Decker: Double Check with Heterogeneous Knowledge for Commonsense Fact Verification

Anni Zou, Zhuosheng Zhang, Hai Zhao*

annie0103@sjtu.edu.cn

Shanghai Jiao Tong University

Background

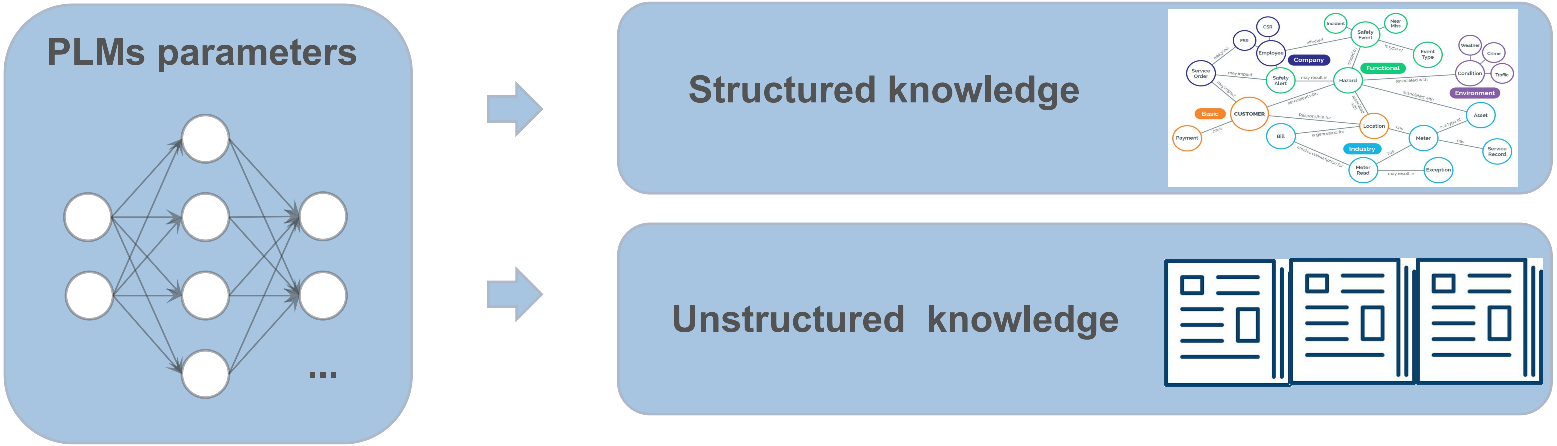
- **Commonsense fact verification**: verify through facts whether a given commonsense claim is correct or not
  - ➢ derive solely from question & implement reasoning on top of it
- Current Methods:
  - ➢ Direct use of knowledge preserved in ore-trained language models (**PLMs**) **parameters**
  - ➢ Resort to external knowledge bases, either **structured or unstructured knowledge**

Question	Answer	Dataset
You cannot be in Vienna and Paris at the same time.	Yes ✅	CSQA2.0
july always happens in the summer around the world?	No ❌	CSQA2.0
Carrots contain large amounts of vitamin A.	True ✅	CREAK
Humans cannot eat fennel because it's poisonous.	False ❌	CREAK

➡ 🤔 **How to effectively seize high-quality commonsense knowledge ?**

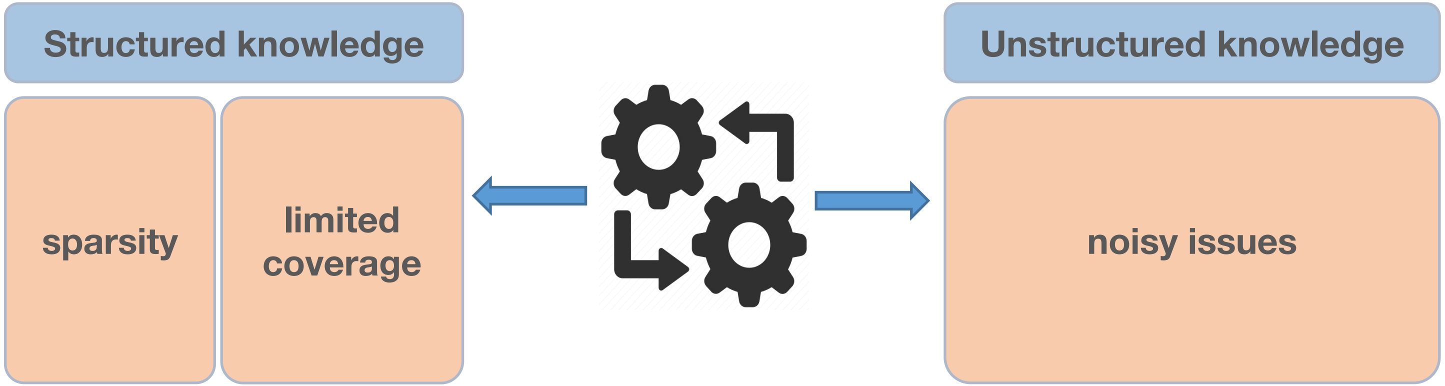
Background

- **Commonsense fact verification**: verify through facts whether a given commonsense claim is correct or not
  - ➢ derive solely from question & implement reasoning on top of it
- Current Methods:
  - ➢ Direct use of knowledge preserved in pre-trained language models **(PLMs) parameters**
  - ➢ Resort to external knowledge bases, either **structured or unstructured knowledge**



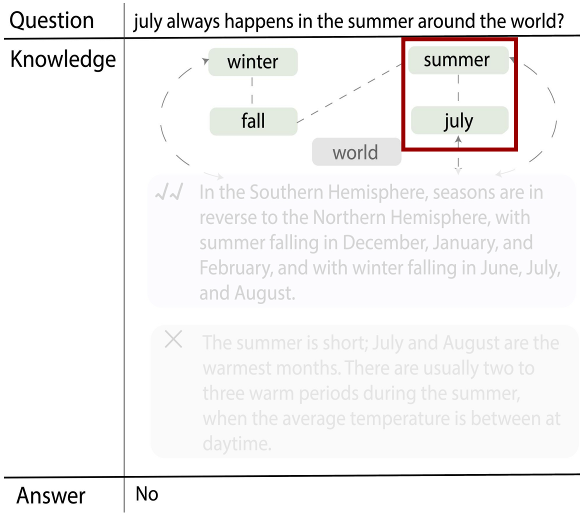
Challenges

- Previous methods heavily rely on **reasoning paths** between question and candidate answers
- Structured knowledge suffers from **sparsity and limited coverage**
- Unstructured knowledge undergoes **noisy issues**



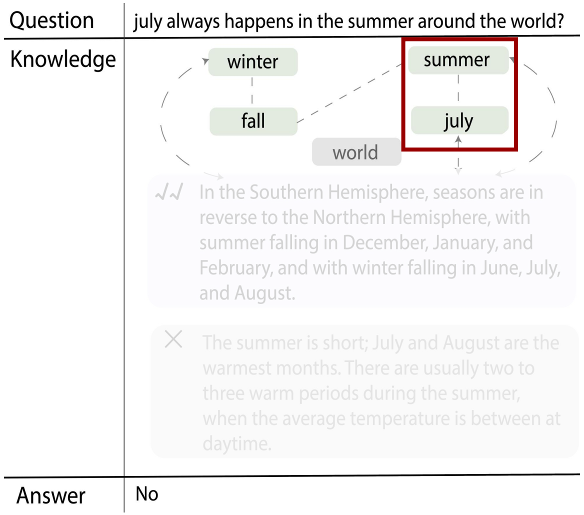
Challenges

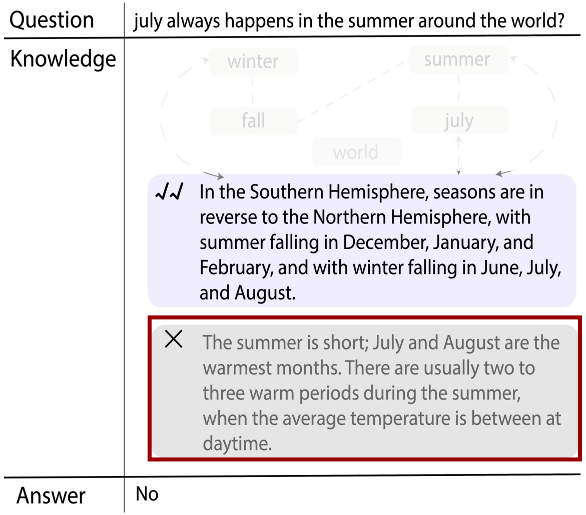
- Previous methods heavily rely on **reasoning paths** between question and candidate answers
- Structured knowledge suffers from **sparsity and limited coverage**
- Unstructured knowledge undergoes **noisy issues**



Challenges

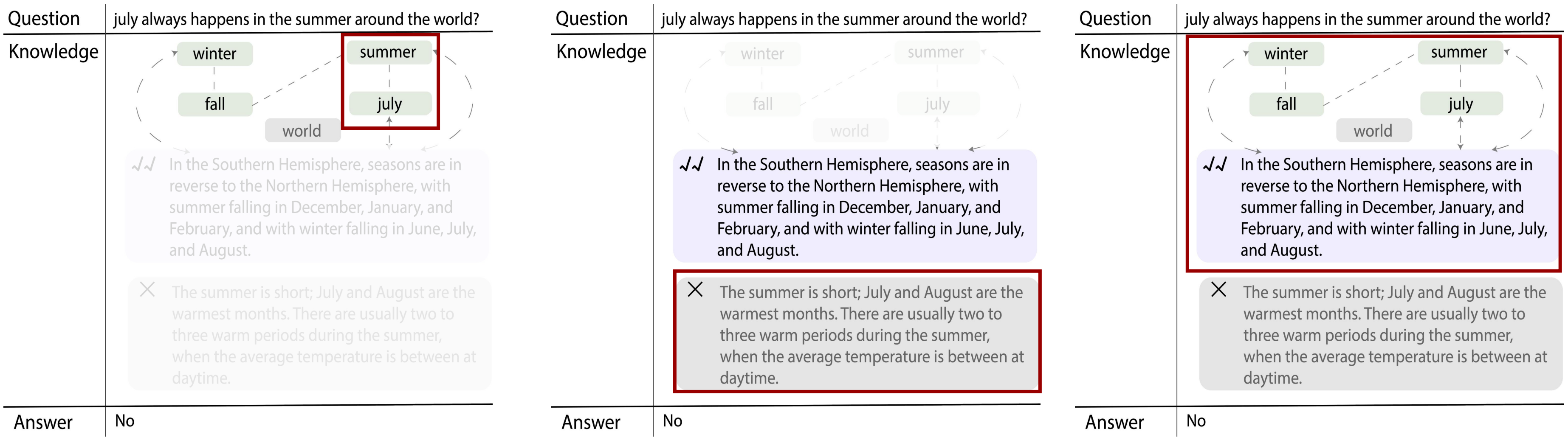
- Previous methods heavily rely on **reasoning paths** between question and candidate answers
- Structured knowledge suffers from **sparsity and limited coverage**
- Unstructured knowledge undergoes **noisy issues**





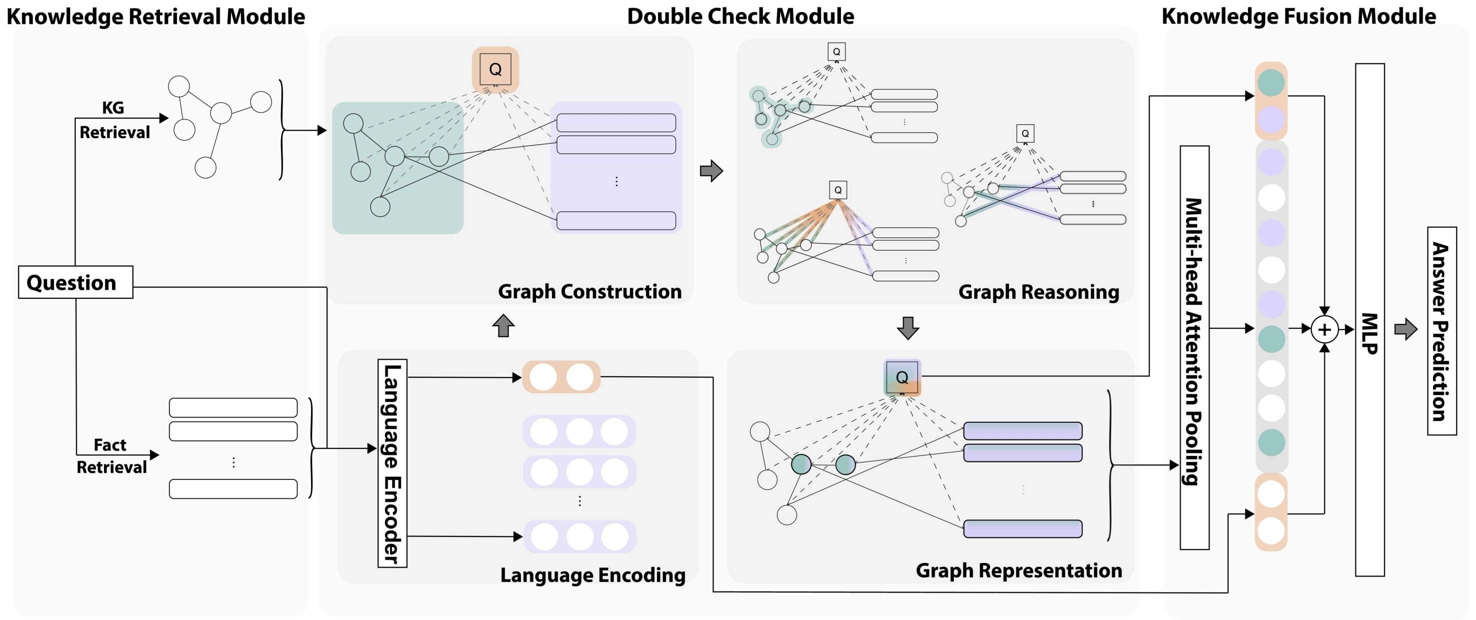
Challenges

- Previous methods heavily rely on **reasoning paths** between question and candidate answers
- Structured knowledge suffers from **sparsity and limited coverage**
- Unstructured knowledge undergoes **noisy issues**

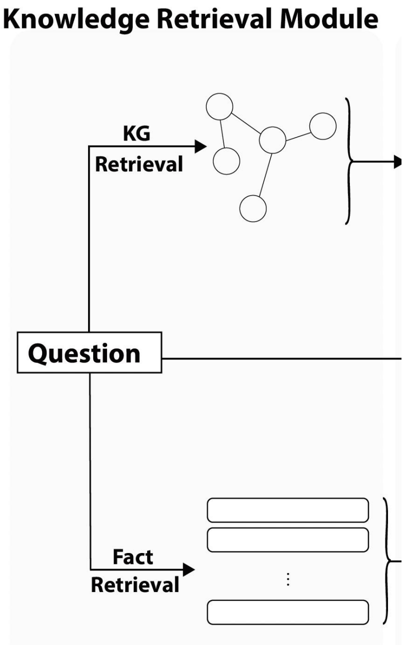


Decker: Overview

- **Knowledge Retrieval Module**: retrieve heterogeneous knowledge based on the input question
- **Double Check Module**: filter and make a double check over the heterogeneous knowledge
- **Knowledge Fusion Module**: obtain a refined knowledge representation and predict the final answer



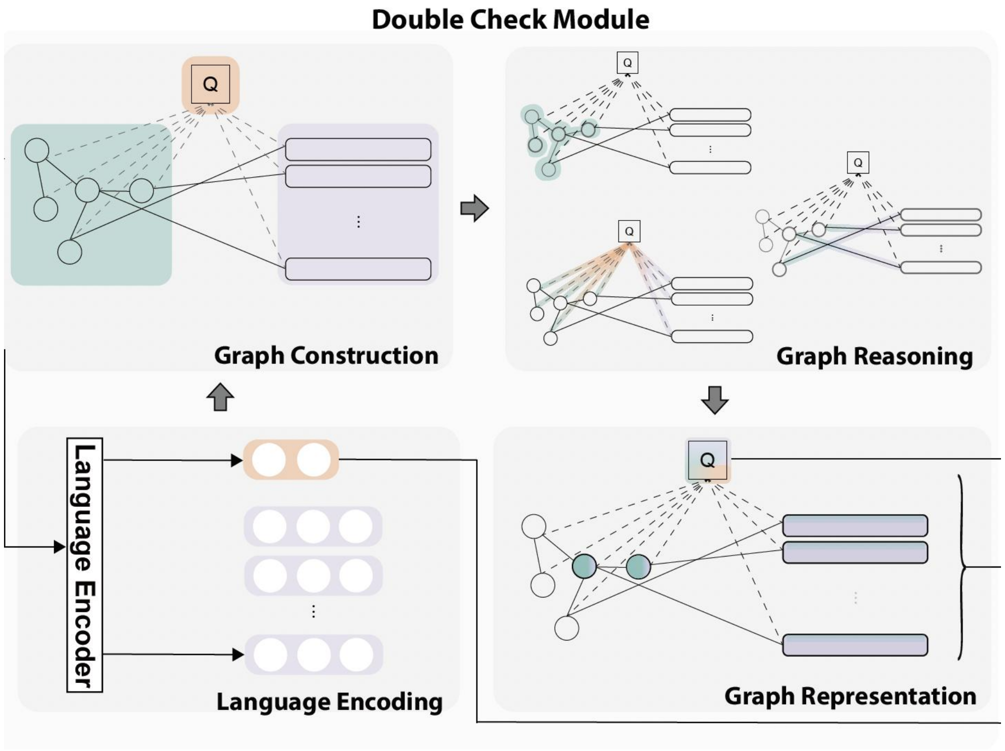
Decker: Knowledge Retrieval Module



- **KG Retriever**:
  - execute entity linking between the question and the pre-defined knowledge graph
  - add any bridge entities that are in a 2-hop path between any two linked entities
  - extract all the edges that join any two nodes
- **Fact Retriever**:
  - employ a pre-trained information retrieval model *Contriever*
  - calculate relevance scores between the question and candidate texts from the pre-defined corpus

G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave Unsupervised Dense Information Retrieval with Contrastive Learning

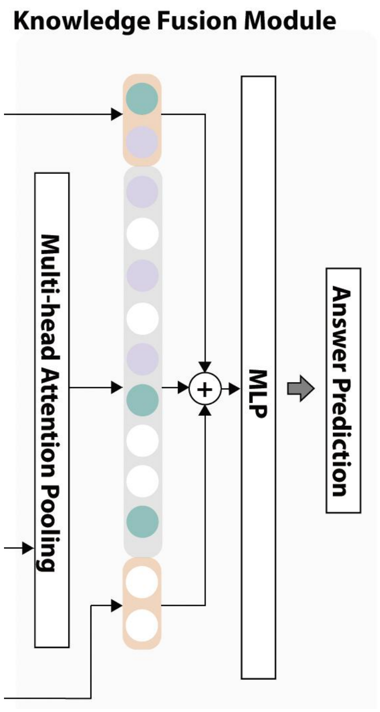
Decker: Double Check Module



- **Language Encoding**: employ a PLM to encode the input question and facts
- **Graph Construction**: construct an *integral graph*
  - four types of edges: concept-to-fact, concept-to-concept, question-to-fact, question-to-concept
  - initialize the node embeddings and align the dimension
- **Graph Reasoning**:
  - adopt relational graph convolutional network (R-GCN)

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$

Decker: Knowledge Fusion Module



- **Multi-head Attention Pooling**:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

$$\mathrm{head}_t = \mathrm{Attn}\left(H_q W_t^Q, H_k W_t^K, H_k W_t^V\right),$$

$$\mathrm{MHA}(H_q, H_k) = [\mathrm{head}_1, \ldots, \mathrm{head}_N]\, W^O,$$

- **Answer Prediction**:

$$l = \mathrm{MLP}\left([q_{enc}; K_a; q_{enc}^{(L)}]\right) \in \mathcal{R},$$

initial question embedding

pooled knowledge representation

enriched question representation

Datasets

- **CSQA2.0**:
  - collected through gamification
  - 14,343 assertions about everyday commonsense knowledge
  - train/dev/test: 9,282/2,544/ 517
- **CREAK**:
  - generated by crowdworkers based on a Wikipedia entity
  - 13,000 assertions about entity knowledge
  - encompass 2,700 entities
  - train/dev/test/contrast: 10,176/1,371/1,371/500

Main Results

- Decker **outperforms** the strong baselines and achieves **comparable** results on CREAK test set
- Decker **surpasses** the current state-of-the-art model on CREAK contrast set
- Decker **exceeds** the billion parameter-level model (3B) with only about **10% of the parameters** (449M)
- Decker enjoys a **lightweight** architecture without mixed data from multiple tasks during training

Model	#Total Params.	Single-task Training	CREAK Test	CREAK Contra	CSQA2.0 Test
Human [27]			-	92.2	-
GreaseLM [47]	~359M	✓	77.5	-	-
UNICORN [25]	~770M	✗	79.5	-	54.9
T5-3B [30]	~ 3B	✗	85.1	70.0	60.2
RACo [43]	≥ 3B	✗	**88.6**	74.4	61.8
Decker (**Ours**)	~449M	✓	88.4	**79.2**	**68.1**

Analysis

- **Combination of heterogeneous knowledge** and the components of Decker are both non-trivial.
- Augmented **interaction with the question** helps refine the enriched knowledge.

Model	Accuracy
DECKER	**89.5**
Knowledge Retrieval	
w/o facts	87.8(↓ 1.7)
w/o knowledge graph	87.9(↓ 1.6)
w/o both	86.1(↓ 3.4)
Graph Construction	
w/o question node	89.3(↓ 0.2)
w/o edge type	87.6(↓ 1.9)
w/o concept-to-fact edges	88.1(↓ 1.4)
w/o question-to-fact edges	88.8(↓ 0.7)
w/o concept-to-concept edges	88.3(↓ 1.2)
w/o question-to-concept edges	89.1(↓ 0.4)

Model	CSQA2.0	CREAK
DeBERTa_{large}	67.9	86.1
DECKER	70.2(↑ 2.3)	89.5(↑ 3.4)

Model	Interaction	Accuracy
$\text{DeBERTa}_{\text{LARGE}}$		86.1
w/ max pooling	✗	87.5
w/ mean pooling	✗	86.7
w/ attention pooling	✓	88.9
w/ MHA pooling	✓	**89.5**

Interpretability: Case Study

Question: Whales can breathe underwater?

1

F1: Some species such as the sperm whale are able to stay submerged for as much as 90 minutes. They have blowholes (modified nostrils) located on top of their heads, through which air is taken in and expelled.

F2: Beluga whales often accompany bowheads, for curiosity and to secure polynya feasible to breathe as bowheads are capable of breaking through ice from underwater by headbutting.

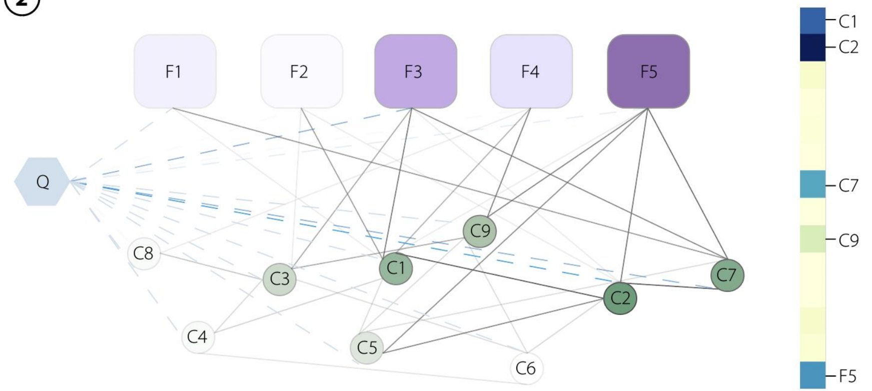
F3: Whales have evolved from land-living mammals. As such whales must breathe air regularly, although they can remain submerged underwater for long periods of time.

F4: Beluga whales swim on the surface between 5% and 10% of the time, while for the rest of the time they swim at a depth sufficient to cover their bodies. They do not jump out of the water like dolphins.

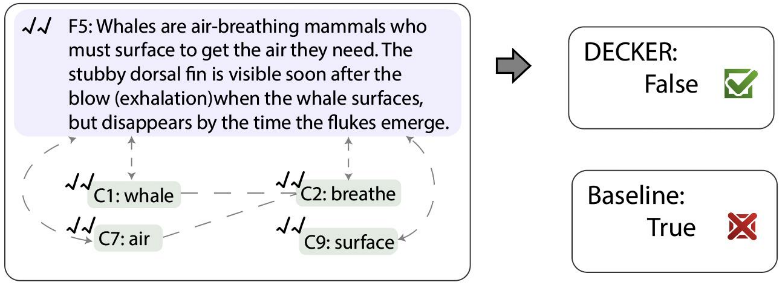
F5: Whales are air-breathing mammals who must surface to get the air they need. The stubby dorsal fin is visible soon after the blow (exhalation) when the whale surfaces, but disappears by the time the flukes emerge.

C1: whale	C2: breathe	C3: underwater
C4: water	C5: blow	C6: dive
C7: air	C8: swim	C9: surface

2



3



Summary

- **Contributions**
  - ✓ Decker **bridges the gap between heterogeneous knowledge** in an **effective** and **intuitive** pattern.
  - ✓ Decker enjoys its strength and superiority in various dimensions, including its **excellent performance, lightweight architecture**, and **favorable interpretability**.
- **Insights**
  - ✨ **Diversity** of knowledge is essential in boosting model capabilities.
  - 💡 **Refinement** of knowledge also plays a vital role.
  - 🤔 How to **merge and refine diverse knowledge in an effective way** remains to be further explored.
- **Sources**
  - Paper: https://arxiv.org/pdf/2305.05921.pdf
  - Code: https://github.com/Anni-Zou/Decker

Q & A

Thanks!