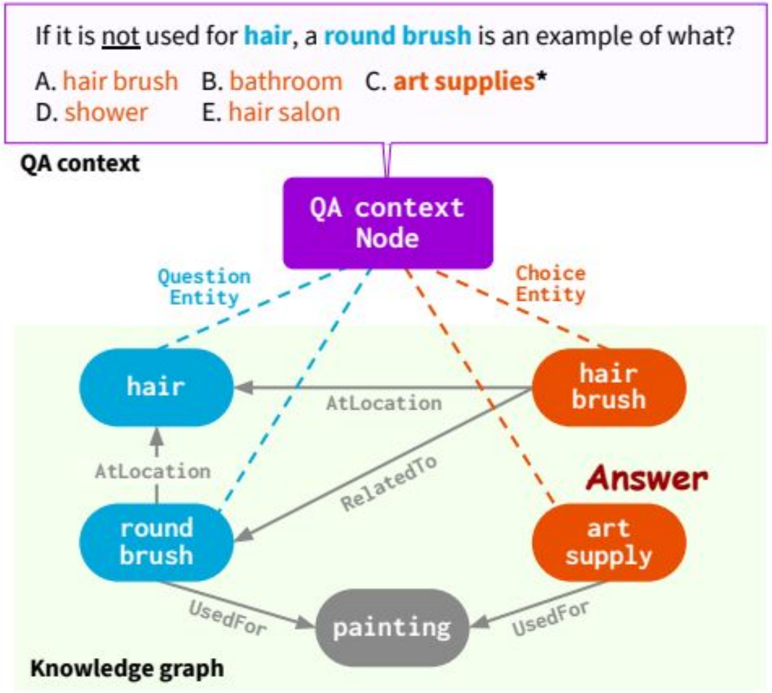
GREASELM: GRAPH REASONING ENHANCED LANGUAGE MODELS FOR QUESTION ANSWERING

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren Percy Liang, Christopher D. Manning, Jure Leskovec

International Conference on Learning Representations (ICLR) 2022.

Motivation

- Answering questions about textual narratives requires reasoning over both stated context and the unstated world knowledge.
- Pretrained language models (LM), do not robustly represent latent relationships between concepts, which is necessary for reasoning.
- Knowledge graphs (KG) are often used to augment LMs with structured representations of world knowledge in a shallow and non-interactive manner.

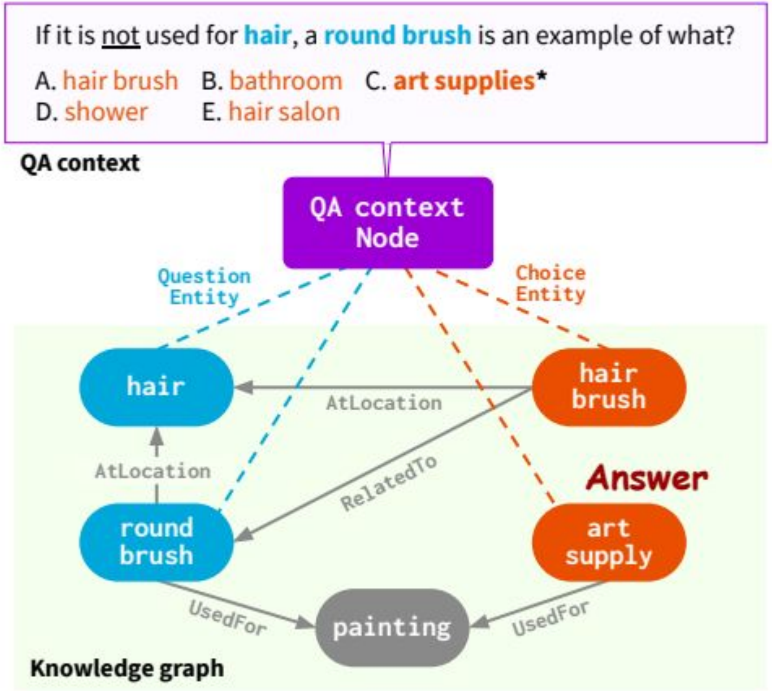


If it is not used for hair, a round brush is an example of what?

A. hair brush B. bathroom C. art supplies*
D. shower E. hair salon

Motivation

- It remains an open question how to fuse and reason over the KG representations and the language context, effectively and interactively to leverage knowledge from both modalities.


If it is not used for hair, a round brush is an example of what?

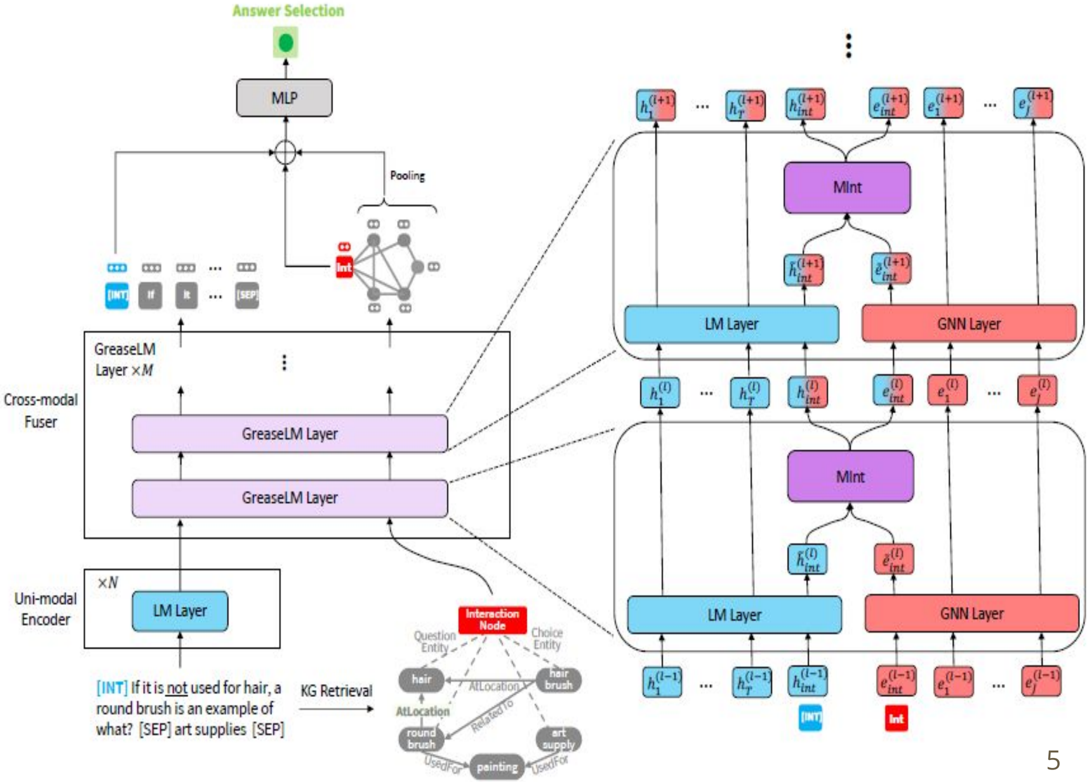
A. hair brush B. bathroom C. art supplies*
D. shower E. hair salon

Related Works

- Prior works (Kagnet, MHGRN, KT-NET) use one modality to ground the other, such as using an encoded representation of a linked KG to augment the textual representation of a QA example.
- QA-GNN (Yasunaga et al., 2021) propose to jointly update the LM and GNN representations via message passing. However, they use a single pooled representation of the LM to seed both modalities.

PROPOSED APPROACH

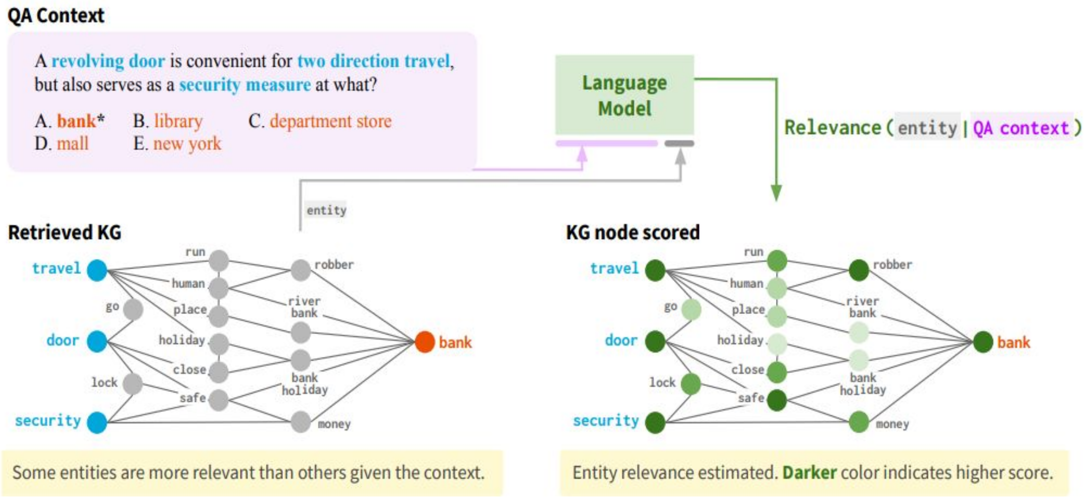
- A set of unimodal LM layers (N) which learn an initial representation of the input tokens.
- A set of upper cross-modal GreaseLM layers (M) which learn to jointly represent the language sequence and linked knowledge graph.
- A special interaction token **w**$_{\textbf{int}}$ is used for LM and a special interaction node **e**$_{\textbf{int}}$ is used for KG.



Answer Selection
MLP
Pooling
[INT]
If
It
...
[SEP]
Int
GreaseLM Layer ×M
Cross-modal Fuser
GreaseLM Layer
GreaseLM Layer
Uni-modal Encoder
×N
LM Layer
[INT] If it is not used for hair, a round brush is an example of what? [SEP] art supplies [SEP]
KG Retrieval
Interaction Node
Question Entity
Choice Entity
hair
hair brush
AtLocation
AtLocation
RelatedTo
round brush
art supply
UsedFor
painting
UsedFor
$h_1^{(l+1)}$ ... $h_T^{(l+1)}$ $h_{int}^{(l+1)}$ $e_{int}^{(l+1)}$ $e_1^{(l+1)}$ ... $e_J^{(l+1)}$
MInt
$\tilde{h}_{int}^{(l+1)}$ $\tilde{e}_{int}^{(l+1)}$
LM Layer
GNN Layer
$h_1^{(l)}$ ... $h_T^{(l)}$ $h_{int}^{(l)}$ $e_{int}^{(l)}$ $e_1^{(l)}$ ... $e_J^{(l)}$
MInt
$\tilde{h}_{int}^{(l)}$ $\tilde{e}_{int}^{(l)}$
LM Layer
GNN Layer
$h_1^{(l-1)}$ ... $h_T^{(l-1)}$ $h_{int}^{(l-1)}$ $e_{int}^{(l-1)}$ $e_1^{(l-1)}$ ... $e_J^{(l-1)}$
[INT]
Int

KG Retrieval

- Node name are concatenated with the context of the QA example, and pass it through a LM.
- The output score of the node name is considered as the relevance score. Top 200 scores nodes are retained.
- A subgraph is formed with all the edges that connect any two nodes.
- Each node in the subgraph is assigned a type according to whether its corresponding entity was linked from the context, question, answer, or from a bridge path.

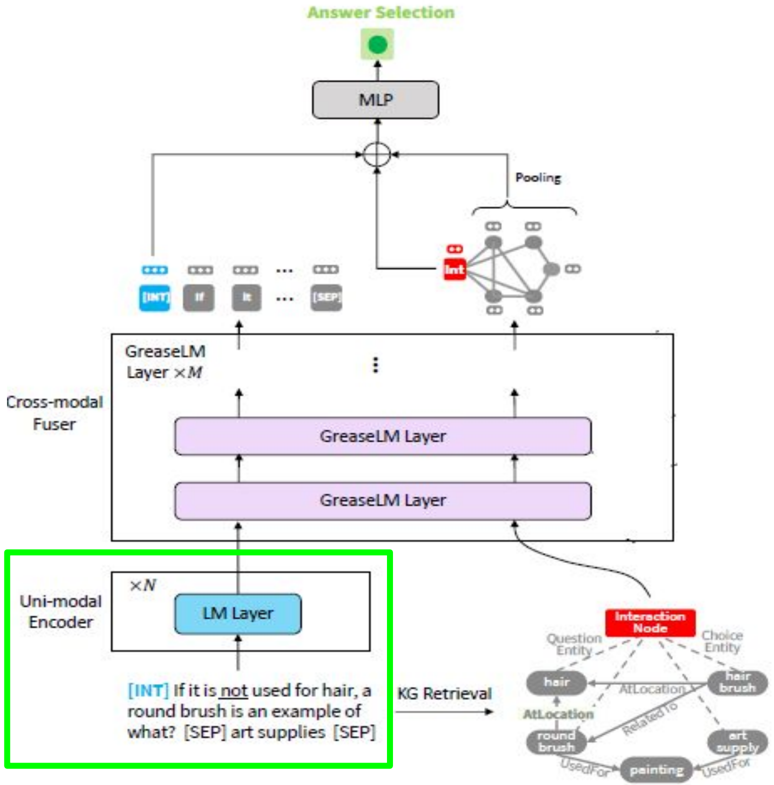

QA Context
A revolving door is convenient for two direction travel, but also serves as a security measure at what?
A. bank* B. library C. department store
D. mall E. new york
Language Model
Relevance (entity | QA context)
entity
Retrieved KG
travel
go
door
lock
security
run
human
place
holiday
close
safe
robber
river bank
bank holiday
money
bank
Some entities are more relevant than others given the context.
KG node scored
travel
go
door
lock
security
run
human
place
holiday
close
safe
robber
river bank
bank holiday
money
bank
Entity relevance estimated. Darker color indicates higher score.

Yasunaga et al. "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering", NAACL, 2021.

Language Pre-Encoding

- In the unimodal encoding component, given the sequence of tokens w_{int}, w_1, …, w_T , an output representation for each layer is computed:

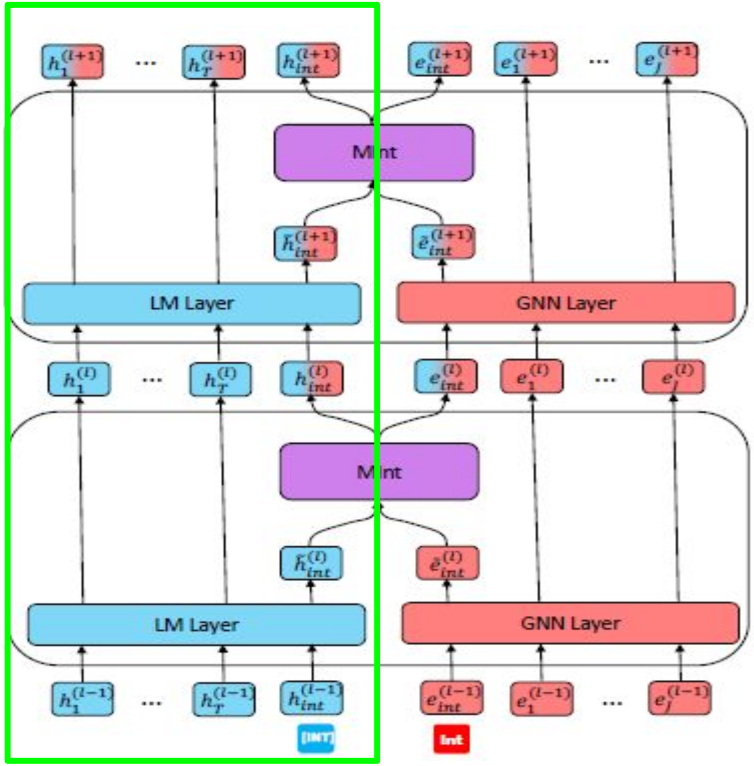
$$\{h_{int}^{(\ell)}, h_1^{(\ell)}, \ldots, h_T^{(\ell)}\} = \text{LM-Layer}(\{h_{int}^{(\ell-1)}, h_1^{(\ell-1)}, \ldots, h_T^{(\ell-1)}\})$$
$$\text{for } \ell = 1, \ldots, N$$



Language Representation

- In each LM layer of GreaseLM, the input token embeddings are fed into additional transformer LM encoder blocks that continue to encode the textual context based on the LM's pretrained representations:

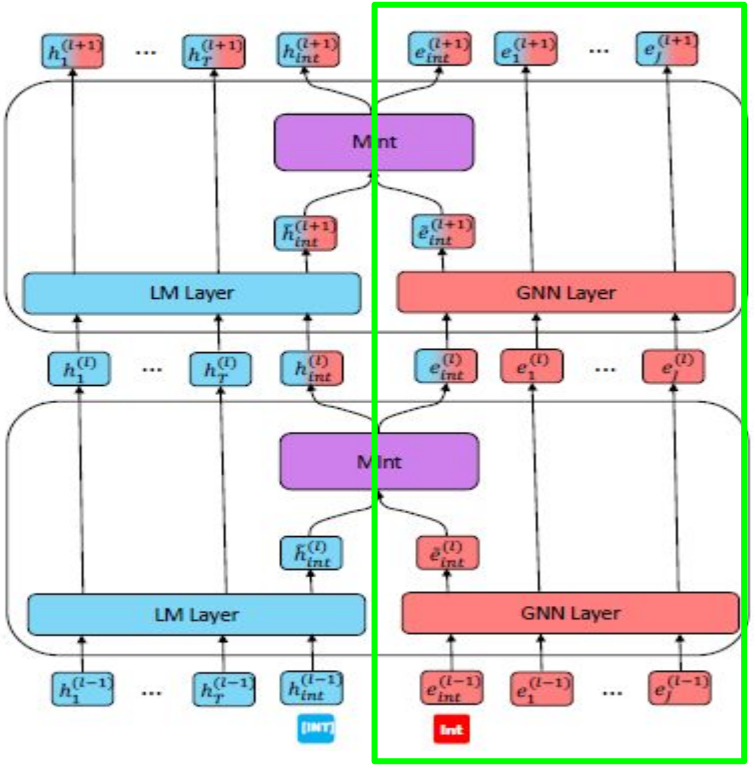
$$\{\tilde{h}_{int}^{(N+\ell)}, \tilde{h}_1^{(N+\ell)}, \ldots, \tilde{h}_T^{(N+\ell)}\} = \text{LM-Layer}(\{h_{int}^{(N+\ell-1)}, h_1^{(N+\ell-1)}, \ldots, h_T^{(N+\ell-1)}\})$$
$$\text{for } \ell = 1, \ldots, M$$



Graph Representation

- in each layer of the GNN, the current representation of the node embeddings is fed into the layer to perform information propagation between nodes in the graph:

$$\{\tilde{e}_{int}^{(\ell)}, \tilde{e}_1^{(\ell)}, \ldots, \tilde{e}_J^{(\ell)}\} = \text{GNN}(\{e_{int}^{(\ell-1)}, e_1^{(\ell-1)}, \ldots, e_J^{(\ell-1)}\})$$
$$\text{for } \ell = 1, \ldots, M$$


$h_1^{(l+1)}$ ... $h_T^{(l+1)}$ $h_{int}^{(l+1)}$
$e_{int}^{(l+1)}$ $e_1^{(l+1)}$ ... $e_J^{(l+1)}$
MInt
$\tilde{h}_{int}^{(l+1)}$
$\tilde{e}_{int}^{(l+1)}$
LM Layer
GNN Layer
$h_1^{(l)}$ ... $h_T^{(l)}$ $h_{int}^{(l)}$
$e_{int}^{(l)}$ $e_1^{(l)}$ ... $e_J^{(l)}$
MInt
$\tilde{h}_{int}^{(l)}$
$\tilde{e}_{int}^{(l)}$
LM Layer
GNN Layer
$h_1^{(l-1)}$ ... $h_T^{(l-1)}$ $h_{int}^{(l-1)}$
$e_{int}^{(l-1)}$ $e_1^{(l-1)}$ ... $e_J^{(l-1)}$
[INT]
Int

Node Embedding Computation

- The GNN computes node representations for each node via message passing between neighbors on the graph using graph attention network:

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

$$\boldsymbol{r}_{sj} = f_r(\tilde{\boldsymbol{r}}_{sj},\, \boldsymbol{u}_s,\, \boldsymbol{u}_j) \qquad m_{sj} = f_m(e_s^{(\ell-1)},\, u_s,\, r_{sj})$$

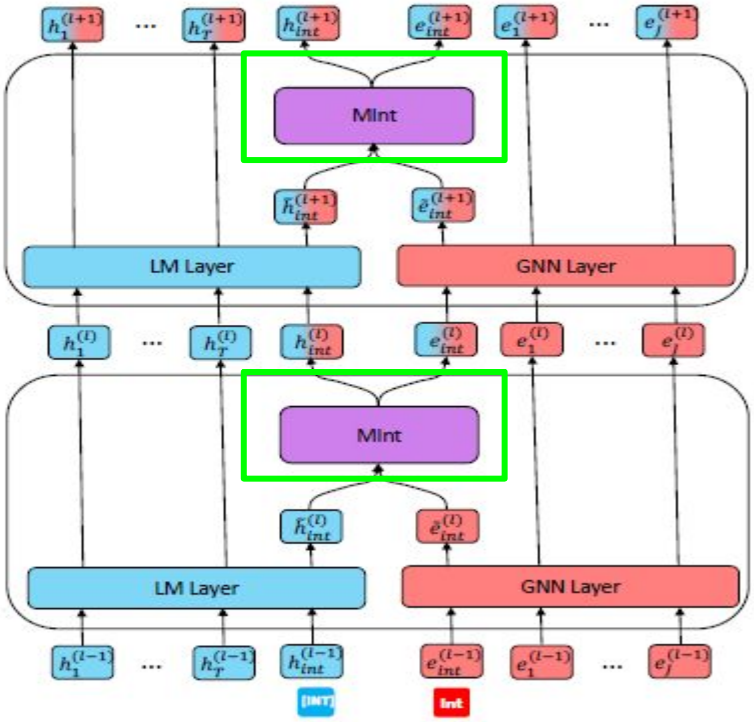
$$q_s = f_q(e_s^{(\ell-1)},\, u_s) \qquad \boldsymbol{k}_j = f_k(\boldsymbol{e}_j^{(\ell-1)},\, \boldsymbol{u}_j,\, \boldsymbol{r}_{sj}) \qquad \gamma_{sj} = \frac{\boldsymbol{q}_s^{\top} \boldsymbol{k}_j}{\sqrt{D}}$$

$$\alpha_{sj} = \frac{\exp(\gamma_{sj})}{\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \exp(\gamma_{sj})}$$

Modality Interaction

- Finally, after using a transformer LM layer and a GNN layer to update token embeddings and node embeddings respectively, a modality interaction layer, **MInt** (a two-layer MLP) is used to let the two modalities fuse information through the bottleneck of the interaction token $\mathbf{w}_{\text{int}}$ and the interaction node $\mathbf{e}_{\text{int}}$:

$$[h_{int}^{(\ell)}; e_{int}^{(\ell)}] = \text{MInt}([\tilde{h}_{int}^{(\ell)}; \tilde{e}_{int}^{(\ell)}])$$



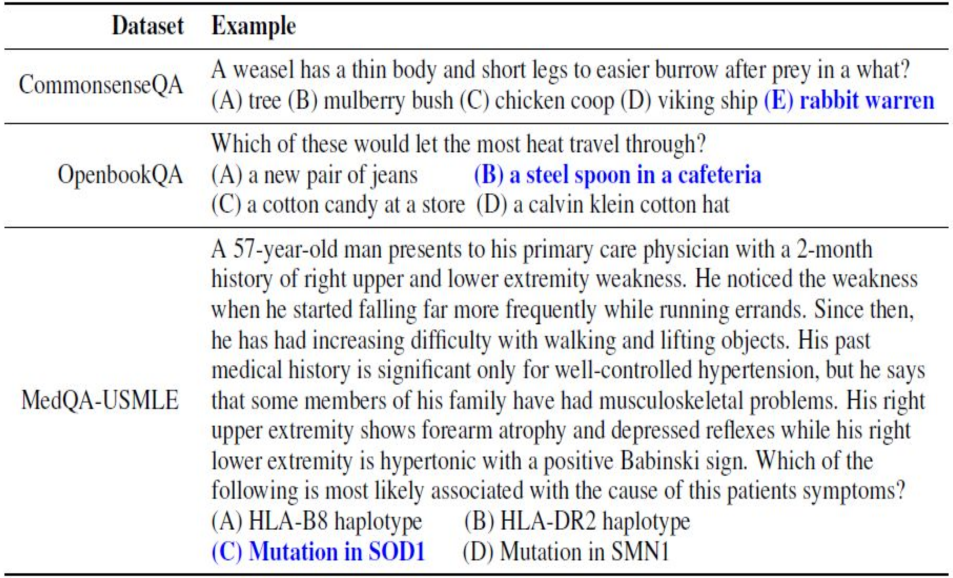
Learning & Inference

- Given a question **q** and an answer **a** from all the candidates **A**, we compute the probability of **a** being the correct answer as:

$$p(a \mid q, c) \propto \exp(\mathrm{MLP}(h_{int}^{(N+M)},\ e_{int}^{(M)},\ g)).$$

Datasets

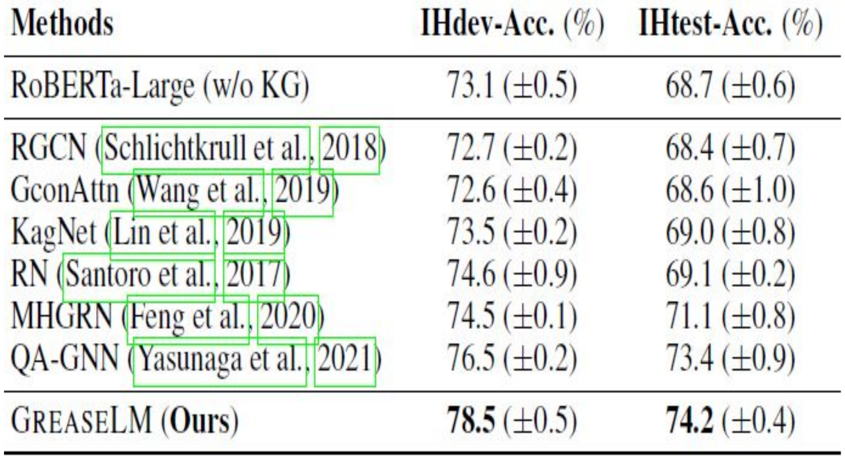
- **CommonsenseQA** is a 5-way multiple-choice question answering dataset that requires background commonsense knowledge beyond surface language understanding.
- **OpenbookQA** is a 4-way multiple-choice question answering dataset that tests elementary scientific knowledge.
- **MedQA-USMLE** a 4-way multiple-choice question answering dataset, which requires biomedical and clinical knowledge.



Dataset	Example
CommonsenseQA	A weasel has a thin body and short legs to easier burrow after prey in a what? (A) tree (B) mulberry bush (C) chicken coop (D) viking ship **(E) rabbit warren**
OpenbookQA	Which of these would let the most heat travel through? (A) a new pair of jeans **(B) a steel spoon in a cafeteria** (C) a cotton candy at a store (D) a calvin klein cotton hat
MedQA-USMLE	A 57-year-old man presents to his primary care physician with a 2-month history of right upper and lower extremity weakness. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had increasing difficulty with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his family have had musculoskeletal problems. His right upper extremity shows forearm atrophy and depressed reflexes while his right lower extremity is hypertonic with a positive Babinski sign. Which of the following is most likely associated with the cause of this patients symptoms? (A) HLA-B8 haplotype (B) HLA-DR2 haplotype **(C) Mutation in SOD1** (D) Mutation in SMN1

Experimental Results

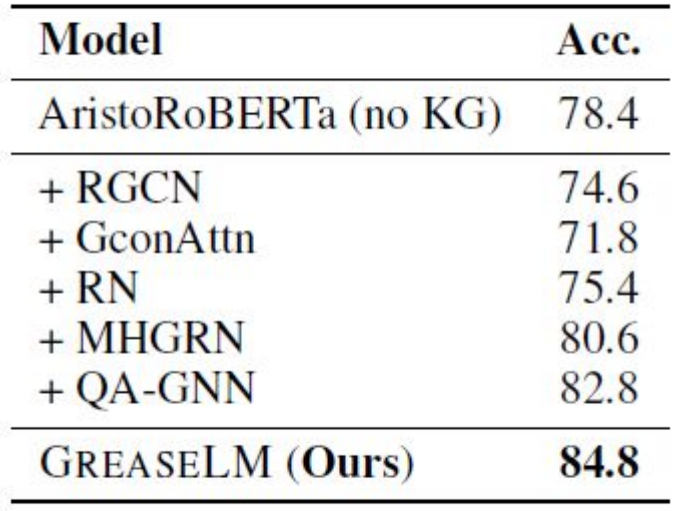
- **CommonsenseQA**
  - test performance improves by 5.5% over fine-tuned LMs.
  - 0.9% improvement over existing LM+KG models.
  - The boost over QA-GNN suggests that GREASELM’s multi-layer fusion component passes more expressive information than other LM+KG methods.



Methods	IHdev-Acc. (%)	IHtest-Acc. (%)
RoBERTa-Large (w/o KG)	73.1 (±0.5)	68.7 (±0.6)
RGCN (Schlichtkrull et al., 2018)	72.7 (±0.2)	68.4 (±0.7)
GconAttn (Wang et al., 2019)	72.6 (±0.4)	68.6 (±1.0)
KagNet (Lin et al., 2019)	73.5 (±0.2)	69.0 (±0.8)
RN (Santoro et al., 2017)	74.6 (±0.9)	69.1 (±0.2)
MHGRN (Feng et al., 2020)	74.5 (±0.1)	71.1 (±0.8)
QA-GNN (Yasunaga et al., 2021)	76.5 (±0.2)	73.4 (±0.9)
GREASELM (**Ours**)	**78.5** (±0.5)	**74.2** (±0.4)

Experimental Results

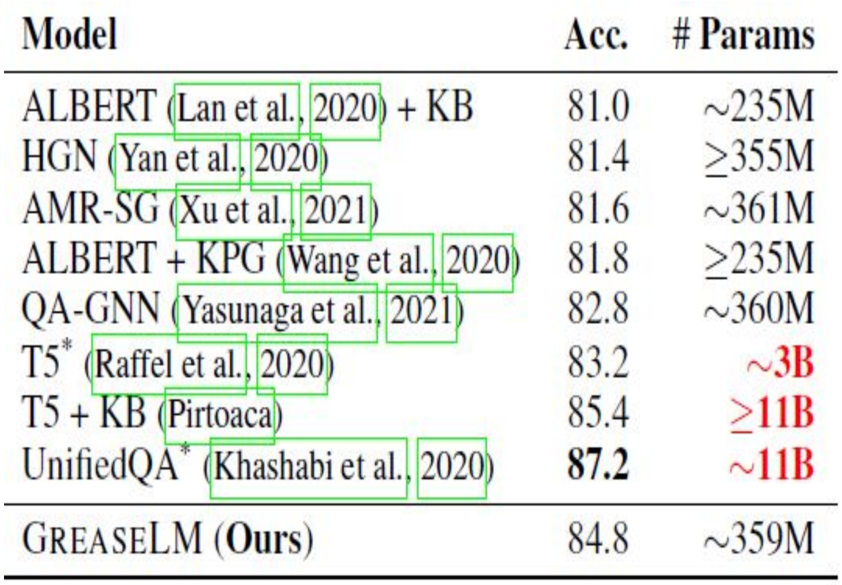
- **OpenbookQA**
  - test performance improves by 6.4% over fine-tuned LMs.
  - 2.0% improvement over existing LM+KG models.



Model	Acc.
AristoRoBERTa (no KG)	78.4
+ RGCN	74.6
+ GconAttn	71.8
+ RN	75.4
+ MHGRN	80.6
+ QA-GNN	82.8
GREASELM (**Ours**)	**84.8**

Experimental Results

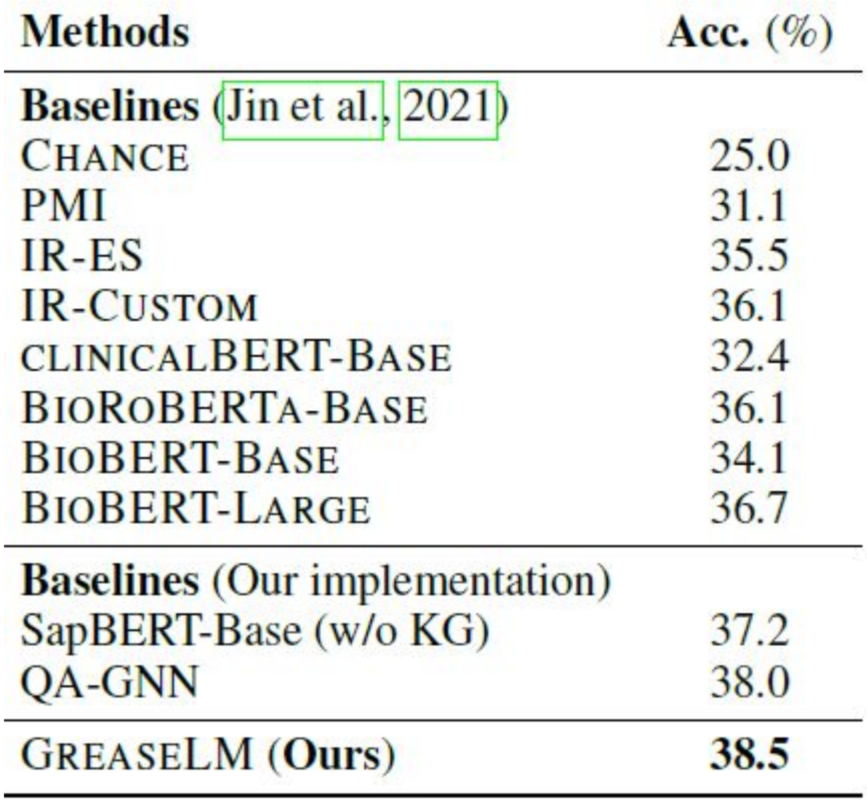
- **OpenbookQA (Public Leaderboard)**
  - competitive results to other systems on the leaderboard of OpenbookQA, posting the third highest score.
  - Parameter-wise 8x efficient.



Model	Acc.	# Params
ALBERT (Lan et al., 2020) + KB	81.0	∼235M
HGN (Yan et al., 2020)	81.4	≥355M
AMR-SG (Xu et al., 2021)	81.6	∼361M
ALBERT + KPG (Wang et al., 2020)	81.8	≥235M
QA-GNN (Yasunaga et al., 2021)	82.8	∼360M
T5* (Raffel et al., 2020)	83.2	∼3B
T5 + KB (Pirtoaca)	85.4	≥11B
UnifiedQA* (Khashabi et al., 2020)	**87.2**	∼11B
GreaseLM (**Ours**)	84.8	∼359M

Domain generality

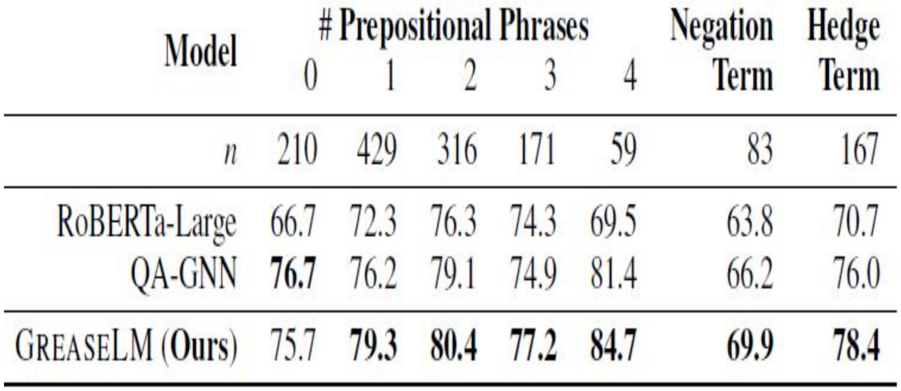
- **MedQA-USMLE**
  - Outperforms SOTA model (SapBERT).
  - LM-agnostic: improvements by GREASELM when it is seeded with other LMs, such as PubmedBERT and BioBERT.
  - GreaseLM is an effective augmentation of pretrained LMs for different domains.



Methods	**Acc.** (%)
Baselines (Jin et al., 2021)	
Chance	25.0
PMI	31.1
IR-ES	35.5
IR-Custom	36.1
ClinicalBERT-Base	32.4
BioRoBERTa-Base	36.1
BioBERT-Base	34.1
BioBERT-Large	36.7
Baselines (Our implementation)	
SapBERT-Base (w/o KG)	37.2
QA-GNN	38.0
GreaseLM (**Ours**)	**38.5**

Quantitative Analysis

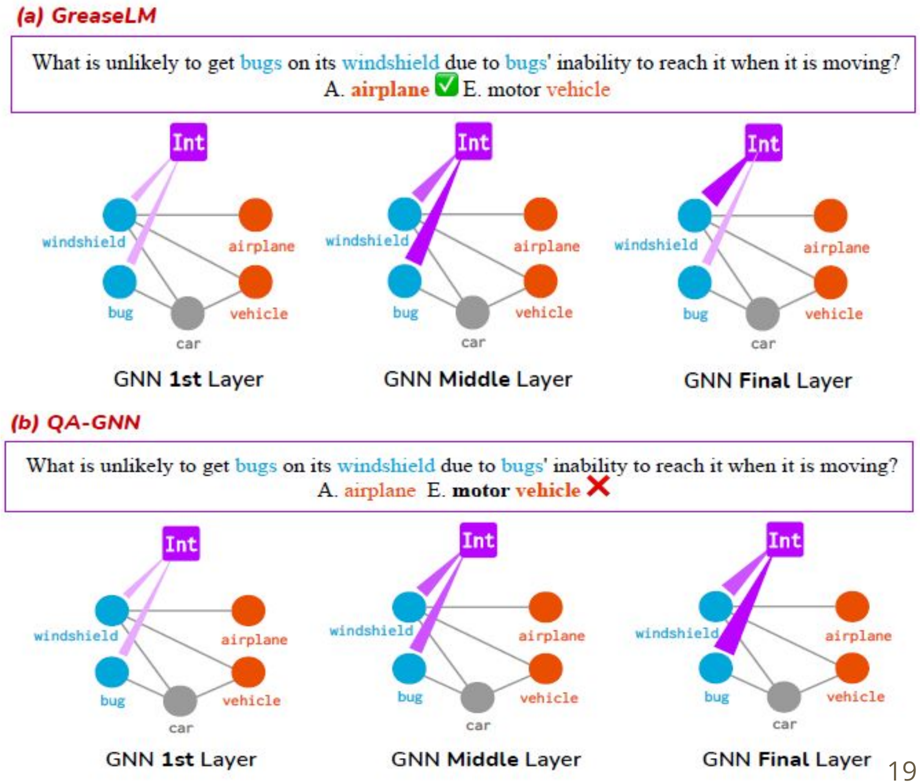
- **Reasoning over complex questions**
  - Number of prepositional phrases.
  - Explicit negation mentions (e.g., no, never).
  - Hedging terms indicating uncertainty (e.g. sometimes; maybe).
- GreaseLM generally outperform RoBERTa-Large and QA-GNN for both questions with negation terms and hedge terms.
- GREASELM performs better than the baselines across all questions with prepositional phrases.
- GreaseLM outperforms QA-GNN where the increasing complexity of questions requires deeper cross-modal fusion between language and knowledge representations.



Model	# Prepositional Phrases					Negation Term	Hedge Term
	0	1	2	3	4		
n	210	429	316	171	59	83	167
RoBERTa-Large	66.7	72.3	76.3	74.3	69.5	63.8	70.7
QA-GNN	**76.7**	76.2	79.1	74.9	81.4	66.2	76.0
GREASELM (Ours)	75.7	**79.3**	**80.4**	**77.2**	**84.7**	**69.9**	**78.4**

Qualitative Analysis

- GNN layers' node-node attention weights are analyzed to examine whether they reflect more expressive reasoning.
- In the example, GreaseLM correctly predicts that the answer is "airplane".
- The attention by the interaction node increases on the "bug" entity in the intermediate GNN layers, but drops again by the final layer, resembling a suitable intuition surrounding the hedge term "unlikely".



(a) GreaseLM

What is unlikely to get bugs on its windshield due to bugs' inability to reach it when it is moving?
A. **airplane** ✅ E. motor vehicle

(b) QA-GNN

What is unlikely to get bugs on its windshield due to bugs' inability to reach it when it is moving?
A. airplane E. **motor vehicle** ✗

Summary

- This work enables interactive fusion through joint information exchange between knowledge from language models and knowledge graphs.
- GreaseLM shows improved capability of modeling questions exhibiting textual nuances, such as negation and hedging.

References

- Zhang et al. “GreaseLM: Graph REASoning Enhanced Language Models for Question Answering”, ICLR 2022. Link: https://arxiv.org/abs/2201.08860
- Yasunaga et al. “QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering”, NAACL 2021. Link: https://arxiv.org/abs/2104.06378

Thank You