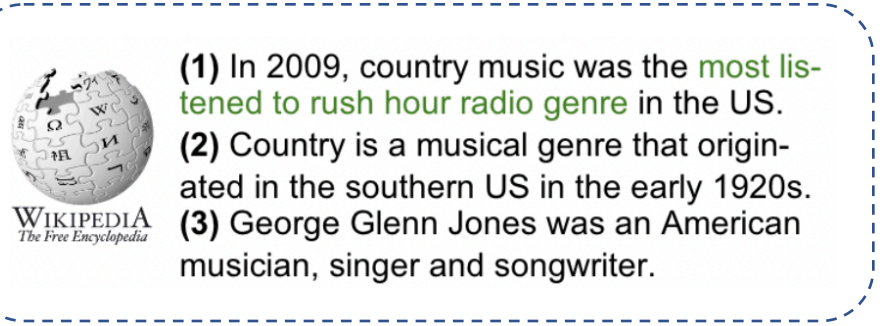
Large language models pretrained over massive text corpora： Wikipedia, Bookcorpus,...



Large language models pretrained over massive text corpora： Wikipedia, Bookcorpus,...

(1) In 2009, country music was the most listened to rush hour radio genre in the US.
(2) Country is a musical genre that originated in the southern US in the early 1920s.
(3) George Glenn Jones was an American musician, singer and songwriter.

Certain knowledge **implicitly** stored in their parameters

- requires ever-larger networks to cover more facts
- pose challenges for dynamically updating learned knowledge

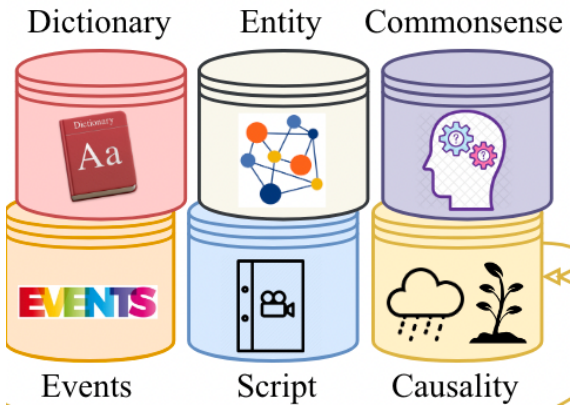
Large language models pretrained over massive text corpora： Wikipedia, Bookcorpus,...

(1) In 2009, country music was the most listened to rush hour radio genre in the US.
(2) Country is a musical genre that originated in the southern US in the early 1920s.
(3) George Glenn Jones was an American musician, singer and songwriter.

Certain knowledge **implicitly** stored in their parameters

- requires ever-larger networks to cover more facts
- pose challenges for dynamically updating learned knowledge

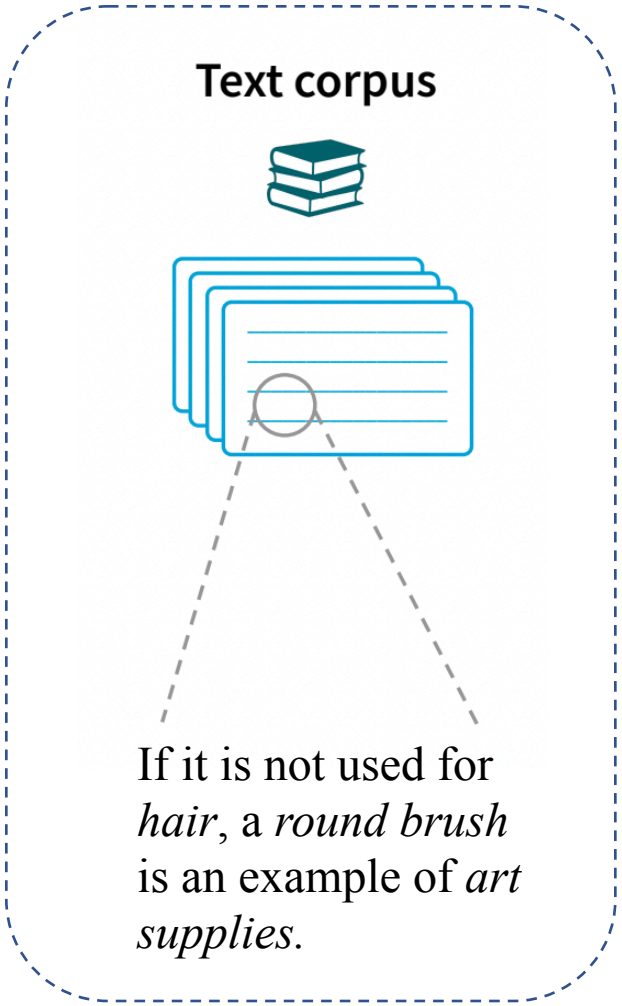
Leverage explicit knowledge



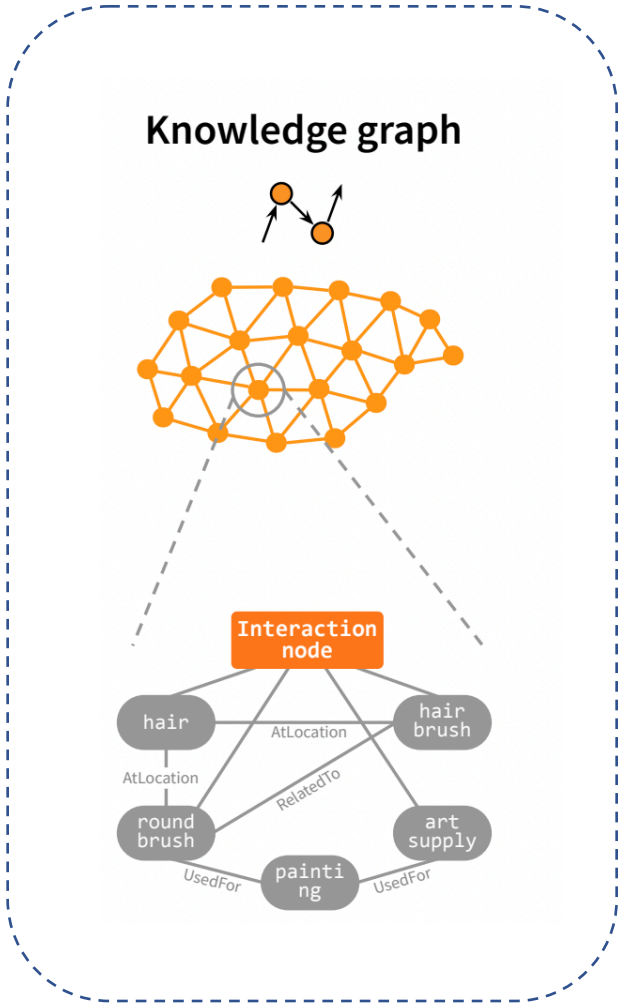
Deep Bidirectional Language-Knowledge Graph Pretraining

Michihiro Yasunaga[1] **Antoine Bosselut**[2] **Hongyu Ren**[1] **Xikun Zhang**[1]
Christopher D Manning[1] **Percy Liang**[1*] **Jure Leskovec**[1*]
[1]Stanford University [2]EPFL [*]Equal senior authorship
{myasu,antoineb,hyren,xikunz2,manning,pliang,jure}@cs.stanford.edu

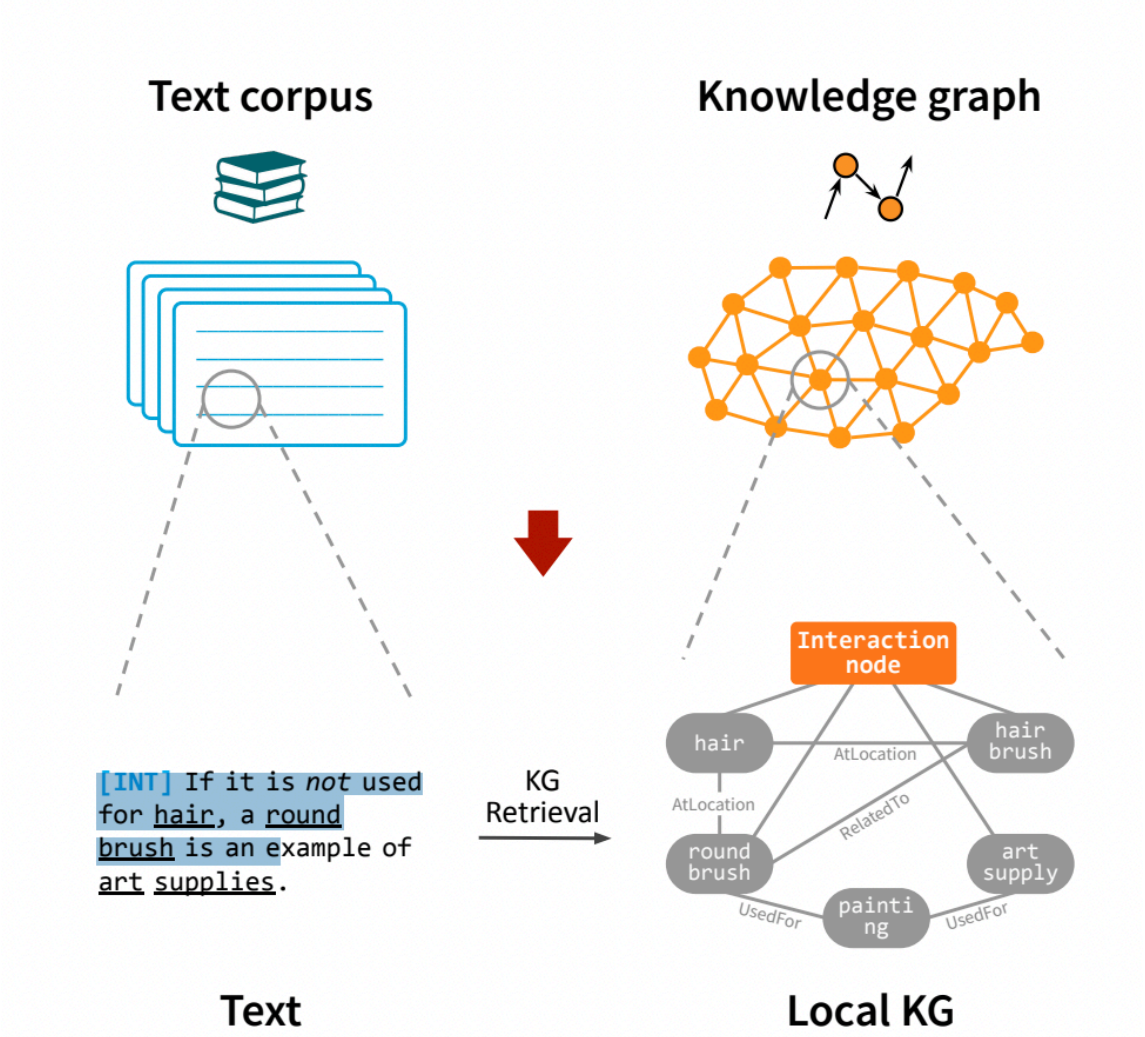
36th Conference on Neural Information Processing Systems (NeurIPS 2022)



contextualized representations of tokens



Complement structured information to text data



Combining text corpus and KG as input data

How to fetch relevant external knowledge?

If it is not used for hair, a round brush is an example of art supplies.

Entity Linking

If it is not used for *hair*, a *round brush* is an example of *art supplies*.

$$V_{el} = \{\ldots, hair, round\ brush, art\ supplies, \ldots\}$$

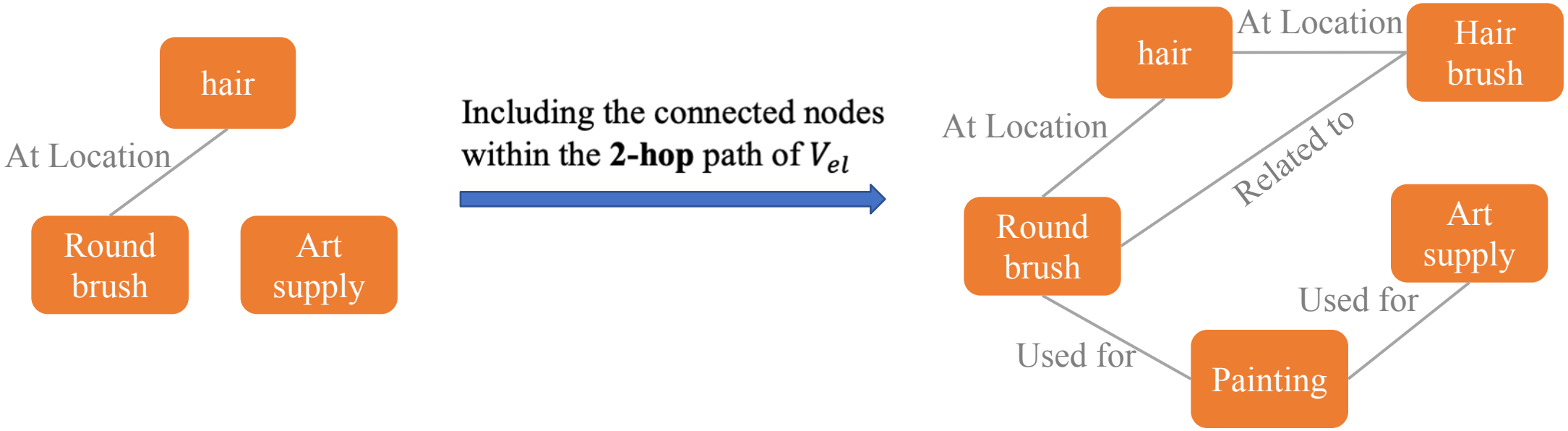
How to fetch relevant external knowledge?

If it is not used for hair, a round brush is an example of art supplies.

Entity Linking

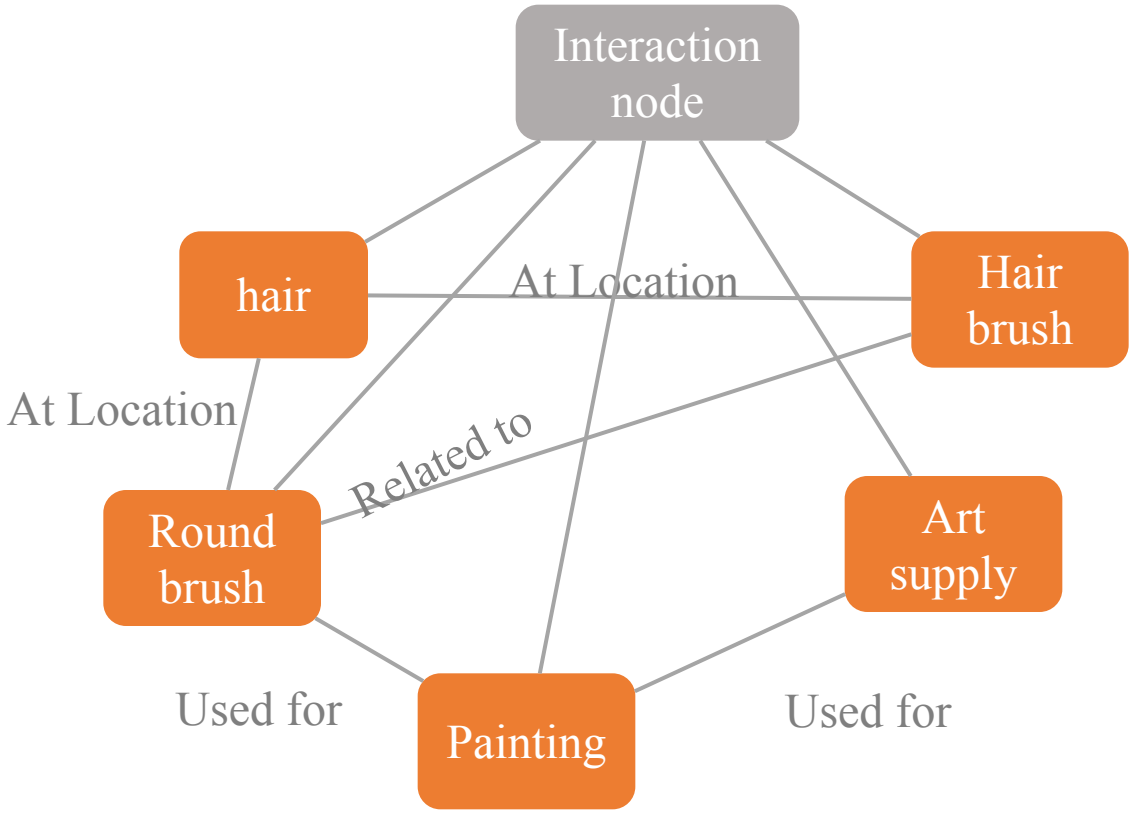
If it is not used for *hair*, a *round brush* is an example of *art supplies*.

$$V_{el} = \{\ldots, hair, round\ brush, art\ supplies, \ldots\}$$

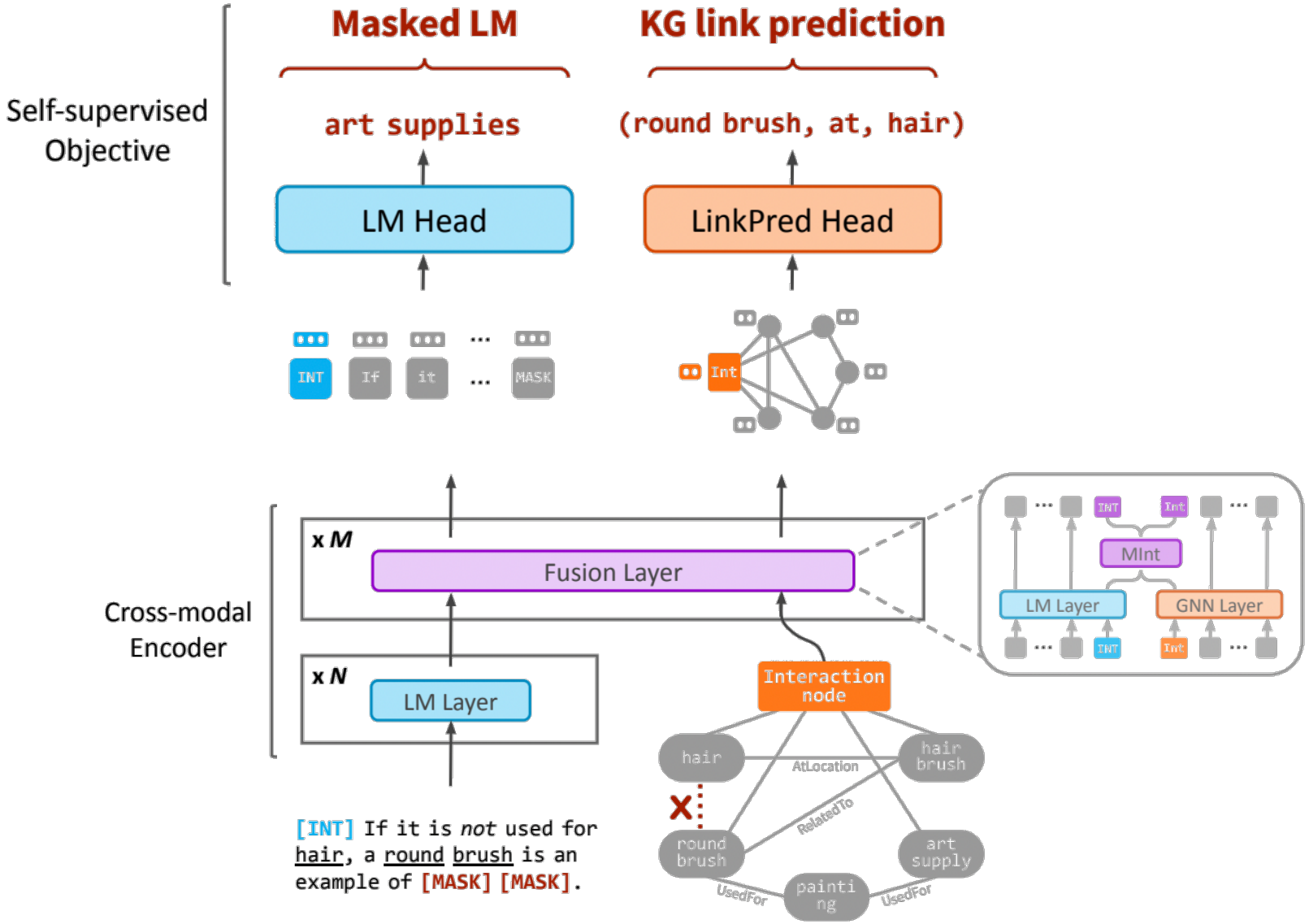


How to fetch relevant external knowledge?

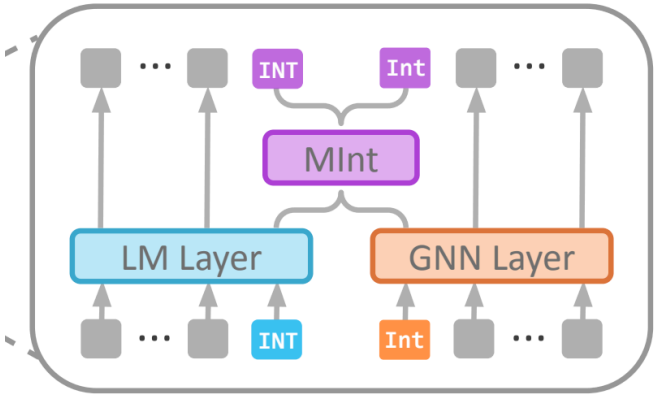
[INT] If it is not used for *hair*, a *round brush* is an example of *art supplies*.



(text, local KG)

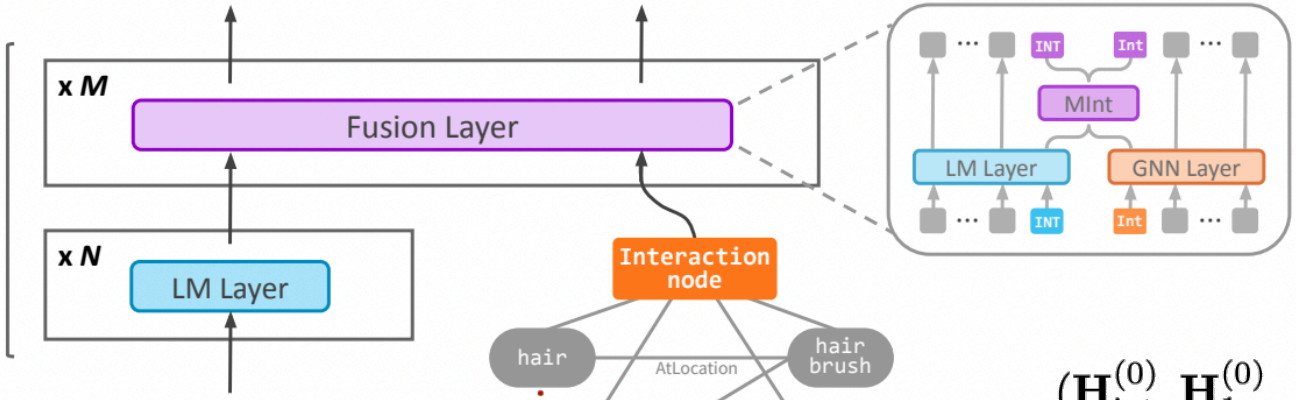

Self-supervised Objective
Masked LM
KG link prediction
art supplies
(round brush, at, hair)
LM Head
LinkPred Head
INT
If
it
MASK
Int
Cross-modal Encoder
x M
Fusion Layer
x N
LM Layer
MInt
LM Layer
GNN Layer
Interaction node
hair
hair brush
AtLocation
RelatedTo
round brush
art supply
painting
UsedFor
UsedFor
[INT] If it is not used for hair, a round brush is an example of [MASK] [MASK].

Knowledge Fusion during feed forward process

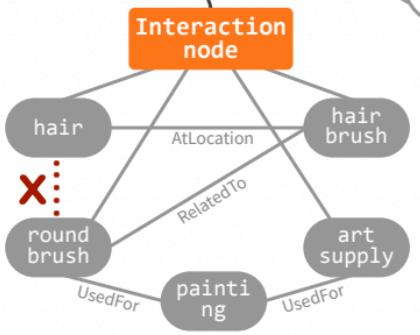


$$(\widetilde{\mathbf{H}}_{\text{int}}^{(\ell)}, \mathbf{H}_1^{(\ell)}, ..., \mathbf{H}_I^{(\ell)}) = \text{LM-Layer}(\mathbf{H}_{\text{int}}^{(\ell-1)}, \mathbf{H}_1^{(\ell-1)}, ..., \mathbf{H}_I^{(\ell-1)}),$$
$$(\widetilde{\mathbf{V}}_{\text{int}}^{(\ell)}, \mathbf{V}_1^{(\ell)}, ..., \mathbf{V}_J^{(\ell)}) = \text{GNN-Layer}(\mathbf{V}_{\text{int}}^{(\ell-1)}, \mathbf{V}_1^{(\ell-1)}, ..., \mathbf{V}_J^{(\ell-1)}),$$

Cross-modal Encoder

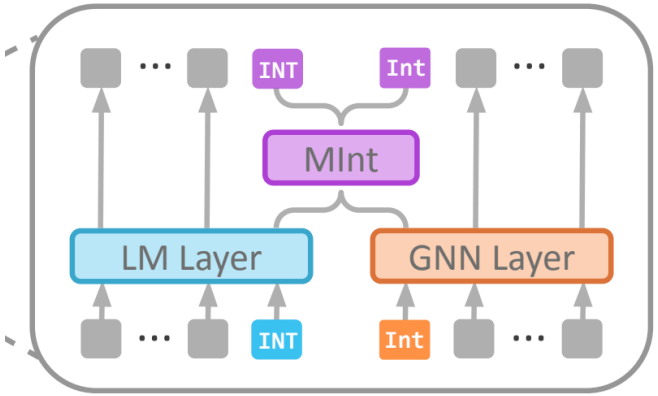


[INT] If it is *not* used for hair, a round brush is an example of [MASK] [MASK].



$$(\mathbf{H}_{\text{int}}^{(0)}, \mathbf{H}_1^{(0)}, ..., \mathbf{H}_I^{(0)}) = \text{LM-Layers}(w_{\text{int}}, w_1 ..., w_I),$$
$$(\mathbf{V}_{\text{int}}^{(0)}, \mathbf{V}_1^{(0)}, ..., \mathbf{V}_J^{(0)}) = \text{Node-Embedding}(v_{\text{int}}, v_1, ..., v_J).$$

Knowledge Fusion during feed forward process



$$(\widetilde{\mathbf{H}}_{\text{int}}^{(\ell)}, \mathbf{H}_1^{(\ell)}, ..., \mathbf{H}_I^{(\ell)}) = \text{LM-Layer}(\mathbf{H}_{\text{int}}^{(\ell-1)}, \mathbf{H}_1^{(\ell-1)}, ..., \mathbf{H}_I^{(\ell-1)}),$$
$$(\widetilde{\mathbf{V}}_{\text{int}}^{(\ell)}, \mathbf{V}_1^{(\ell)}, ..., \mathbf{V}_J^{(\ell)}) = \text{GNN-Layer}(\mathbf{V}_{\text{int}}^{(\ell-1)}, \mathbf{V}_1^{(\ell-1)}, ..., \mathbf{V}_J^{(\ell-1)}),$$

Graph Attention Network

$$\tilde{\boldsymbol{e}}_j^{(\ell)} = f_n\left(\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \alpha_{sj} \boldsymbol{m}_{sj}\right) + \boldsymbol{e}_j^{(\ell-1)}$$

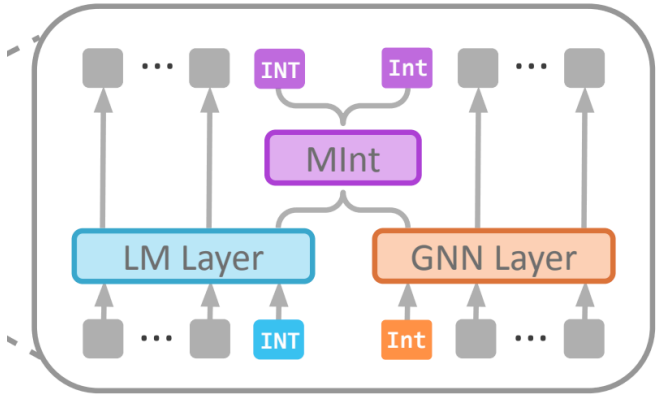
$$\boldsymbol{r}_{sj} = f_r(\tilde{\boldsymbol{r}}_{sj}, \boldsymbol{u}_s, \boldsymbol{u}_j)$$

$$\boldsymbol{m}_{sj} = f_m(\boldsymbol{e}_s^{(\ell-1)}, \boldsymbol{u}_s, \boldsymbol{r}_{sj})$$

$$\boldsymbol{q}_s = f_q(\boldsymbol{e}_s^{(\ell-1)}, \boldsymbol{u}_s) \quad (7) \qquad \boldsymbol{k}_j = f_k(\boldsymbol{e}_j^{(\ell-1)}, \boldsymbol{u}_j, \boldsymbol{r}_{sj}) \quad (8)$$

$$\gamma_{sj} = \frac{\boldsymbol{q}_s^\top \boldsymbol{k}_j}{\sqrt{D}} \quad (9) \qquad \alpha_{sj} = \frac{\exp(\gamma_{sj})}{\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \exp(\gamma_{sj})} \quad (10)$$

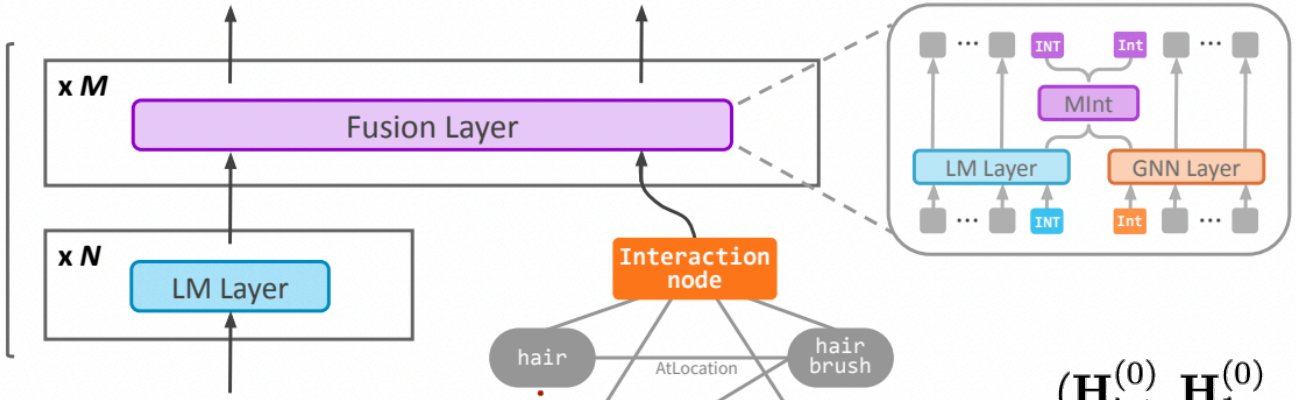
Knowledge Fusion during feed forward process



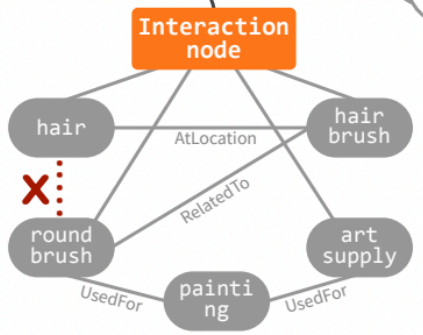
$$(\widetilde{\mathbf{H}}_{\text{int}}^{(\ell)}, \mathbf{H}_1^{(\ell)}, ..., \mathbf{H}_I^{(\ell)}) = \text{LM-Layer}(\mathbf{H}_{\text{int}}^{(\ell-1)}, \mathbf{H}_1^{(\ell-1)}, ..., \mathbf{H}_I^{(\ell-1)}),$$
$$(\widetilde{\mathbf{V}}_{\text{int}}^{(\ell)}, \mathbf{V}_1^{(\ell)}, ..., \mathbf{V}_J^{(\ell)}) = \text{GNN-Layer}(\mathbf{V}_{\text{int}}^{(\ell-1)}, \mathbf{V}_1^{(\ell-1)}, ..., \mathbf{V}_J^{(\ell-1)}),$$
$$[\mathbf{H}_{\text{int}}^{(\ell)}; \mathbf{V}_{\text{int}}^{(\ell)}] = \text{MInt}([\widetilde{\mathbf{H}}_{\text{int}}^{(\ell)}; \widetilde{\mathbf{V}}_{\text{int}}^{(\ell)}]).$$

Text info ⟺ *local KG info*

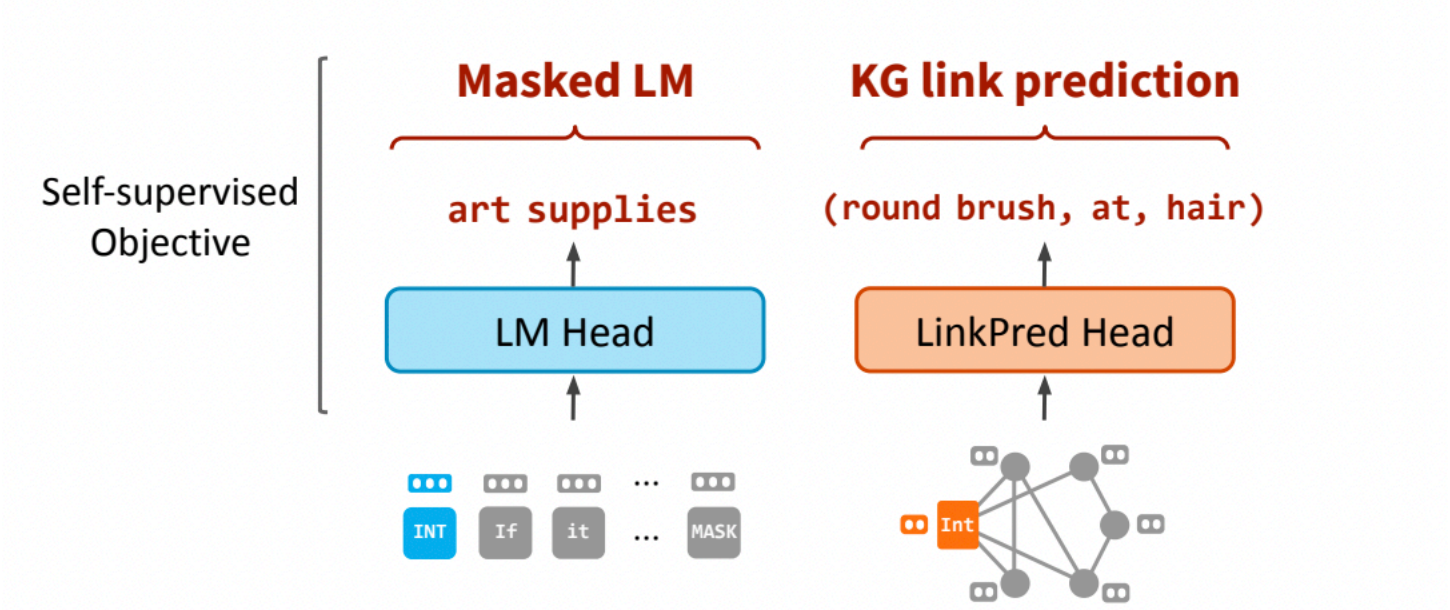
Cross-modal Encoder



[INT] If it is *not* used for hair, a round brush is an example of [MASK] [MASK].



$$(\mathbf{H}_{\text{int}}^{(0)}, \mathbf{H}_1^{(0)}, ..., \mathbf{H}_I^{(0)}) = \text{LM-Layers}(w_{\text{int}}, w_1 ..., w_I),$$
$$(\mathbf{V}_{\text{int}}^{(0)}, \mathbf{V}_1^{(0)}, ..., \mathbf{V}_J^{(0)}) = \text{Node-Embedding}(v_{\text{int}}, v_1, ..., v_J).$$

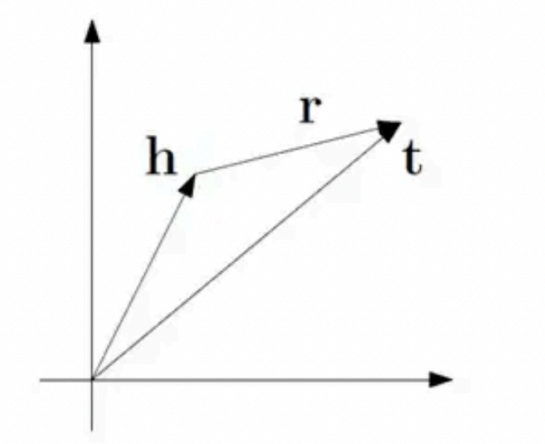


$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log p(w_i \mid \mathbf{H}_i).$$

$$\mathcal{L}_{\text{LinkPred}} = \sum_{(h,r,t) \in S} \left(-\log \sigma(\phi_r(\mathbf{h}, \mathbf{t}) + \gamma) + \frac{1}{n} \sum_{(h',r,t')} \log \sigma(\phi_r(\mathbf{h}', \mathbf{t}') + \gamma) \right),$$

Scoring function

Scoring function of predicted triplets



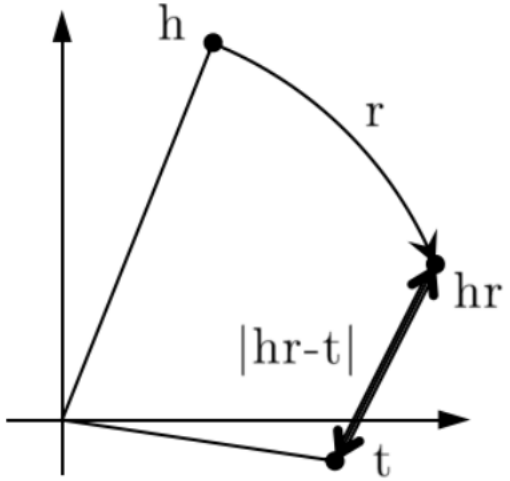
Triple: (head, relation, tail)

$head + relation \approx tail$

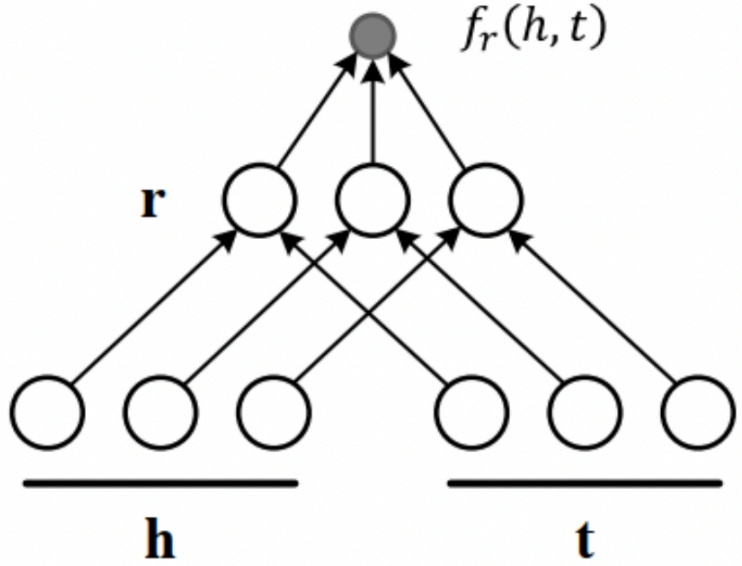
TransE-based : $\phi_r(h,t) = -\|h + r - t\|$

- A **higher** $\phi_r(h,t)$ indicates a higher chance of (h, r, t) being a **positive** triplet (edge) instead of negative (no edge).

Scoring function of predicted triplets



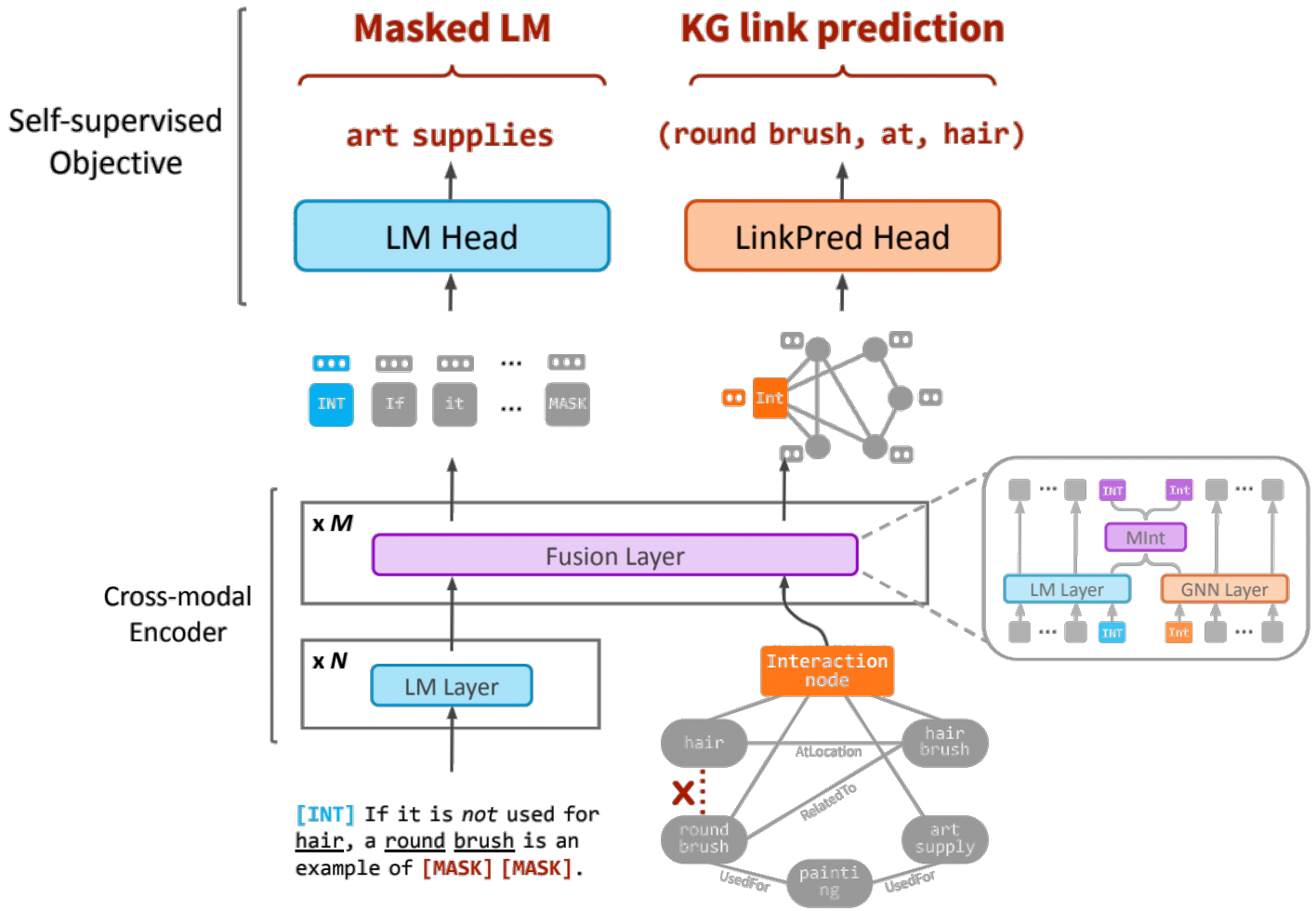
RotateE-based $\phi_r(h,t) = -\|h \odot r - t\|$



DisMult-based $\phi_r(h,t) = h^T M_r t$

RotateE models r as rotation in complex plane.

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log p(w_i \mid \mathbf{H}_i). \qquad \mathcal{L}_{\text{LinkPred}} = \sum_{(h,r,t) \in S} \left(-\log \sigma(\phi_r(\mathbf{h}, \mathbf{t}) + \gamma) + \frac{1}{n} \sum_{(h',r,t')} \log \sigma(\phi_r(\mathbf{h}', \mathbf{t}') + \gamma) \right),$$

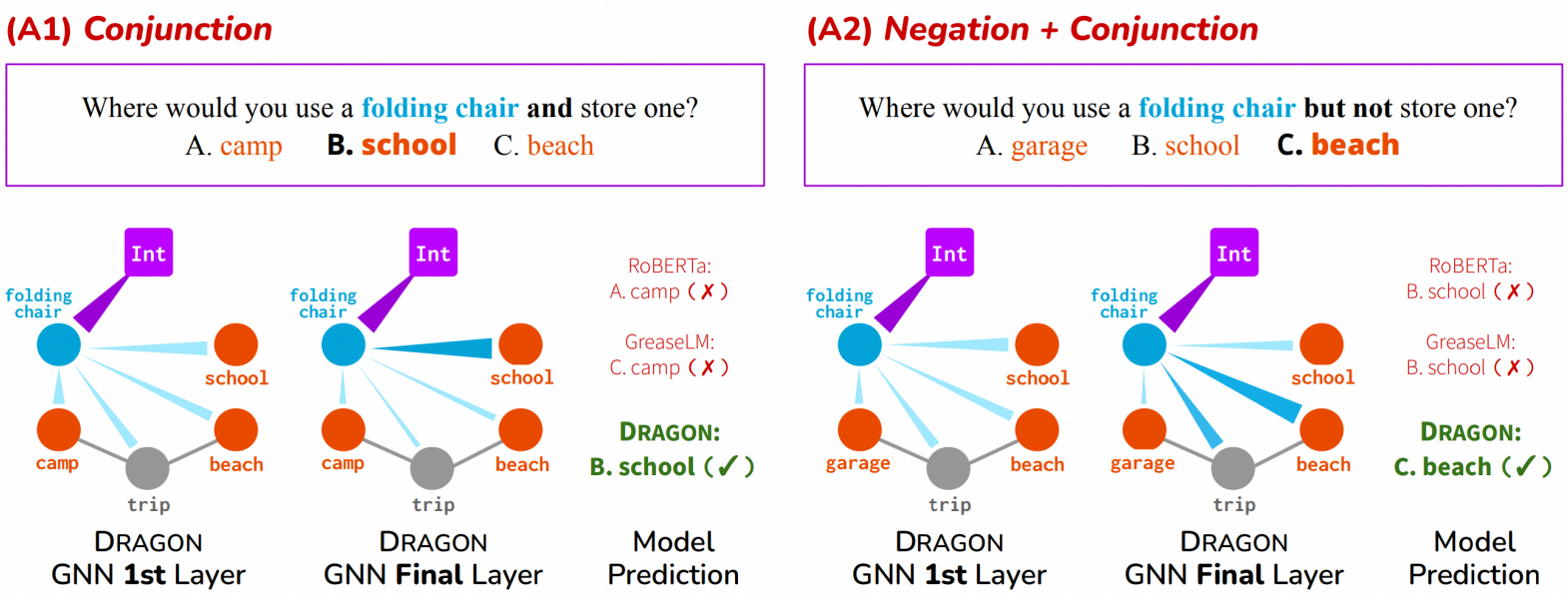


$$L = L_{MLM} + L_{LinkPred}$$

Experiments

Ablation Type	Ablation	CSQA	OBQA
Pretraining objective	MLM + LinkPred (**final**)	**76.0**	**72.0**
	MLM only	74.3	67.2
	LinkPred only	73.8	66.4
LinkPred head	DistMult (**final**)	**76.0**	**72.0**
	TransE	75.7	71.4
	RotatE	75.8	71.7
Cross-modal model	Bidirectional interaction (**final**)	**76.0**	**72.0**
	Concatenate at end	74.5	68.0
KG structure	Use graph (**final**)	**76.0**	**72.0**
	Convert to sentence	74.7	70.1

Experiments



	Negation	Conjunction	Hedge	# Prepositional Phrases 0	1	2	3	# Entities >10
RoBERTa	61.7	70.9	68.6	67.6	71.0	71.1	73.1	74.5
QAGNN	65.1	74.5	74.2	72.1	71.6	75.6	71.3	78.6
GreaseLM	65.1	74.9	76.6	75.6	73.8	74.7	73.6	79.4
DRAGON (**Ours**)	**75.2**	**79.6**	**77.5**	**79.1**	**78.2**	**77.8**	**80.9**	**83.5**

Table 2: Accuracy of DRAGON on *CSQA* + *OBQA* dev sets for **questions involving complex reasoning** such as negation terms, conjunction terms, hedge terms, prepositional phrases, and more entity mentions. DRAGON consistently outperforms the existing LM (RoBERTa) and KG-augmented QA models (QAGNN, GreaseLM) in these complex reasoning settings.

KNOWLEDGE-IN-CONTEXT: TOWARDS KNOWLEDGE-ABLE SEMI-PARAMETRIC LANGUAGE MODELS

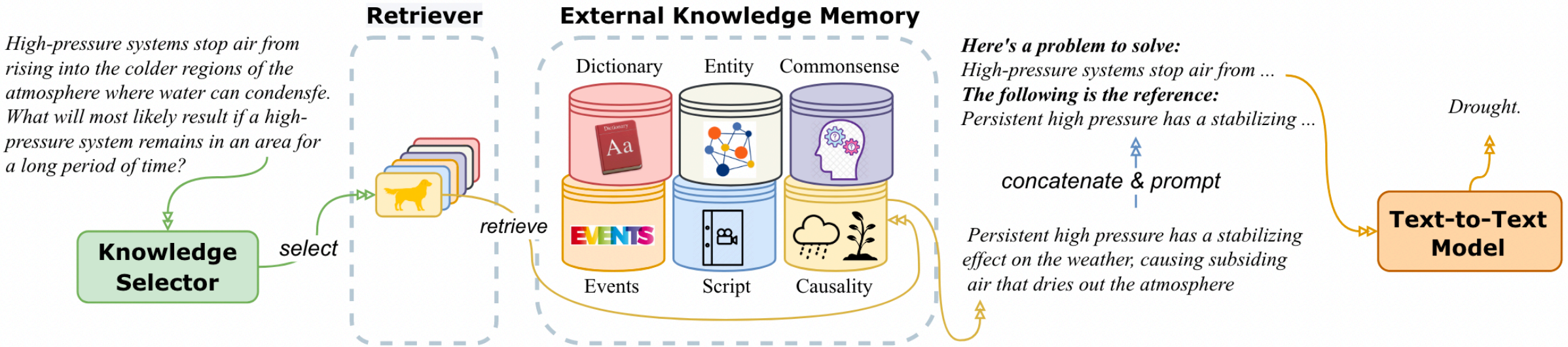
Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu & Jianshu Chen *
Tencent AI Lab, Bellevue, WA 98004, USA

Challenge

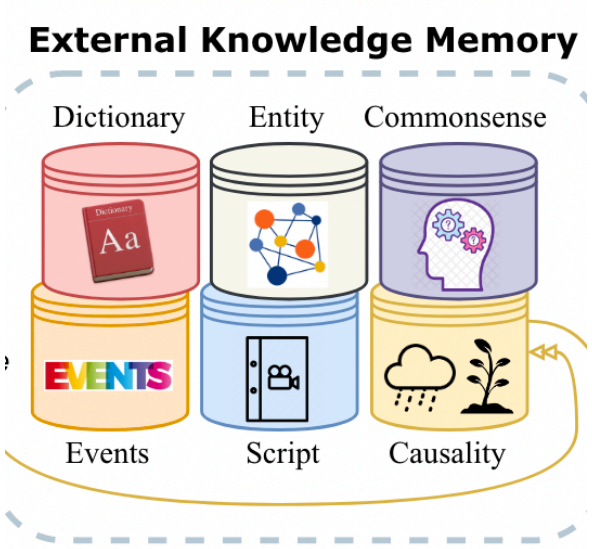
- useful knowledge pieces are generally sparsely distributed over a large textual corpus

Question	*High-pressure systems stop air from rising into the colder regions of the atmosphere where water can condense. What will most likely result if a high-pressure system remains in an area for a long period of time?*
Answer	*Drought*
CausalBank (structured)	*Persistent high pressure has a stabilizing effect on the weather, causing subsiding air that dries out the atmosphere.*
Wikipedia (plain text)	*High-pressure systems are alternatively referred to as anticyclones. On English - language weather maps, high-pressure centers are identified by the letter H in English, within the isobar with the highest pressure value. On constant pressure upper level charts, it is located within the highest height line contour.*

Instance-adaptive knowledge augmentation



Construction of External Knowledge Memory



- Dictionary: Wikitionary
- Commonsense: ConceptNet
- Event: auto-extracted event knowledge graph
- Entity Knowledge: Wikipedia and Wikidata
- Scripts: triples in the form of < verbal information, context, nonverbal information >

triplets

- Causality: Casual Bank

because-mode sentences

each knowledge piece is in the form of **< subject, relation, object >** triplet

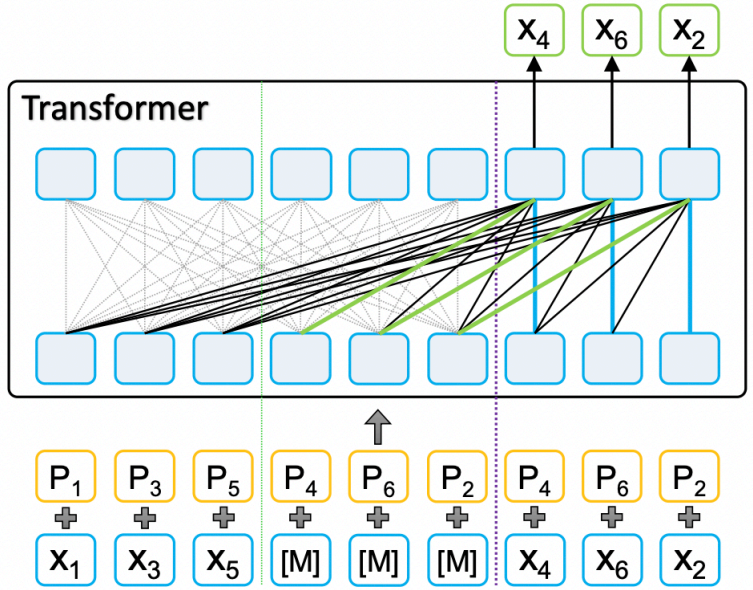
Construction of External Knowledge Memory

triples

< subject (s), relation (r), object (o) >

key-value pairs

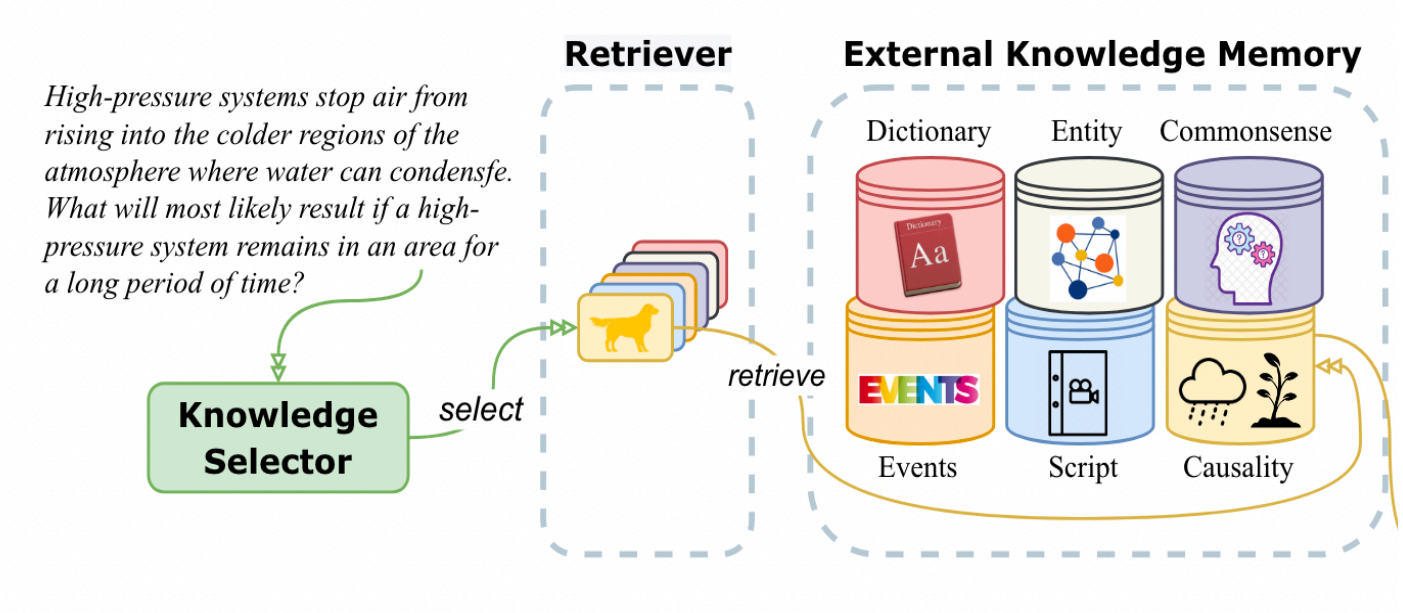
	Key	Value
Dictionary	s	o
Commonsense	s $s \oplus o$ $s \oplus r \oplus o$	$s \oplus r \oplus o$ $s \oplus r \oplus o$ $s \oplus r \oplus o$
Entity	s o	o o
Event	s $s \oplus o$ $s \oplus r \oplus o$	$s \oplus r \oplus o$ $s \oplus r \oplus o$ $s \oplus r \oplus o$
Script	s o	r r
Causality	$s \oplus o$ $o \oplus s$	$s \oplus o$ $o \oplus s$



- Encode the keys into dense vectors
- Retain the corresponding values in textual forms

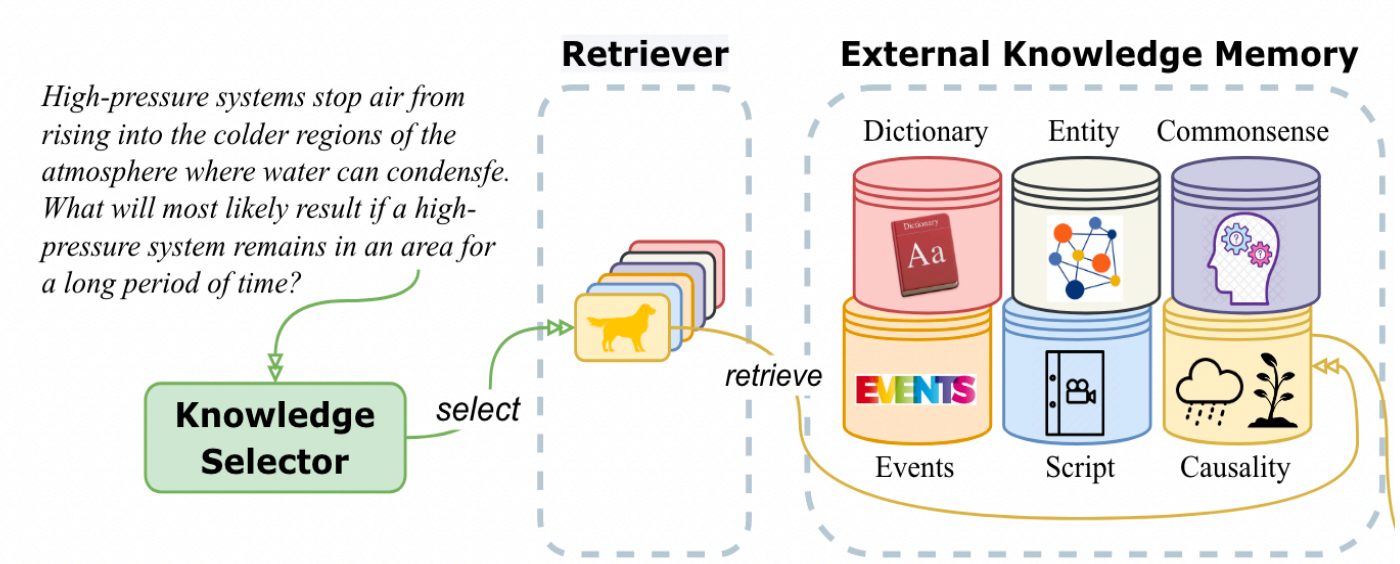
generate **key-value pairs** from the original knowledge pieces

Fetch Relevant Knowledge

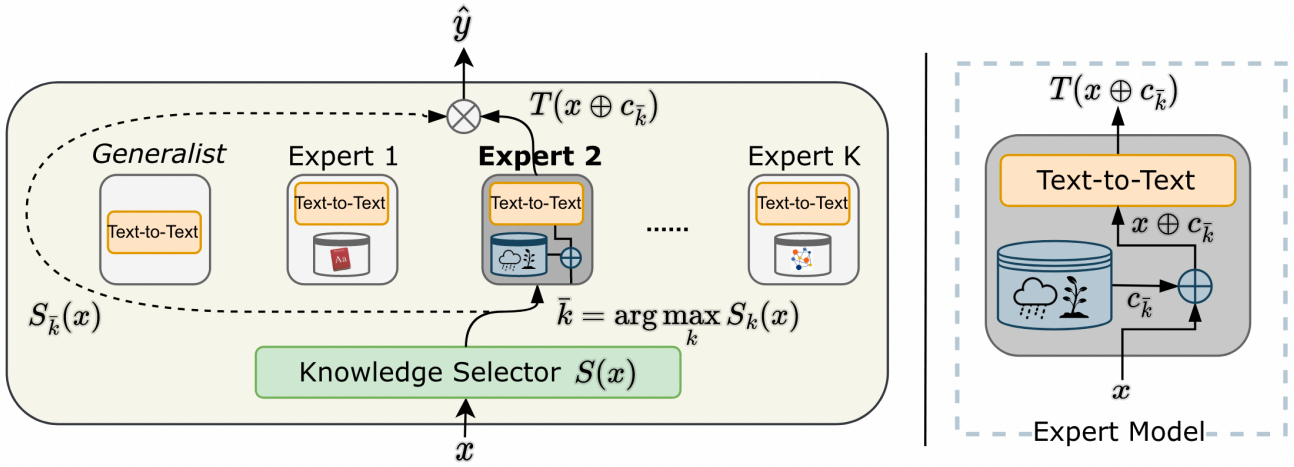


- Knowledge Selector: (K+1) class linear classifier
  - K: number of knowledge classes
  - +1: no external knowledge is needed

Fetch Relevant Knowledge



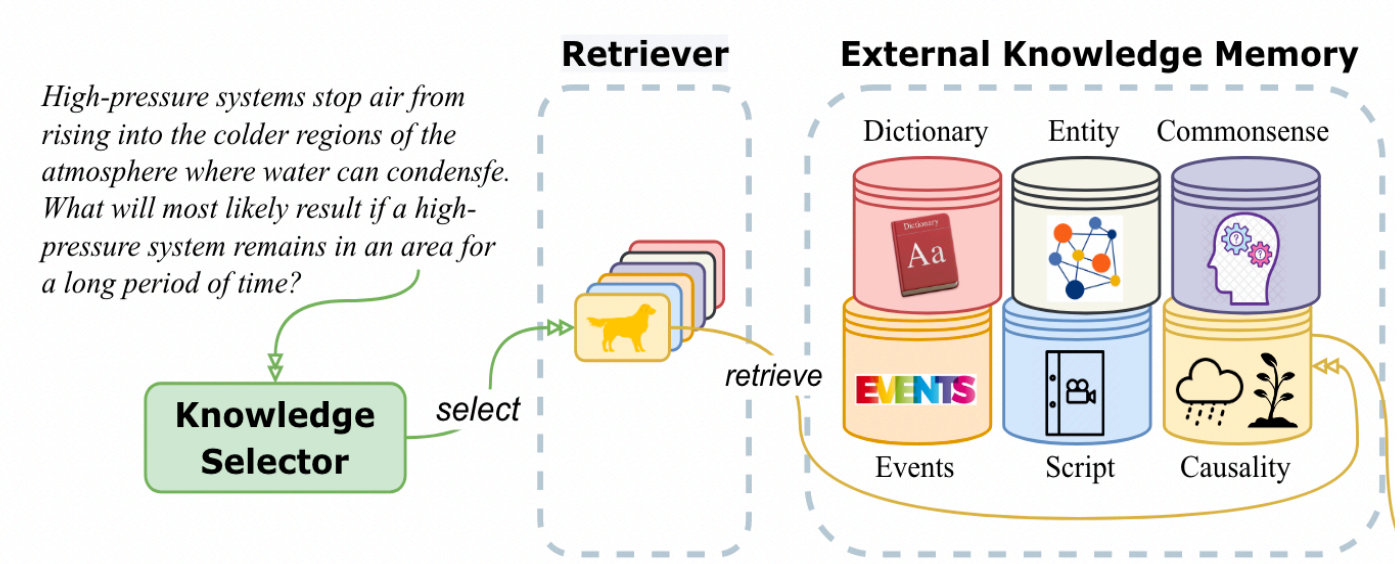
- Knowledge Selector: (K+1) class linear classifier
  - K: number of knowledge classes
  - +1: no external knowledge is needed



$$\bar{k} = \arg\max_k S_k(x)$$

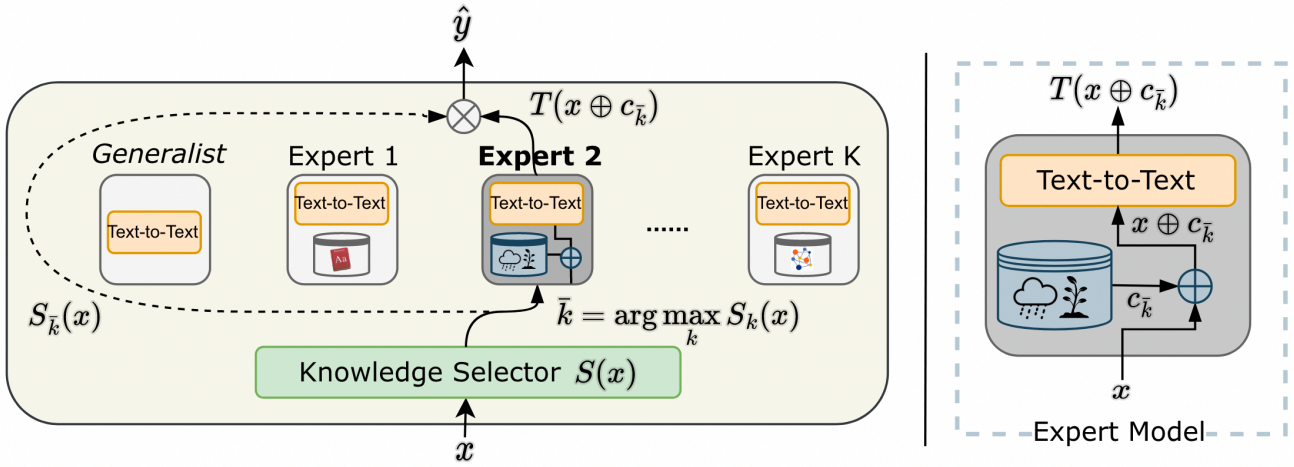
$$\hat{y} = T(x \oplus c_{\bar{k}}) \cdot S_{\bar{k}}(x)$$

Fetch Relevant Knowledge

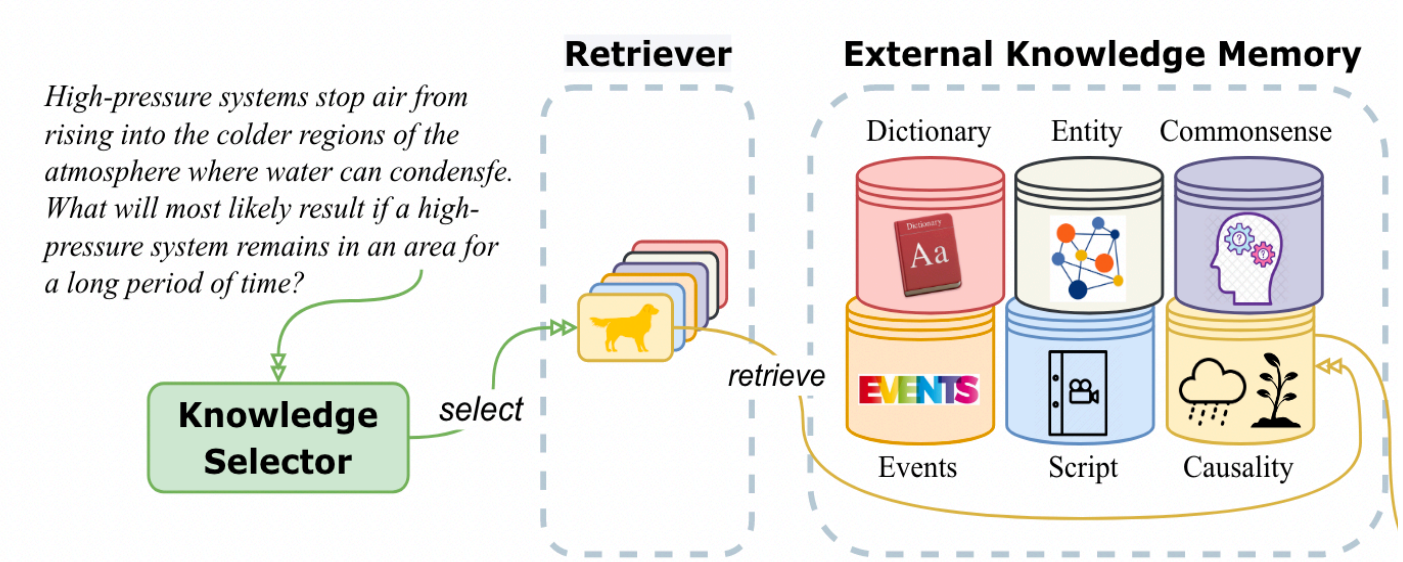


$$\bar{k} = \arg\max_k S_k(x)$$

$$\hat{y} = T(x \oplus c_{\bar{k}}) \cdot S_{\bar{k}}(x)$$



Fetch Relevant Knowledge



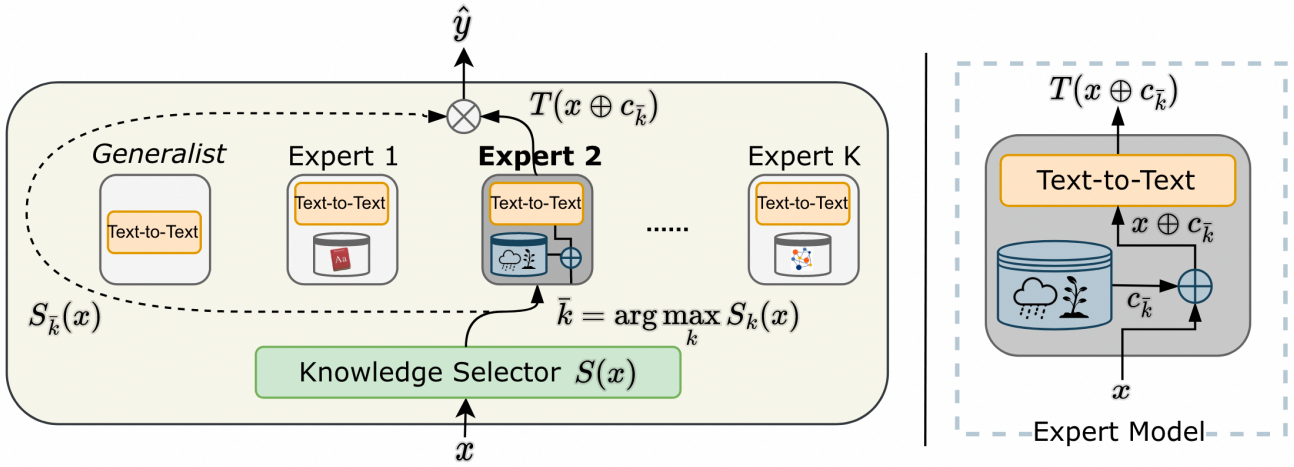
$$\bar{k} = \arg\max_k S_k(x)$$

$$\hat{y} = T(x \oplus c_{\bar{k}}) \cdot S_{\bar{k}}(x)$$

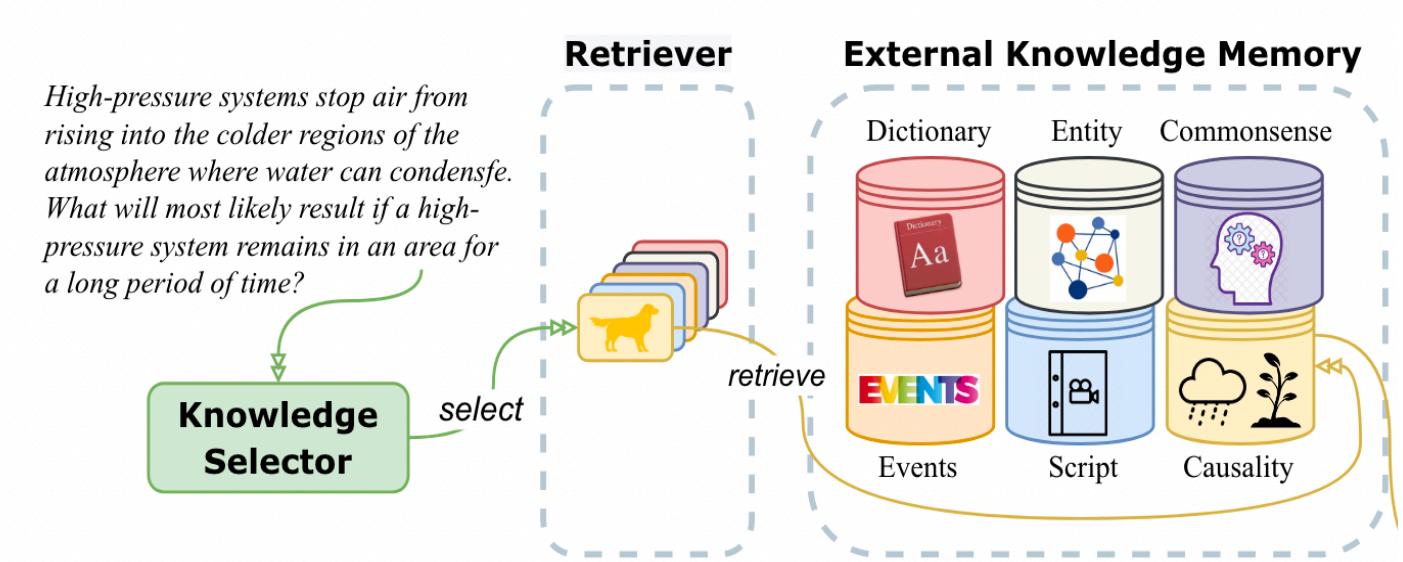
$c_{\bar{k}}$: retrieved knowledge piece

Maximum Inner Product Search

$$c_k = \arg\max_{c_k \in C} C^T x$$



Fetch Relevant Knowledge



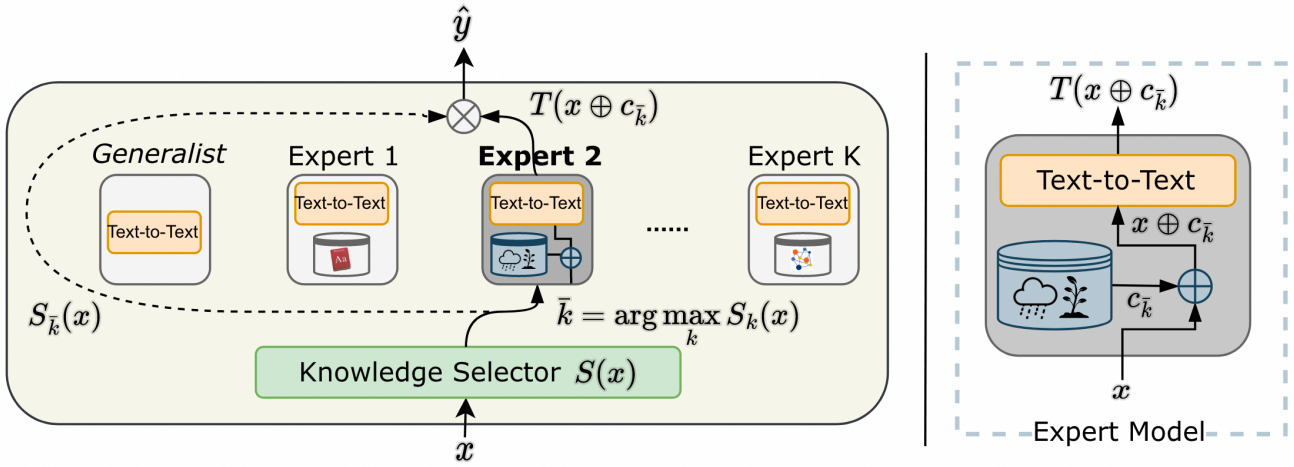
$$\bar{k} = \arg\max_k S_k(x)$$

$$\hat{y} = T(x \oplus c_{\bar{k}}) \cdot S_{\bar{k}}(x)$$

$c_{\bar{k}}$: retrieved knowledge piece

Maximum Inner Product Search

$$c_k = \arg\max_{c_k \in C} C^T x$$



$$\mathcal{L}(x, y) = \sum_{t=1}^{T} \texttt{CrossEntropy}(\hat{y}_t, y_t) + \alpha \cdot \texttt{Balancing}(S(x))$$

$$\texttt{Balancing}(S(x)) = (K+1) \cdot \sum_{i=0}^{K+1} f_i \cdot P_i,$$

Experimental Results

Category	Task	Task Reference	None	ENT	DIC	COM	EVT	SCR	CAU
Coreference	WSC	Levesque et al. (2012)	60.3	49.5	62.0	53.1	**64.7**	62.0	63.4
	Wino. debiased	Sakaguchi et al. (2021)	**59.1**	53.5	57.3	56.9	58.3	58.5	54.1
	Wino. xl	Sakaguchi et al. (2021)	63.5	63.0	64.2	**64.5**	64.3	63.8	63.5
NLI	CB	De Marneffe et al. (2019)	87.5	87.5	85.9	87.5	85.9	**90.6**	84.4
	RTE	Wang et al. (2019)	77.1	76.2	79.0	**79.2**	76.6	76.9	71.3
Paraphrase	MRPC	Dolan & Brockett (2005)	82.9	80.5	**87.7**	77.9	84.9	84.4	82.0
	QQP	Wang et al. (2018)	89.4	89.1	**89.5**	89.5	89.2	89.3	89.4
	PAWS	Zhang et al. (2019a)	**94.6**	94.2	94.3	94.4	94.4	94.5	94.2
Closed QA	ARC-Easy	Clark et al. (2018)	52.8	52.6	53.1	51.7	56.1	51.7	**64.6**
	ARC-Challenge	Clark et al. (2018)	30.9	36.2	30.9	33.5	34.2	37.2	**39.5**
	WikiQA	Yang et al. (2015)	**96.2**	95.6	95.8	95.9	95.7	95.7	96.2
Extr. QA	ReCoRD	Zhang et al. (2018)	53.9	53.9	53.2	54.0	**54.1**	53.9	53.5
Sentiment	IMDB	Maas et al. (2011)	94.8	**94.9**	94.7	94.9	94.7	94.7	94.8
	Rotten Tomatoes	Pang & Lee (2005)	90.2	89.6	**90.3**	89.9	90.0	90.0	89.6
Completion	HellaSwag	Zellers et al. (2019)	49.8	49.3	50.6	51.8	52.0	49.8	**53.7**
	COPA	Roemmele et al. (2011)	58.0	58.5	59.8	54.5	58.9	56.2	**62.0**

Experimental Results

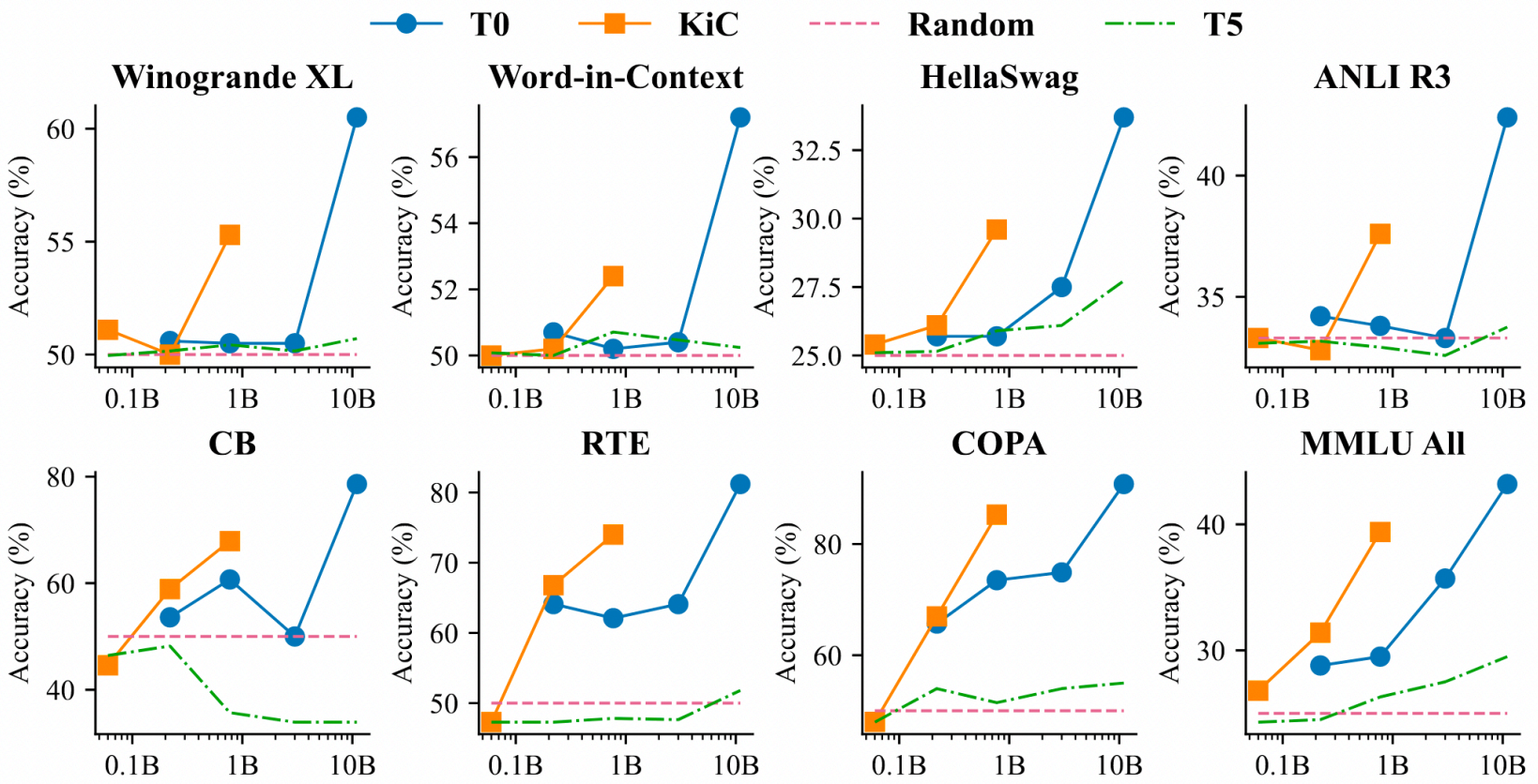


Figure 3: Emerging behaviors of T5, T0 and KiC models. Our KiC model shows emerging behavior at a much smaller model scale (when it increases from 0.22B to 0.77B) compared to T0.

Thanks for your listening