KNOWLEDGE REASONING EMPOWERED LANGUAGE MODEL (OREOLM)

IMPLEMENTATION

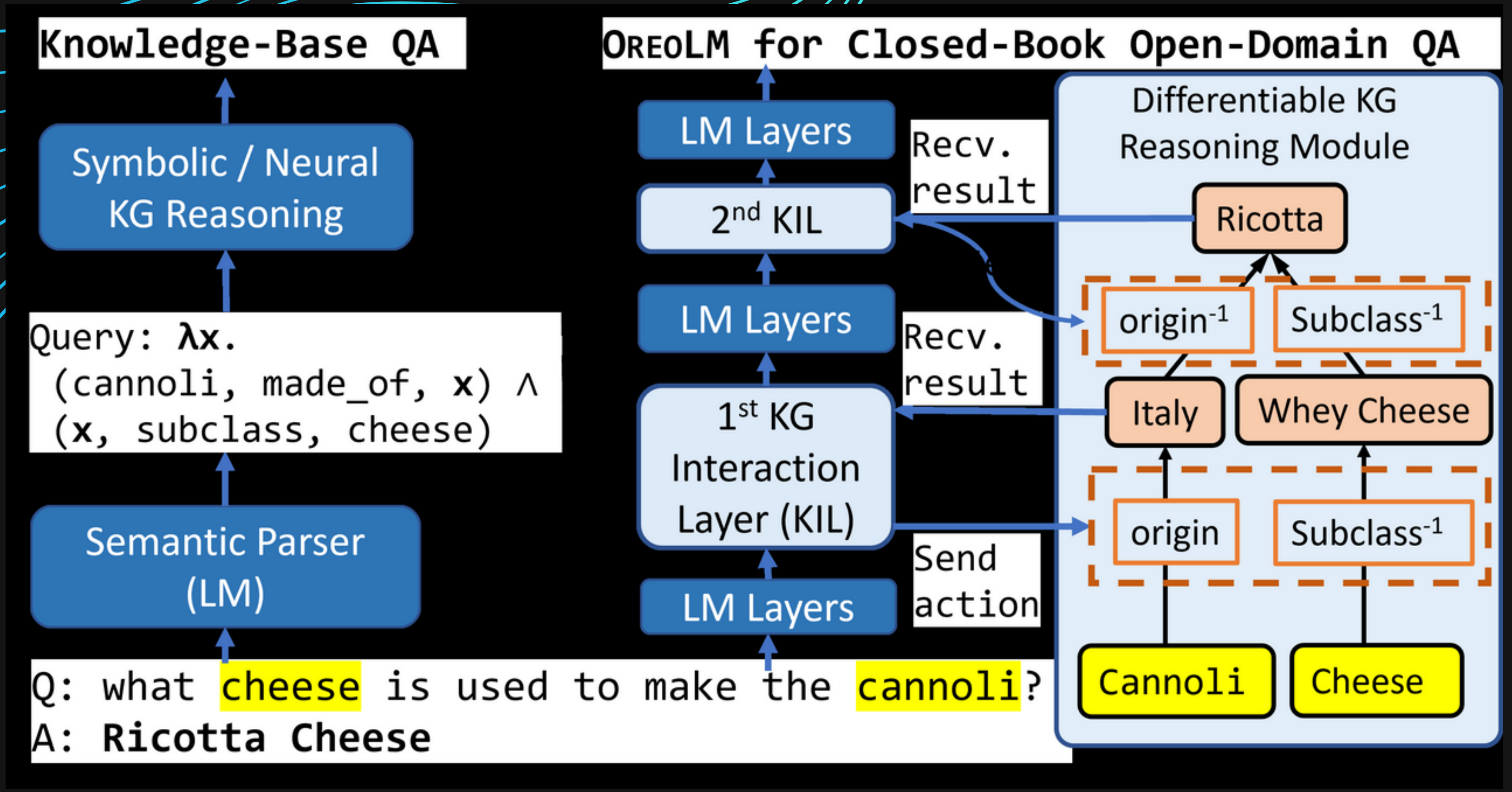

Knowledge-Base QA
Symbolic / Neural KG Reasoning
Query: λx. (cannoli, made_of, x) ∧ (x, subclass, cheese)
Semantic Parser (LM)
OREOLM for Closed-Book Open-Domain QA
LM Layers
Recv. result
2nd KIL
LM Layers
Recv. result
1st KG Interaction Layer (KIL)
Send action
LM Layers
Differentiable KG Reasoning Module
Ricotta
origin-1
Subclass-1
Italy
Whey Cheese
origin
Subclass-1
Cannoli
Cheese
Q: what cheese is used to make the cannoli?
A: Ricotta Cheese

Figure 1: An illustrative figure of OREOLM compared with previous KBQA.

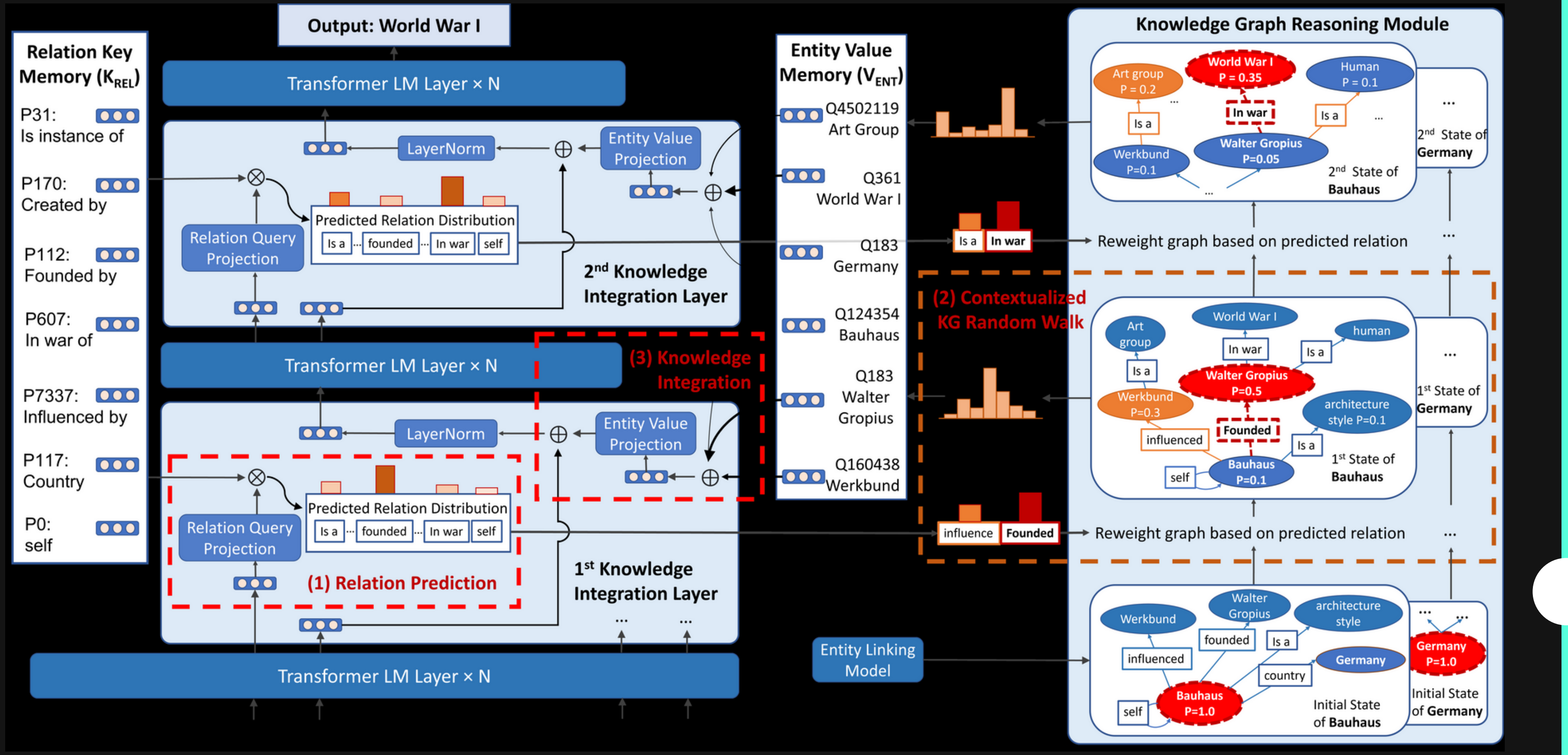


Figure 2: All the light blue locks are the added components to support KG reasoning, while the dark blue Transformer layers are knowledge-injected LM.

OreoLM can be applied to Transformer-based LMs to improve *Closed-Book* ODQA. The key component is the Knowledge Interaction Layers (KIL) inserted amid LM layers.
By stacking *T* layers of KIL, it can retrieve entities that are T-hop away from in-context entities.

QA problem can be decomposed into two steps:
1) KG Reasoning, which autoregressively walks through the graph to get a path *pi* starting from each entity mention *mi.* The relation path *pi* requires the selection of next entity at each step, so it decomposes into two further step: predict the next-hop relation based on the current state and context, and predict the next-hop entity based on the KG and the predicted relation;
2) knowledge-injected LM benefits from the reasoning paths to obtain the out-context knowledge for answer prediction.

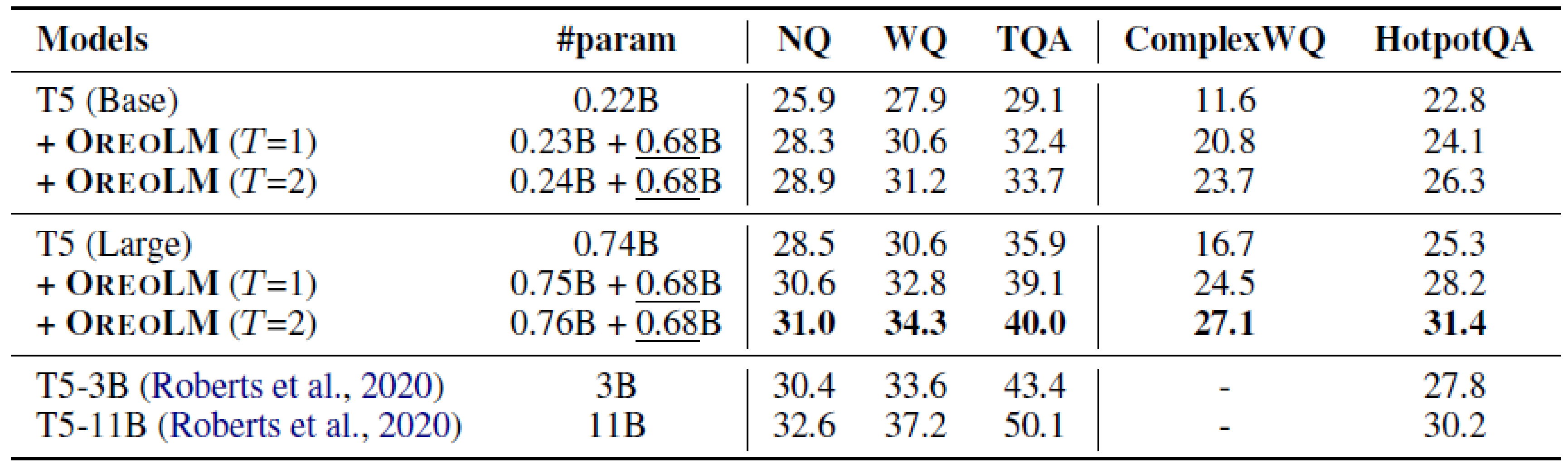
To avoid that KIL makes a random prediction, and the retrieved entities are likely to be unrelated to the question, there are two self-supervised guidance that helps: Entity Linking Loss and Weakly Supervised Relation Path Loss.

RESULTS

We evaluate OREOLM in *Closed-Book* QA, 3 single-hop ODQA (NQ, WQ, TQA) and 2 multi-hop QA (Complex WQ, HotpotQA)

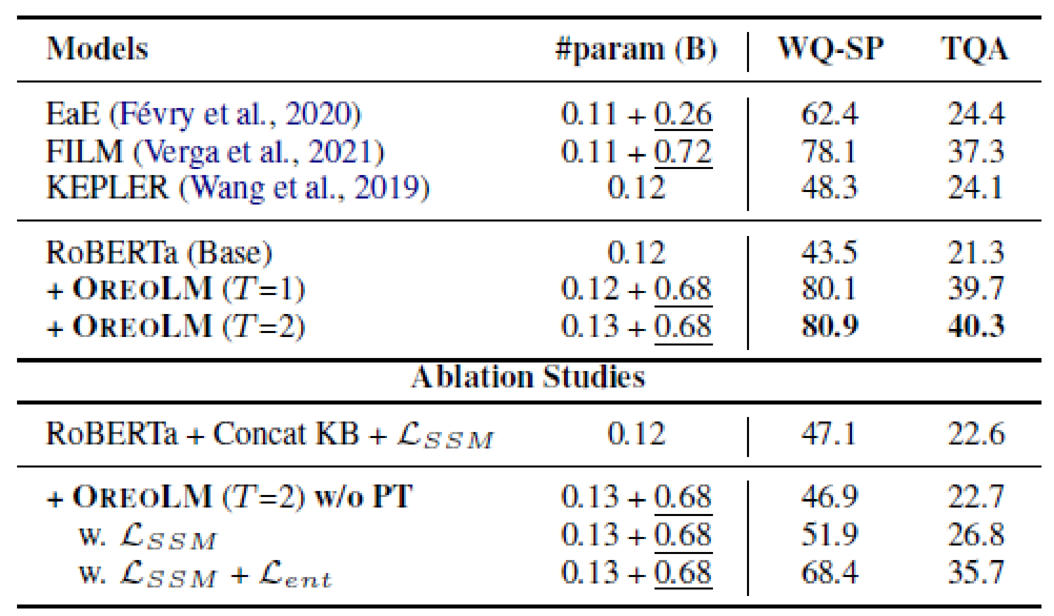
We evaluate it also in *Closed-Book Entity Prediction.* The evaluation dataset are WQ-SP and TQA and finetune the RoBERTa model augmented by OREOLM. We compare it with EaE, FILM and KEPLER.
There are some ablation studies: remove the KG, Lssm and Lent

Thought OREOLM is designed for Closed-Book QA, it can be used as backbone for Open-Book QA. We take Fid and DPR models as baseline. We can see OREOLM improve the results.
Some recent models dedicated for Open-Book QA are better, but we could use OREOLM to improve their performance.



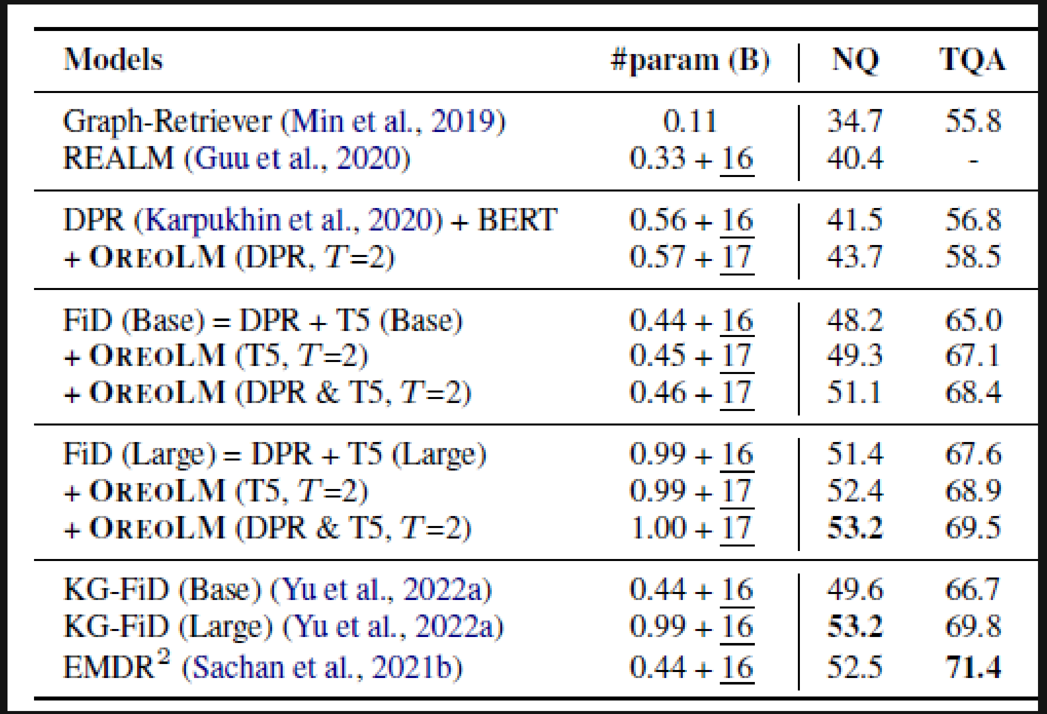
Models	#param	NQ	WQ	TQA	ComplexWQ	HotpotQA
T5 (Base)	0.22B	25.9	27.9	29.1	11.6	22.8
+ OREOLM (T=1)	0.23B + 0.68B	28.3	30.6	32.4	20.8	24.1
+ OREOLM (T=2)	0.24B + 0.68B	28.9	31.2	33.7	23.7	26.3
T5 (Large)	0.74B	28.5	30.6	35.9	16.7	25.3
+ OREOLM (T=1)	0.75B + 0.68B	30.6	32.8	39.1	24.5	28.2
+ OREOLM (T=2)	0.76B + 0.68B	**31.0**	**34.3**	**40.0**	**27.1**	**31.4**
T5-3B (Roberts et al., 2020)	3B	30.4	33.6	43.4	-	27.8
T5-11B (Roberts et al., 2020)	11B	32.6	37.2	50.1	-	30.2

Table 1: Closed-book QA



Models	#param (B)	WQ-SP	TQA
EaE (Févry et al., 2020)	0.11 + 0.26	62.4	24.4
FILM (Verga et al., 2021)	0.11 + 0.72	78.1	37.3
KEPLER (Wang et al., 2019)	0.12	48.3	24.1
RoBERTa (Base)	0.12	43.5	21.3
+ OREOLM (T=1)	0.12 + 0.68	80.1	39.7
+ OREOLM (T=2)	0.13 + 0.68	**80.9**	**40.3**
Ablation Studies			
RoBERTa + Concat KB + $\mathcal{L}_{SSM}$	0.12	47.1	22.6
+ OREOLM (T=2) w/o PT	0.13 + 0.68	46.9	22.7
w. $\mathcal{L}_{SSM}$	0.13 + 0.68	51.9	26.8
w. $\mathcal{L}_{SSM}$ + $\mathcal{L}_{ent}$	0.13 + 0.68	68.4	35.7

Table 2: Closed-book Entity Prediction



Models	#param (B)	NQ	TQA
Graph-Retriever (Min et al., 2019)	0.11	34.7	55.8
REALM (Guu et al., 2020)	0.33 + 16	40.4	-
DPR (Karpukhin et al., 2020) + BERT	0.56 + 16	41.5	56.8
+ OREOLM (DPR, T=2)	0.57 + 17	43.7	58.5
FiD (Base) = DPR + T5 (Base)	0.44 + 16	48.2	65.0
+ OREOLM (T5, T=2)	0.45 + 17	49.3	67.1
+ OREOLM (DPR & T5, T=2)	0.46 + 17	51.1	68.4
FiD (Large) = DPR + T5 (Large)	0.99 + 16	51.4	67.6
+ OREOLM (T5, T=2)	0.99 + 17	52.4	68.9
+ OREOLM (DPR & T5, T=2)	1.00 + 17	**53.2**	69.5
KG-FiD (Base) (Yu et al., 2022a)	0.44 + 16	49.6	66.7
KG-FiD (Large) (Yu et al., 2022a)	0.99 + 16	**53.2**	69.8
EMDR^2 (Sachan et al., 2021b)	0.44 + 16	52.5	**71.4**

Table 3: Open-book QA

LIMITATIONS

In the experiment, T=2 is better to T=1 on one-hop and multi-hop QA dataset.
But T=3 didn't get better results for the following reasons:
1) bigger T means more noise is included.
2) in Transformer model the representation space in lower and upper layer might be very different. Currently OREOLM adopt a MLP projection head, that maps integrated knowledge into the same space, but it might have many flaws.

Current design requires to learn a huge entity embedding table through additional supervision and could not directly fine-tune to downstream tasks.

It requires an additional step of entity linking for incoming questions, and then add special tokens as interface.

Though there are lots of potential reasoning tasks, it mainly focus on path-based relational reasoning, and it should not work for other reasoning tasks at current stage.

We assume that reasoning paths starting from different entities should be indipendent. This is not always correct, especially for questions that require logical reasoning, so the current methods might not work for those complex QA with logical dependencies.