VIS 2023

CommonsenseVIS: Visualizing and Understanding Commonsense Reasoning Capabilities of Natural Language Models

Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, Huamin Qu

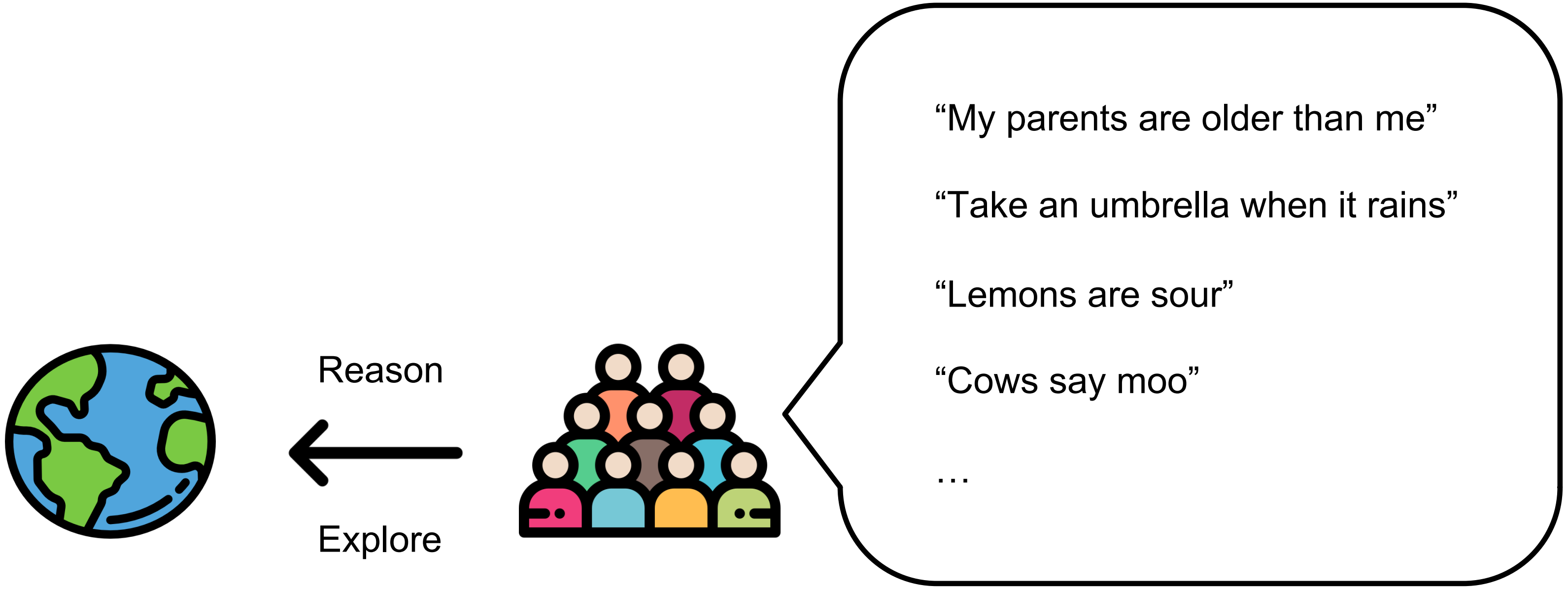
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Cornell University

Introduction

Commonsense knowledge and reasoning

Commonsense knowledge describes the general facts and beliefs about the world that are obvious and intuitive to most humans

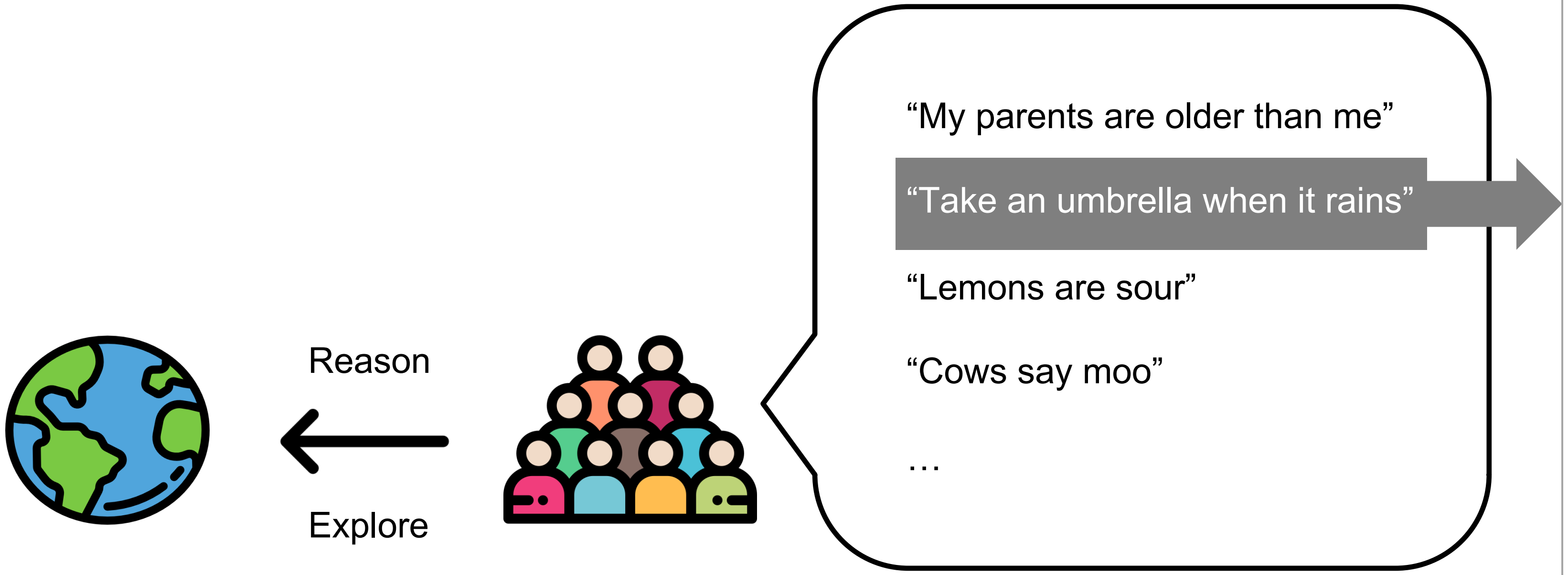


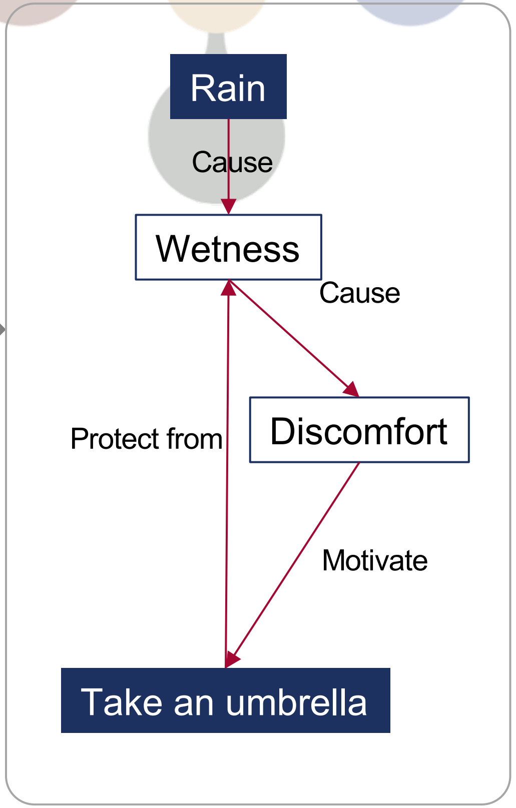
Introduction

Commonsense knowledge and reasoning

Commonsense knowledge describes the general facts and beliefs about the world that are obvious and intuitive to most humans

- It can be generally presented as **graphs**

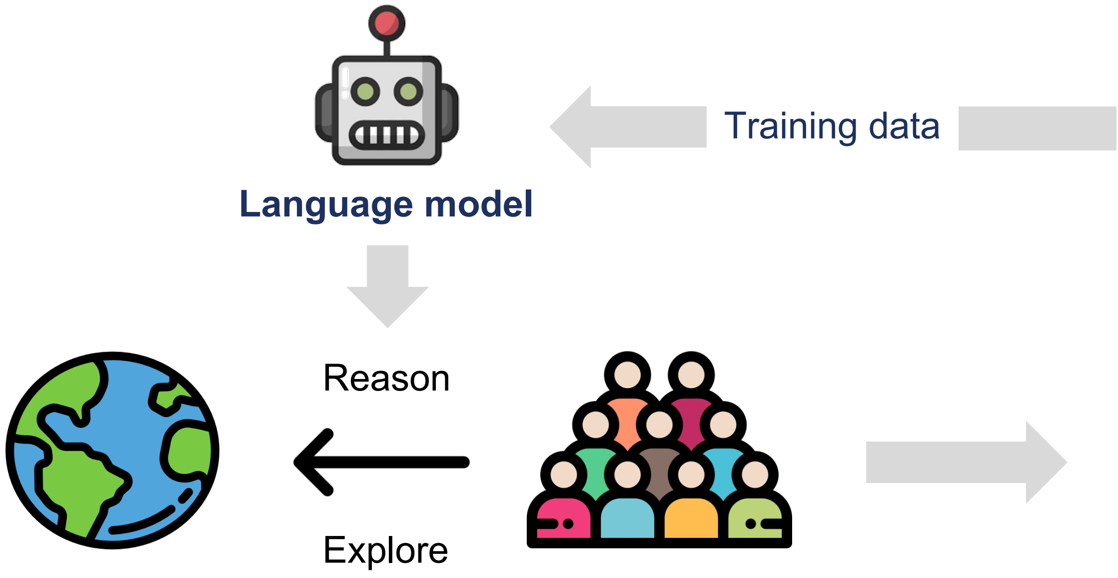




Introduction

Commonsense knowledge and reasoning

Equipping machines with humanlike commonsense reasoning abilities is a long-standing challenging topic in NLP. Researchers have constructed commonsense QA benchmarks for developing language models



Social IQA

In the school play, Robin played a hero in the struggle to the death with the angry villain. How would others feel afterwards?

A. sorry for the villain
B. Hopeful that Robin will succeed
C. Like Robin should lose

Commonsense QA

Where on a river can you hold a cup upright to catch water on a sunny day

A. waterfall, B. bridge, C. valley, D. pebble, E. mountain

SWAG

On stage, a woman takes a seat at the piano. She

A. sits on a bench as her sister plays with the doll.
B. smiles with someone as the music plays.
C. is in the crowd, watching the dancers.
D. nervously sets her fingers on the keys.

Introduction

Commonsense knowledge and reasoning

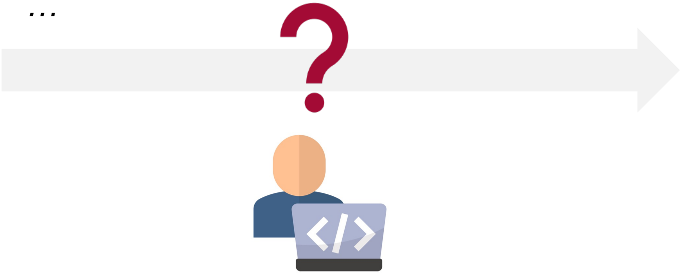
Current language models with billions of parameters achieve impressive results on commonsense benchmarks. However, they lack interpretability and transparency, which hinders model debugging, development, and deployment

Language models

GPT-4
Llama
GPT-3
PaLM
ERNIE
FLAN
ELMO
BERT
LaMDA
RoBERTa
Megatron-LM
T5

- *Do LMs know properties of a concept?*
- *Do LMs merely explore spurious correlation?*

…



NLP experts

Commonsense benchmarks

CommonsenseQA
SWAG
Com2Sense
hellaSWAG
Winograd Schema Challenge (WSC)
Social IQA

What **commonsense knowledge** language models have **learned and used** in the reasoning process?

Motivation

Understanding language models' reasoning process

Dataset	**Acc.**
A	73%
B	68%
C	70%

Input
↓
Model → Auxiliary classifier → POS, dependency relations
↓
Output

What is the relations between A & B
↓
Model

Zero/few-shot accuracy
Lack detailed analysis

Auxiliary classifier
Focus on linguistic knowledge

Direct prompting
Many NLP models cannot be easily prompted; subject to hallucinations

Motivation

Understanding language models' reasoning process

Feature attributions quantify the importance of input features (e.g., words and phrases) to the model outputs

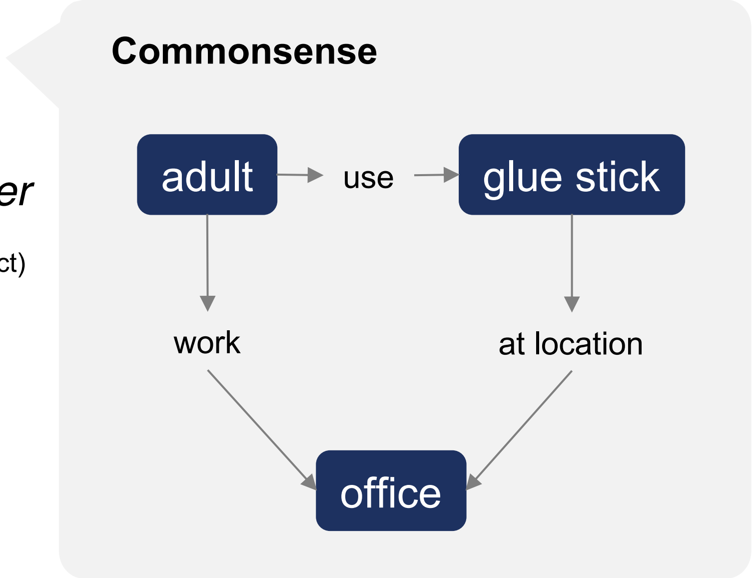
Input:

Where do adults use glue sticks?
A. classroom; B. office; C. desk drawer

(**Red**: positive impact)

Output:

office



Limitations

1. Cannot reveal models' relational reasoning over concepts in different contexts

*Q1: do LMs know the **relations** between **adults** and **office***

*Q2: do LMs know the **context** of using **glue sticks***

Motivation

Understanding language models’ reasoning process

Feature attributions quantify the importance of input features (e.g., words and phrases) to the model outputs

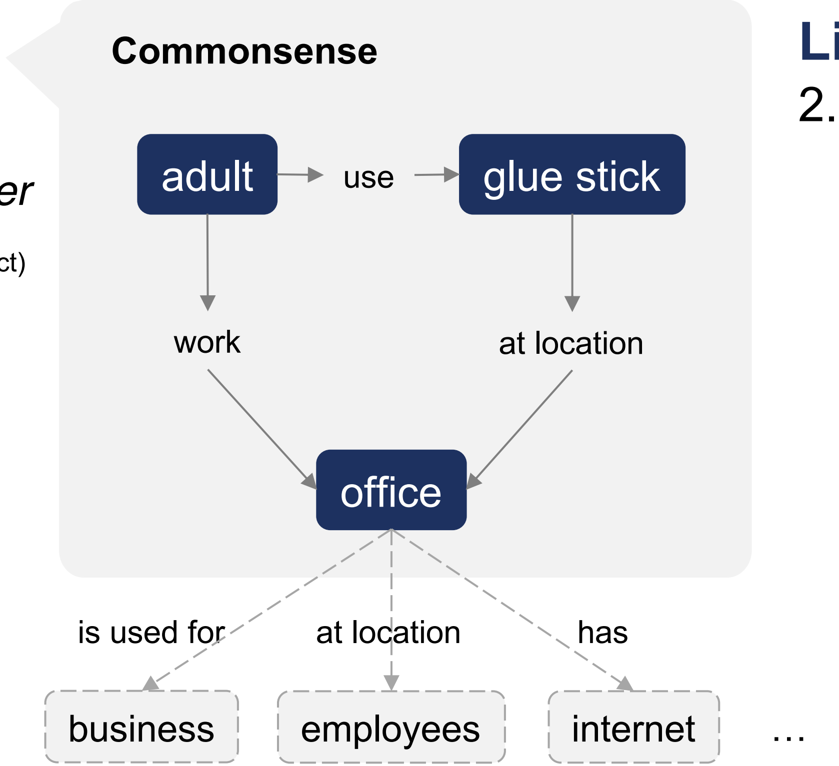
Input:

Where do adults use glue sticks?
A. classroom; B. office; C. desk drawer

(**Red**: positive impact)

Output:

office



Limitations

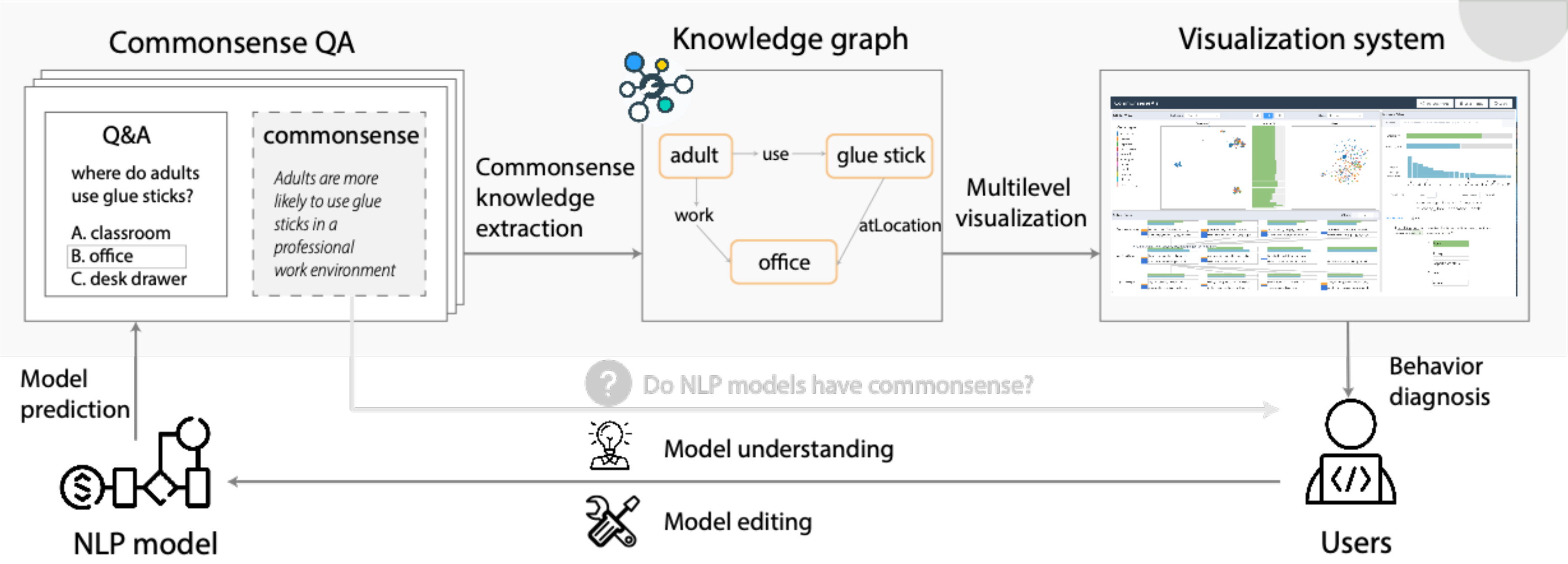
2. Difficult to scale up to efficiently support high-level abstractions of model behavior

Complexity and vastness of commonsense knowledge space

Our solution

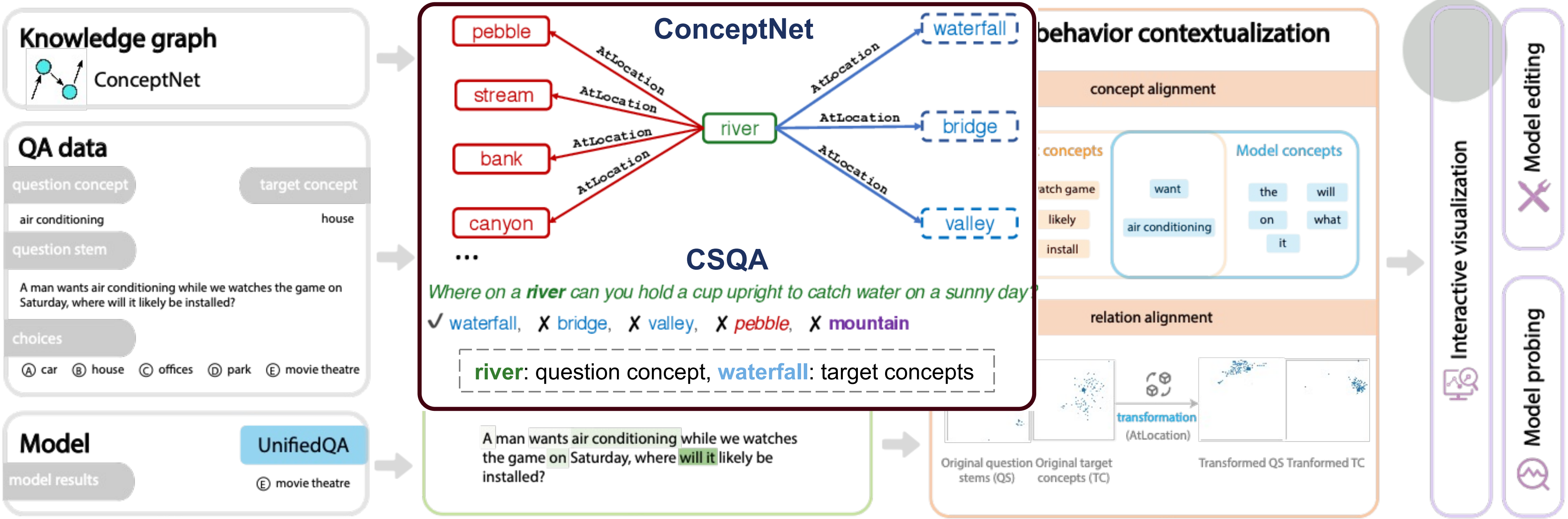
Model contextualization via knowledge graph

We employ a **knowledge graph** to derive implicit commonsense in QA instances as contexts. Then, we use it to align model behavior with human reasoning through **multi-level interactive visualizations**.



CommonsenseVIS
System overview

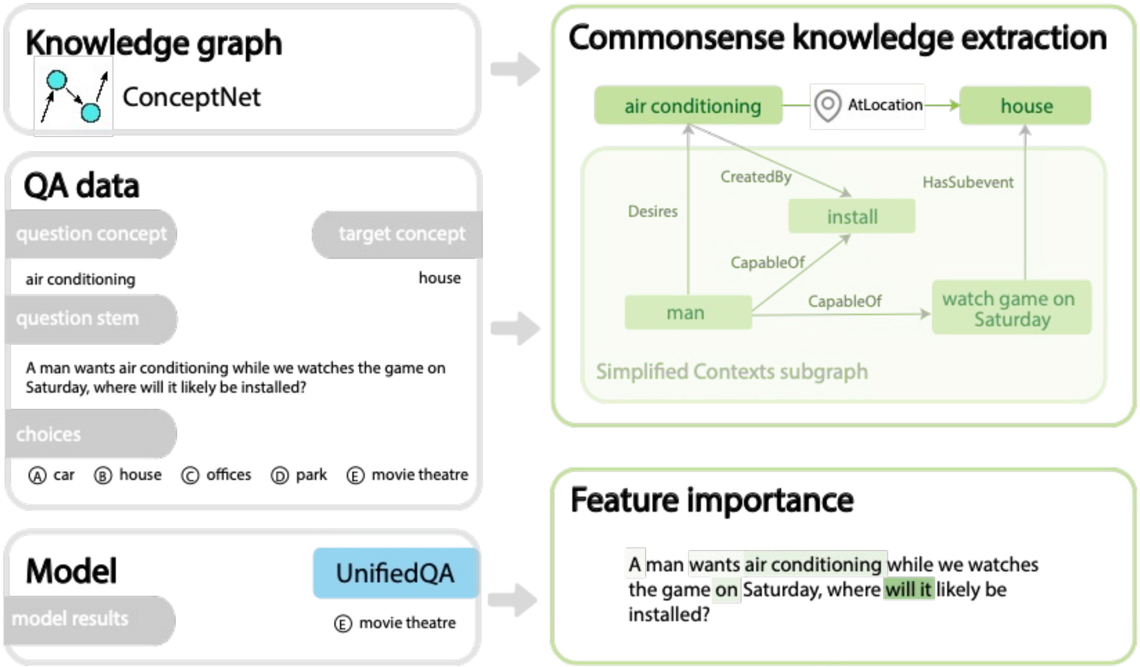
We use ConceptNet to contextualize model behavior on concepts and relations in commonsense QA (CSQA) dataset



CommonsenseVIS

System overview

Step 1: extract relevant commonsense knowledge



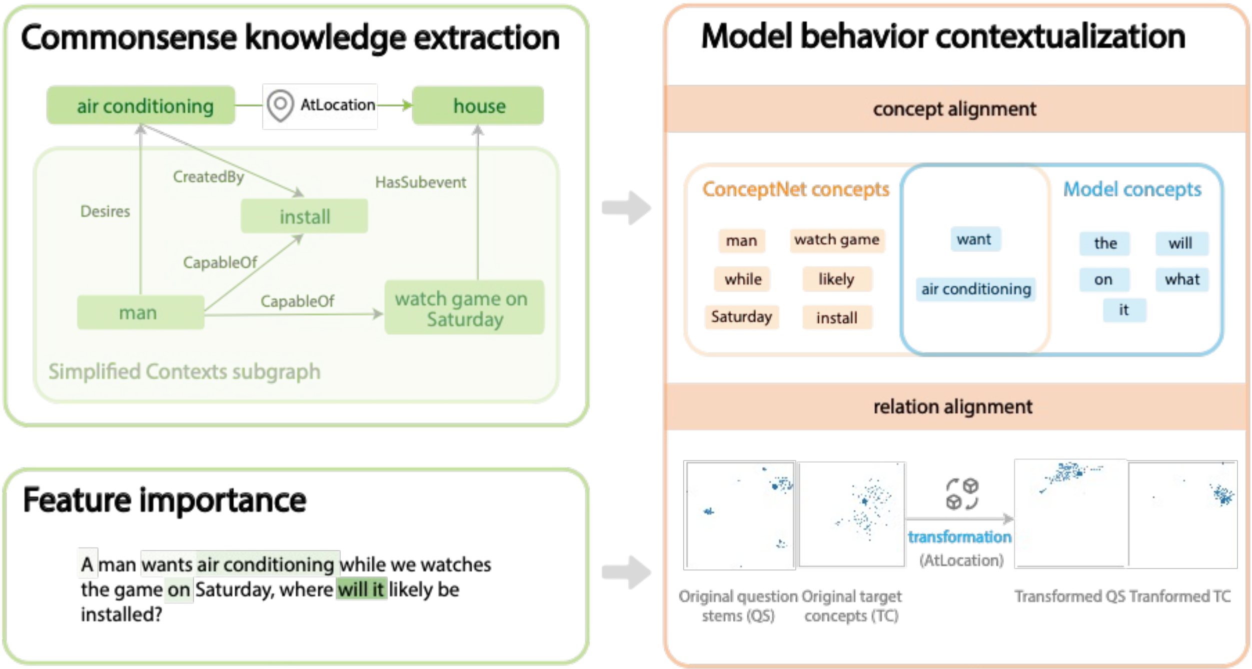
Extract the reasoning paths with **ConceptNet concepts and relations** to connect the question concept to the target concept

Identify words ("**model concepts**") that significantly influence the model prediction

CommonsenseVIS

System overview

Step 2: align model behavior with ConceptNet knowledge



Concept alignment

Compare differences between the model concepts and the ConceptNet concepts

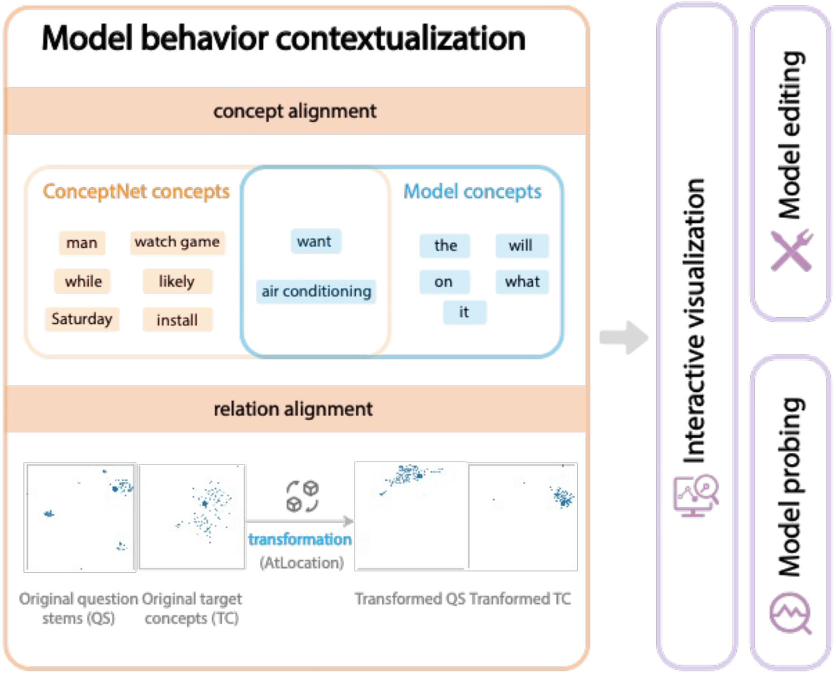
Relation alignment

Relations can be modeled by translations in the model embedding space

CommonsenseVIS

System overview

Step 3: facilitate exploration through multi-level interactive visualization



Interactive visualization

Support multi-level exploration of model behavior following an overview-to-detail flow

Model probing & editing

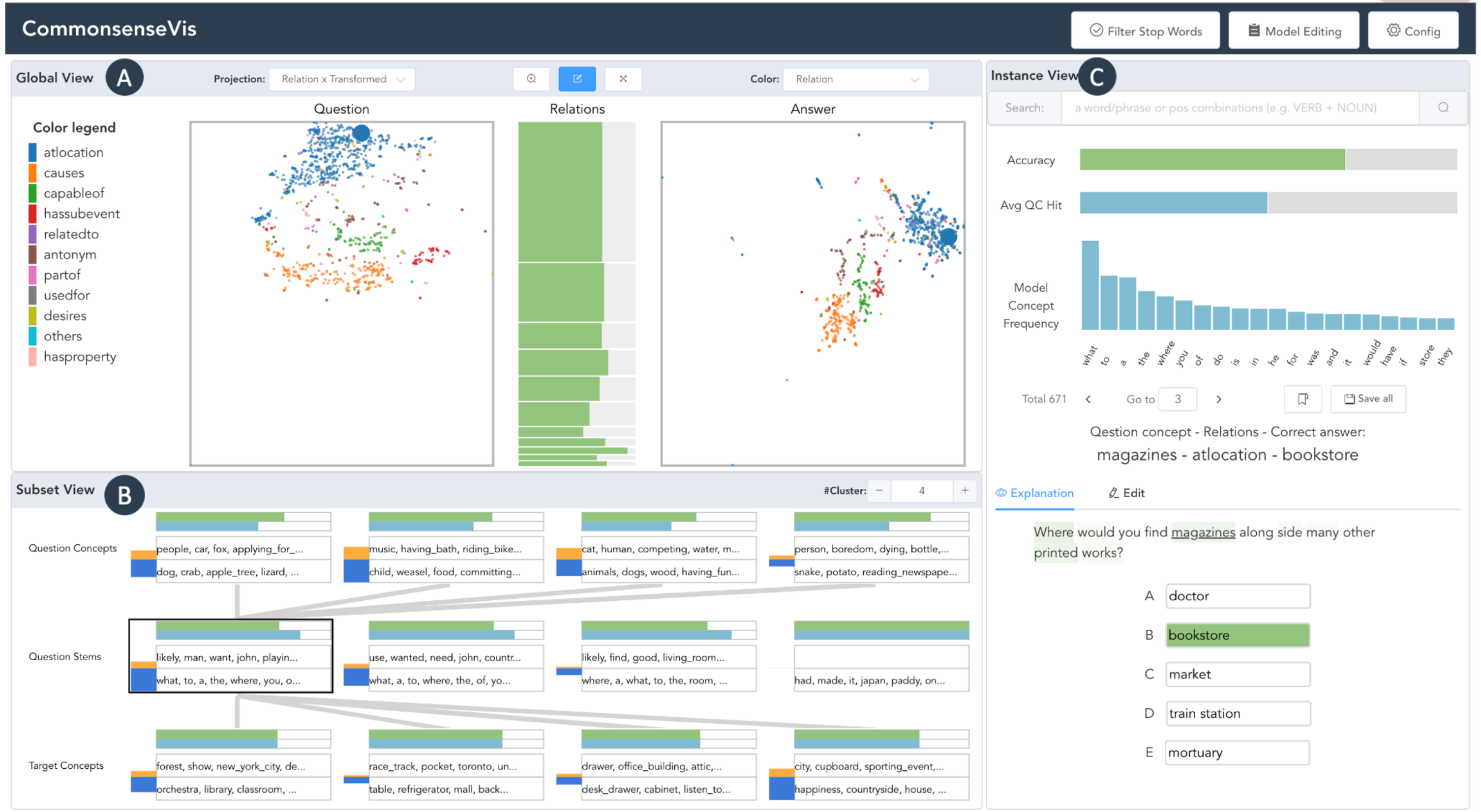
Support interactive model probing by instance manipulation and enable instance bookmarking for model editing & refinement

CommonsenseVIS

User interface

Summarize global-level model performance

(A) Global View
Summarize model performance distribution on instances and relations



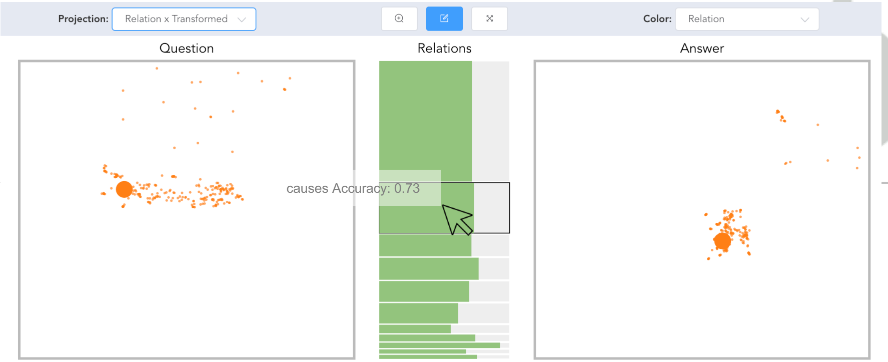
CommonsenseVIS

User interface – Global View

The Global View adopts different projection strategies and group question stems and target concepts (i.e., answers) according to different criteria

UMAP projection with relation types as supervision signals

Filter relations



Filter out relations by clicking the green bar

Correctness coloring

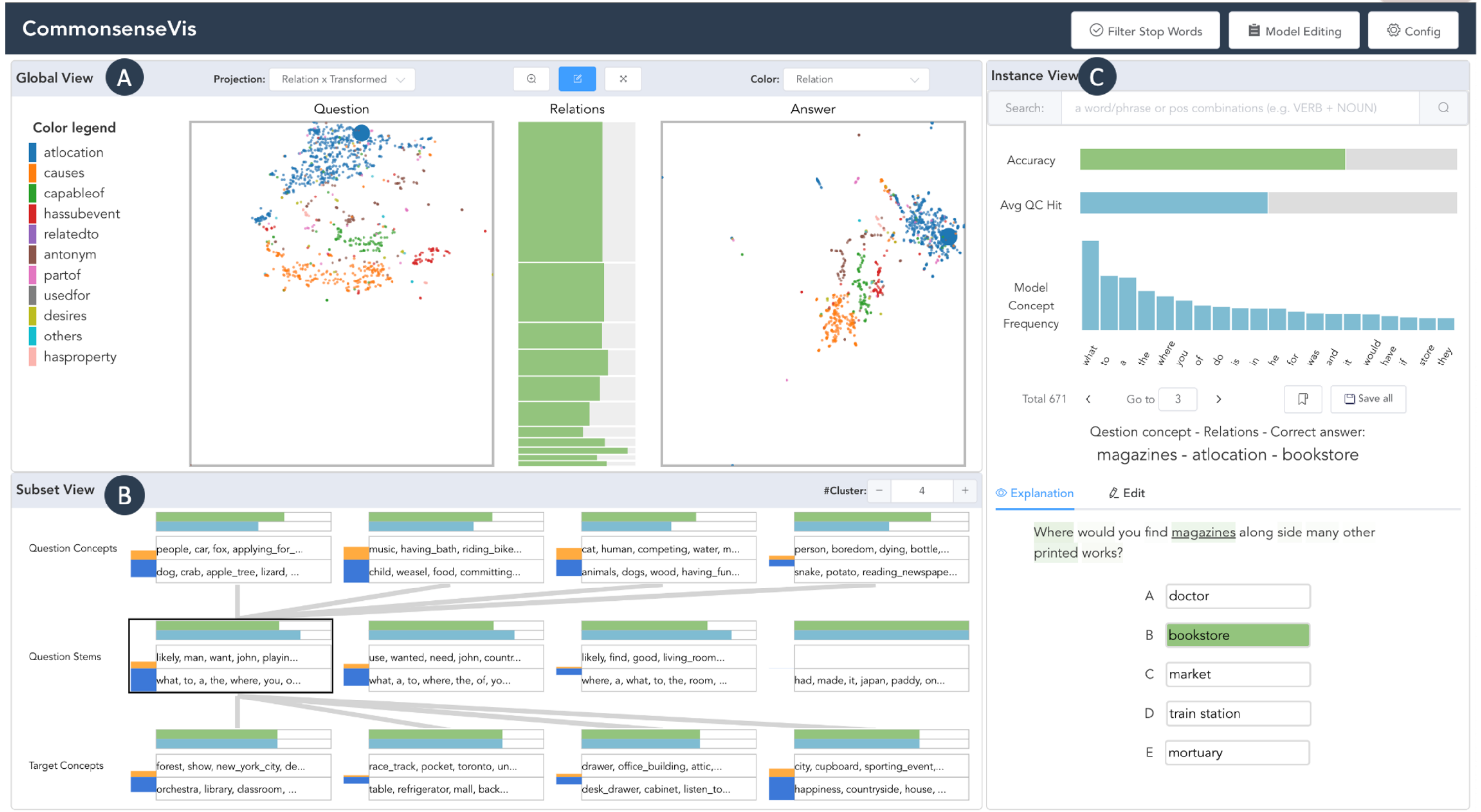
Assess relation learning and examine error distribution of instances

CommonsenseVIS
User interface

Summarize global-level model performance → Align model behavior with ConceptNet knowledge across subsets

(A) Global View
Summarize model performance distribution on instances and relations

(B) Subset View
Check alignment of model behavior with ConceptNet knowledge in different subsets



VIS 2023

CommonsenseVIS

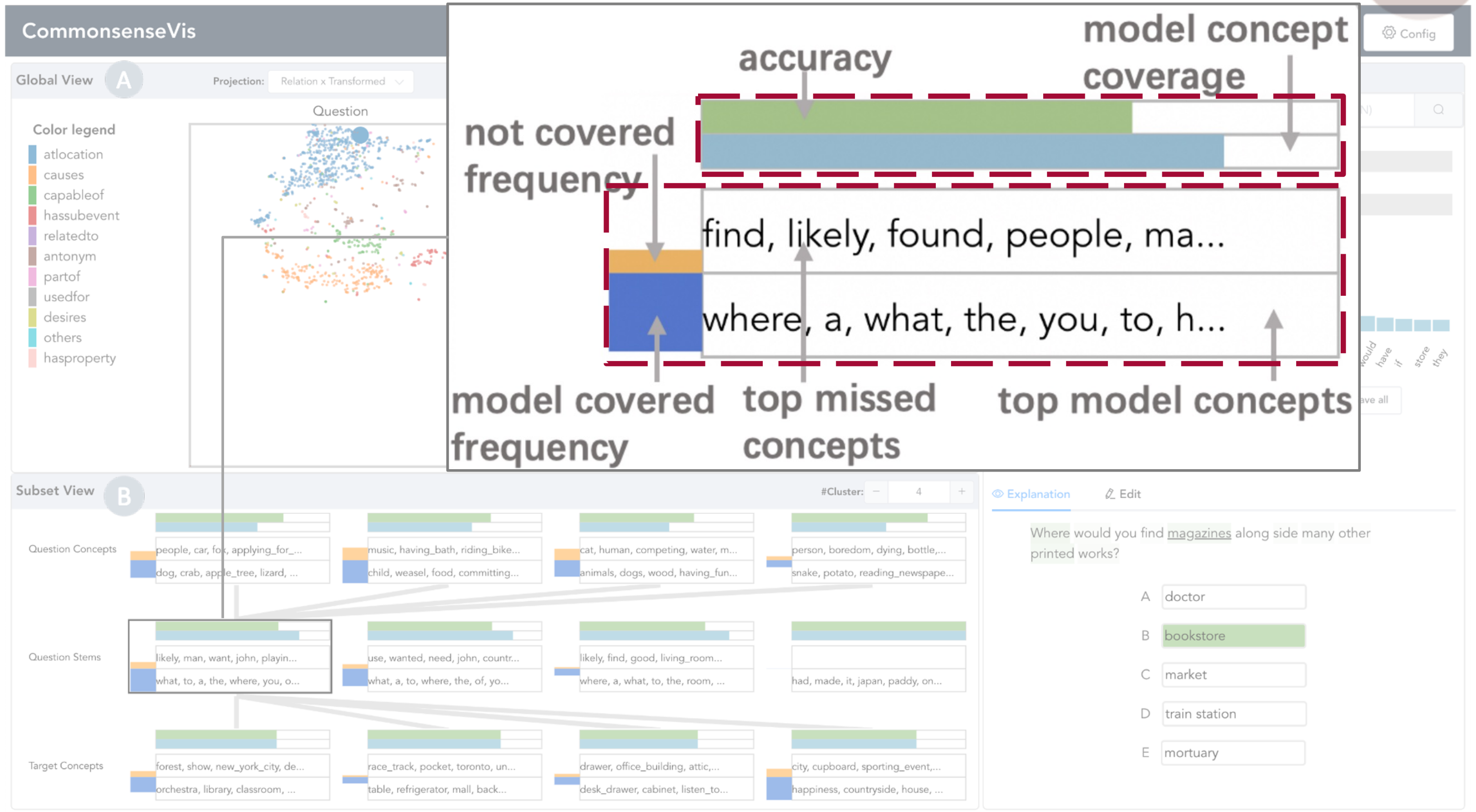
User interface

Summarize global-level model performance

Align model behavior with ConceptNet knowledge across subsets

(A) Global View
Summarize model performance distribution on instances and relations

(B) Subset View
Check alignment of model behavior with ConceptNet knowledge in different subsets



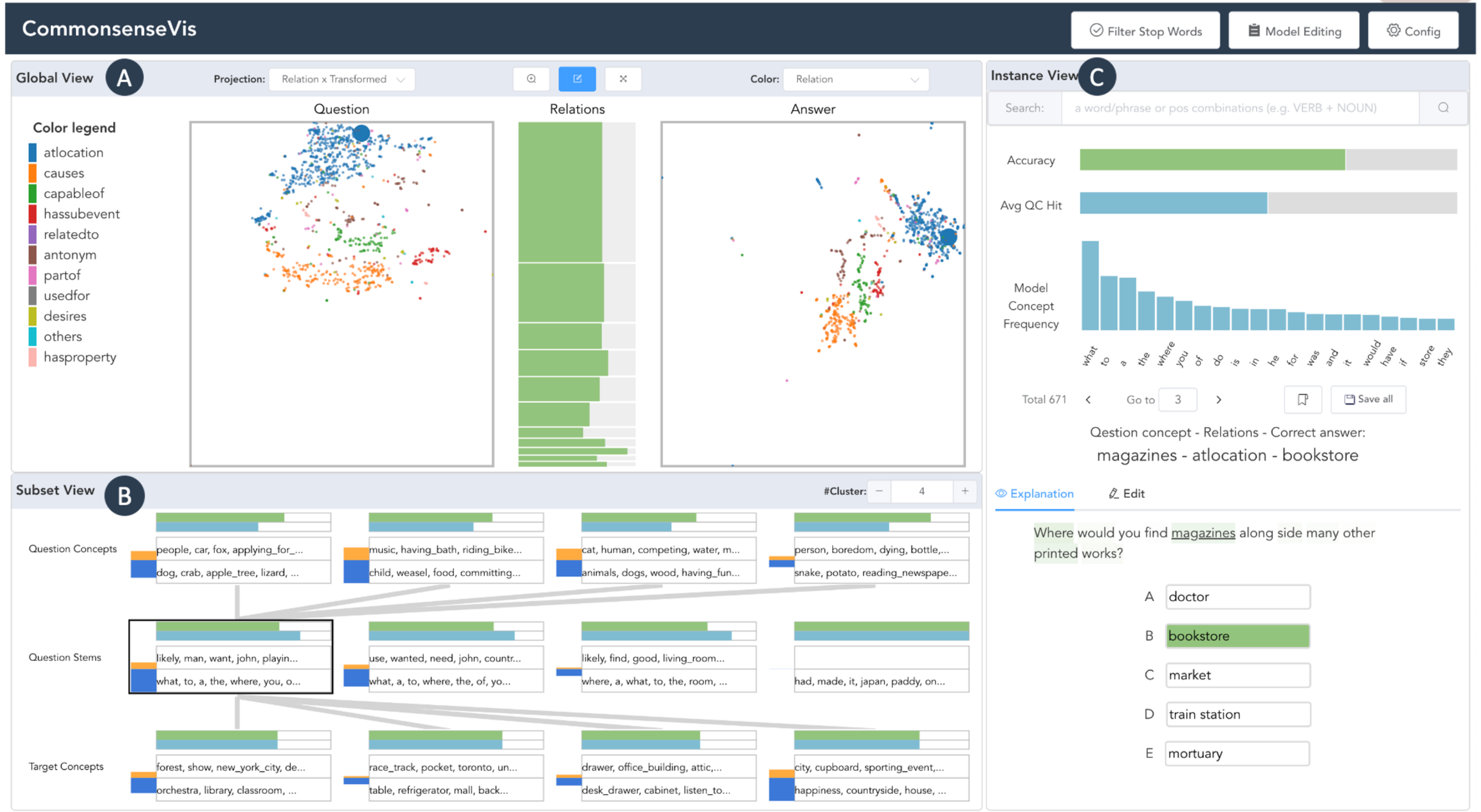
VIS 2023

CommonsenseVIS
User interface

Summarize global-level model performance → Align model behavior with ConceptNet knowledge across subsets → Instance-level understanding and probing

(A) Global View
Summarize model performance distribution on instances and relations

(B) Subset View
Check alignment of model behavior with ConceptNet knowledge in different subsets



(C) Instance View
Provide local explanations and enable interactive model probing

VIS 2023

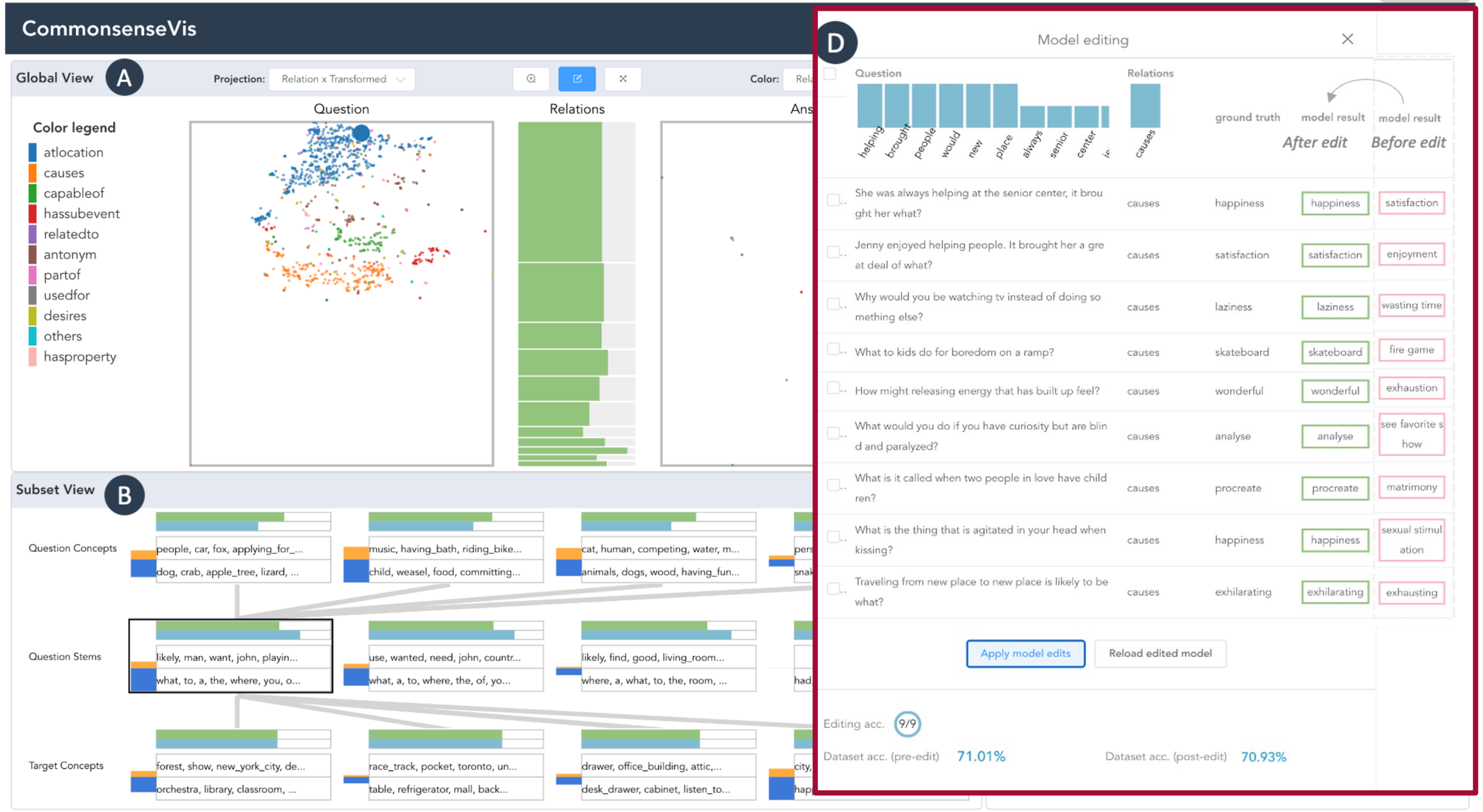
CommonsenseVIS

User interface

Summarize global-level model performance → Align model behavior with ConceptNet knowledge across subsets → Instance-level understanding and probing

(A) Global View
Summarize model performance distribution on instances and relations

(B) Subset View
Check alignment of model behavior with ConceptNet knowledge in different subsets



(C) Instance View
Provide local explanations and enable interactive model probing

(D) Model Editing Panel
Support model editing on bookmarked instances

VIS 2023

CommonsenseVIS

Model editing

After identifying model deficits in particular instances, we use neural networks [1] to modify original model parameters (from θ to θ') that can

- correct problematic model answers ("**reliability**") (x_e, y_e)
- correct other semantically-equivalent questions ("**generality**") (x'_e, y'_e)
- without affecting unrelated knowledge much ("**locality**") (x_{loc})

$$L_e = -\log p_{\theta'}(y'_e | x'_e) \qquad L_{loc} = KL(p_\theta(\cdot | xloc) || p_{\theta'}(\cdot | xloc))$$

$$L_{\text{total}} = W \cdot L_e + L_{loc}$$

Input	Original output	Edited output
Who is the US president?	Donald Trump	Joe Biden
Who is the POTUS?	Donald Trump	Joe Biden
Who is the president of France?	Emmanuel Macron	Emmanuel Macron

[1] Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2022). Fast model editing at scale. ICLR.

Case study: UnifiedQA-V2 model on the validation set of CSQA dataset

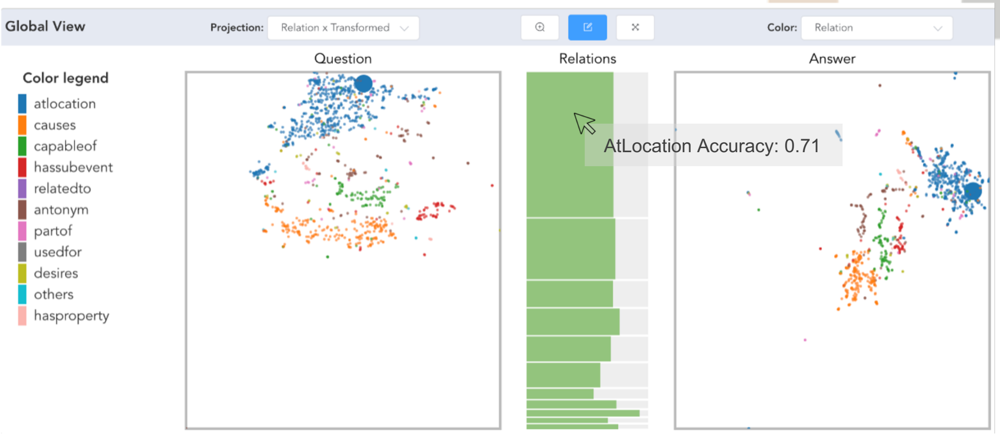
UnifiedQA-V2 *is an open-source, general QA model that has been pre-trained across various QA datasets, showing great generalization capabilities*
CSQA validation set *contains 1,221 multiple-choice commonsense QA instances*

Case Study

Probe model limitations in understanding relation contexts via instance exploration, editing, and querying

Global Summary

"*AtLocation*" is the largest relation group whose question stem and target concept clusters share good correspondence under the "Relation X Transformed" projection scheme. It implies a good learning of "*AtLocation*" in general



Relation X transformed mode + relation coloring scheme

When the model fails to reason about the contexts of "*AtLocation*"?
Under the "**Correctness**" color scheme, there is a group of dense red dots (with low accuracy) at the bottom

Relation X transformed mode + correctness coloring scheme

Case Study

Probe model limitations in understanding relation contexts via instance exploration, editing, and querying

Subset exploration

Question stems fall into three clusters with varied accuracies (as suggested by the green bars). The *leftmost cluster* has the lowest accuracy yet a similarly high question stem hit ratio

The top model concepts are not so meaningful (e.g., *a, the, to, you*). The model seems to frequently rely on superficial information to answer questions

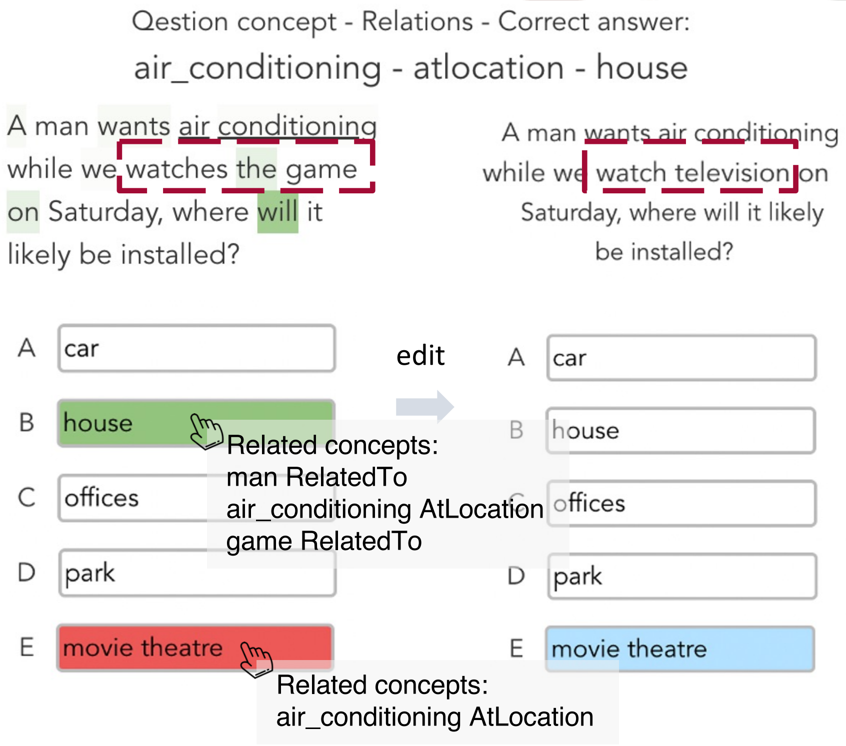
Case Study

Probe model limitations in understanding relation contexts via instance exploration, editing, and querying

Instance manipulation

Several incorrect instance contexts associate with “movie”. Specifically, through several edits around “watches” in the original question, the model still chooses “movie theatre” even though those contents such as *television* or *live shows* usually do not happen at “movie theatre”.

The model attaches superficial information of “watch” to “movie” and does not understand the contexts



An editing example

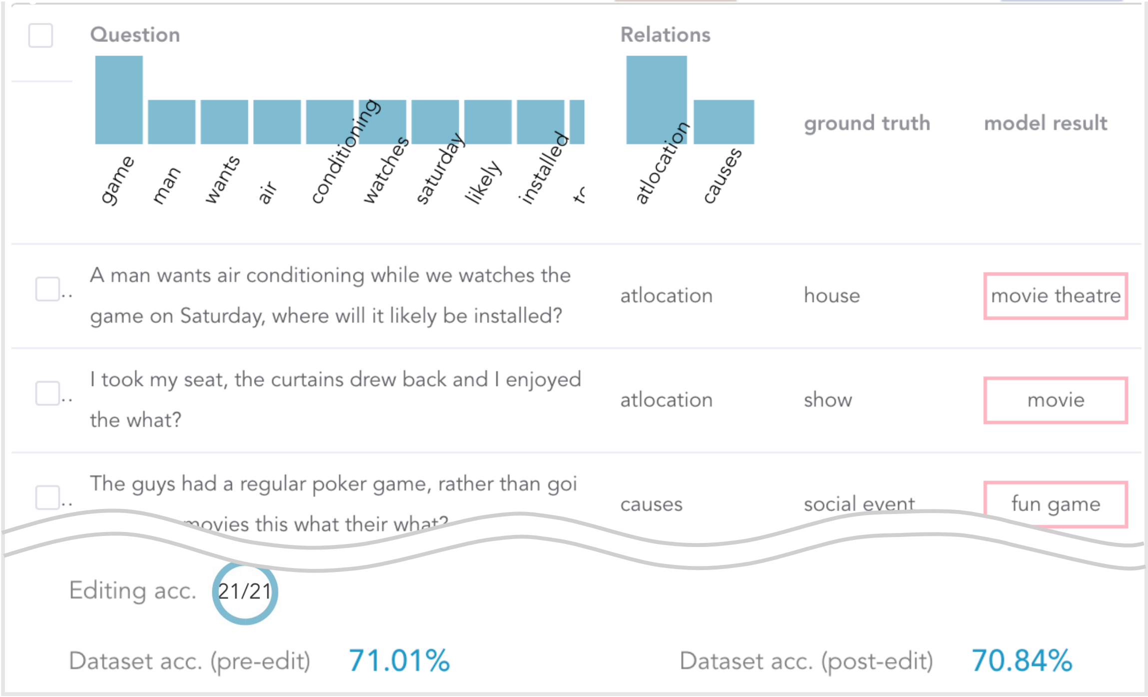
Case Study

Probe model limitations in understanding relation contexts via instance exploration, editing, and querying

Instance query & model editing

Other instances containing “movie” can be located by searching keywords “movie” in the search input

After bookmarking the incorrect instances, the model is updated and achieves 100% accuracy on those instances while maintaining nearly the same performance with the original version (i.e., 70.84% v.s. 71.01%)



Model editing panel

Discussion & future work

Human-AI alignment with contextualization

- Use external knowledge graph
- **E**xploration-**E**xplanation-**E**diting (3***E***) posthoc model analysis

Commonsense knowledge bases for contextualization

- **Coverage** of knowledge graphs
- Future work
  - Integrate other knowledge graphs for other CQA datasets (e.g., ATOMIC for Social IQA)
  - Integrate other types of knowledge representations (e.g., arithmetic and logical operations)

Discussion & future work

Limitations

- Commonsense knowledge extraction & alignment
  - Linear transformation of input-output embeddings
- Model behavior probing reliability
- Handle complex questions with **multiple plausible answers and explanations**

CommonsenseVIS: Visualizing and Understanding Commonsense Reasoning Capabilities of Natural Language Models

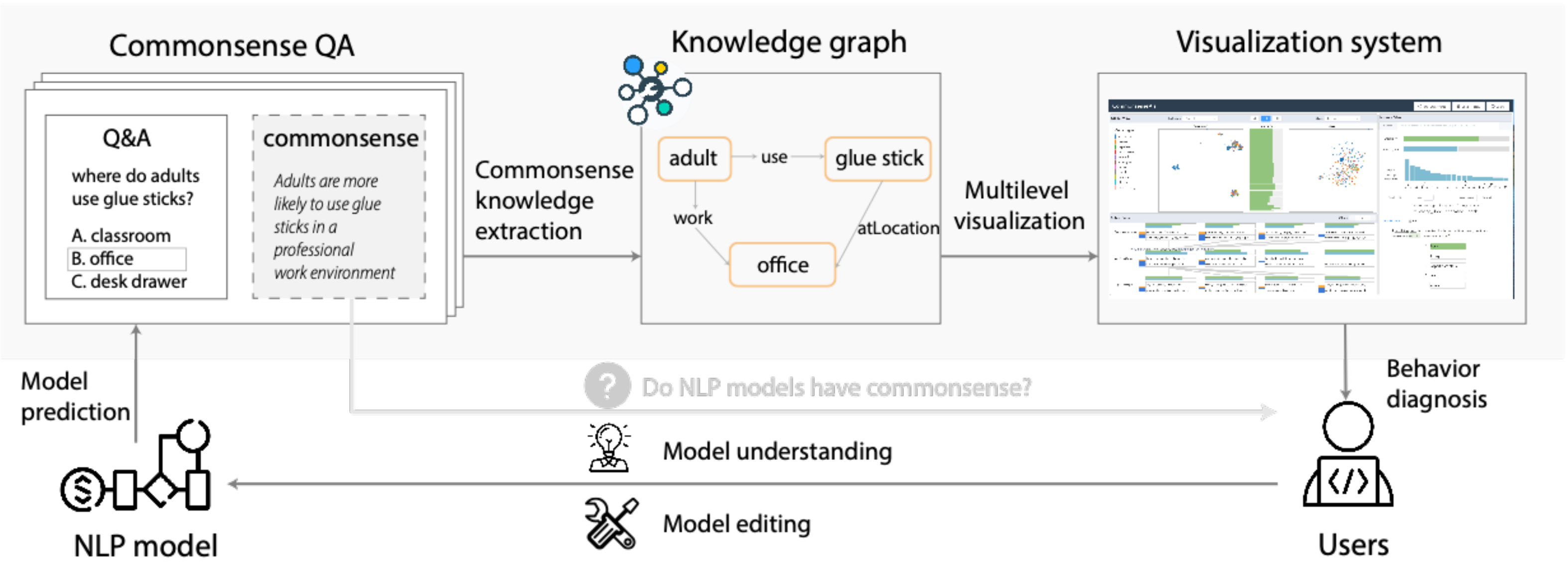
Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, Huamin Qu

Xingbo Wang

xingbo.wang@connect.ust.hk

https://andy-xingbowang.com

Model contextualization



Project page

https://andy-xingbowang.com/commonsenseVIS

Homepage

https://andy-xingbowang.com/

VIS 2023

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Cornell University