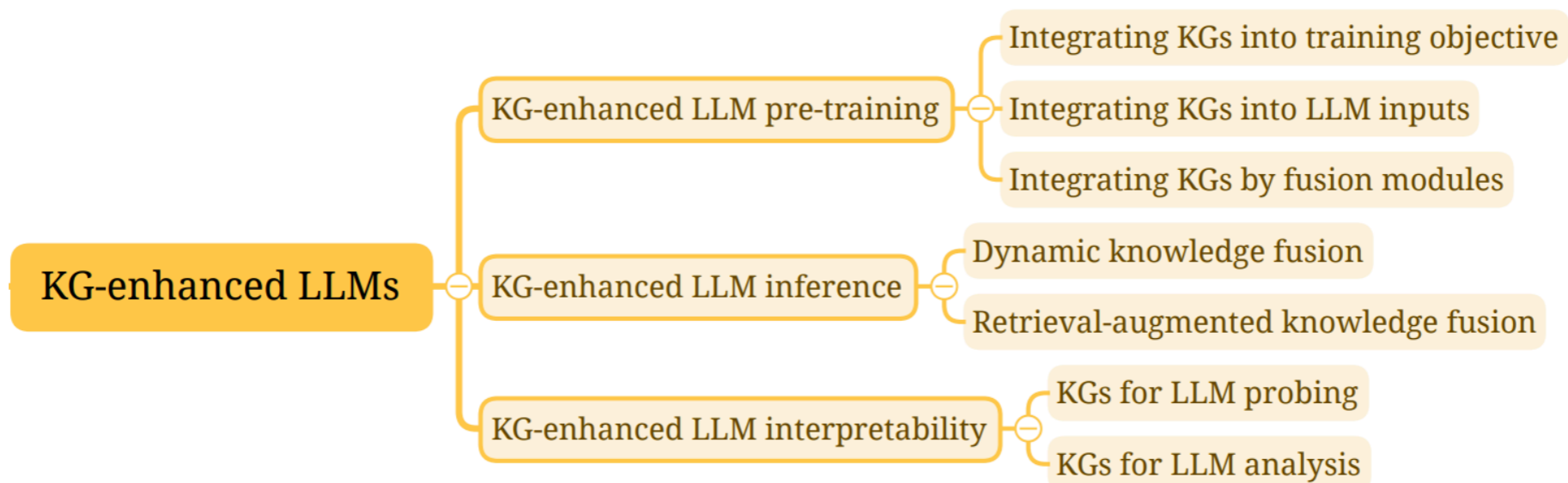


c. Synergized LLMs + KGs

Fine-grained Categorization of Existing and Ongoing Research



KG-enhanced LLMs

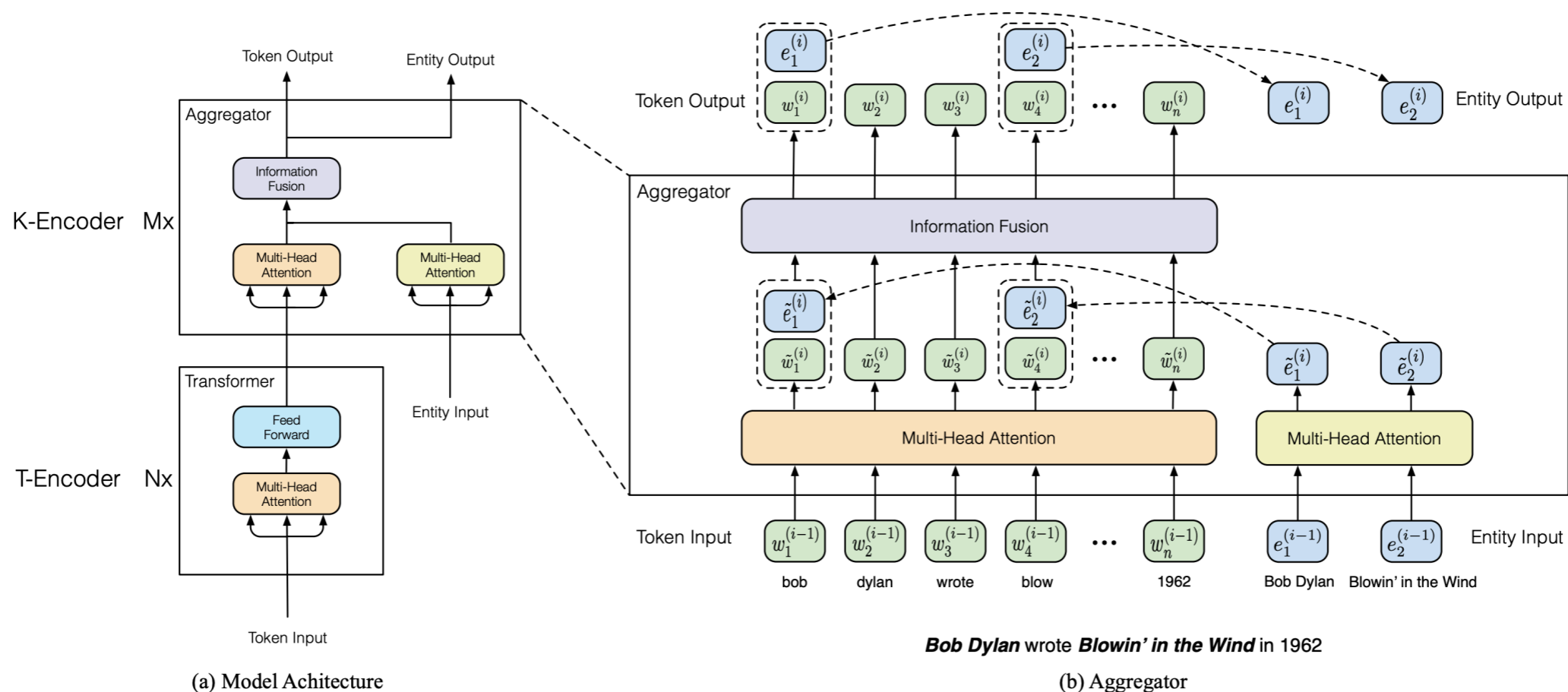


KG-enhanced LLM Pretraining

- **ERNIE: Denoising auto-encoder**

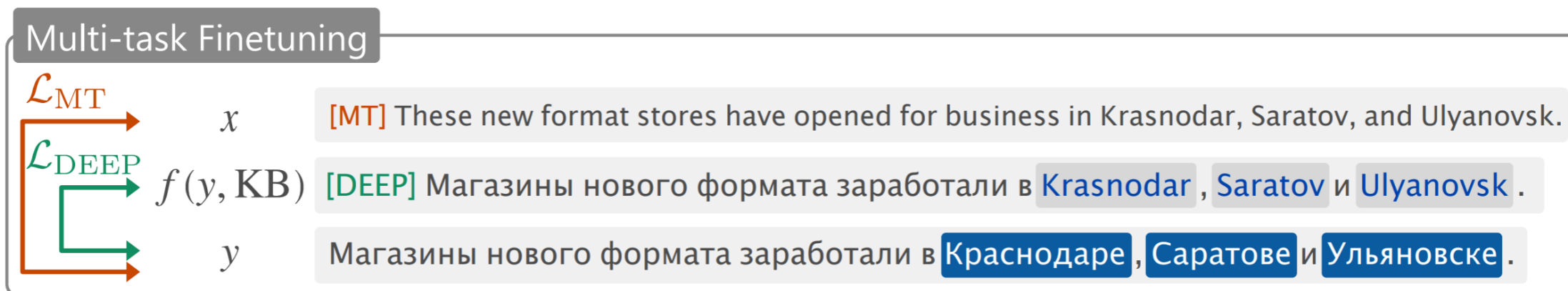
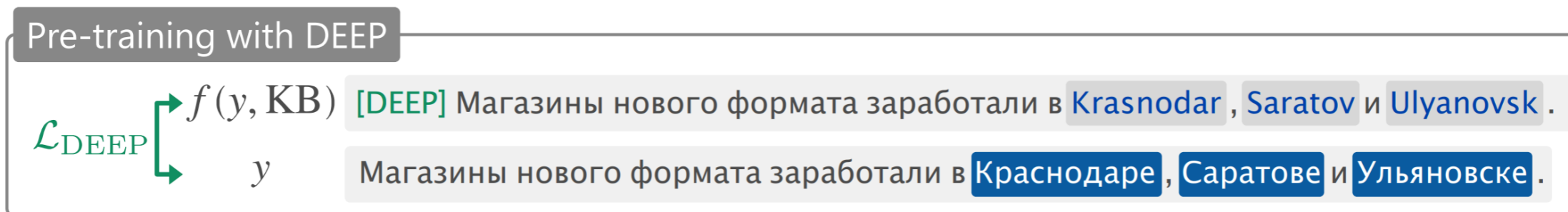
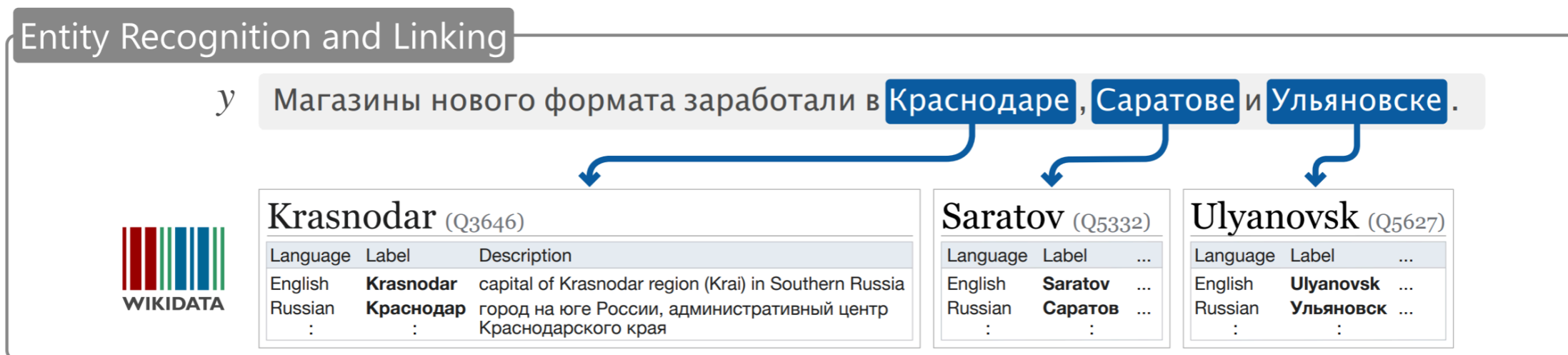
- 5% of time, replace the entity with a random entity, and train the model to correct the errors
- 15% of time, mask the token-entity alignment, train the model to predict the entity given the word
- 80% of time, keep token-entity alignments unchanged, train the model to integrate entity info into token embedding for better language modeling.

$$p(e_j|w_i) = \frac{\exp(\text{linear}(\mathbf{w}_i^o) \cdot \mathbf{e}_j)}{\sum_{k=1}^m \exp(\text{linear}(\mathbf{w}_i^o) \cdot \mathbf{e}_k)}$$



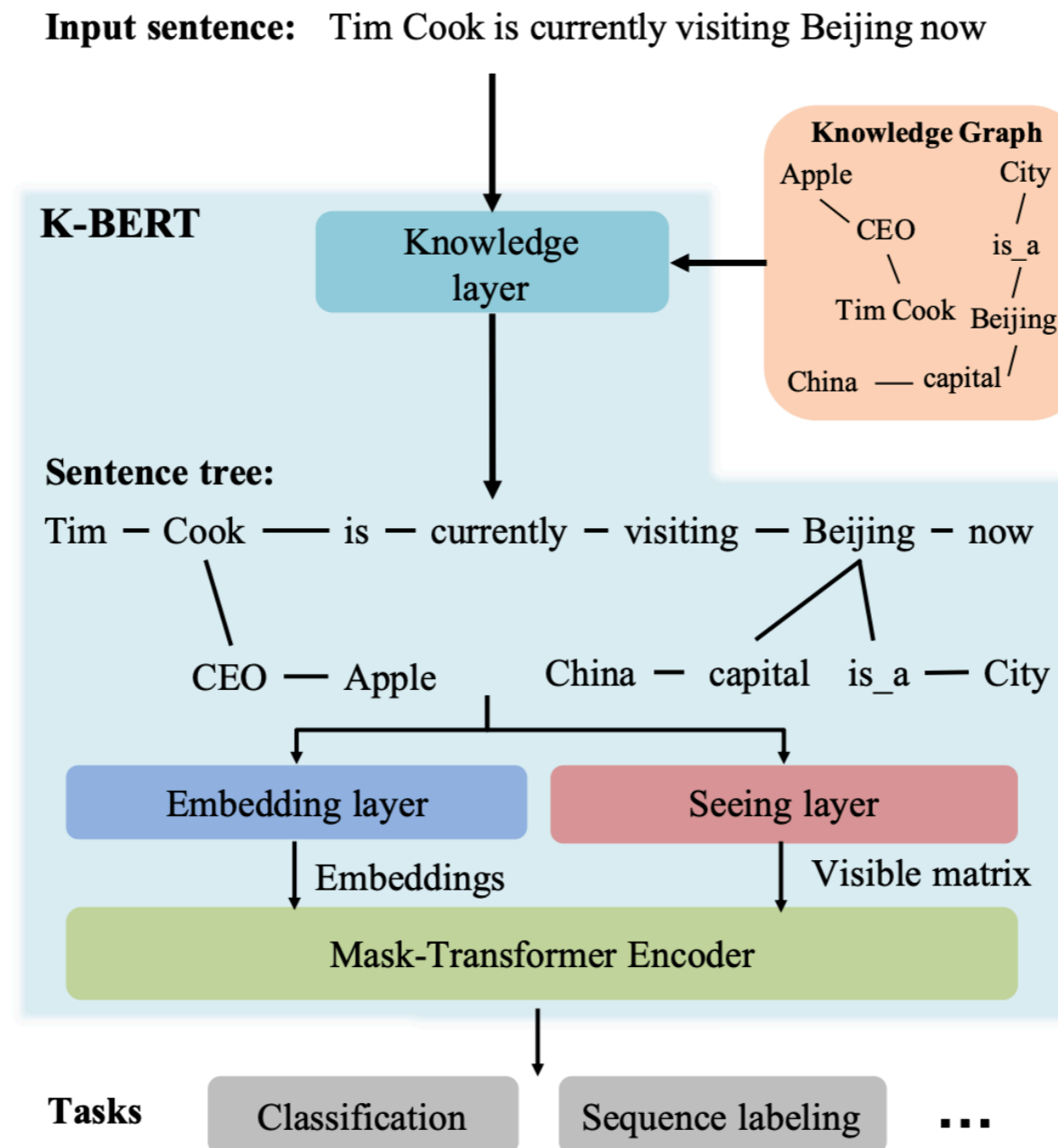
KG-enhanced LLM Pre-training

- **DEEP:** Integrate multilingual KG data to multilingual LLM pre-training:
 - Perform entity linking to detect mentions and link them to a KG
 - Obtain their multilingual translational information, add them to the pre-training data
 - Train an encoder-decoder model by machine translation and denoising auto-encoding objectives (multi-task learning)



K-BERT

- Add some relation triplets to entity mentions in the sentence
 - (Tim Cook, CEO, Apple), (Beijing, capital, China), (Beijing, is_a, City)



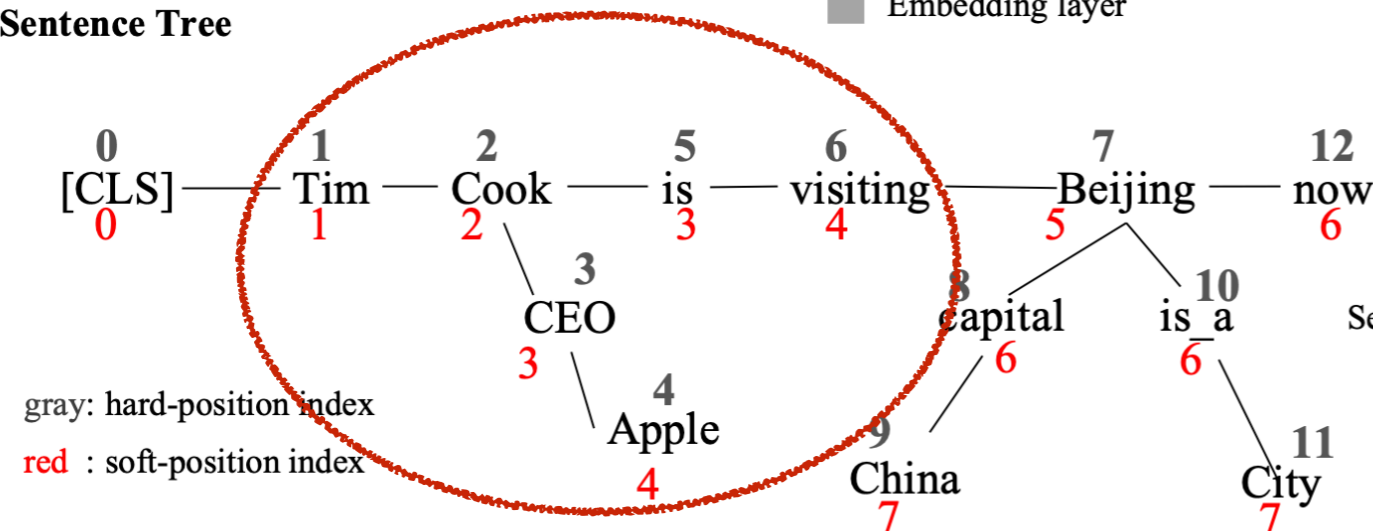
K-BERT

- Modify the position embedding of the added path
- Modify the attention masks so that only the entity mention (for example, **Tim Cook**) can attend to its corresponding relation triplet (**Tim Cook, CEO, Apple**).

Embedding Representation

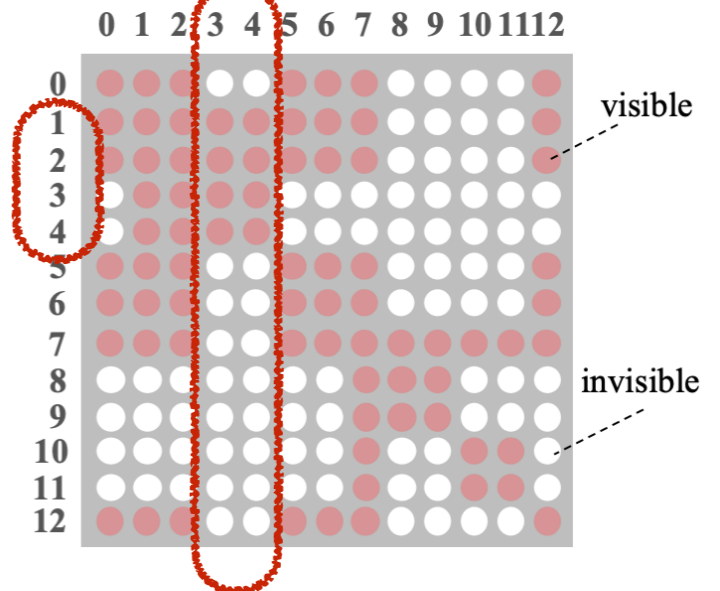
Token embedding	[CLS]	Tim	Cook	CEO	Apple	is	visiting	Beijing	capital	China	is_a	City	now
	+	+	+	+	+	+	+	+	+	+	+	+	+
Soft-position embedding	0	1	2	3	4	3	4	5	6	7	6	7	6
	+	+	+	+	+	+	+	+	+	+	+	+	+
Segment embedding	A	A	A	A	A	A	A	A	A	A	A	A	A

Sentence Tree



Embedding layer

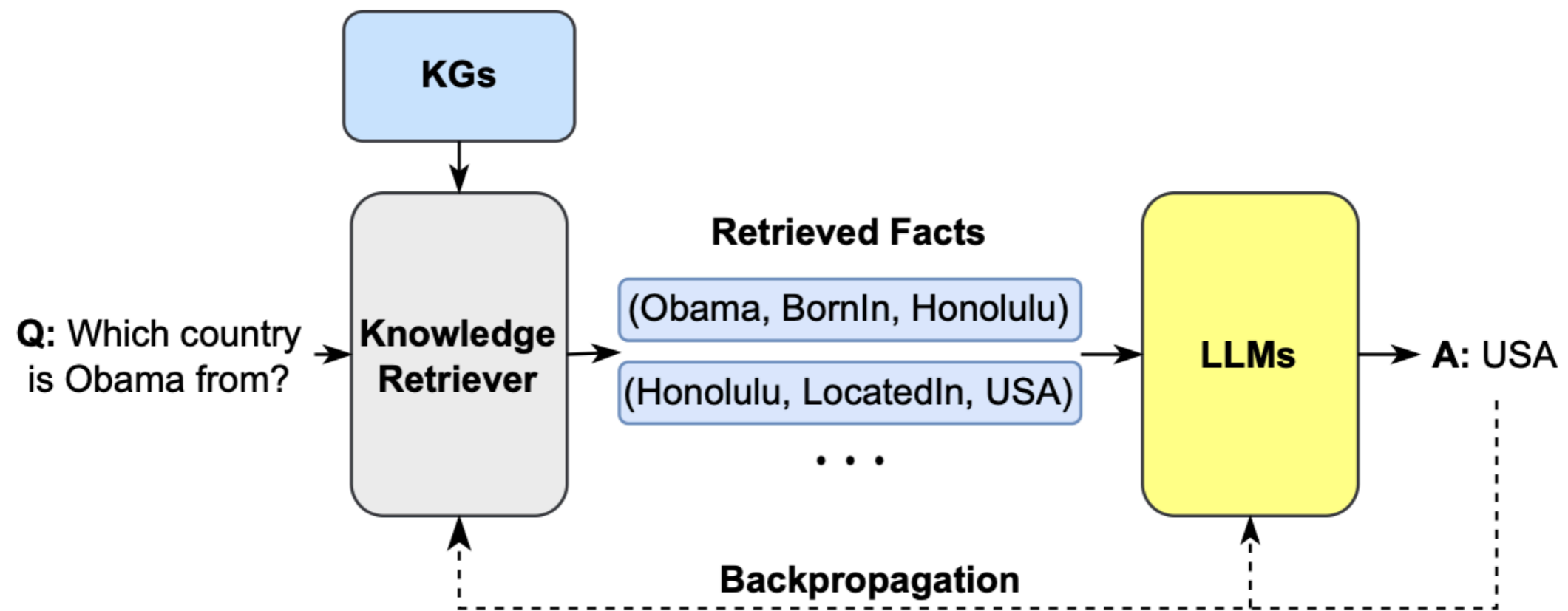
Visible Matrix



Seeing layer

KG-enhanced LLM Inference

- At test time, retrieve related KG information to augment the LLM for prediction
- Most methods focus on OpenQA tasks, as the model requires up-to-date knowledge



(Dotted line: Backpropagation is optional)

SeqGraph: Integrate KGs to prompt LLMs

- Given a text query q , we have access to a search engine to retrieve a set of related text passages $c_{1:n} = \{c_1, \dots, c_n\}$ and an entity graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ constructed from the retrieved unstructured texts. We aim to use a LLM to estimate:

Reasoning probability $p(r|q, c_{1:n}, \mathcal{G})$

Answer probability $p(a|q, c_{1:n}, \mathcal{G}, r)$

Q: Who is the executive producer of the film that has a score composed by Jerry Goldsmith?

C1: The iconic, avant-garde score to the film "Alien" was composed by Jerry Goldsmith.

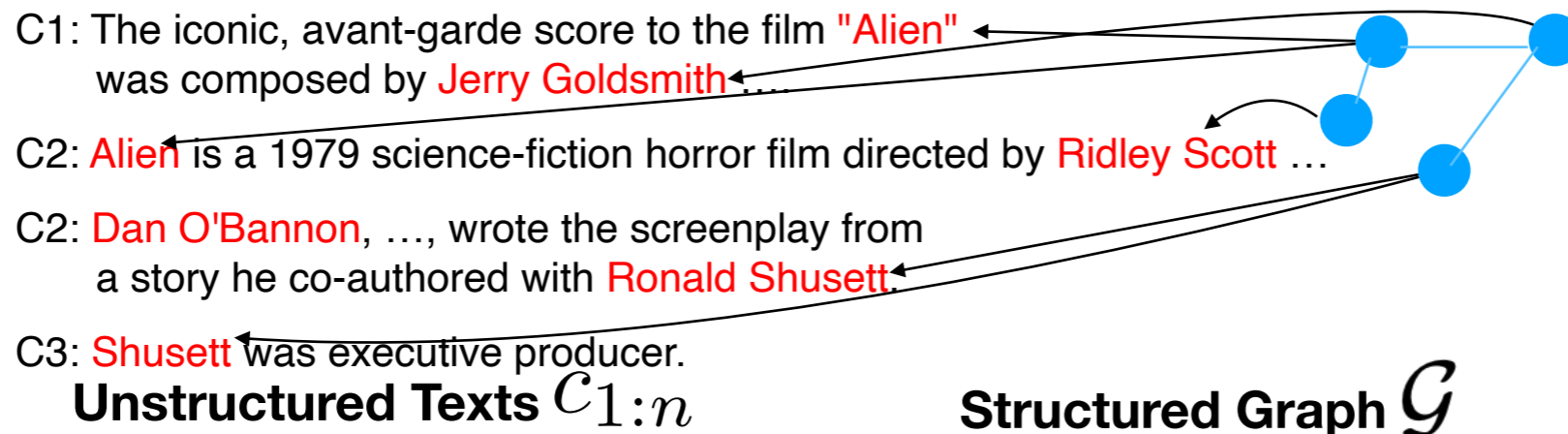
C2: Alien is a 1979 science-fiction horror film directed by Ridley Scott ...

C2: Dan O'Bannon, ..., wrote the screenplay from a story he co-authored with Ronald Shusett.

C3: Shusett was executive producer.

Unstructured Texts $c_{1:n}$

Structured Graph \mathcal{G}



SeqGraph Workflow

- Combine retrieved context passages with the question as input to LLMs

Question:

"An American Werewolf in Paris was a partial sequel to the comedy film starring whom?"

Passage 1: An American Werewolf in Paris

[f1] It follows the ... An American Werewolf in London
[f2] The film is a... the United States and France.

Input Sequence 1

Question: An American ... whom? **Title:** An American Werewolf ... **Context:** [f1]: It follows ...
[f2] The film is a ... the United States and France.

Entity span
Passage title span

SeqGraph Workflow

- Identify entity mentions from the retrieved context passages, link them to WikiData, and construct a local knowledge graph

Question:

"An American Werewolf in Paris was a partial sequel to the comedy film starring whom?"

Passage 1: An American Werewolf in Paris

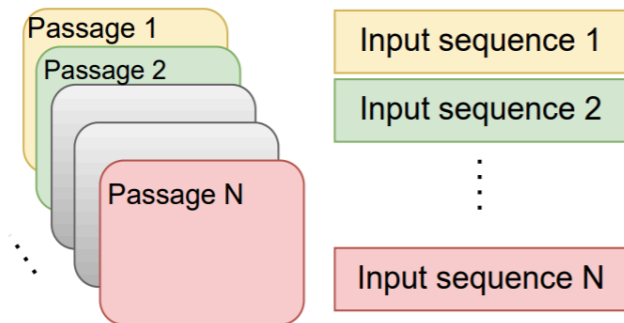
[f1] It follows the ... An American Werewolf in London
[f2] The film is a... the United States and France.

Input Sequence 1

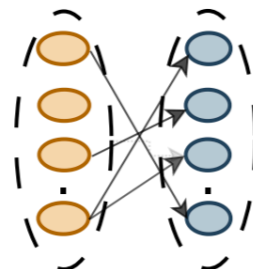
Question: An American ... whom? **Title:** An American Werewolf ... **Context:** [f1]: It follows ... [f2] The film is a ... the United States and France.

Entity span

Passage title span

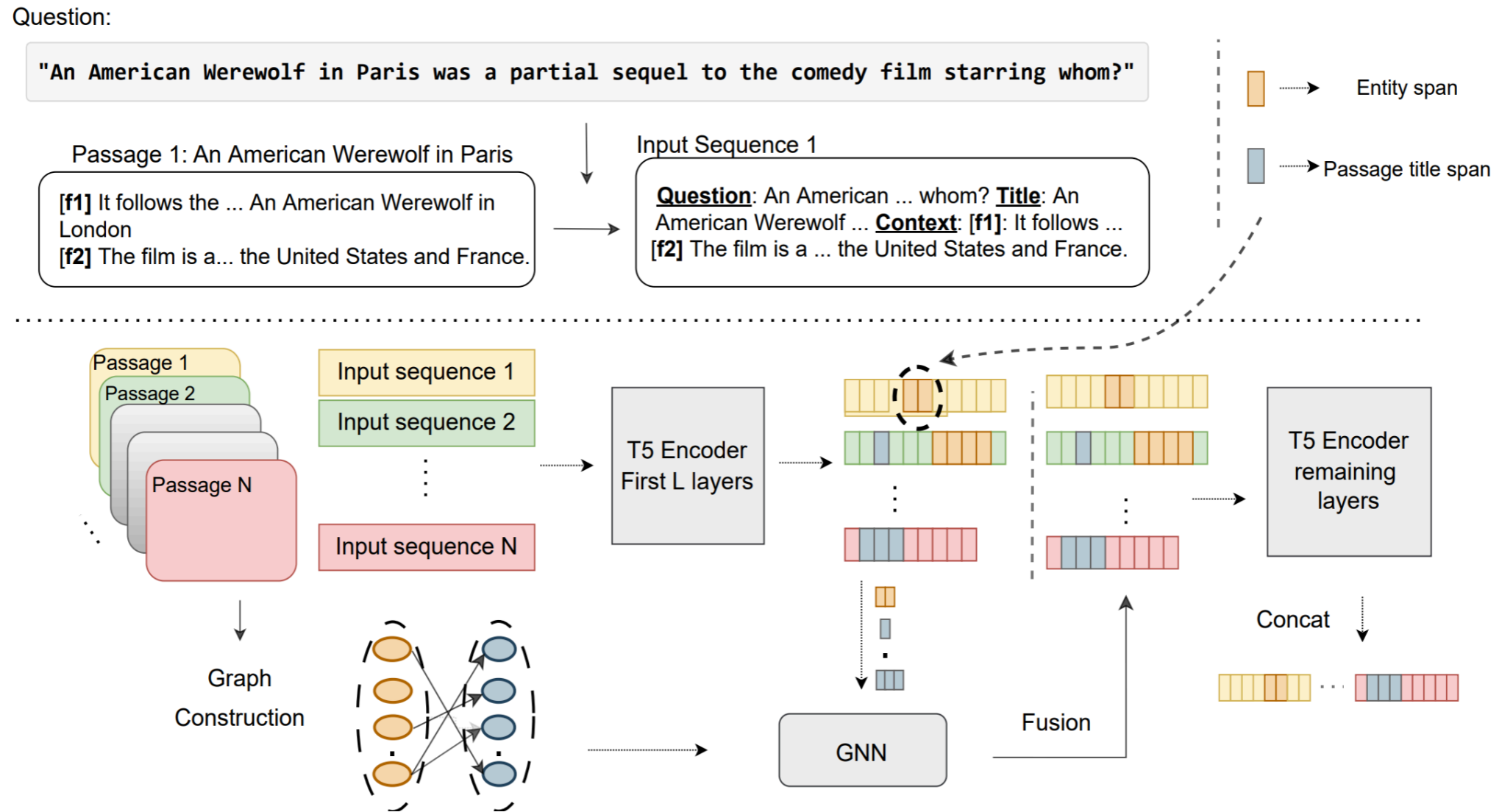


Graph Construction



SeqGraph Workflow

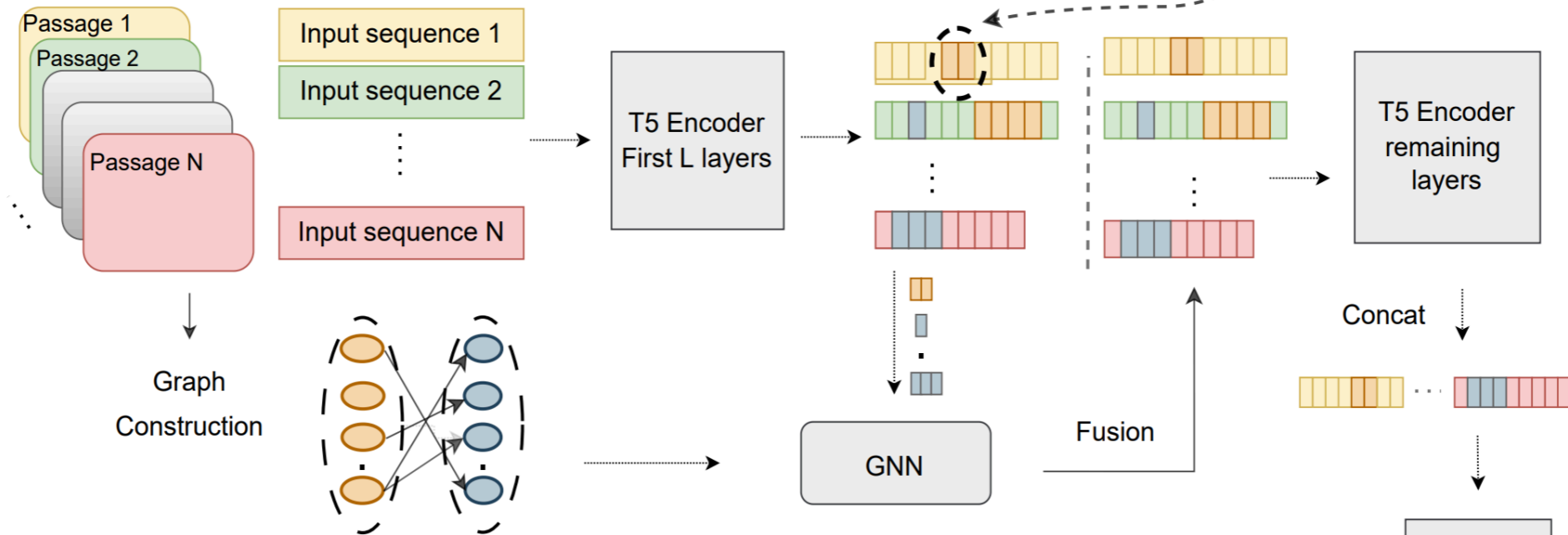
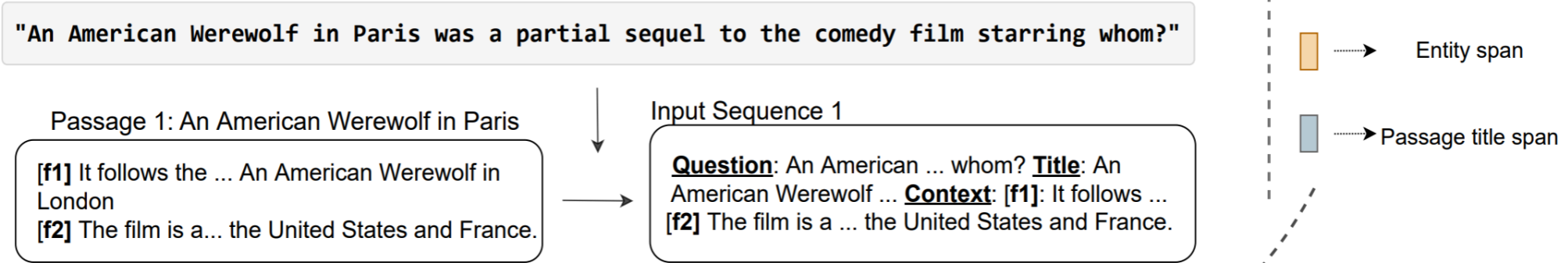
- Use a LLM to encode text, and a GNN to encode KG, and fuse their representations



SeqGraph Workflow

- Feed the fused representations to a LLM decoder to generate both reasoning path, and the final answer to a multi-hop question

Question:



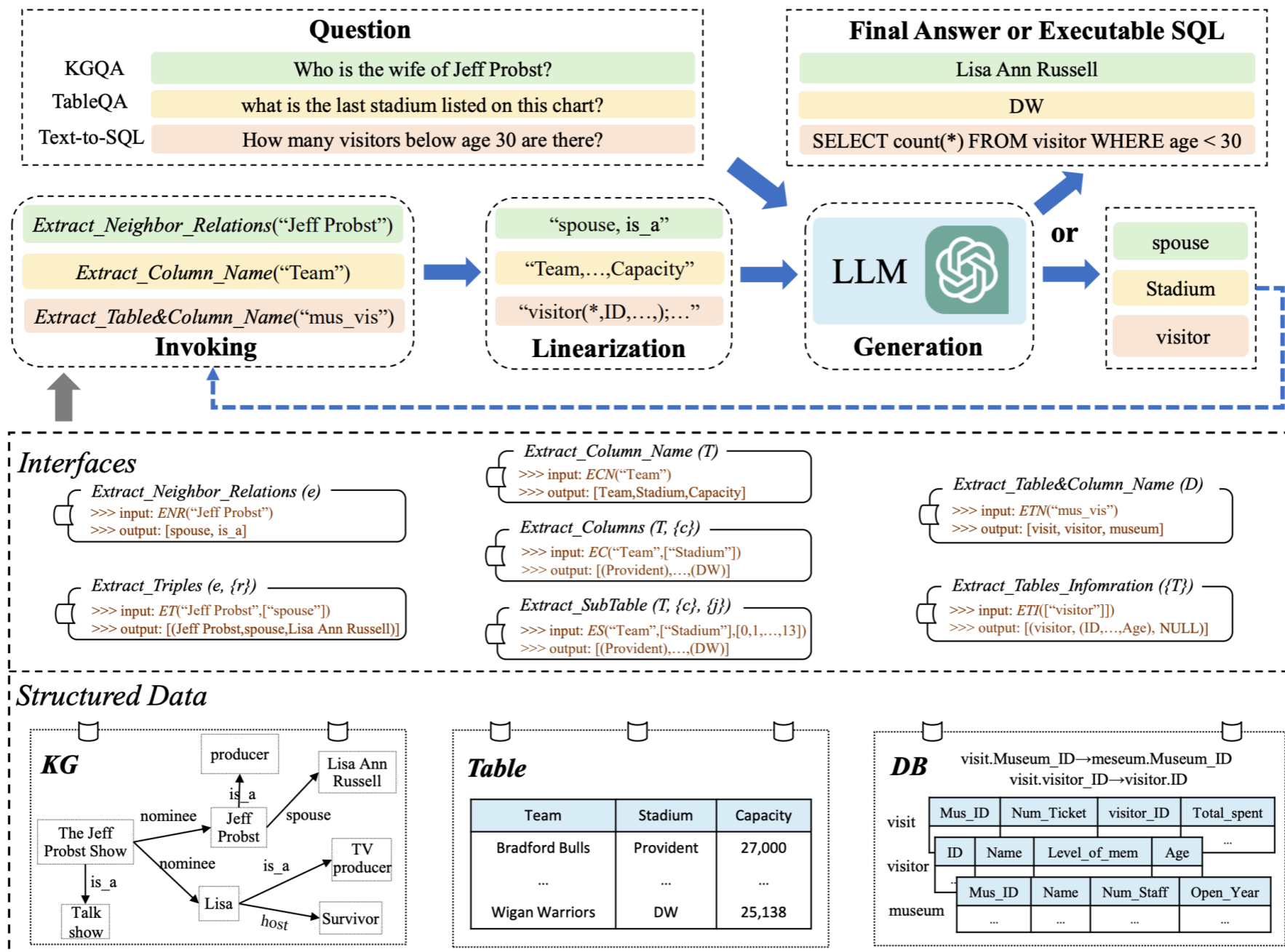
Reasoning path

[title-1] An American Werewolf in Paris [facts-1] [f1] [title-2] An American Werewolf in London [facts-2] [f2] [answer] David Naughton, Jenny Agutter and Griffin Dunne"

StructGPT

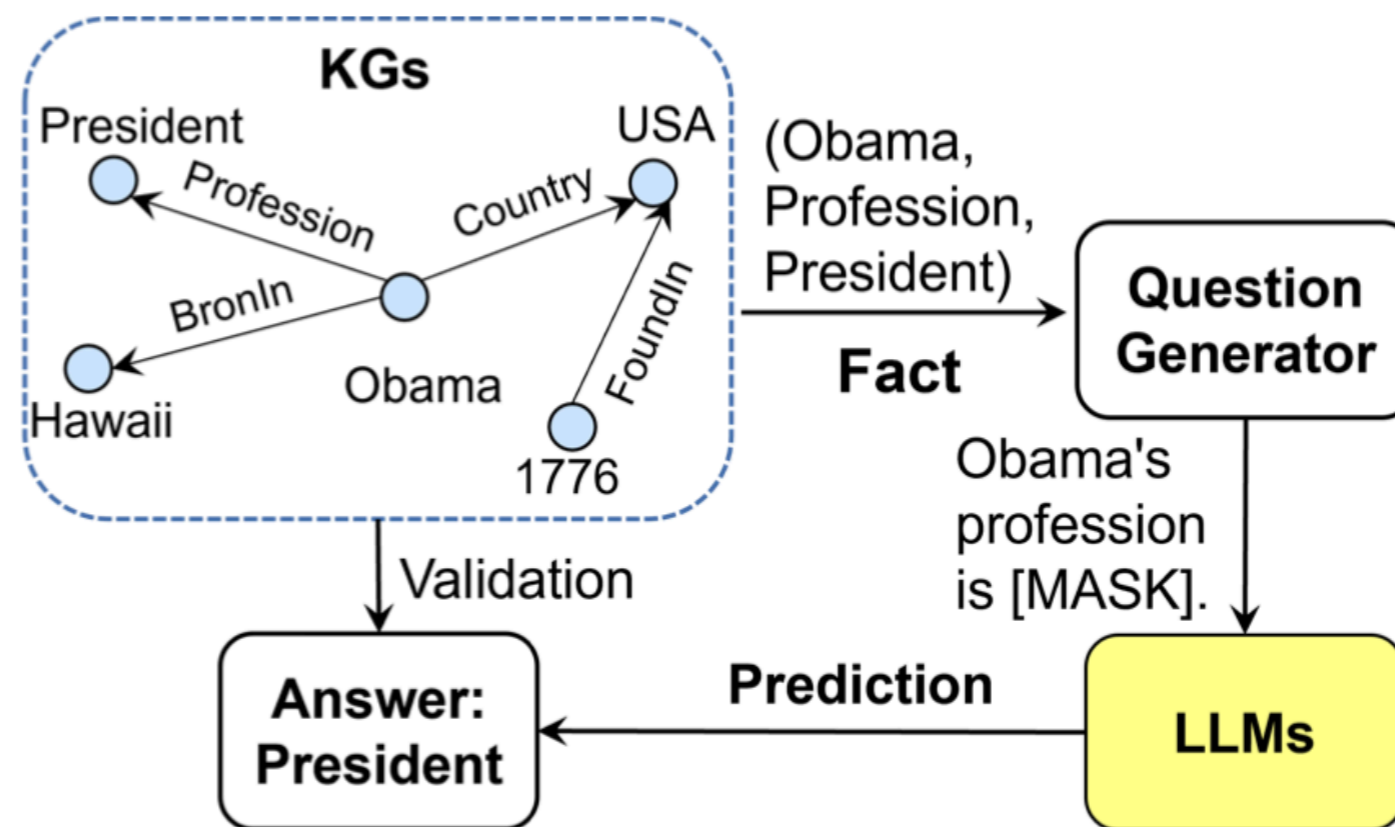
- **Prompt engineering without fine-tuning**

- Invoking: retrieve related information from structured data
- Linearization: Convert retrieved information into a text sequence
- Generate: ask LLMs to generate either an answer or an executable SQL comment



KG-enhanced LLM Interpretability

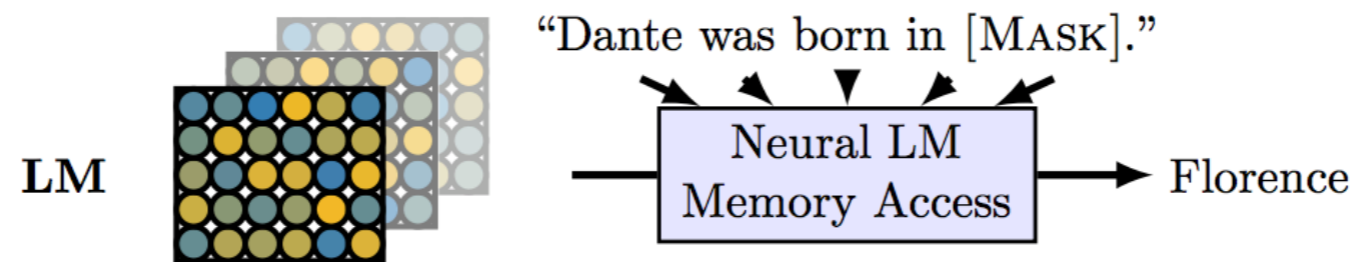
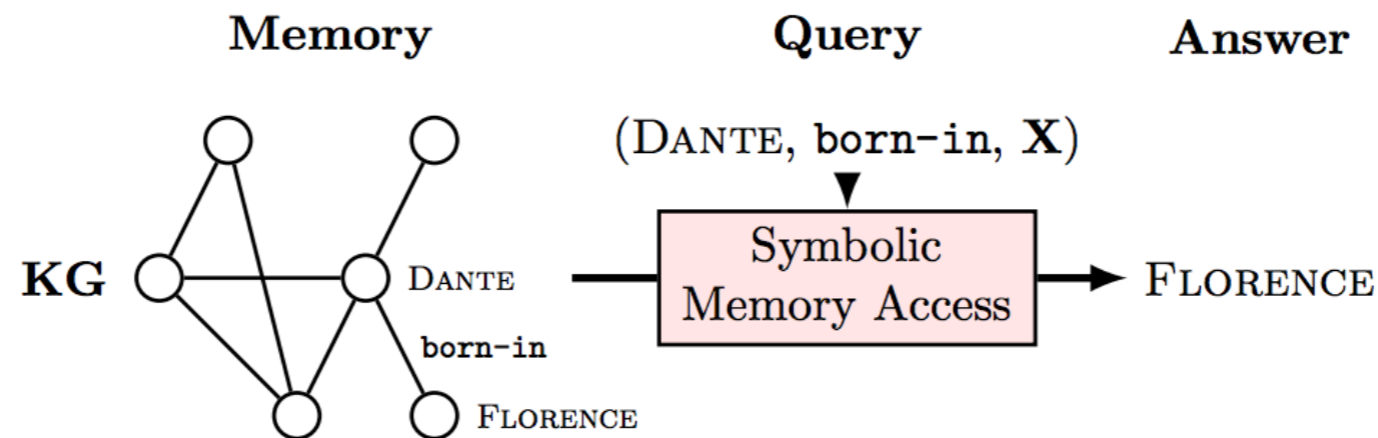
- **KGs for LLM Probing:** Use KGs to understand LLMs' embedded knowledge. LAMA is the first study, and follow-up works include LPAQA, BioLAMA, MedLAMA
 - Sample some relation triplets, and construct some simple questions with a [MASK] token
 - Ask the LLMs to predict the [MASK] token
 - Use the ground-truth relation triplets to validate the accuracy



LAMA: LMs as KBs?

(Petroni et al. 2019)

- Structured queries (e.g., SQL) to query KBs.
- Natural language prompts to query LMs.



e.g. ELMo/BERT

LMs as KBs?

(Petroni et al. 2019)

- LAMA benchmark
 - Manual prompts for 41 relations: “[X] was founded in [Y].”
 - Fill in subjects and have LMs (e.g., BERT) predict objects: “Bloomberg L.P. was founded in [MASK].”
 - Accuracy: ELMo 7.1%, Transformer-XL 18.3%, BERT-base 31.1%

Mask 1 Predictions:

5.2% **Chicago**

4.1% **London**

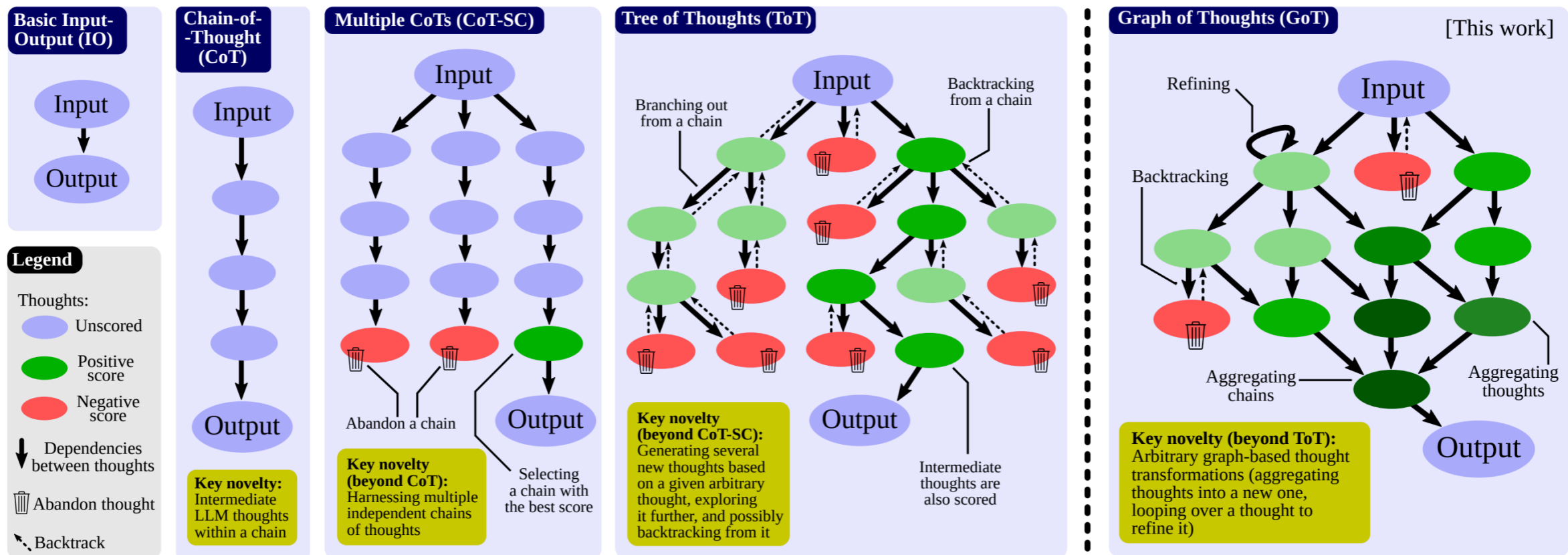
2.8% **Toronto**

2.3% **c**

1.6% **India**

Structured Prompting Strategies

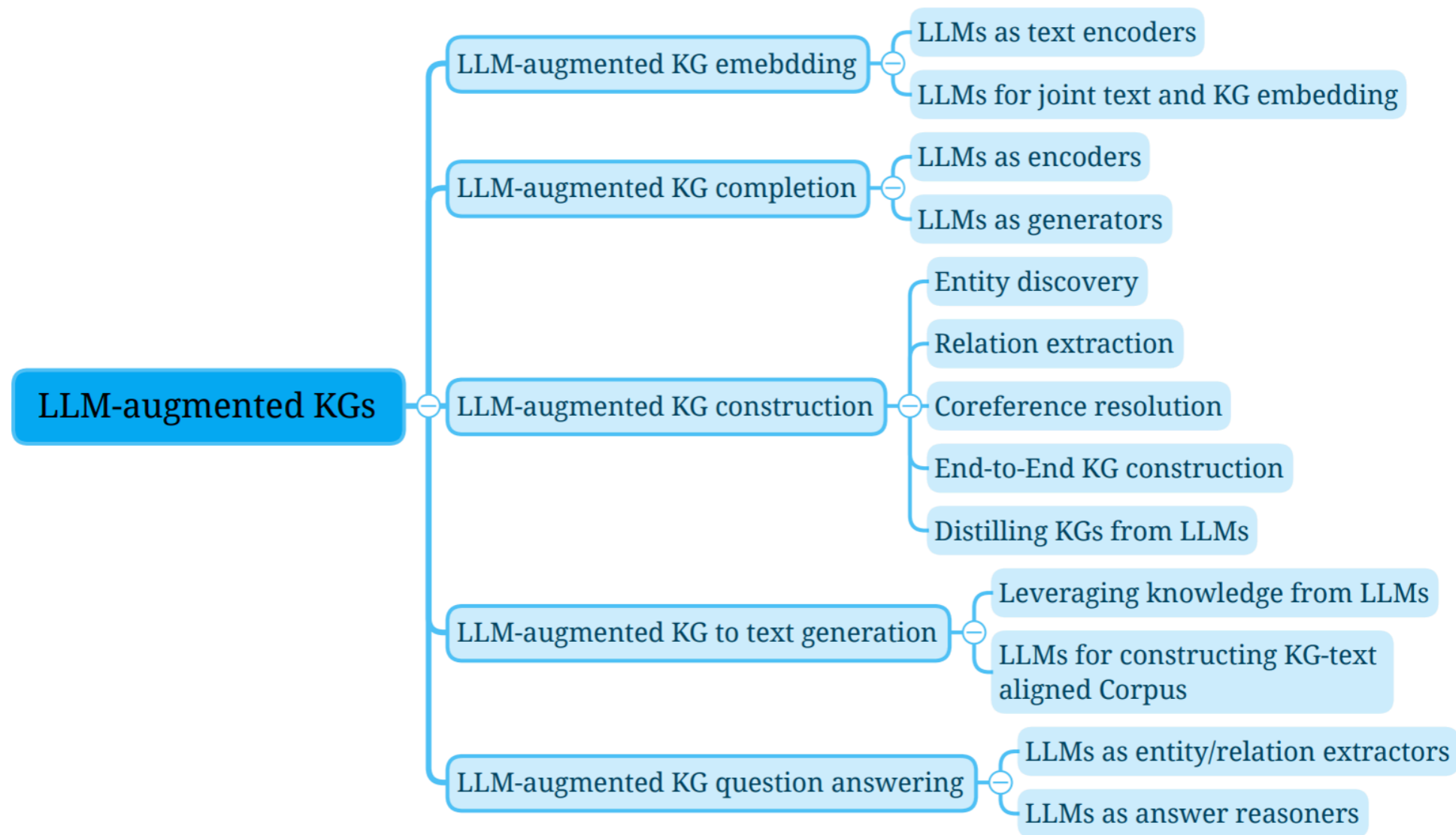
- In addition to prompting LLMs to generate an entity in a relation triplet (LAMA), recent structured prompting methods prompt the LLMs to generate a long sequence in a linear chain (Chain-of-thought), a tree (Tree-of-thought) or a graph (Graph-of-thought).
- Essentially, **probe the LLMs to recover some paths in a KG**



LLM-augmented KGs

LLM-augmented KGs

- Use LLM to improve KG tasks:

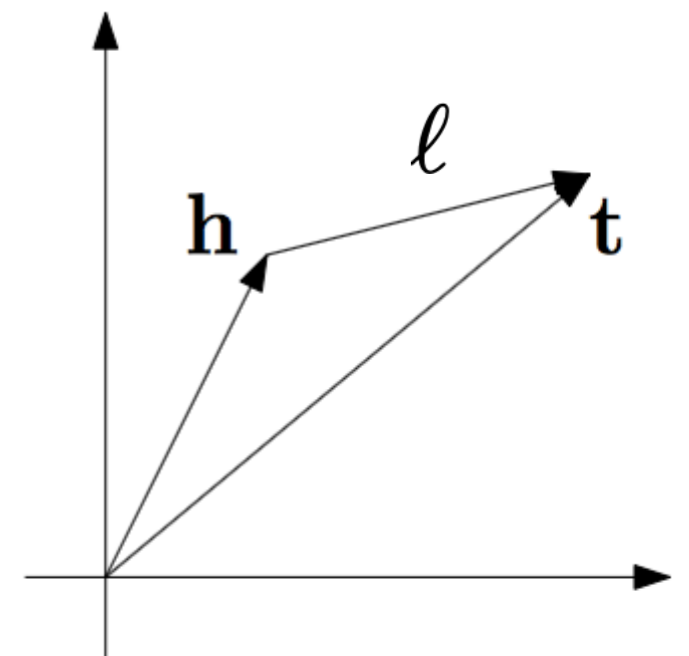


Task 1: Learning KG embeddings — before LLMs

- **Learning node/edge embeddings for a KG before LLMs**: rely more on **structural** information of KGs than **semantic** information to optimize a scoring function computed from embeddings
 - Examples: TransE, and DisMult
 - Limitations: unseen entities and long-tailed relations due to their limited structural connectivity

- **TransE**: express triples as additive transformation
- **Objective**: minimize the distance of existing triples with a margin-based loss that

$$\sum_{(h,\ell,t) \in S} \sum_{(h',\ell,t') \in S'_{(h,\ell,t)}} [\gamma + d(\mathbf{h} + \boldsymbol{\ell}, \mathbf{t}) - d(\mathbf{h}' + \boldsymbol{\ell}, \mathbf{t}')]_+$$



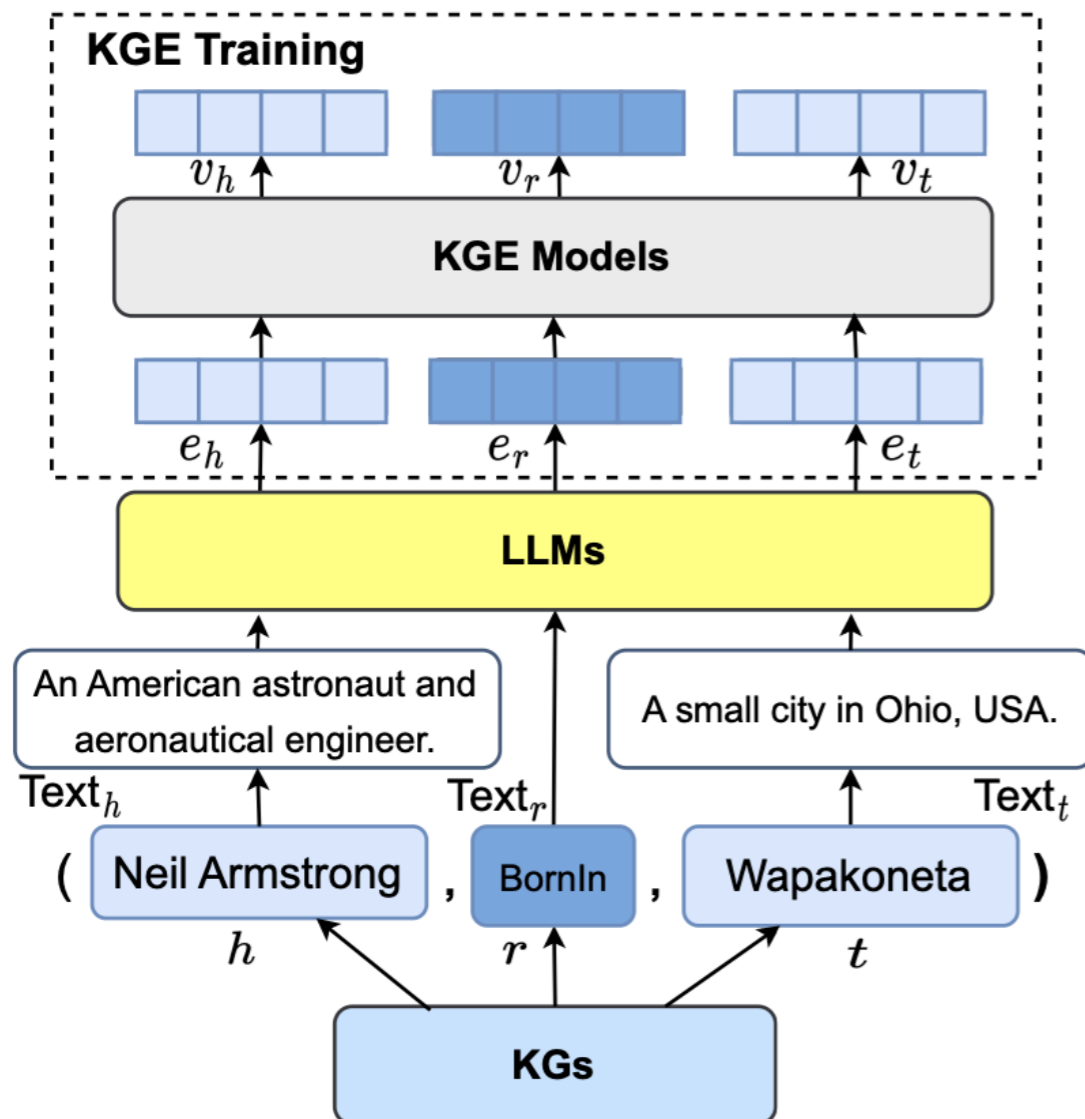
(a) TransE

- Note: one vector for each relation, additive modification only

Bordes et al. NIPS 2013. Translating Embeddings for Modeling Multi-relational Data

LLMs as encoder for KG Embeddings

- **Pretrain-KGE**: capture contextualized semantic meanings for nodes in a KG
 - Use LLMs to encode the text description of entities as the initial KG embeddings
 - Add an additional KGE model to further learn fine-grained representations.



Margin-based contrastive loss

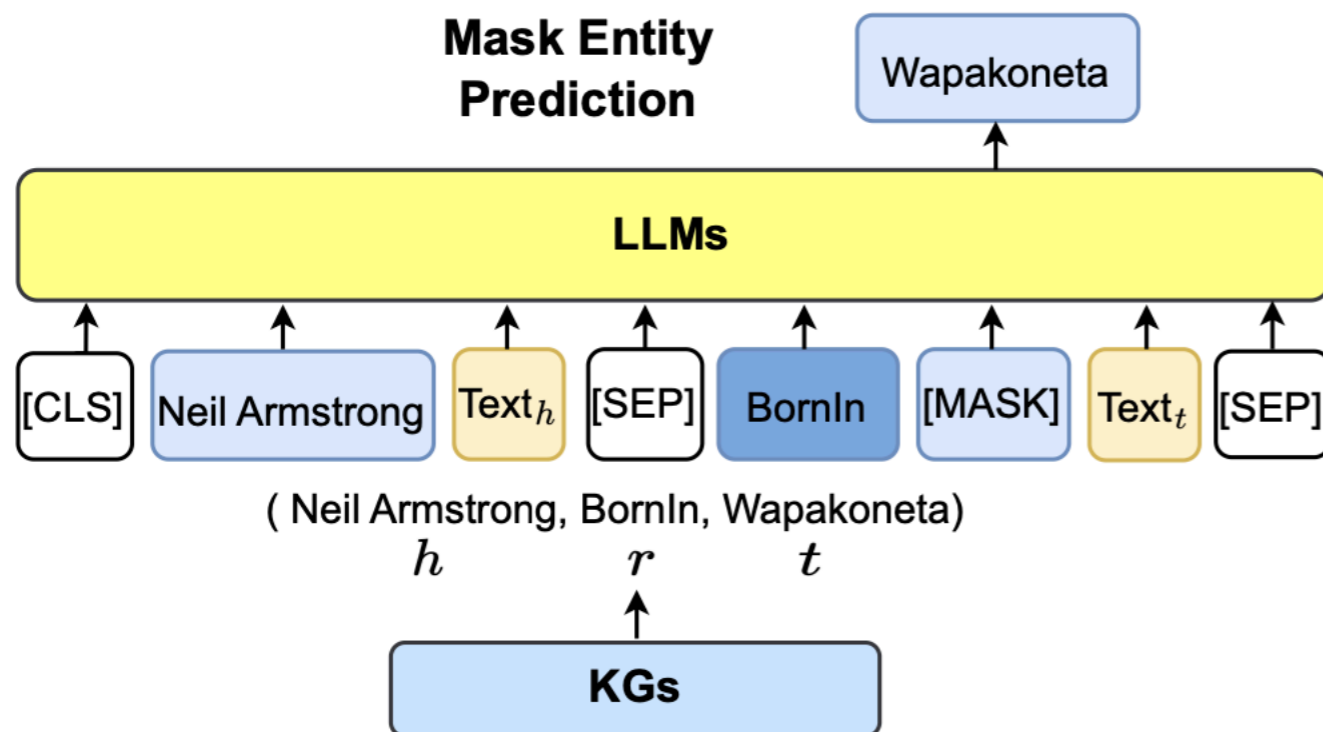
$$\mathcal{L} = [\gamma + f(v_h, v_r, v_t) - f(v'_h, v'_r, v'_t)],$$

$$e_h = \text{LLM}(\text{Text}_h), e_t = \text{LLM}(\text{Text}_t), e_r = \text{LLM}(\text{Text}_r),$$

LLMs for Joint Text and KG Embeddings

- **kNN-KGE:**

- treats the entities and relations as special tokens in the LLM.
- During training, convert each relation triplet (h, r, t) into a sentence
- Asks the model to predict the tail node given the head and relation.



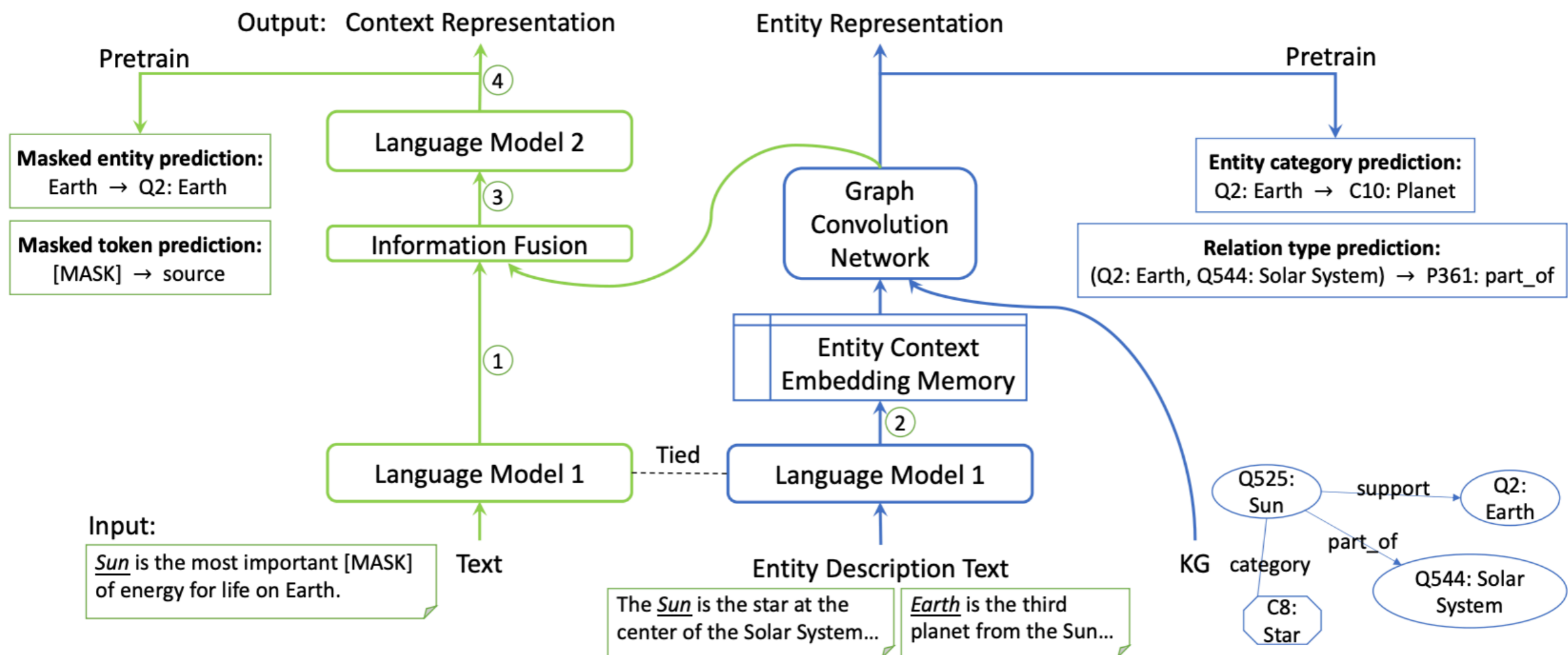
$$P_{LLM}(t|h, r) = P([\text{MASK}] = t | x, \Theta),$$

$$x = [\text{CLS}] h \text{Text}_h [\text{SEP}] r [\text{SEP}] [\text{MASK}] \text{Text}_t [\text{SEP}]$$

LLM +GNN fusion for KG Embeddings

(JAKET, Yu et al. 2022)

- Instead of using only LLM as encoder, use **GNN to capture structure information**
- Self-supervised tasks: **Entity category prediction and relation type prediction**



Yu et al. 2022. JAKET: Joint Pre-training of Knowledge Graph and Language Understanding

Task 2: KG completion

— before LLMs

- **Information extraction (IE):** extracting relation triplets from text
 - **Schema-based IE:** Pre-define a set of relations (a.k.a. schema) that we could extract for pairs of entities from text.
 - **OpenIE (schema free):** predicts an open-set of relations (mostly based on linguistic syntax)

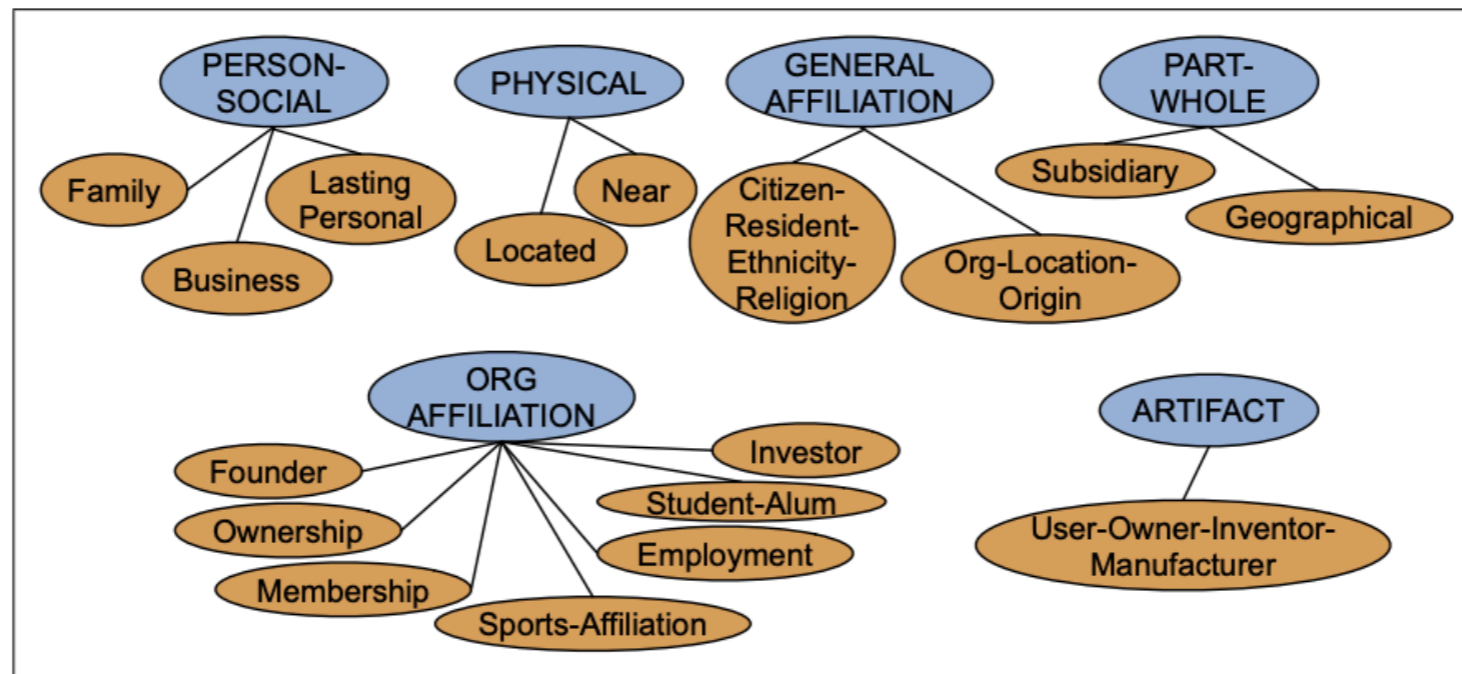


Figure 17.1 The 17 relations used in the ACE relation extraction task.

Relations	Types	Examples
Physical-Located	PER-GPE	He was in Tennessee
Part-Whole-Subsidiary	ORG-ORG	XYZ , the parent company of ABC
Person-Social-Family	PER-PER	Yoko 's husband John
Org-AFF-Founder	PER-ORG	Steve Jobs , co-founder of Apple...

Figure 17.2 Semantic relations with examples and the named entity types they involve.

Supervised Relation Extraction Baseline

- **Training:**
 - Labeled dataset: a KB triplet $t = \langle e1, r, e2 \rangle$ on a sentence s
 - Supervised training of models (e.g., logistic regression, NNs)
- **Test:**
 - Find any pairs of entities in a sentence
 - Apply the relation classifier on all entity pairs

function FINDRELATIONS(*words*) **returns** *relations*

relations \leftarrow nil

entities \leftarrow FINDENTITIES(*words*)

forall entity pairs $\langle e1, e2 \rangle$ **in** *entities* **do**

if RELATED?(*e1*, *e2*)

relations \leftarrow *relations* + CLASSIFYRELATION(*e1*, *e2*)

Distant Supervision for Relation Extraction (Mintz et al. 2009)

- **Motivation:** Supervised baseline is still limited to the labeled data size.
- Given an KB triplet $t = \langle e_1, r, e_2 \rangle$, extract all sentences (s_1, s_2, \dots, s_N) that matches these two entities $\langle e_1, e_2 \rangle$, and use these texts to train the relation classifier

S_1 *[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story.*

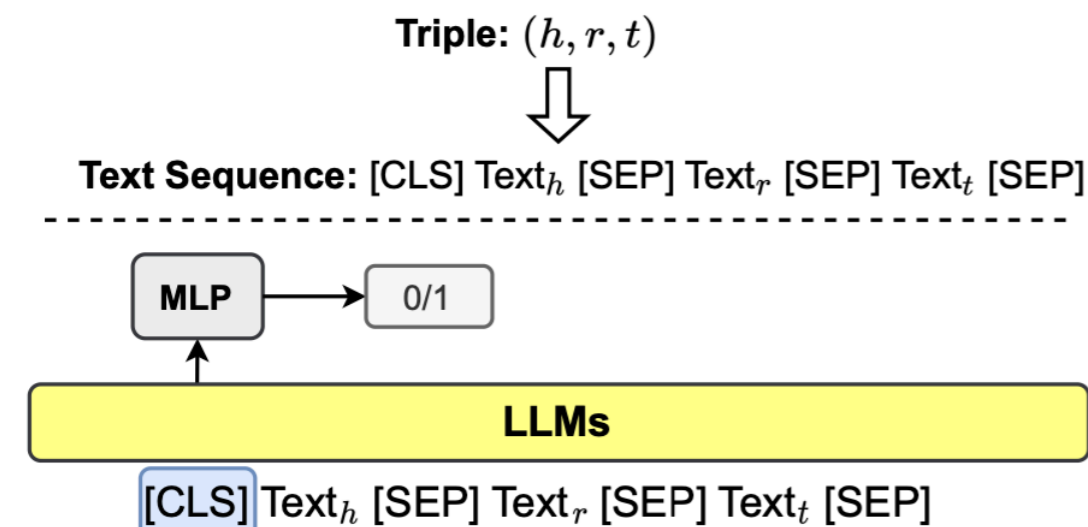
S_2 *Allison co-produced the Academy Award-winning [Saving Private Ryan], directed by [Steven Spielberg]...*

$$\text{classifier}(s_i, e_1, e_2) \rightarrow r, \quad \forall e_1 \in s_i \wedge e_2 \in s_i$$

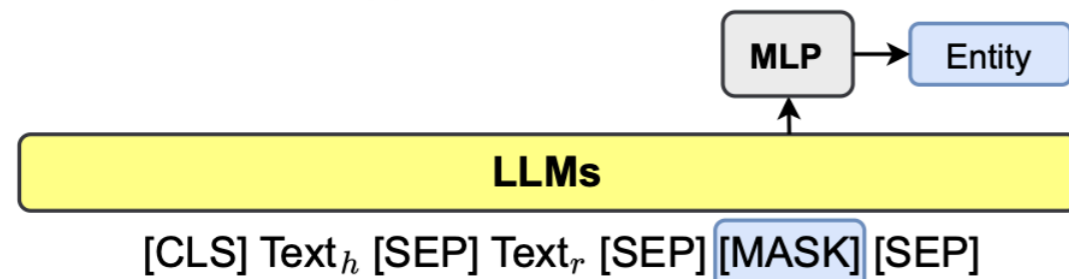
LLMs for KG Completion

- **Fine-tune** LLMs on fewer labeled data to predict more relation triplets

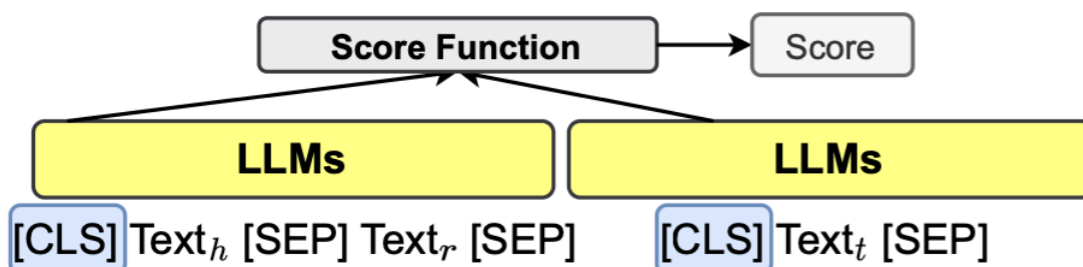
- **Prompt** LLMs w/o updating the model to predict relation triplets



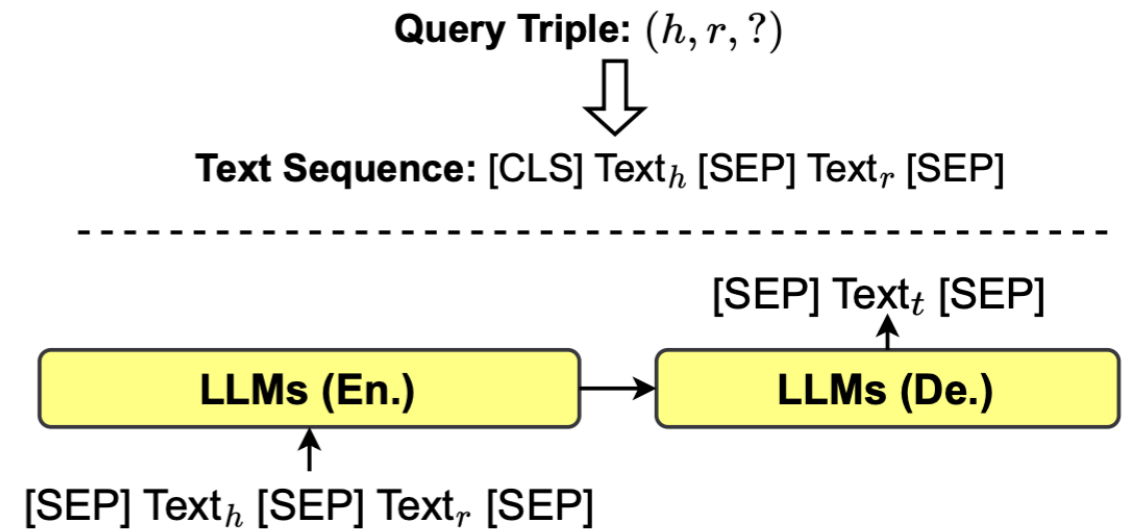
(a) Joint Encoding



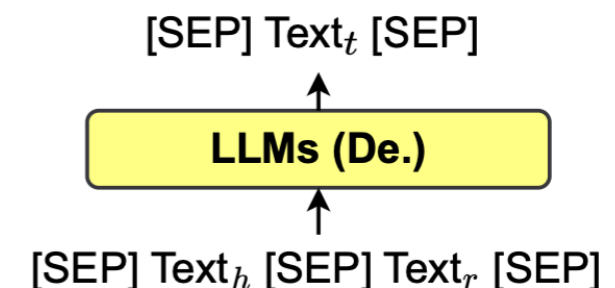
(b) MLM Encoding



(c) Separated Encoding



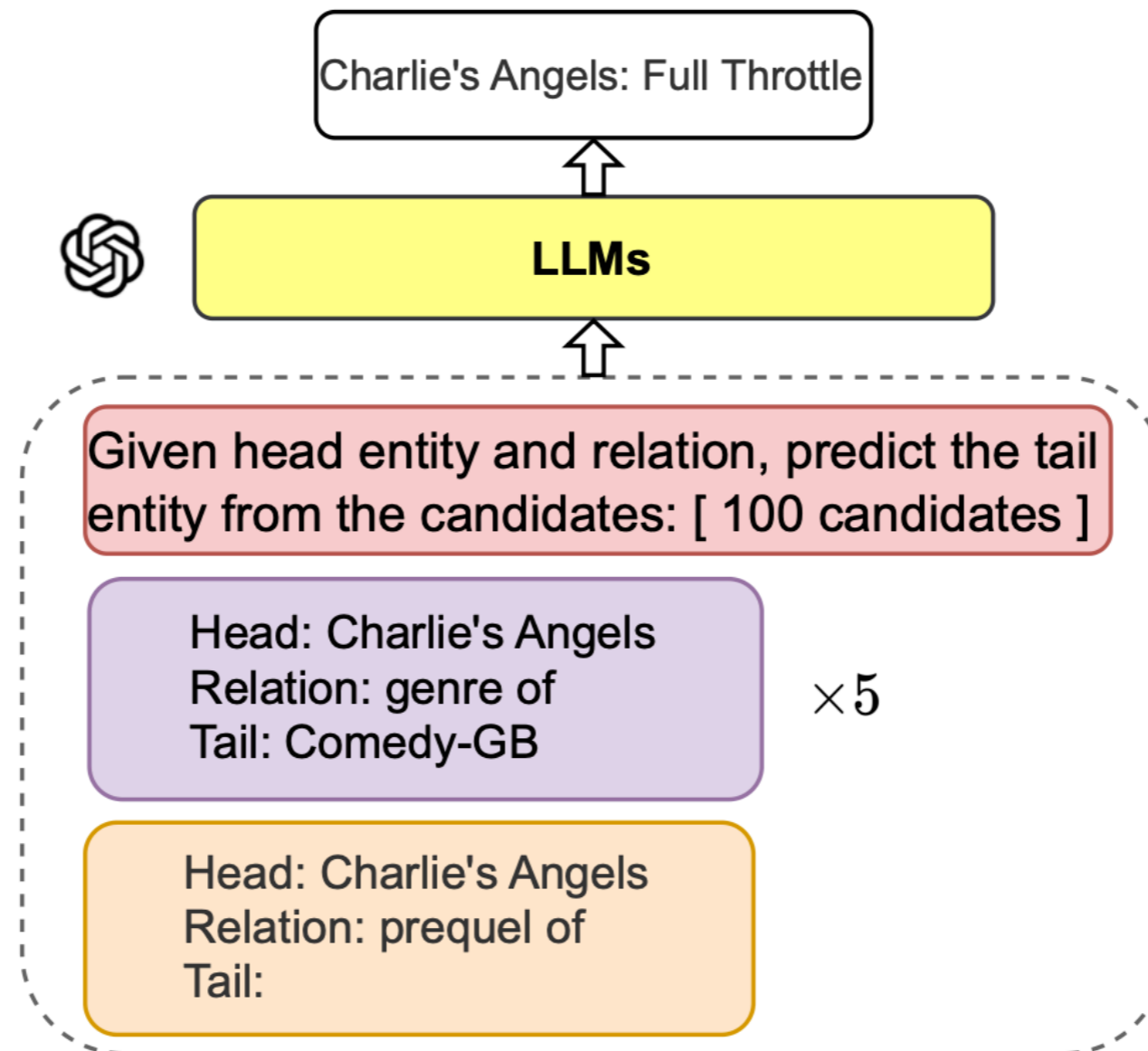
(a) Encoder-Decoder PaG



(a) Decoder-Only PaG

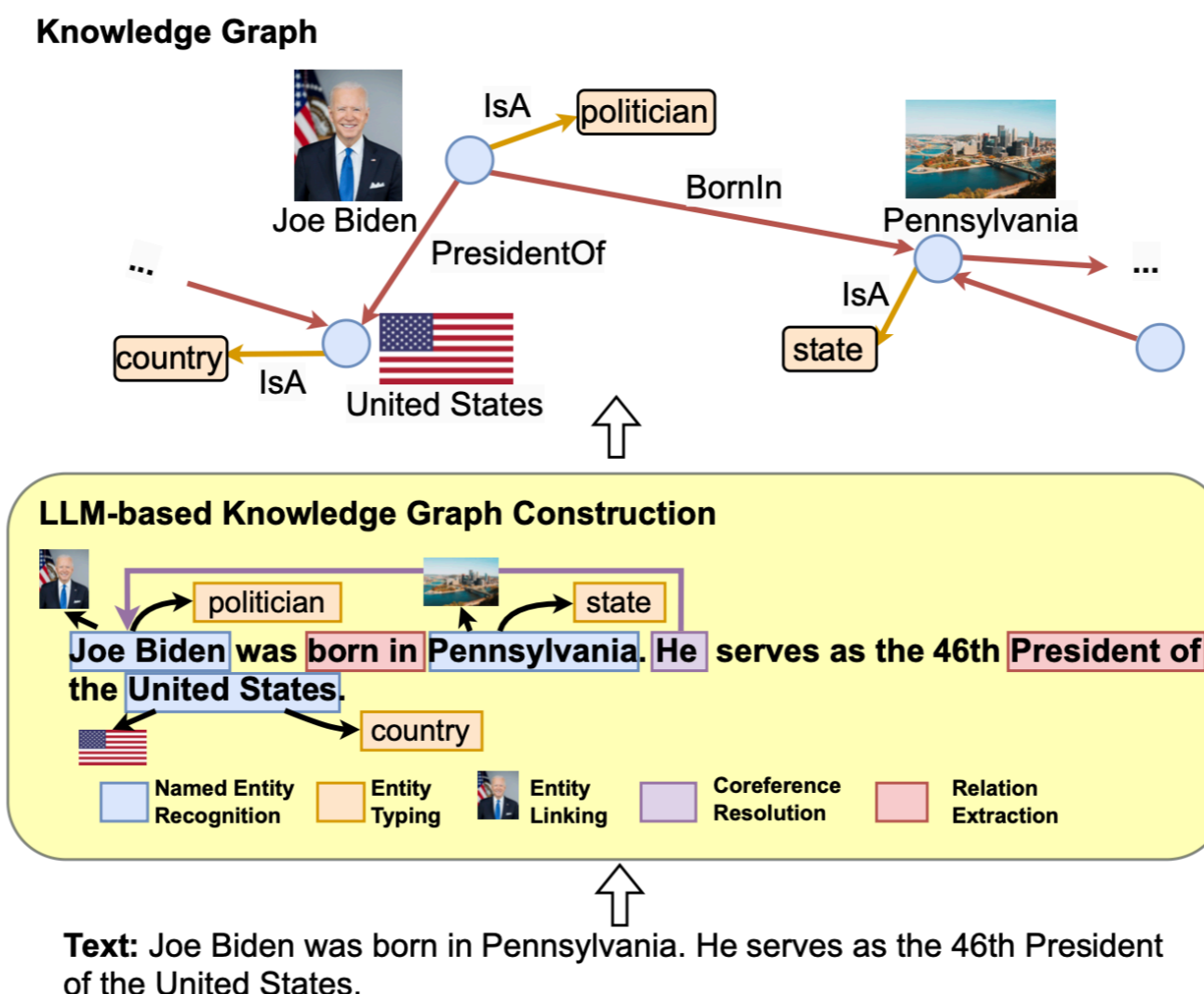
LLMs for KG Completion

- Given an existing KG, predict **missing relation triplets from more text data**
- **In-context prediction for a head entity:** Given a head node, add its 5 relation triplets in the prompt, and ask LLMs to predict the tail entity from the 100 candidates



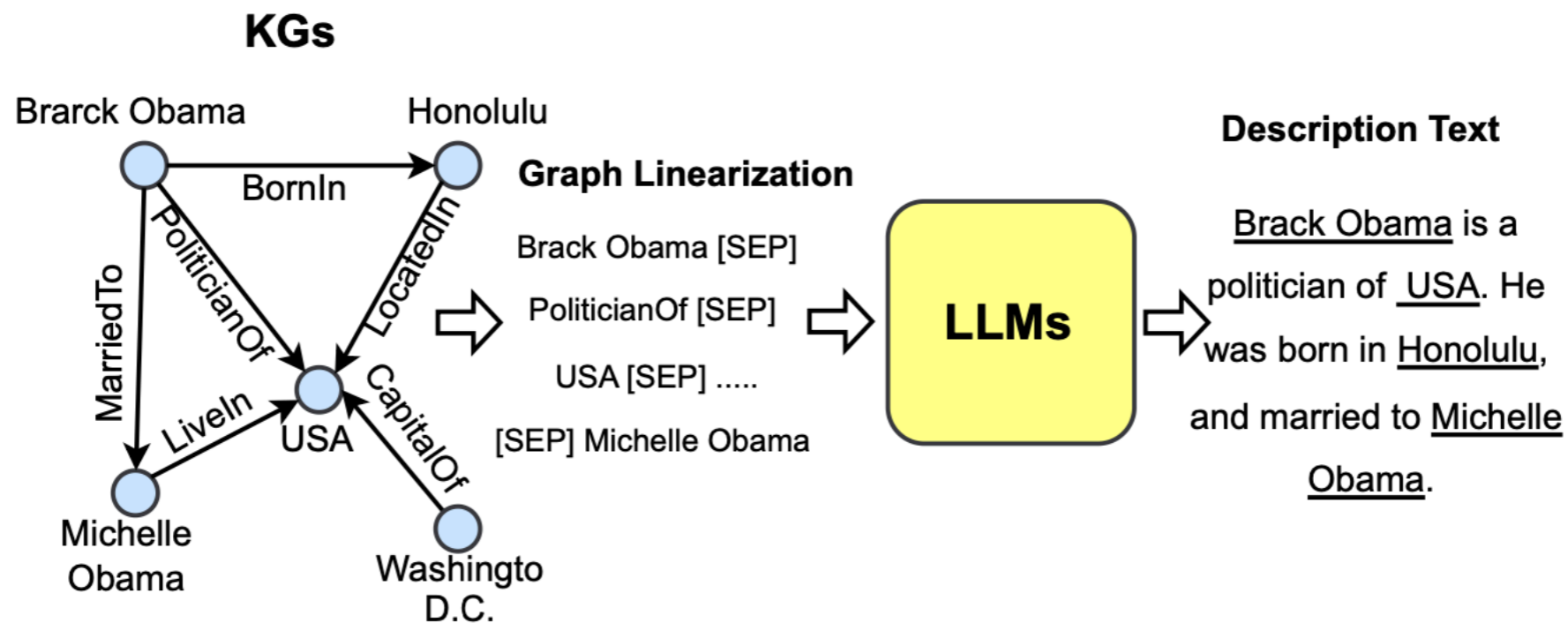
Task 3: LLMs for KG Construction

- Construct a KG from texts by a LLM from scratch
- **Predict all relation triplets in a sentence:** Prompt the LLM to predict a KG directly, which involves multiple tasks such as NER, Entity linking/typing, Coreference resolution, Relation prediction.
- **Before LLMs**, we build **a pipeline of multiple models** to construct KGs.



Task 4: LLM-augmented KG-to-text Generation

- Goal: generate high-quality texts that describe the input KG.
- Applications: include storytelling, KG-grounded dialogue.
- Challenge: collect large graph-text parallel data. So many studies use weakly-aligned graph-text pairs



Task 5: LLM-augmented KG for Question Answering

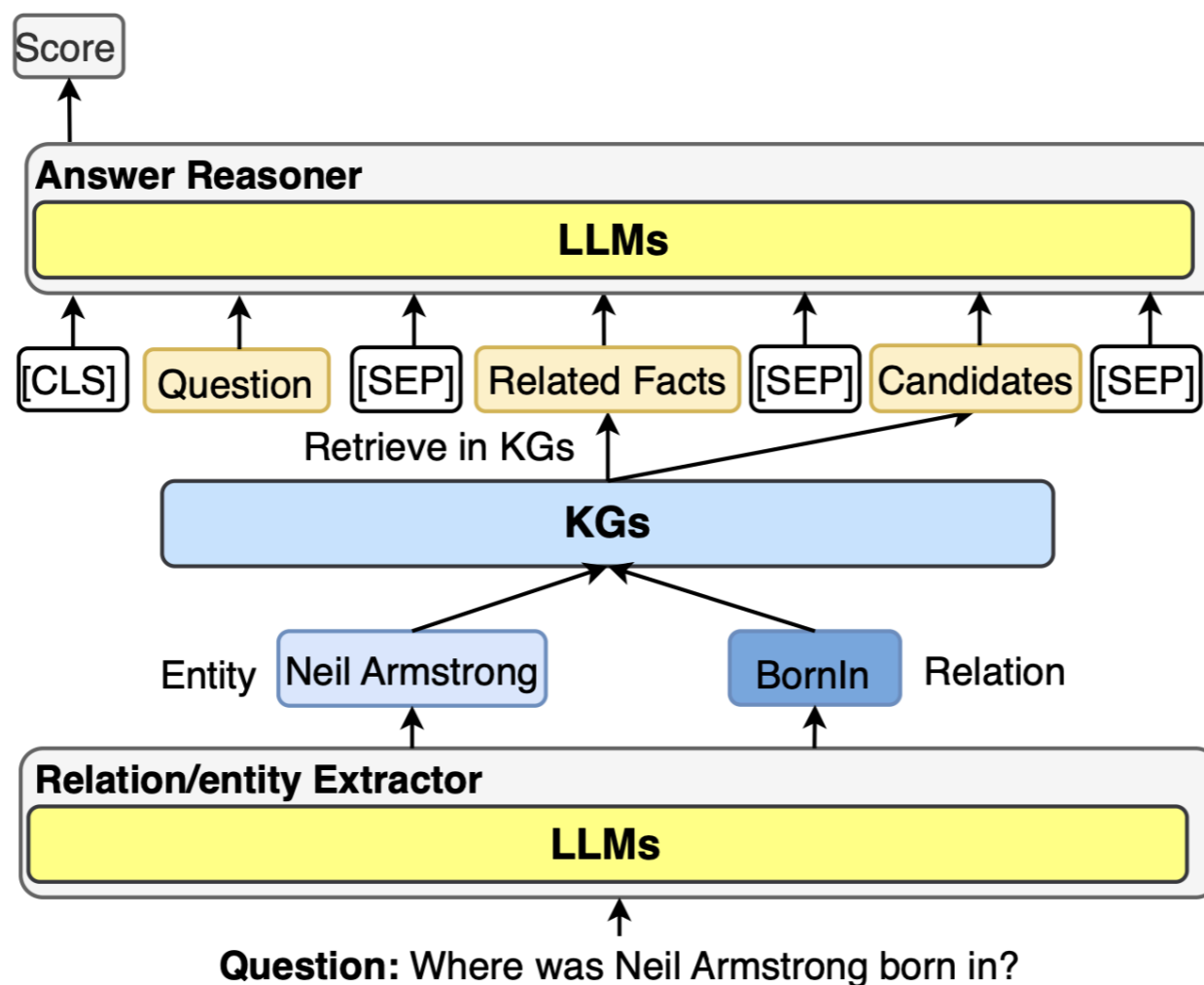
- **KGQA**: find answers to a question based on structured facts stored in KGs. The key challenge is to retrieve related facts and predict the answer based on the retrieved facts.
 - **Symbolic method (semantic parsing)**: convert a natural language query into a structured format (e.g., SQL, Python) to query the KG
 - **Neural symbolic method**: embed a natural language query and the knowledge base information into an embedding space, learn some integration modules to combine information, and make answer prediction

Knowledge Base Question Answering (KBQA)

- Construct a KB from texts or other resources either manually or automatically
- **Symbolic method (semantic parsing)**: convert a natural language query into a structured format (e.g., SQL) to query the KB
- **Neural symbolic method**: embed a natural language query and the knowledge base information into an embedding space, learn some integration modules to combine information, and make answer prediction

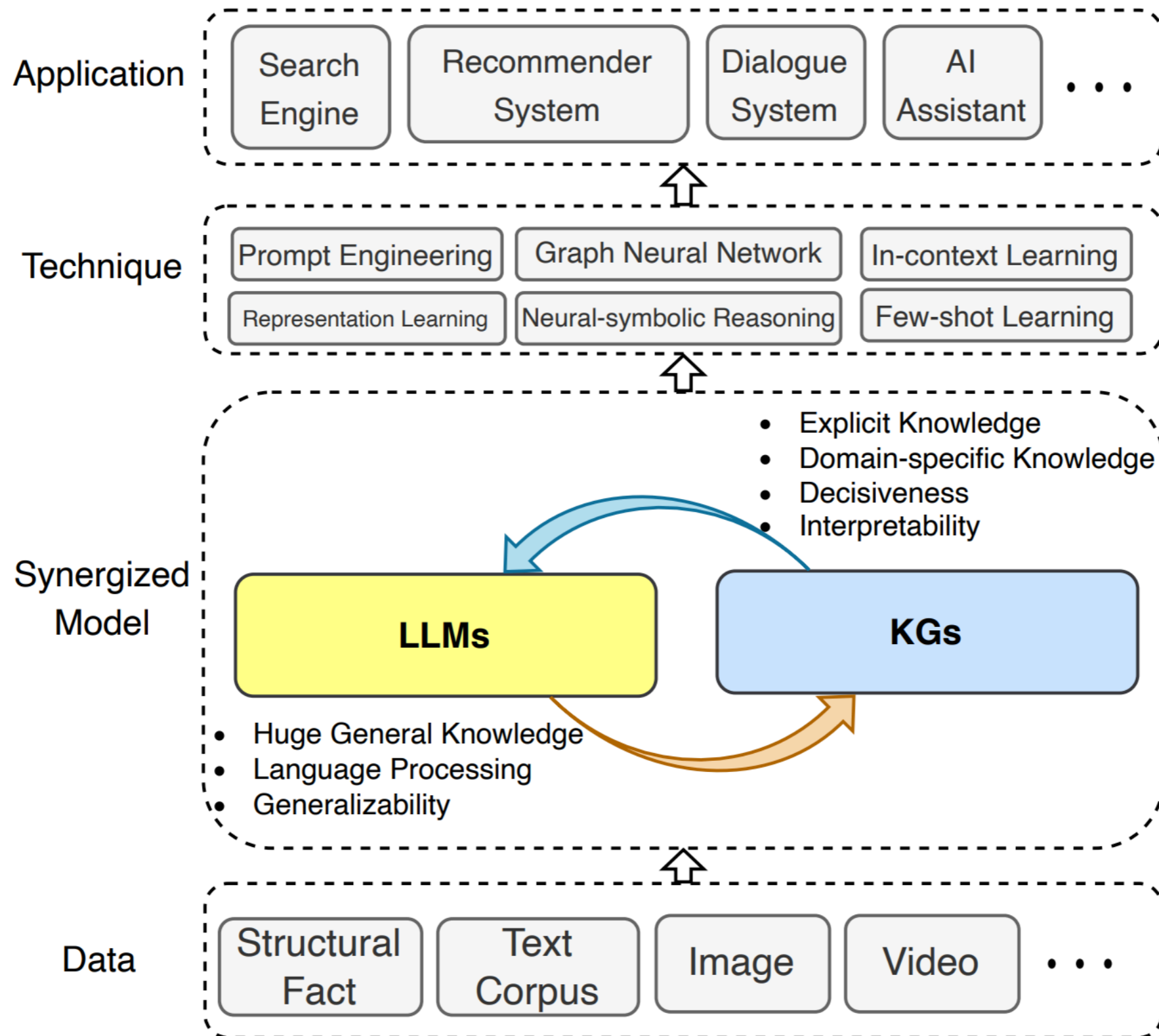
LLM-augmented KG for Question Answering

- **DEKCOR**: Use LLMs in two ways
 - Relation extractor: extract a head entity and a relation, and use them to retrieve related facts from KGs
 - Answer reasoner: concatenate a question with all related KG facts and a candidate answer as a sequence input to the LLM that outputs a score for this answer.
- **Limitation**: this assumes that we know all candidate answers beforehand, and also requires testing all candidate answers

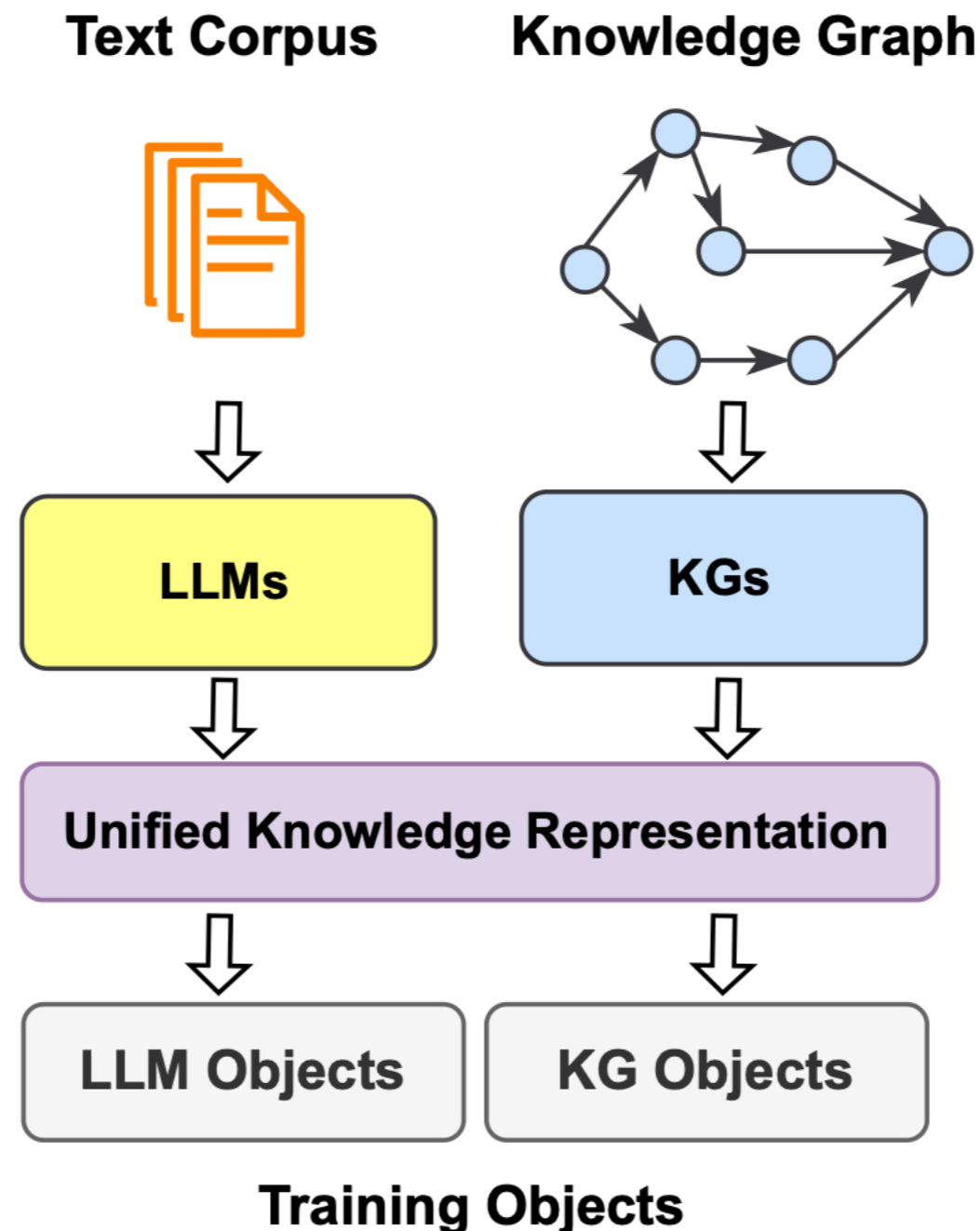


Unifying LLMs and KGs: Future directions

Synergized LLMs + KGs



Knowledge Representation of both unstructured and structured data

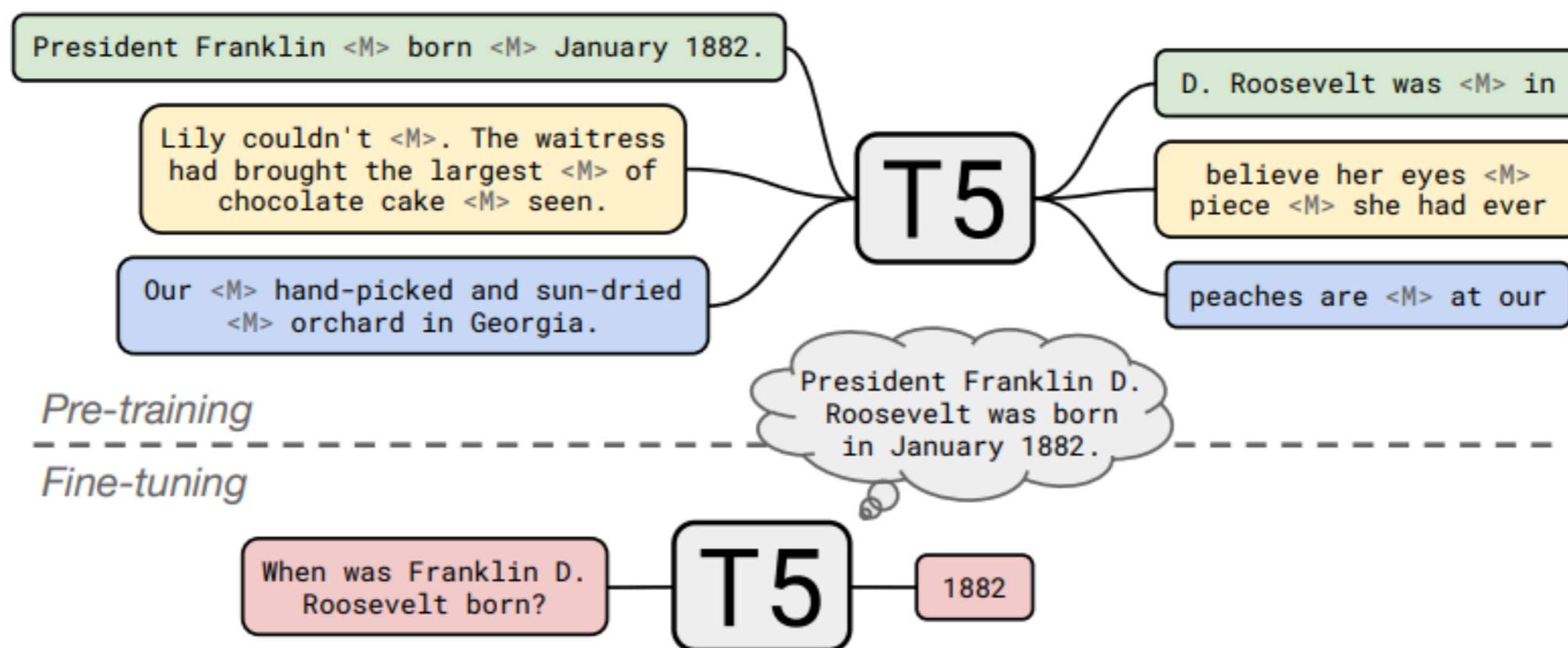


Comparison

- KBQA
 - Low coverage of knowledge
 - Faithful and interpretable
 - Dense structured
- TextQA
 - Wide coverage of knowledge
 - Misinformation
 - Massive raw texts
 - Enhanced with a text retrieval model
- LM-QA
 - Wide coverage of knowledge
 - Misinformation & out-dated information
 - Large model size
 - Black-box, not controllable

We need LLMs as backbone

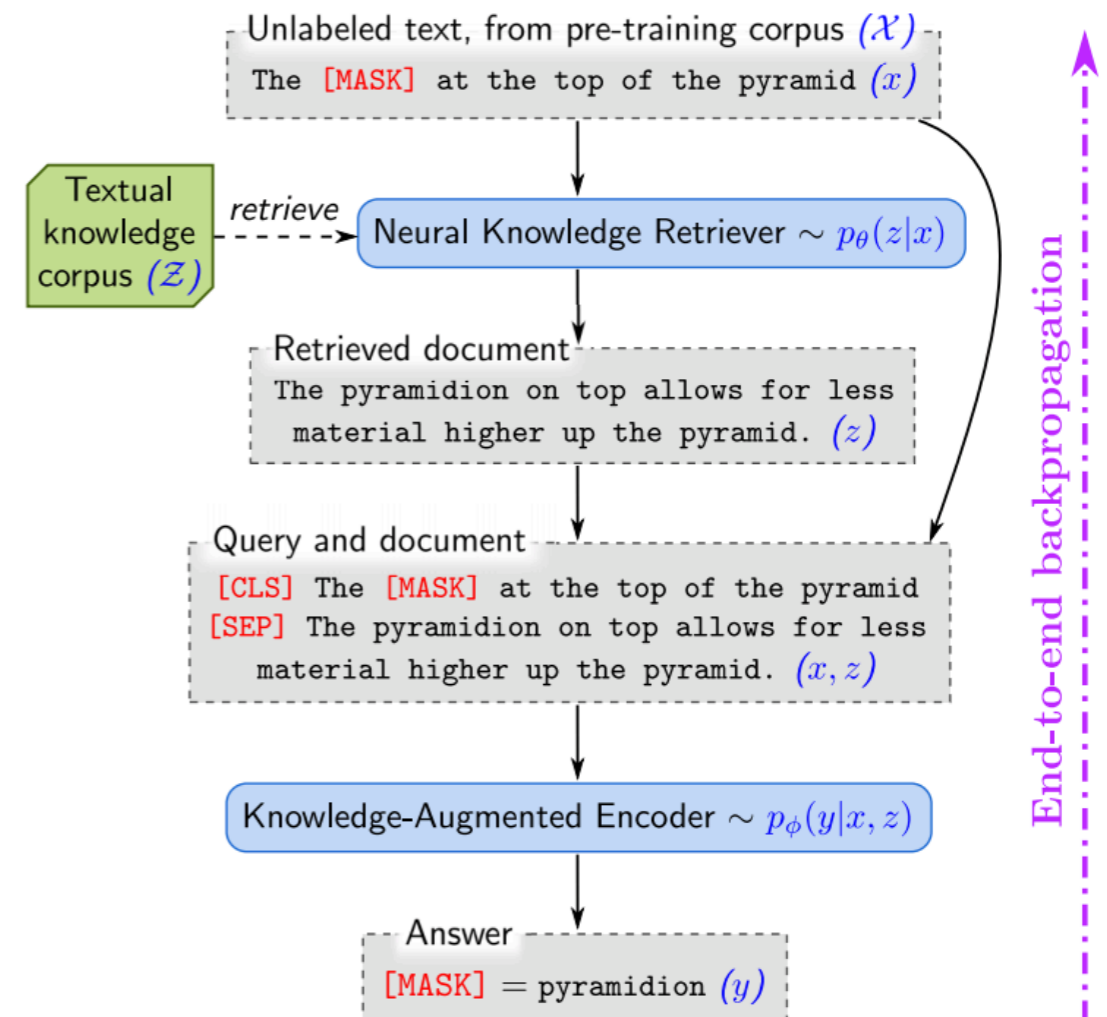
- Pre-trained LLMs can generalize to understand even unseen questions/tasks.



We need a retrieval component!

- For knowledge-intensive tasks like QA, nonparametric models (w/ retrieved context) outperform parametric models (w/o context) by a large margin.
- For example, REALM (Guu et al. 2020), RAG (Lewis et al. 2020) on the NaturalQuestion datasets.

Close-book T5	34.5
REALM	40.4
RAG	44.5



Future research directions in the era of LLMs

- **Representations:** How to encode structure data for LLMs?
 - Sequentialize KGs to text
 - Fuse KG embeddings to text embeddings
- **Decoding:** How to decode data that aligns with KGs?
 - Output a text sequence that follows a path in KG
 - Reuse LLM to generate KG directly
- **Alignment & Hallucination:** How to align LLM outputs to KG paths and detect hallucinations?
 - Add a validation step to check the correctness of LLM outputs against a KG
- **Multimodality:** How to encode multimodality data (images/videos) from KGs?
 - Align multimodal data representations into a shared embedding space
- **Interpretability:** How to edit (add/update/delete) knowledge in LLMs with KGs?
 - Understand what knowledge from KGs is stored in LLMs, and what are still missing or inconsistent.

Questions?