Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation

Mrigank Raman, **Aaron Chan***, Siddhant Agarwal*, Peifeng Wang, Hansen Wang, Sungchul Kim, Ryan Rossi, Handong Zhao, Nedim Lipka, Xiang Ren

INK Lab @ USC

USC University of Southern California

Adobe Research

KG-Augmented Neural Models

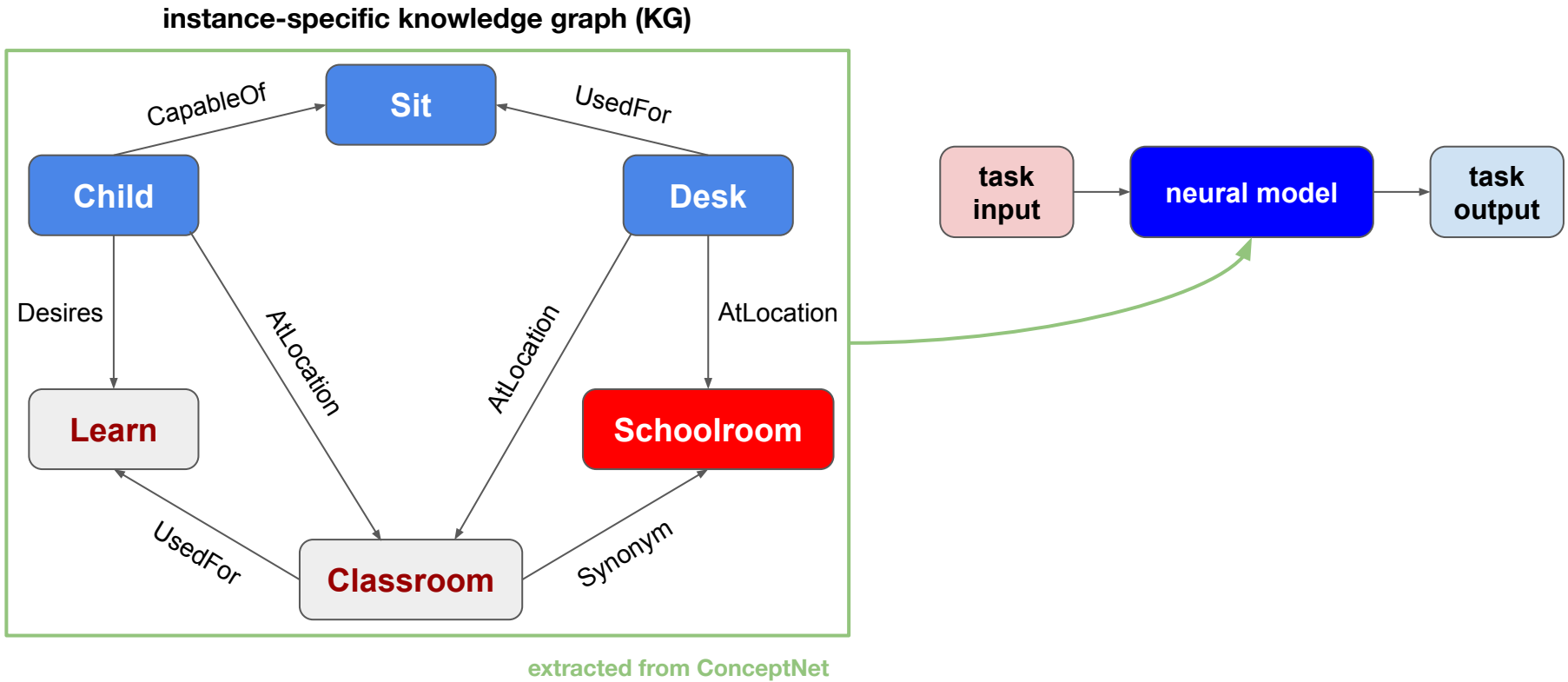
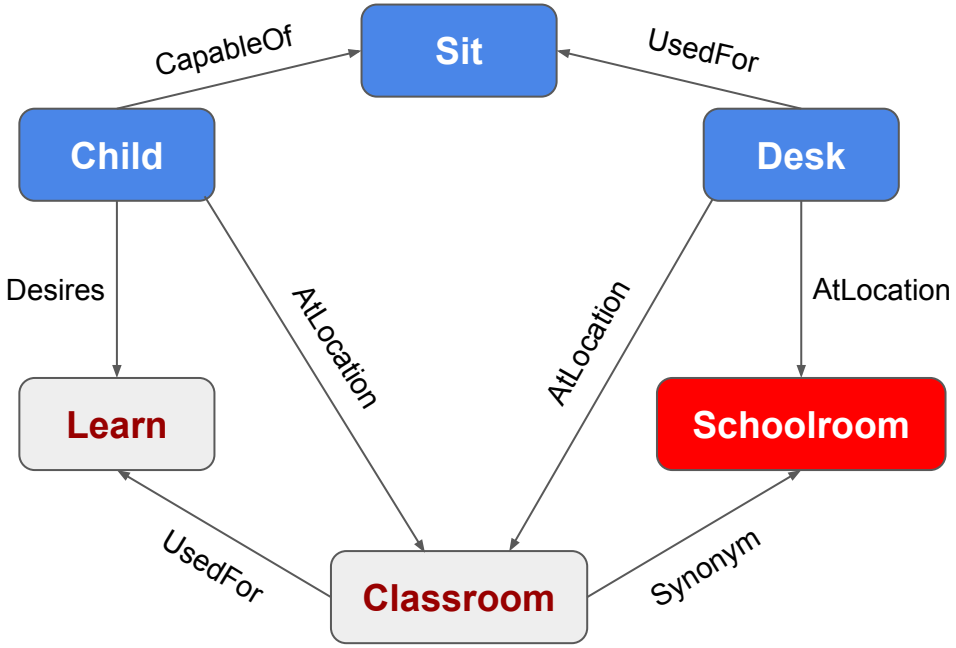


Figure Credit: Feng, et al., 2020

KG-Augmented **Commonsense QA**

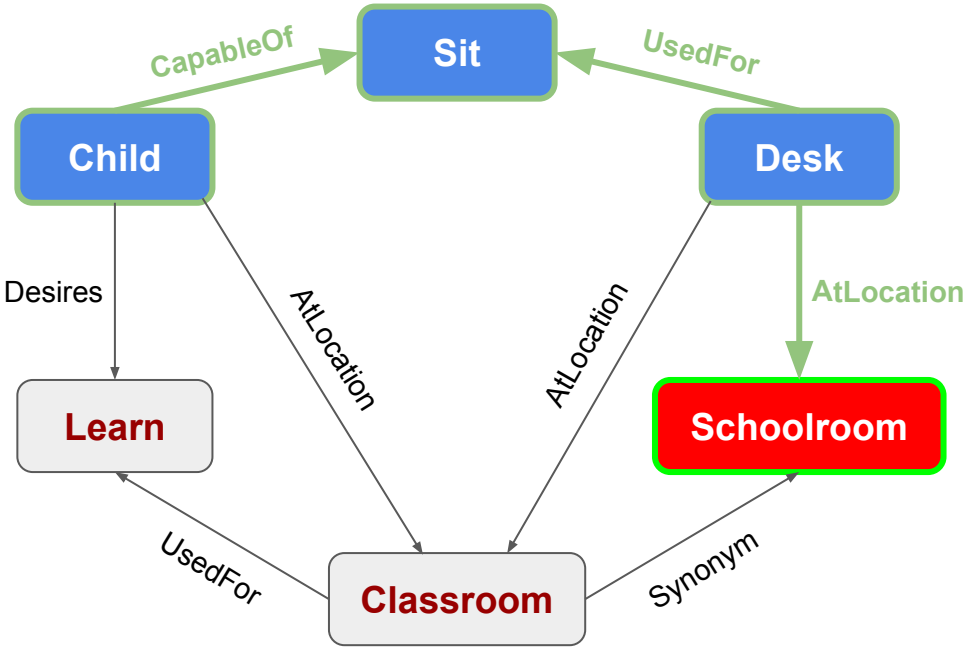


Where does a child likely sit at a desk?

A. Schoolroom
B. Furniture store
C. Patio
D. Office building
E. Library

Figure Credit: Feng, et al., 2020

KG-Augmented **Commonsense QA**



Where does a child likely sit at a desk?

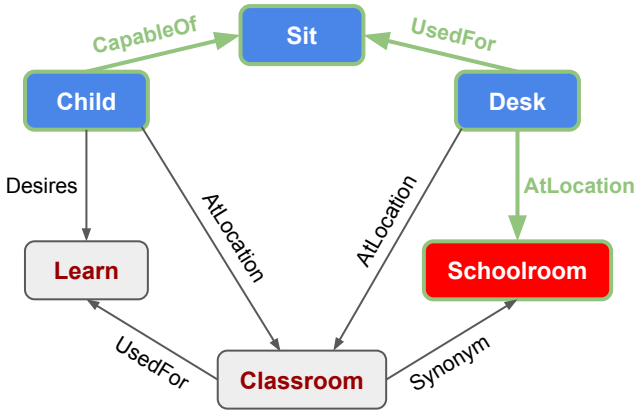
A. Schoolroom
B. Furniture store
C. Patio
D. Office building
E. Library

Figure Credit: Feng, et al., 2020

Question:
Do KG-augmented models use KG info in a way that makes sense to humans?

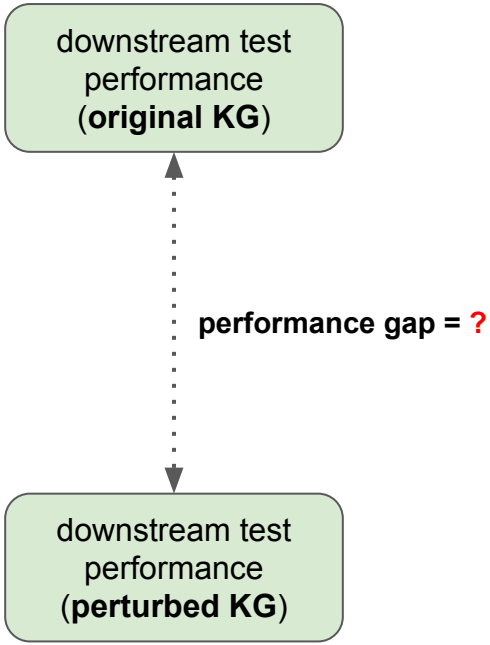
Where does a child likely sit at a desk?

A. Schoolroom
B. Furniture store
C. Patio
D. Office building
E. Library

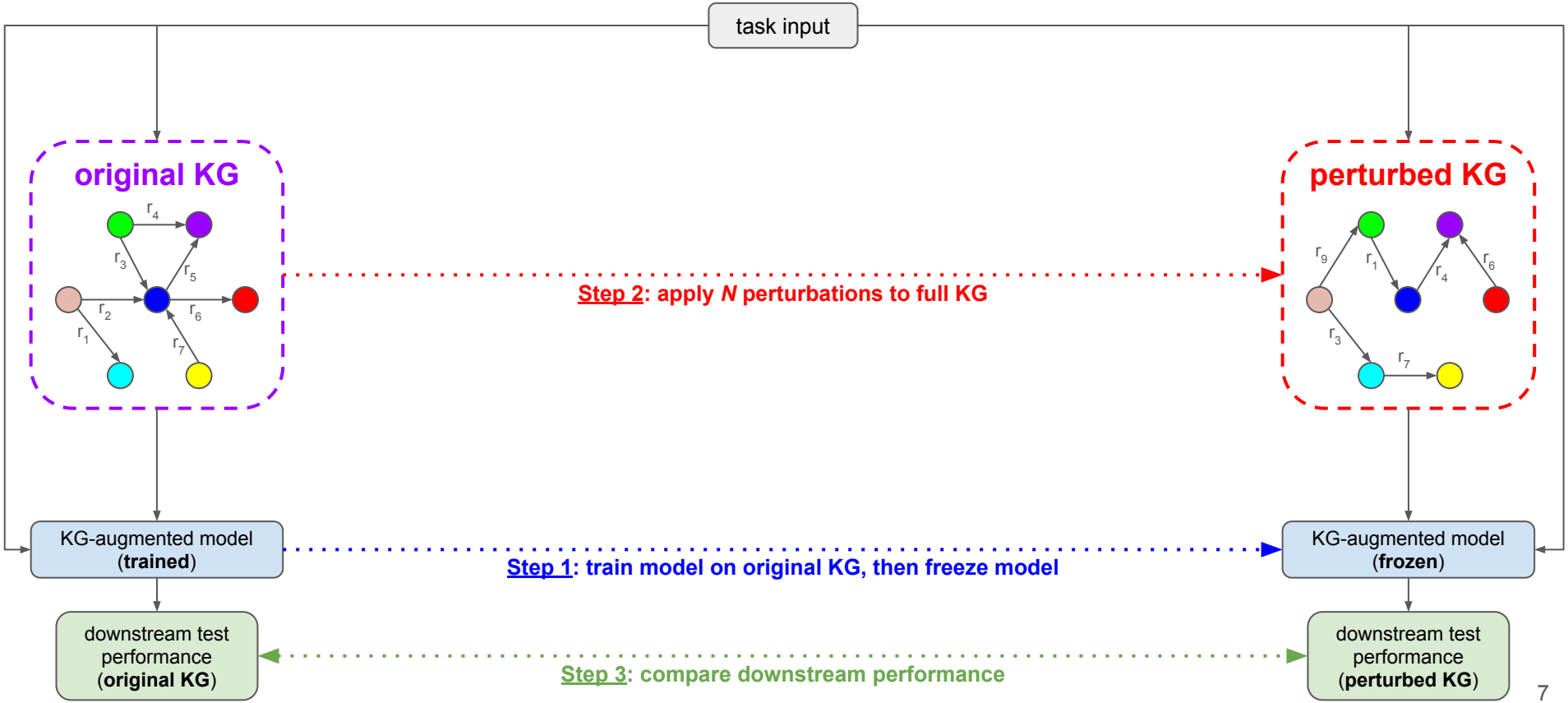


Raman et al., *Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation*, ICLR 2021

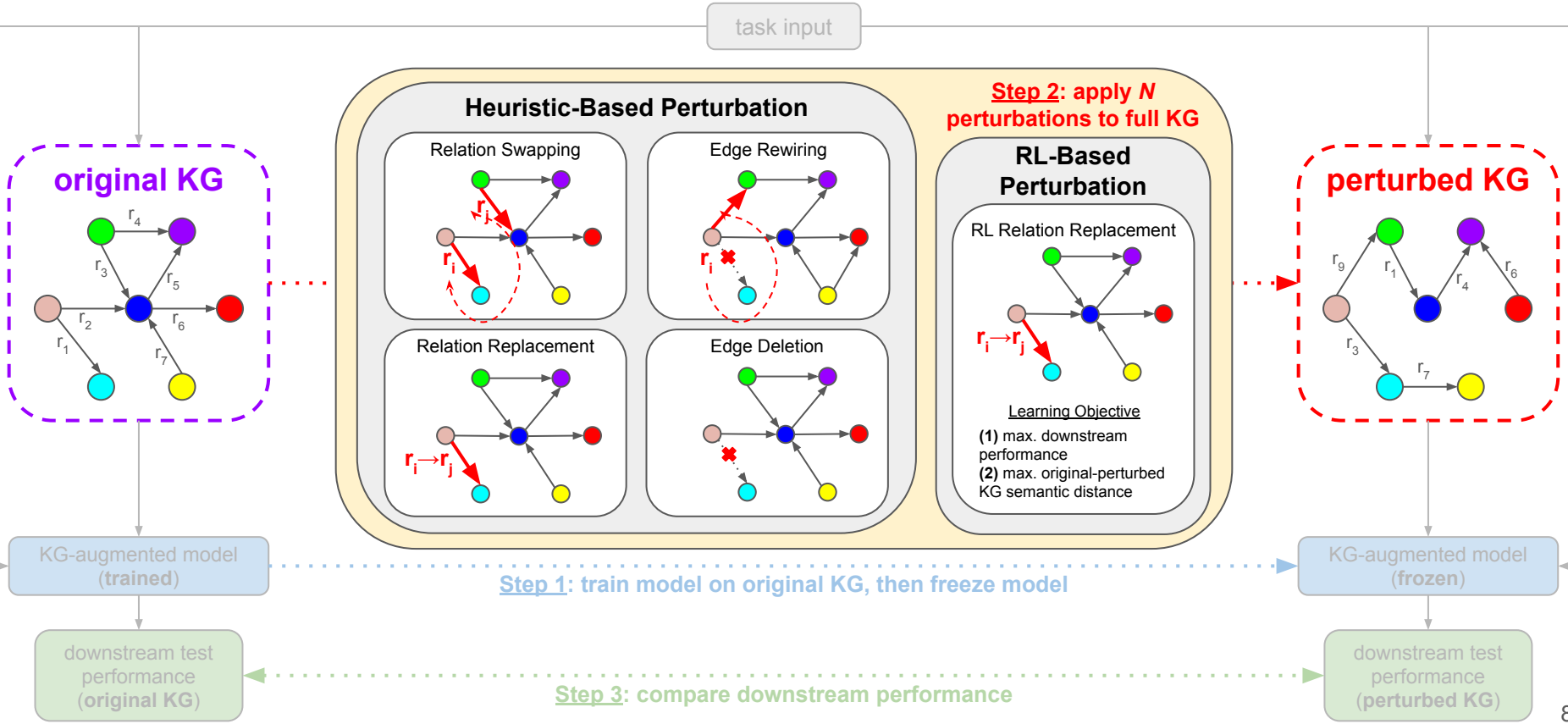
Experiment 1: Measure how well KG-augmented models perform using perturbed KGs.



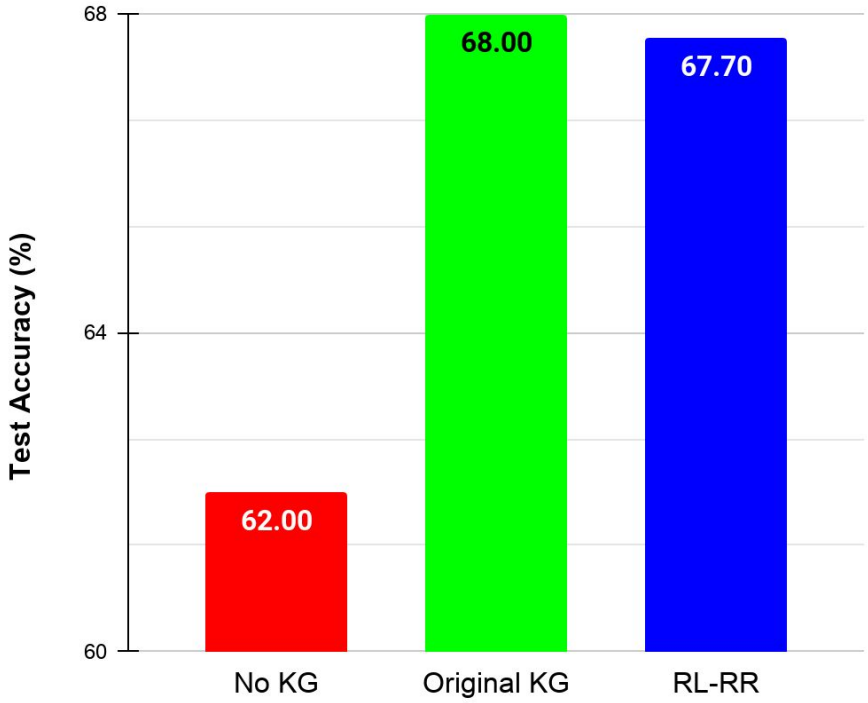
Problem Setting



KG Perturbation Methods



Results: **Commonsense QA**



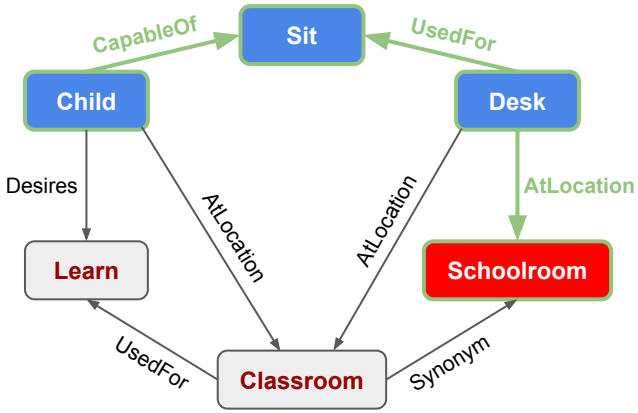
Performance on the **OBQA** dataset across various perturbation methods, using the **MHGRN** model (Feng, et al., 2020).

Surprisingly, RL-RR performs roughly as well as Original KG!

Experiment 2:
Ask humans to rate KG explanation paths that KG-augmented models found helpful.

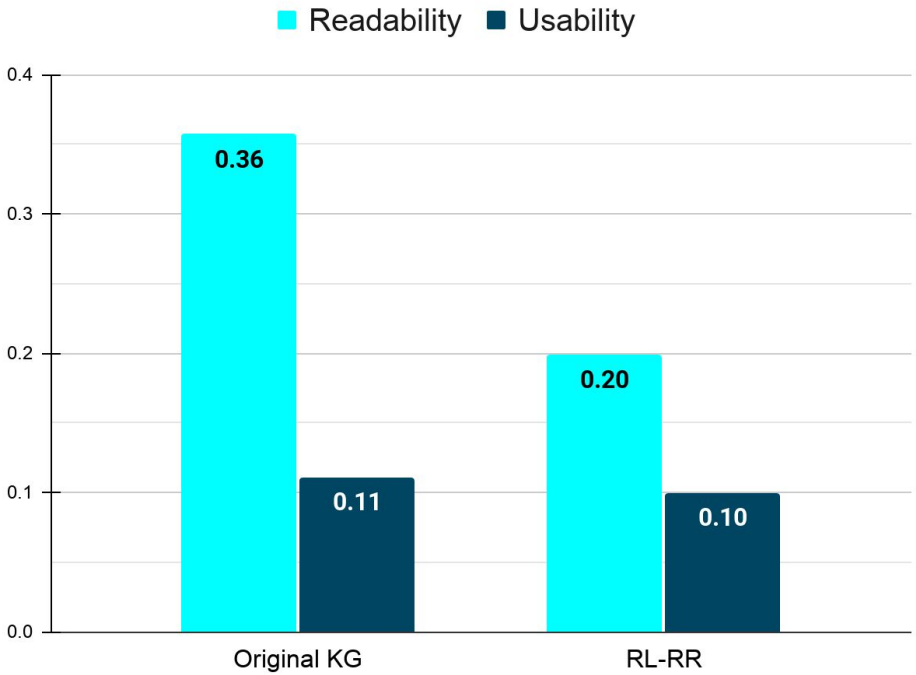
Where does a child likely sit at a desk?

A. **Schoolroom**
B. Furniture store
C. Patio
D. Office building
E. Library



Raman et al., *Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation*, ICLR 2021

Results: **Human Evaluation**



Readability/Usability of *top-scoring* paths from original KG and RL-RR, using **MHGRN** on **OBQA**.

Both readability and usability are measured on a **[0, 1] scale**.

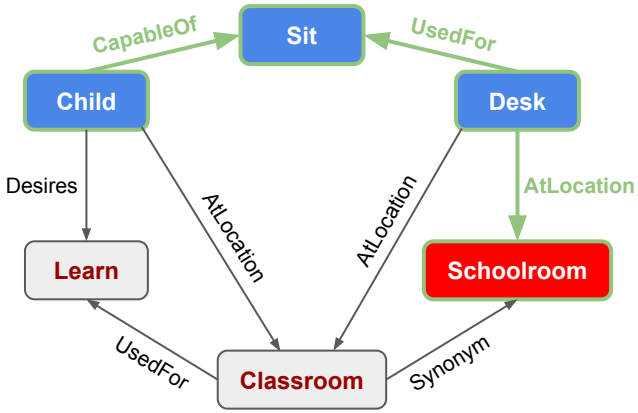
Humans **struggle to read/use** explanation paths that were helpful for MHGRN.

Conclusion:

No, KG-augmented models do *not* use KG info in a way that makes sense to humans.

Where does a child likely sit at a desk?

A. Schoolroom
B. Furniture store
C. Patio
D. Office building
E. Library



Raman et al., *Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation*, ICLR 2021

Future Work

Future Work

- Further analyze **how existing KG-augmented models use KG info**

Future Work

- Further analyze **how existing KG-augmented models use KG info**

- Design KG-augmented models that **use KG info more effectively** for downstream tasks

Future Work

- Further analyze **how existing KG-augmented models use KG info**
- Design KG-augmented models that **use KG info more effectively** for downstream tasks
- Design KG-augmented models that **produce KG explanations** which are more:
  - **plausible**: convincing to *humans*
  - **faithful**: reflective of the *model's reasoning process*

Thank You!