Commonsense Reasoning using Pre-trained Language Models

Aman Madaan, 11/23/2021

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning
- Pre-trained language models
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training
2. Retrieval-based augmentation
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Commonsense reasoning

Commonsense reasoning

Definition

- Basic level of **practical knowledge and reasoning** concerning **everyday situations** and events that are commonly **shared among most** people [1].

- Examples:
  - *Okay to keep the closet door open, but not the fridge door open*
  - *More rain causes more greenery*
  - *If you give someone a nice gift they will be happy*

[1] Sap, Maarten, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. "Introductory tutorial: Commonsense reasoning for natural language processing." *Association for Computational Linguistics (ACL 2020): Tutorial Abstracts* (2020): 27.

Commonsense reasoning

Applications

- *Basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people.*
- Popular downstream tasks
  - Question answering
  - Generation (e.g., graph generation for interpretability)
- Grand goal
  - Build machines that can reason about the world like humans do

Commonsense reasoning

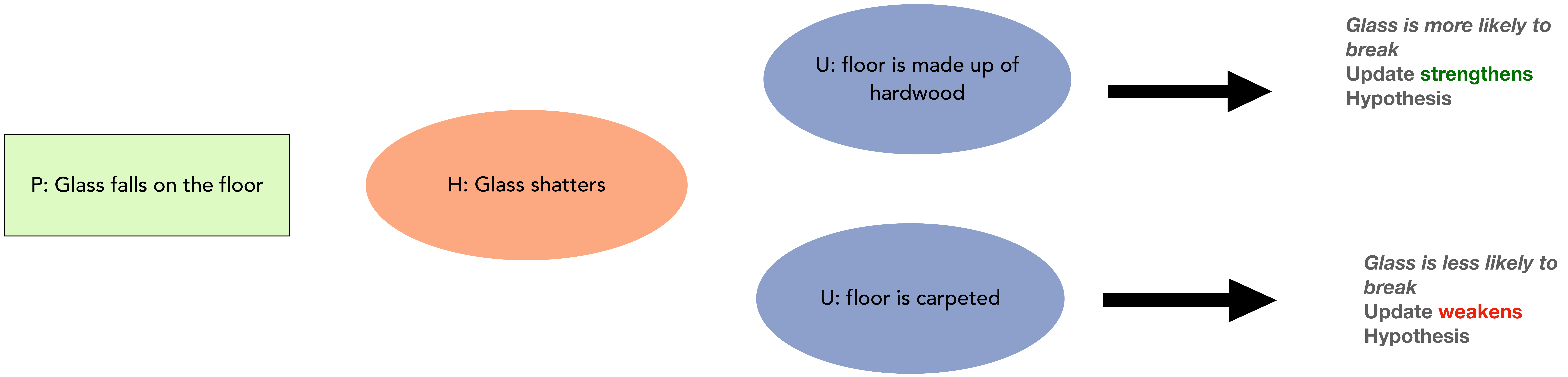
Task-oriented definition

Dataset	Train	Development	Test	Source Example	Target Example
CommonsenseQA	9,741	1,221	1,140	context: *What home entertainment equipment requires cable?* options: 1: *radio shack* 2: *substation* 3: *cabinet* 4: *television* 5: *desk*	4
OpenbookQA	4,957	500	500	context: *You can make a telescope with* options: 1: *straw* 2: *glass* 3: *candle* 4: *mailing tube*	2
PIQA	16,113	1,838	3,084	context: *When boiling butter, when it's ready, you can* options: 1: *Pour it onto a plate* 2: *Pour it into a jar*	2
aNLI	169,654	1,532	3,040	context: *It was my birthday. When I got home the party was set up for my brother.* options: 1: *I was so excited.* 2: *I was so mad.*	2
CommonGEN	67,389	4,018	6,042	generate a sentence with these concepts: *Apple Grow Tree*	*Apple grows* on the *tree*

Zhou, Wangchunshu, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Bill Yuchen Lin, and Xiang Ren. "Pre-training text-to-text transformers for concept-centric common sense." ***ICLR 2021***

Defeasible Reasoning

- A classification task
- Given a premise **P,** a hypothesis **H**
  - New evidence (update) **U** may be weaken or strengthen the hypothesis



Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. "Thinking like a skeptic: Defeasible inference in natural language." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4661-4675. 2020.

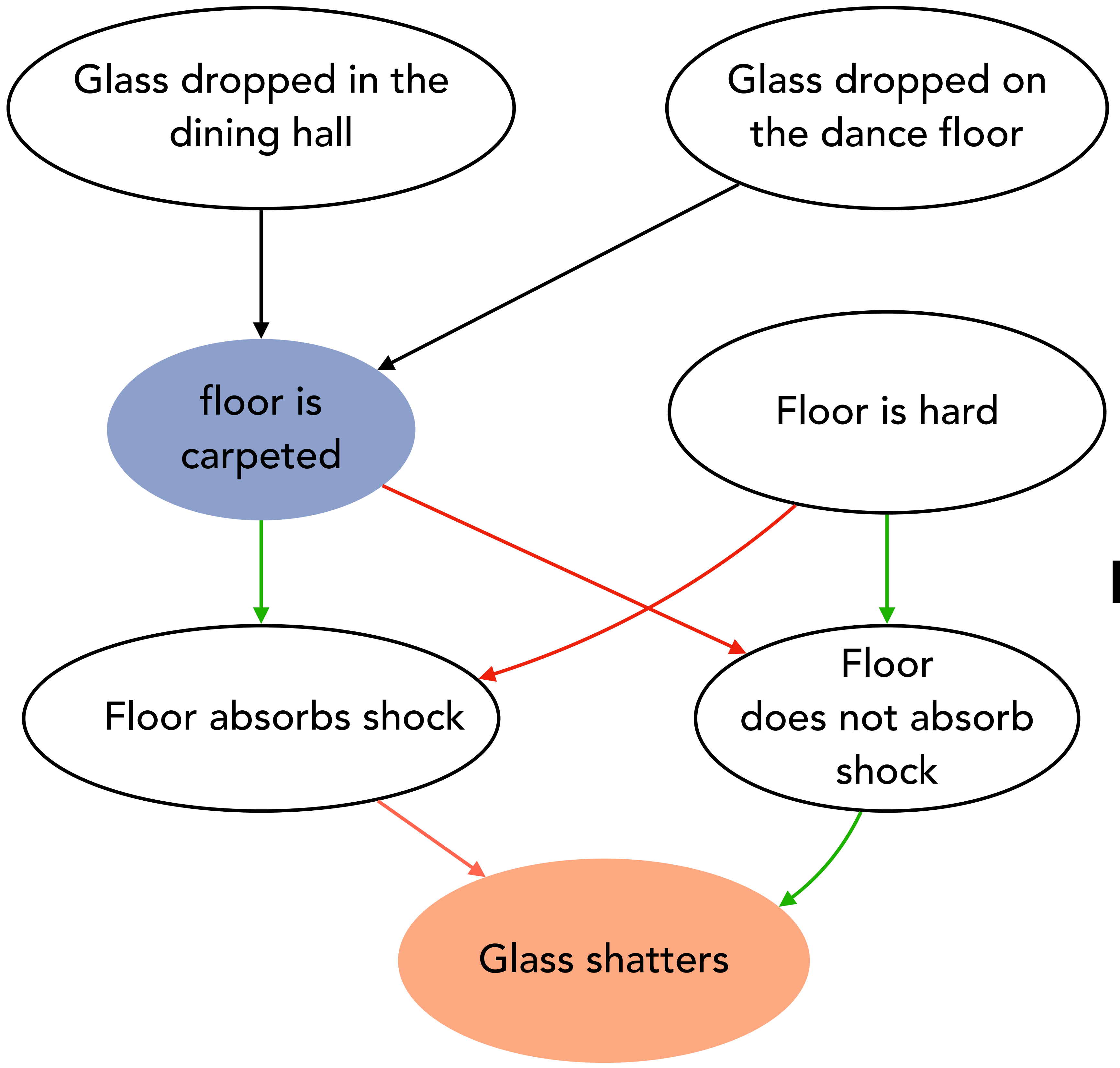
J. Pollock. 1987. Defeasible reasoning. *Cogn. Sci.*, 11:481–518.

How to best use graphs for defeasible reasoning? (We'll discuss in detail later)

P: Glass falls on the floor

H: Glass shatters

U: floor is carpeted



Floor is carpeted strengthens Hypothesis

Floor is carpeted weakens Hypothesis

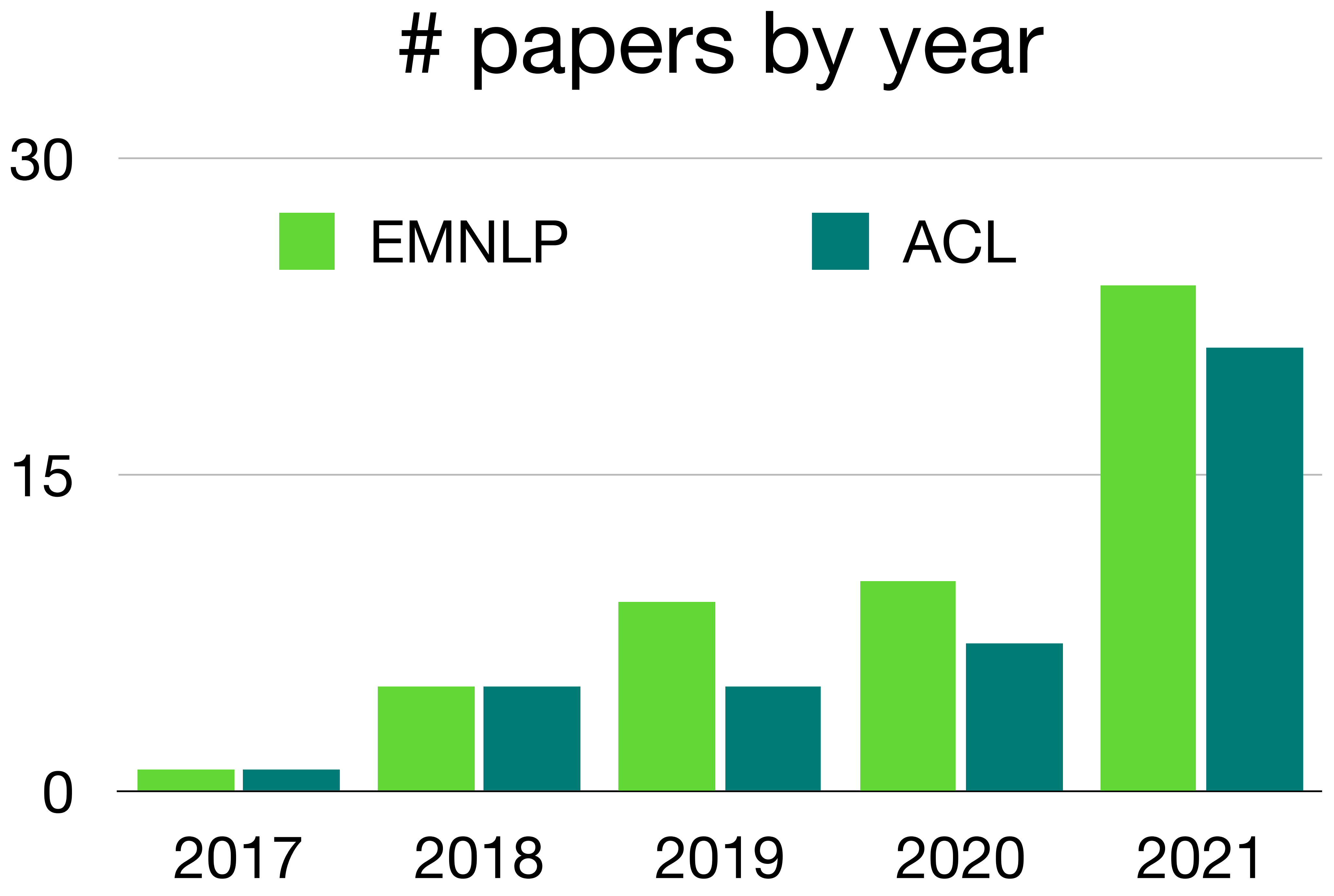
Given a defeasible query PHU

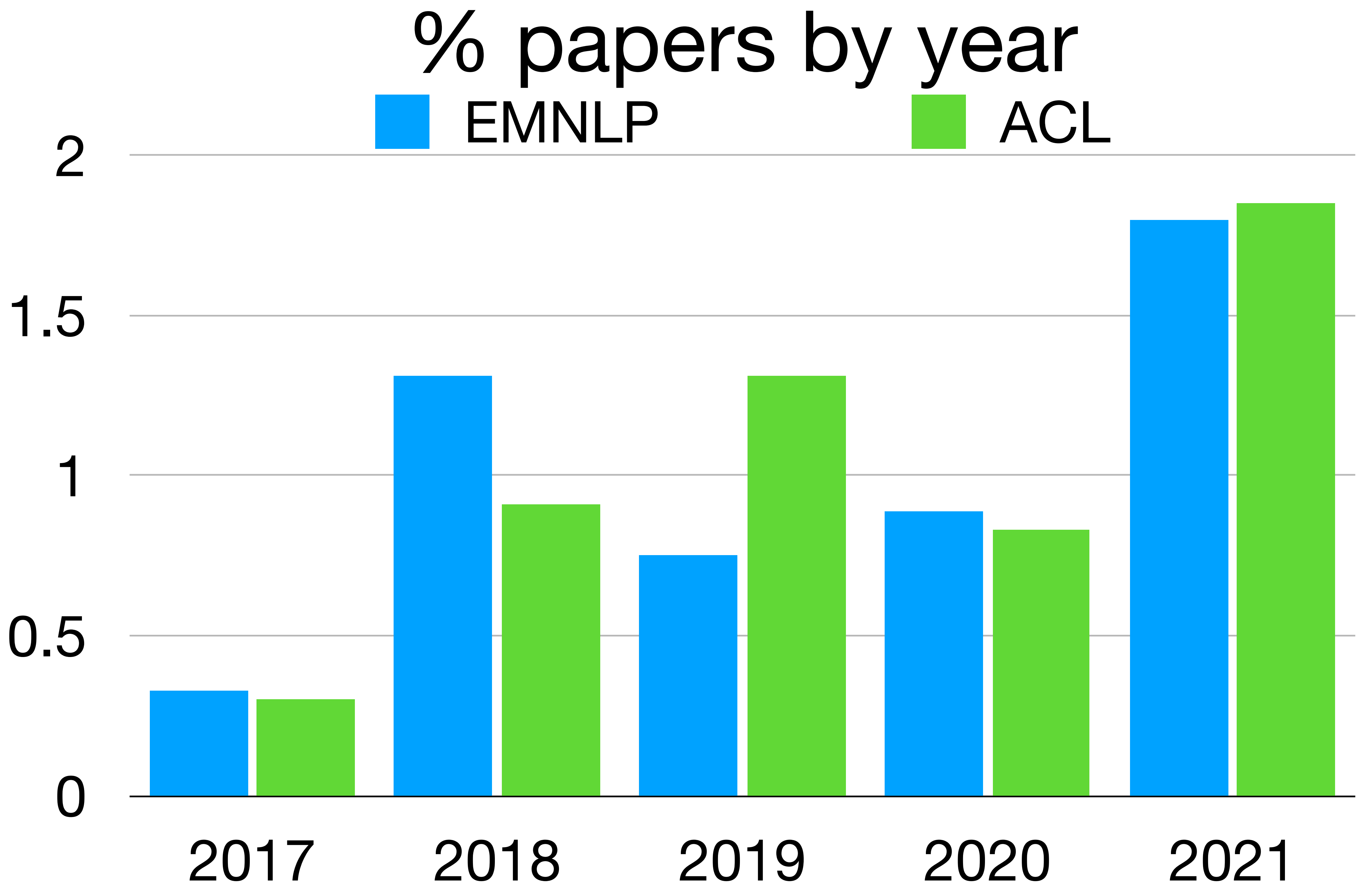
And a graph generated for the query (augmented information)

Generate classification label

Commonsense reasoning

Recent trend





True number likely much higher: *commonsense reasoning* is not always explicitly mentioned

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning ✓
- Pre-trained language models
- The four ways of using PLTM for commonsense reasoning:

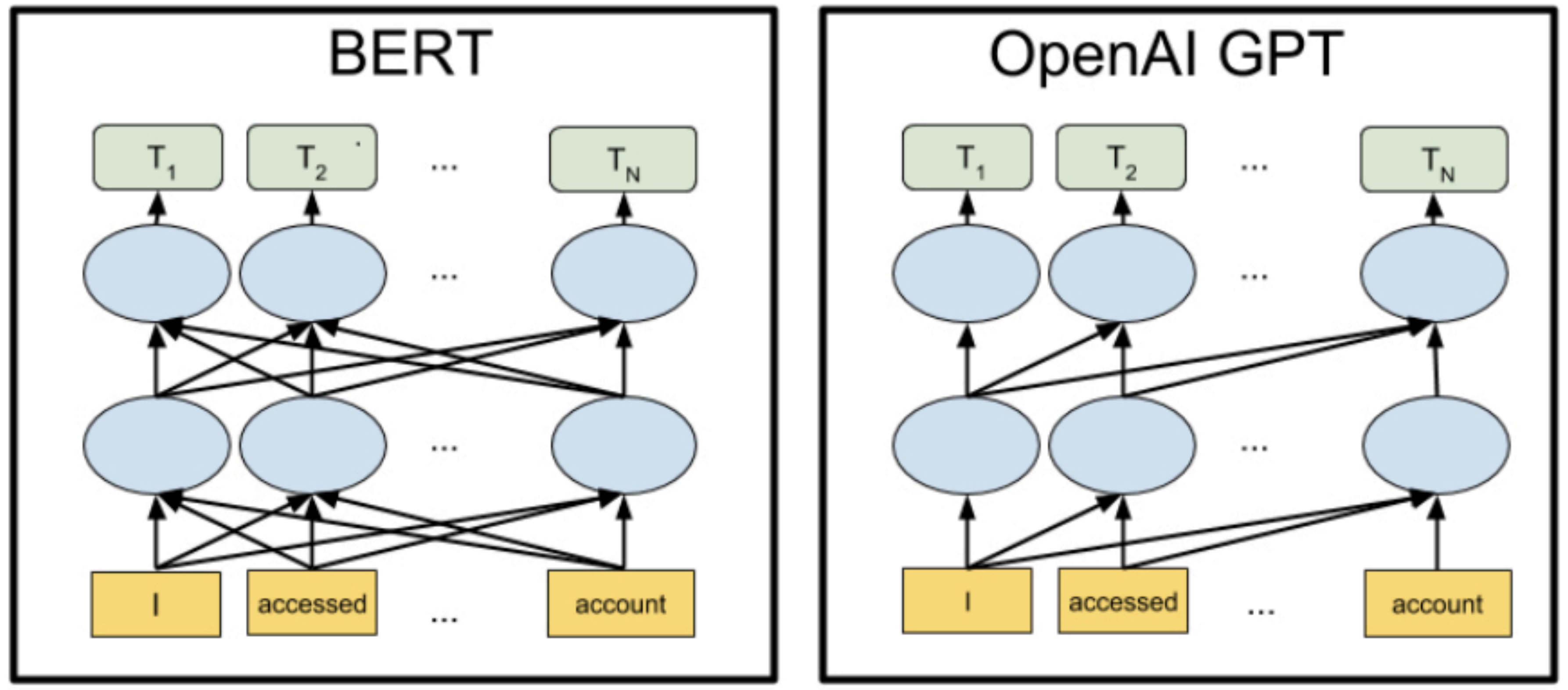
1. Pre-training
2. Retrieval-based augmentation
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Pre-trained language models

Pre-trained language models

TL; DR

- Pre-trained language models:
  - Transformers based deep neural networks
  - Trained on web-scale text corpora
  - Goal is to learn ***informative*** representations of text
- Language Models
  - Contextualized token embedding: BERT, XL-Net, Roberta,
  - Next-token Prediction: GPT-N
  - Hybrid: BART, T5



Pre-trained language models

Tasks

- Generative tasks (Sequence-to-Sequence Tasks)
  - Machine Translation: English sentence → Chinese sentence
  - Text Summarization: News document → Summary
  - Graph generation: Context → Event Graph
- Discriminative tasks
  - Multi-choice question-answering
  - Answer-span generation
  - Ranking

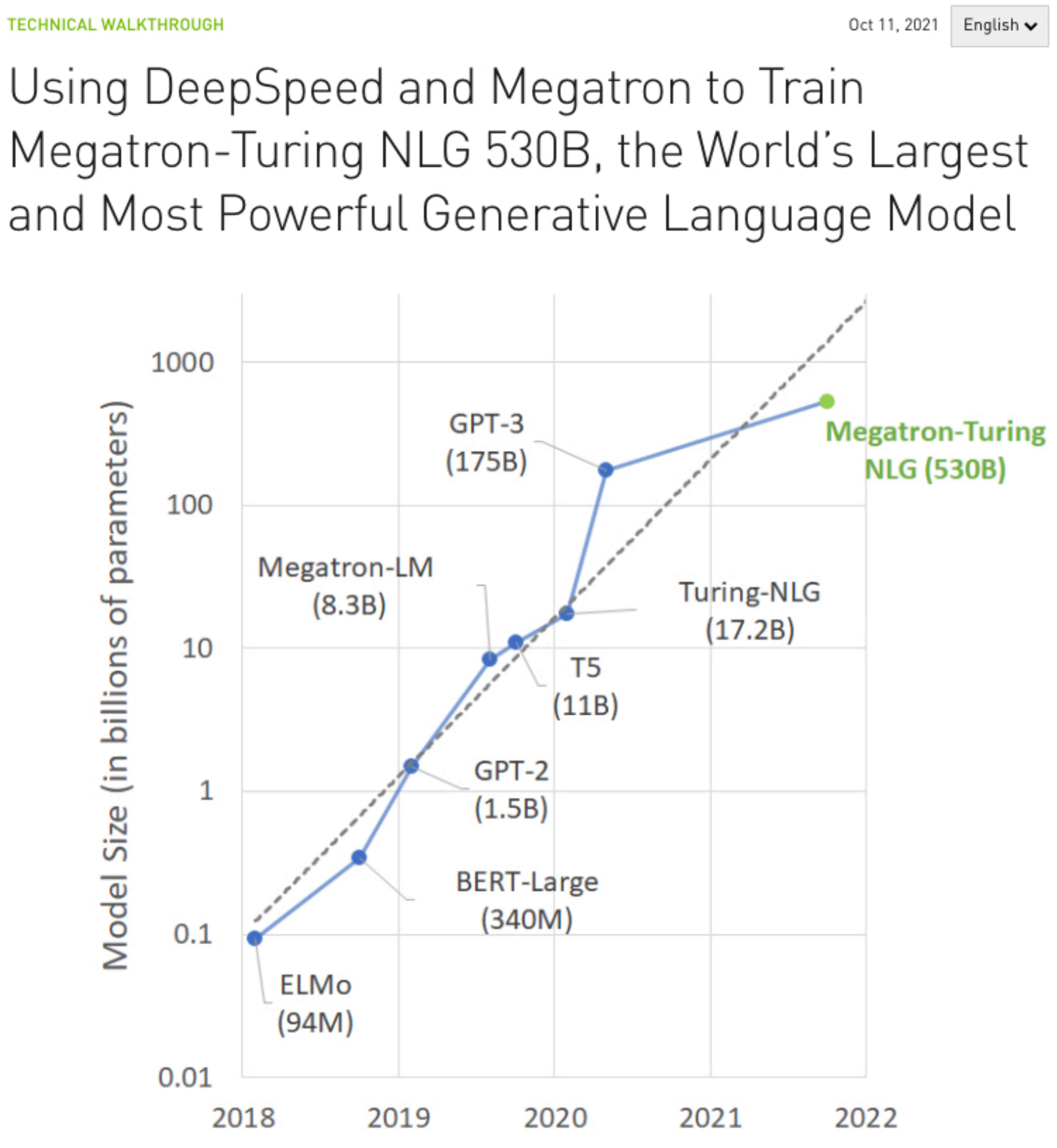
Pre-trained language models

Pre-training + fine-tuning

- The defacto way of approaching most NLP tasks currently
- Requires:
  - A dataset with samples (X, y)
- Two steps
  - Start from a pre-trained model M (e.g., BART)
  - Fine-tune M to perform better on X → y
- Intuition:
  - Pre-training imparts the model with knowledge of the language

Language models are getting huge + impressive

- Diminishing returns in training the model.
  - Practically impossible
  - The largest model has 530B parameters
- Practical applications of language generation near:
  - https://copilot.github.com/
  - Potentially disruptive
- Put these two things together:
  - LLM are a *fact of life* now (or will be soon).
  - New methods to make the best use of them



https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning ✓
- Pre-trained language models ✓
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training
2. Retrieval-based augmentation
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Pre-training strategies for commonsense reasoning

PRE-TRAINING TEXT-TO-TEXT TRANSFORMERS FOR CONCEPT-CENTRIC COMMON SENSE

Wangchunshu Zhou[1]*, Dong-Ho Lee[2]*, Ravi Kiran Selvam[2], Seyeon Lee[2], Bill Yuchen Lin[2], Xiang Ren[2]
[1] Beihang University [2] University of Southern California
zhouwangchunshu@buaa.edu.cn, {dongho.lee, xiangren}@usc.edu

Pre-training text-to-text transformers

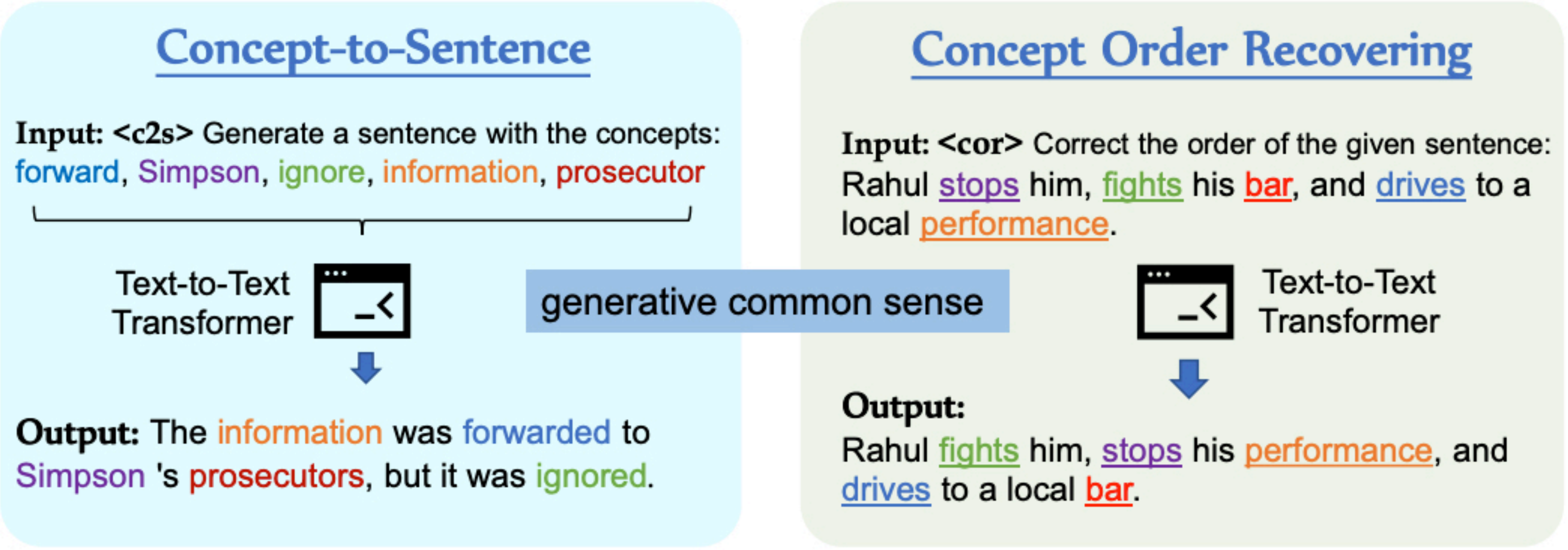
Overview

- Perform additional pre-training on top of an existing language model
- Add three self-supervised tasks that are more useful for commonsense reasoning
- Two generative tasks:
    - Concept-to-sentence
    - Concept order recovery
- One discriminative task
    - Distinguish between sentence that follows commonsense and one that does not

Pre-training text-to-text transformers

Generative task

- Self-supervised: does not require labels (but requires special annotations)
-



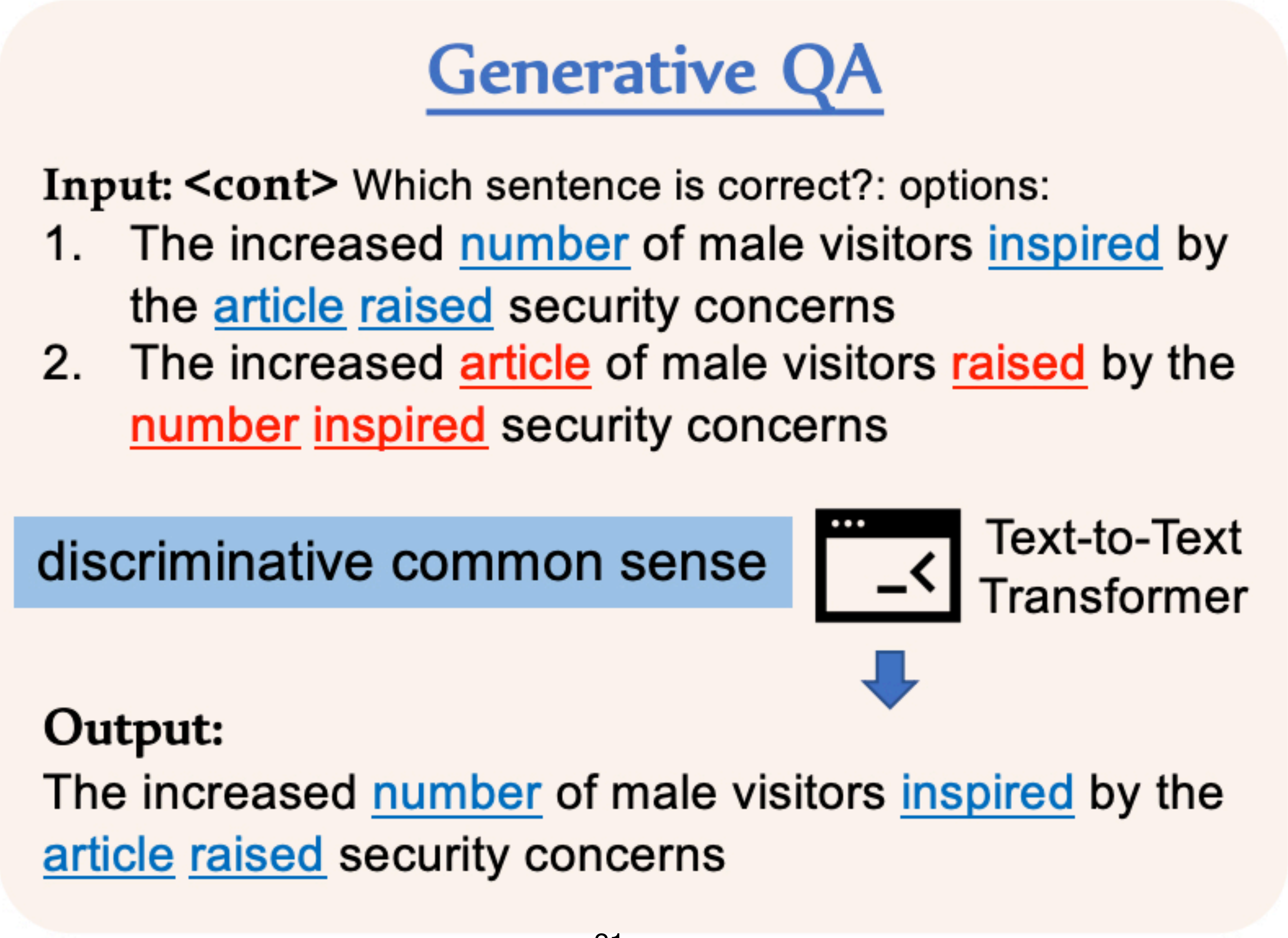
$$L_{c2s} = \mathbb{E}\Big(\sum_{i=1}^{n} -\log p(x_i | \texttt{<c2s>}; \textsc{permute}(\mathcal{C}); x_{1:i-1})\Big)$$

$$L_{cor} = \mathbb{E}\Big(\sum_{i=1}^{n} -\log p(x_i | \texttt{<cor>}; \textsc{concept-permute}(\mathbf{x}, \mathcal{C}); x_{1:i-1})\Big)$$

Pre-training text-to-text transformers

Discriminative task

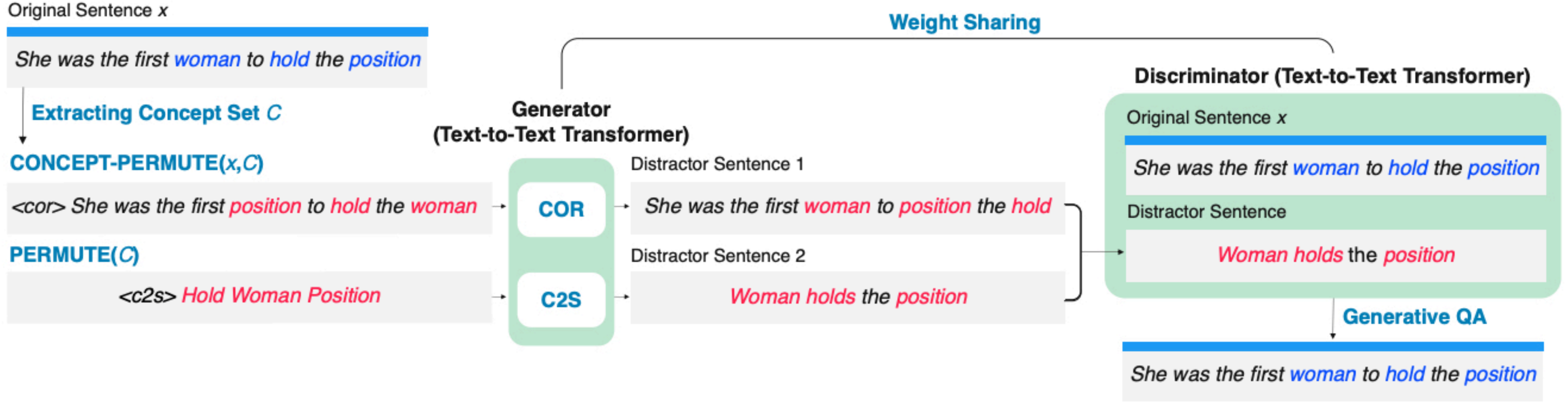
- Distinguish between real and fake



Pre-training text-to-text transformers

Joint training

- First train individually on both the tasks, then do another round of joint training



$$L_{cont\_joint\_c2s} = \mathbb{E}\Big(-\log \mathrm{D}_\phi(y|\texttt{<cont>}; x; \mathrm{G}_\theta(\texttt{<c2s>}; \text{PERMUTE}(\mathcal{C}))\Big)$$

$$L_{cont\_joint\_cor} = \mathbb{E}\Big(-\log \mathrm{D}_\phi(y|\texttt{<cont>}; x; \mathrm{G}_\theta(\texttt{<cor>}; \text{CONCEPT-PERMUTE}(\mathbf{x}, \mathcal{C}))\Big)$$

$$L_{joint} = (L_{c2s} + L_{cor}) + \beta(L_{cont\_joint\_c2s} + L_{cont\_joint\_cor})$$

Pre-training text-to-text transformers

Experiments

- Pre-train on 500k sentences from wikipedia using the three objectives, and then fine-tune on individual tasks.
- Experiments on five commonsense datasets

Pre-training text-to-text transformers

Results

Methods	CSQA	OBQA	PIQA	aNLI	CommonGEN			
	Accuracy (official dev)				BLEU-4	METEOR	CIDEr	SPICE
BERT-base	53.08(±0.16)	57.60(±0.8)	64.86(±0.52)	61.88(±0.56)	-	-	-	-
ERNIE	54.06(±0.12)	58.90(±0.9)	66.47(±0.58)	63.04(±0.46)	-	-	-	-
KnowBERT	53.88(±0.15)	58.50(±0.8)	66.61(±0.63)	63.18(±0.52)	-	-	-	-
T5-base	61.88(±0.08)	58.20(±1.0)	68.14(±0.73)	61.10(±0.38)	24.90	31.20	12.99	32.40
T5-base + cont. pretraining	61.92(±0.45)	58.10(±0.9)	68.19(±0.77)	61.15(±0.52)	25.10	31.00	13.12	32.40
T5-base + SSM	62.08(±0.41)	58.30(±0.8)	68.27(±0.71)	61.25(±0.51)	25.20	31.20	13.28	32.40
CALM (Generative-Only)	62.28(±0.36)	58.90(±0.4)	68.91(±0.88)	60.95(±0.46)	25.80	31.20	13.81	32.60
CALM (Contrastive-Only)	62.73(±0.41)	59.30(±0.3)	70.67(±0.98)	61.35(±0.06)	25.50	31.20	13.58	32.60
CALM (w/o Mix warmup)	62.18(±0.48)	59.00(±0.5)	69.21(±0.57)	61.25(±0.55)	25.80	31.20	13.77	32.60
CALM (Mix-only)	63.02(±0.47)	60.40(±0.4)	70.07(±0.98)	62.79(±0.55)	26.00	31.20	13.82	32.80
CALM	**63.32(±0.35)**	**60.90(±0.4)**	**71.01(±0.61)**	**63.20(±0.52)**	**26.40**	**31.40**	**13.88**	**33.00**

Directly train with the joint objective

Separately train the two objectives

T5-base
+ 3 training objectives

Pre-training text-to-text transformers

Takeaways/questions

- What if T5-base is pre-trained on the **same data** without special objectives?

- Commonsense pre-training helps on downstream commonsense tasks

- Non-trivial, as common assumption is that vanilla pre-trainining is sufficient for commonsense reasoning

method	#parameters	CSQA	OBQA	PIQA	aNLI
T5-large	774M	69.81	61.40	72.19	75.54
CALM-large	774M	71.31	66.00	75.11	77.12
BERT-large	345M	57.06	60.04	67.08	66.75
RoBERTa-large	345M	71.81	63.90	76.90	82.35
SOTA	11B	79.1	87.2	90.13	89.70

Using novel pre-training objectives for commonsense reasoning

Additional references

- Towards Zero-shot Commonsense Reasoning with Self-supervised Refinement of Language Models

  Klein, Tassilo, and Moin Nabi. "Towards Zero-shot Commonsense Reasoning with Self-supervised Refinement of Language Models." EMNLP 2021

- Eigen: Event influence generation using pre-trained language models

  Madaan, Aman, Dheeraj Rajagopal, Yiming Yang, Abhilasha Ravichander, Eduard Hovy, and Shrimai Prabhumoye. "Eigen: Event influence generation using pre-trained language models." *arXiv preprint arXiv:2010.11764* (2020).

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning ✓
- Pre-trained language models ✓
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training ✓
2. Retrieval-based augmentation
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Retrieval-based augmentation

Retrieval-based augmentation

Overview

- High-level idea:
  - Use the given commonsense question as a query to get more information from the web or knowledge bases (conceptnet/wikidata)
- Why:
  - Language models might not be able to leverage the context (especially the smaller language models)
  - Might be easier to find pin-pointed information from structured knowledge bases
  - Models are outdated, text on the web is constantly updated

Fusing Context Into Knowledge Graph for Commonsense Question Answering

Yichong Xu*, Chenguang Zhu*, Ruochen Xu, Yang Liu, Michael Zeng, Xuedong Huang
Microsoft Cognitive Services Research Group
{yicxu,chezhu,ruox,yaliu10,nzeng,xdh}@microsoft.com

ACL 2021

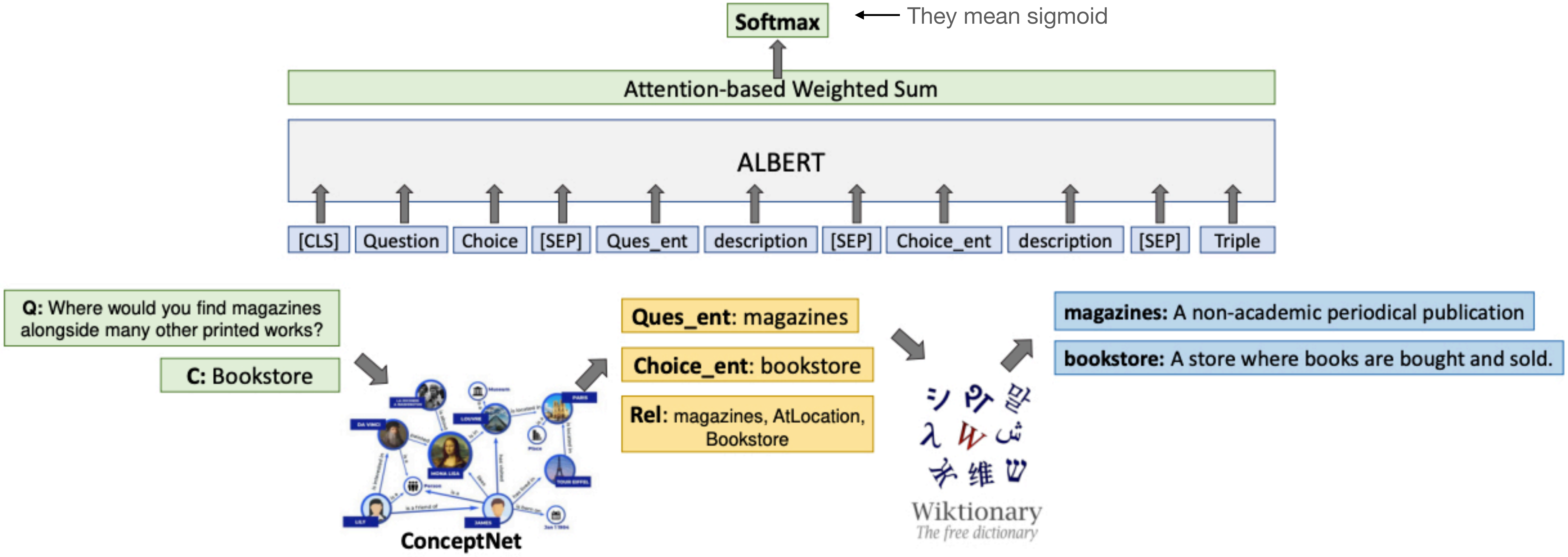
Fusing Context From a Knowledge Graph for Commonsense Question Answering

Overview

- Given a multiple choice commonsense question:
    - Identify entities in the question and choice
    - Identify triples from conceptnet that connect question and answer.
    - Use wiktionary to retrieve definition of all the concepts mention in the question and answer choices
- Feed the question and the choices individually to ALBERT, and classify

Fusing Context Into Knowledge Graph for Commonsense Question Answering

Idea



Fusing Context Into Knowledge Graph for Commonsense Question Answering

Results

Methods	Single	Ensemble
BERT+OMCS	62.5	-
RoBERTa	72.1	72.5
RoBERTa+HyKAS	73.2	-
XLNet+DREAM	-	73.3
RoBERTa+KE	73.3	-
RoBERTa+KEDGN	-	74.4
XLNet+GraphReason	75.3	-
ALBERT	-	76.5
RoBERTa+MHGRN	75.4	76.5
ALBERT+PG-Full	75.6	78.2
T5	78.1	-
ALBERT+KRD	78.4	-
UnifiedQA	79.1	-
ALBERT+KCR	79.5	-
DEKCOR (ours)	**80.7**	**83.3**

Commonsense QA

Methods	Accuracy
BERT + Careful Selection	72.0
AristoRoBERTa	77.8
ALBERT + KB	81.0
ALBERT + PG-Full	81.8
TTTTT (T5-3B)	83.2
UnifiedQA (T5-11B)	**87.2**
DEKCOR (ours)	<u>82.4</u>

OpenBook QA

Retrieval Enhanced Model for Commonsense Generation

Han Wang[1*], **Yang Liu**[2], **Chenguang Zhu**[2], **Linjun Shou**[3],
Ming Gong[3], **Yichong Xu**[2], **Michael Zeng**[2]
[1]New York University
[2]Microsoft Cognitive Services Research Group
[3]STCA NLP Group, Microsoft, Beijing, China
hwang@nyu.edu
{yaliu10,chezhu,lisho,migon,yicxu,nzeng}@microsoft.com

ACL 2021

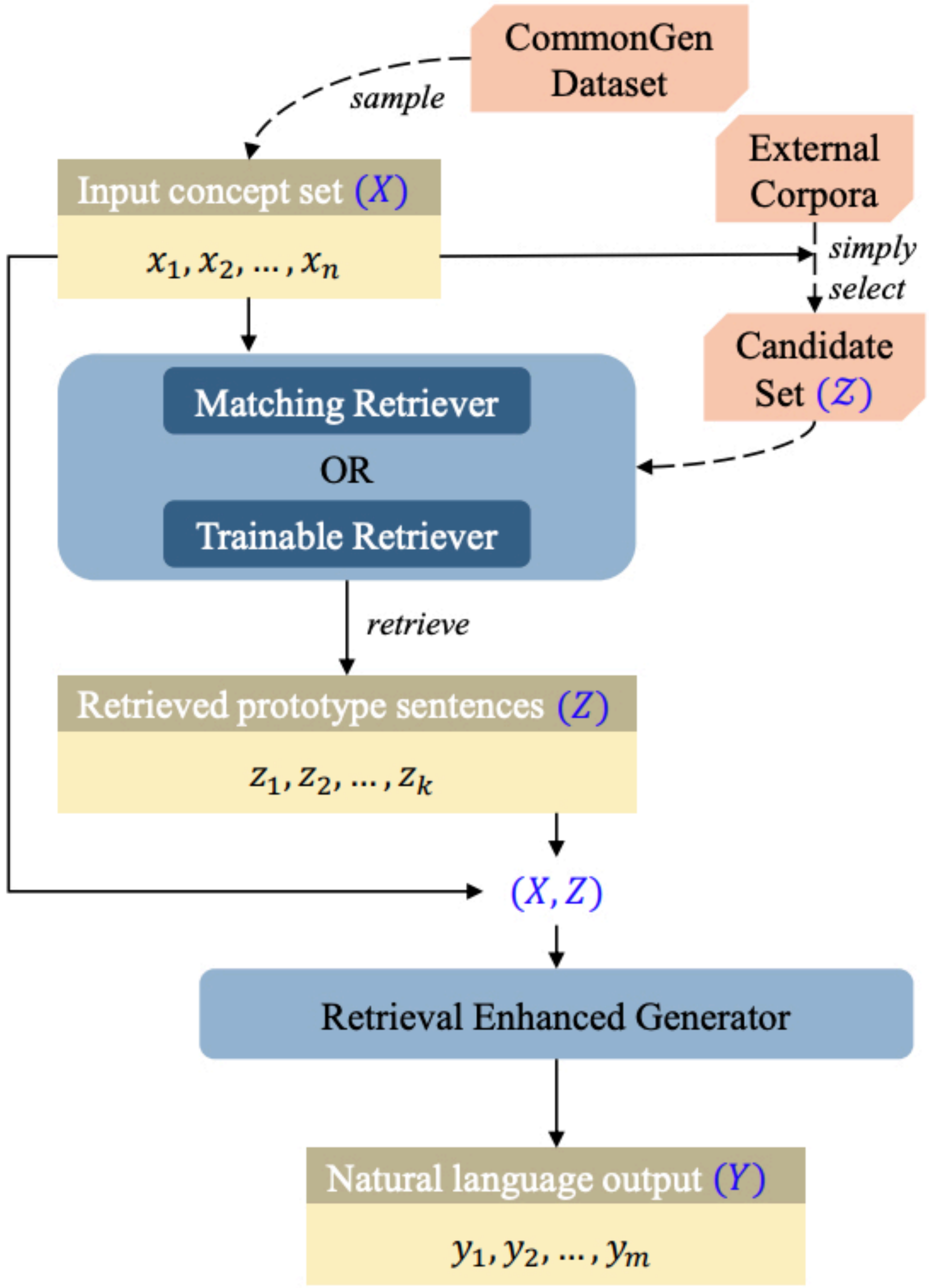
Retrieval Enhanced Model for Commonsense Generation

Overview

- Task: commongen
  - {tree, apple, grow} $\longrightarrow$ Apples grow on tree.
- Method:
  - For a given set of input concepts, retrieve sentences that contain them.
  - Re-rank the retrieved sentences.
- Also do CALM style pre-training

Retrieval Enhanced Model for Commonsense Generation

Method



- Trainable retriever: train BERT to rank the true sentence the highest (binary classification task).

  score(y) > score(z_i)

Retrieval Enhanced Model for Commonsense Generation

Example

Concept Set: trailer shirt side sit road
T5: A man sits on the side of a trailer and a shirt.
Matching Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Two men, one wearing a straw cone hat, blue shirt, talking with a guy in a tan sunhat, red plaid shirt, both with baskets in front of them, sitting on the side of a dirt road. (3)An older guy with a tan shirt and hat sitting on the side of a road with bricks all around him and a small green bowl on the side. **RE-T5(matching retriever)**: a man in a tan shirt sits on the side of a road.
Trainable Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Teenagers in matching shirts stand at the side of the road holding trash bags. (3)A man in a white shirt and black pants standing at the side or the road. **RE-T5(trainable retriever)**: a man in a white shirt and black pants sits on the side of a trailer on the road.

Retrieval Enhanced Model for Commonsense Generation

Results

Model	BLEU-4	CIDEr	**SPICE**	SPICE(v1.0)
GPT-2 (Radford et al., 2019)	26.833	12.187	23.567	25.90
BERT-Gen (Bao et al., 2020)	23.468	12.606	24.822	27.30
UniLM (Dong et al., 2019)	30.616	14.889	27.429	30.20
BART (Lewis et al., 2020)	31.827	13.976	27.995	30.60
T5-base (Raffel et al., 2020)	18.546	9.399	19.871	22.00
T5-large (Raffel et al., 2020)	31.962	15.128	28.855	31.60
EKI-BART (Fan et al., 2020)	35.945	16.999	29.583	32.40
KG-BART (Liu et al., 2021)	33.867	16.927	29.634	32.70
CALM(T5-base) (Zhou et al., 2021)	-	-	-	33.00
RE-T5 (ours)	**40.863**	**17.663**	**31.079**	**34.30**

Model	**SPICE**
Retrieve (only)	29.60
T5	30.80[3]
T5 + *MR*	33.60
T5 + *MR* + pretrain	33.90
RE-T5 (T5 + *TR* + pretrain)	**34.30**

Retrieval-based augmentation

- KFCNet:

Li, Haonan, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. "KFCNet: Knowledge Filtering and Contrastive Learning Network for Generative Commonsense Reasoning." ***EMNLP 2021***

- Differentiable open-ended commonsense reasoning

Lin, Bill Yuchen, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. "Differentiable Open-Ended Commonsense Reasoning." In ***Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*****, pp. 4611-4625. 2021.**

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning ✓
- Pre-trained language models ✓
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training ✓
2. Retrieval-based augmentation ✓
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Model-based augmentation

Model-based augmentation

Overview

- Conceptnet and knowledge bases contain information about a fixed set of entities
- How do we generate information to augment when open-domain events are involved?
  - What happens when you smash a glass on a wooden floor?
  - If someone is wearing sunglasses, is it more likely that rain is falling?
- Language models as knowledge bases
- Our recent works

Carnegie Mellon University
Language Technologies Institute

Explainable defeasible reasoning over graphs using Mixture-of-experts

Aman Madaan, Yiming Yang

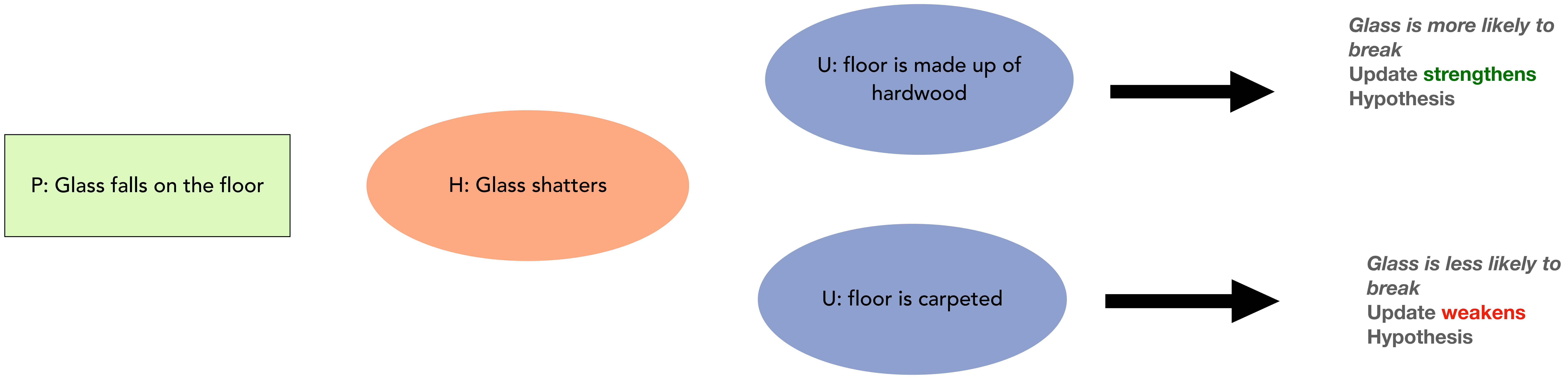
Joint work with Niket Tandon, Dheeraj Rajagopal, Peter Clark, Eduard Hovy

https://github.com/madaan/thinkaboutit

Defeasible Reasoning

Defeasible Reasoning

- A classification task
- Given a premise **P,** a hypothesis **H**
    - New evidence (update) **U** may be weaken or strengthen the hypothesis



Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. "Thinking like a skeptic: Defeasible inference in natural language." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4661-4675. 2020.

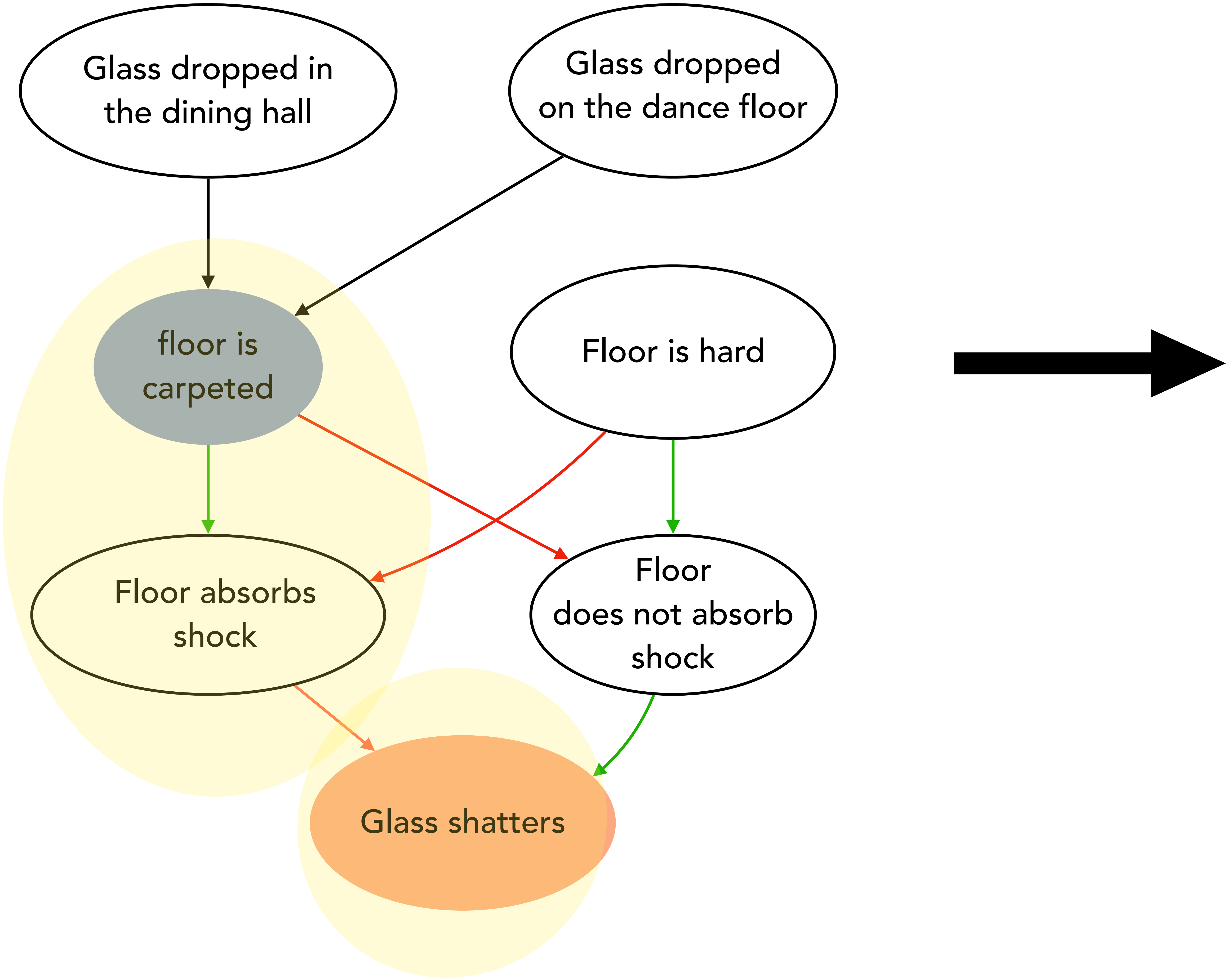
J. Pollock. 1987. Defeasible reasoning. *Cogn. Sci.*, 11:481–518.

Defeasible reasoning requires implicit background knowledge

P: Glass falls on the floor

H: Glass shatters

U: floor is carpeted

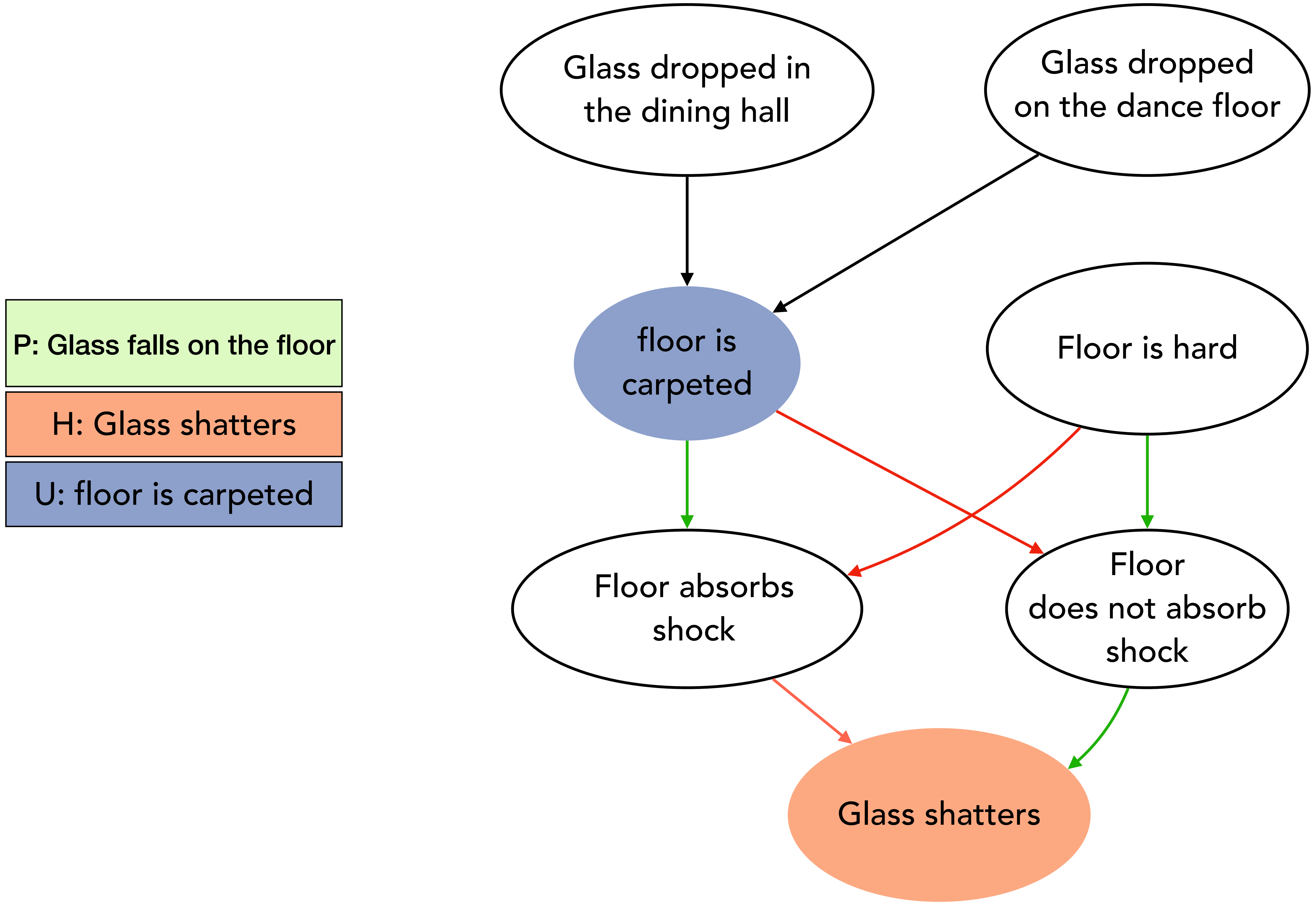


Glass is less likely to break

Update weakens Hypothesis

Dataset

- A large dataset of defeasible reasoning queries and such graphs are available
- Dataset of defeasible queries [1]
    - Manually created, spans three domains:
        - ATOMIC (Commonsense, 43K)
        - SOCIAL-CHEM-101 (Social norms, 95K)
        - SNLI (NLI, 92K)
- A lot of implicit knowledge is used for answering these queries
    - Dataset of graphs generated using transfer learning [2]
    - For each defeasible query, the graph captures additional context that can be useful
    - Think about the question scenario before answering



[1] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. "Thinking like a skeptic: Defeasible inference in natural language." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4661-4675. 2020.

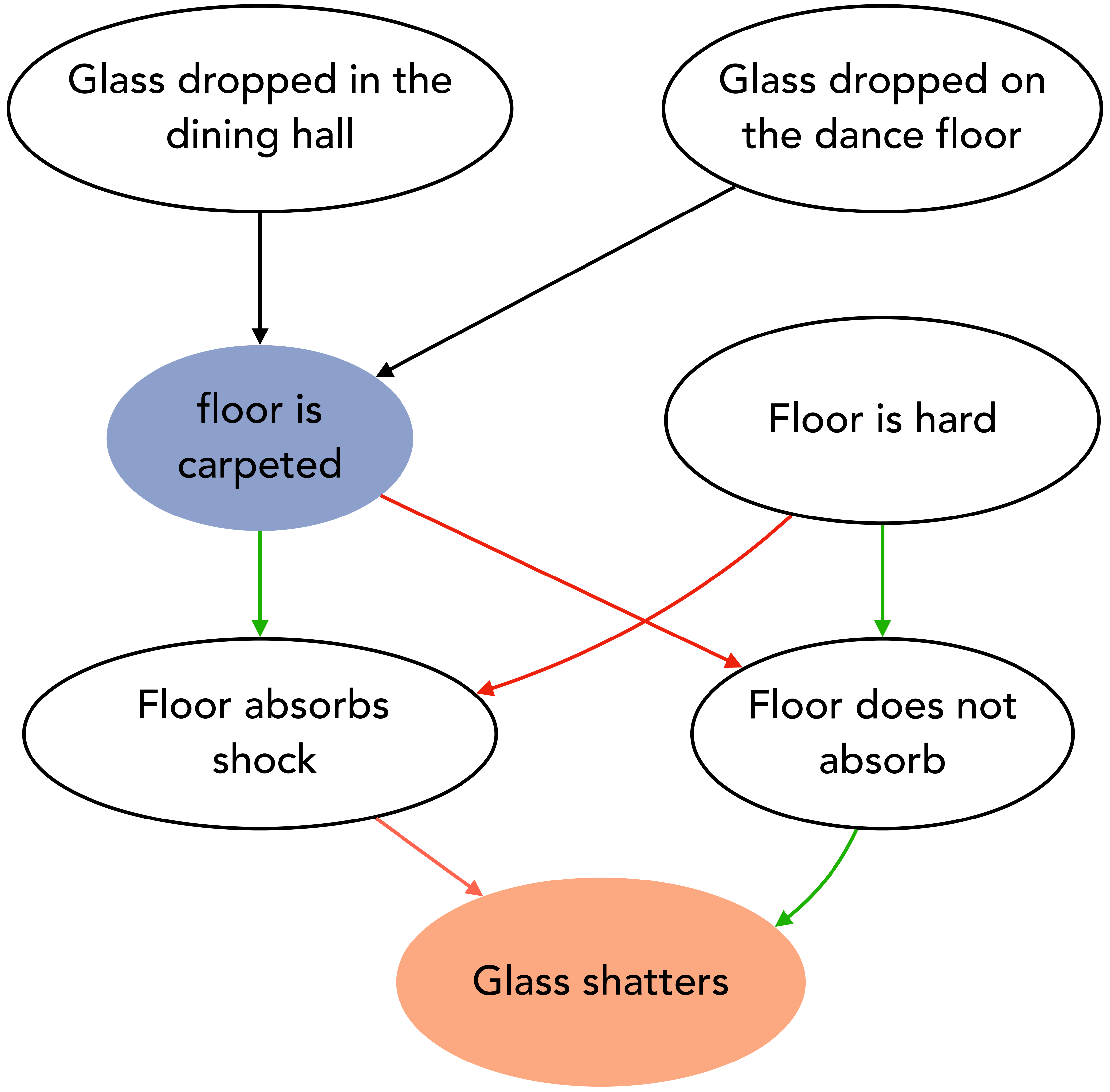
[2] Aman Madaan, Dheeraj Rajagopal, Niket Tandon, Yiming Yang and Eduard H. Hovy. “Could you give me a hint ? Generating inference graphs for defeasible reasoning.” *ACL FINDINGS* (2021).

How to best use graphs for defeasible reasoning?

P: Glass falls on the floor

H: Glass shatters

U: floor is carpeted



Floor is ca strengthens Hypothesis

Update weakens Hypothesis

Given a defeasible query PHU

And a graph generated for the query (augmented information)

Generate classification label

Baselines

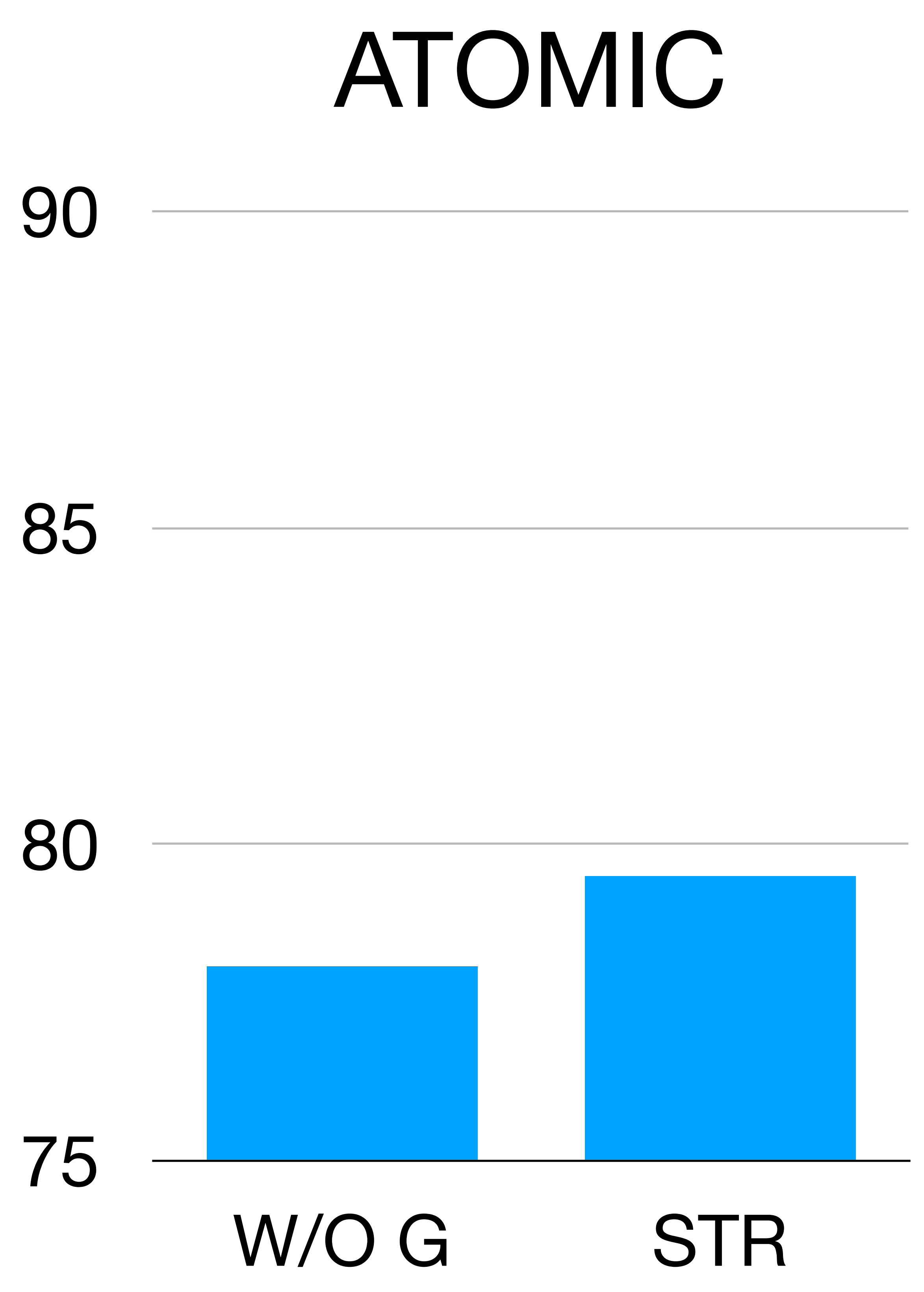
- **W/O G:** Concatenate PHU as a single string and fine-tune (Rudinger et al. 2020)

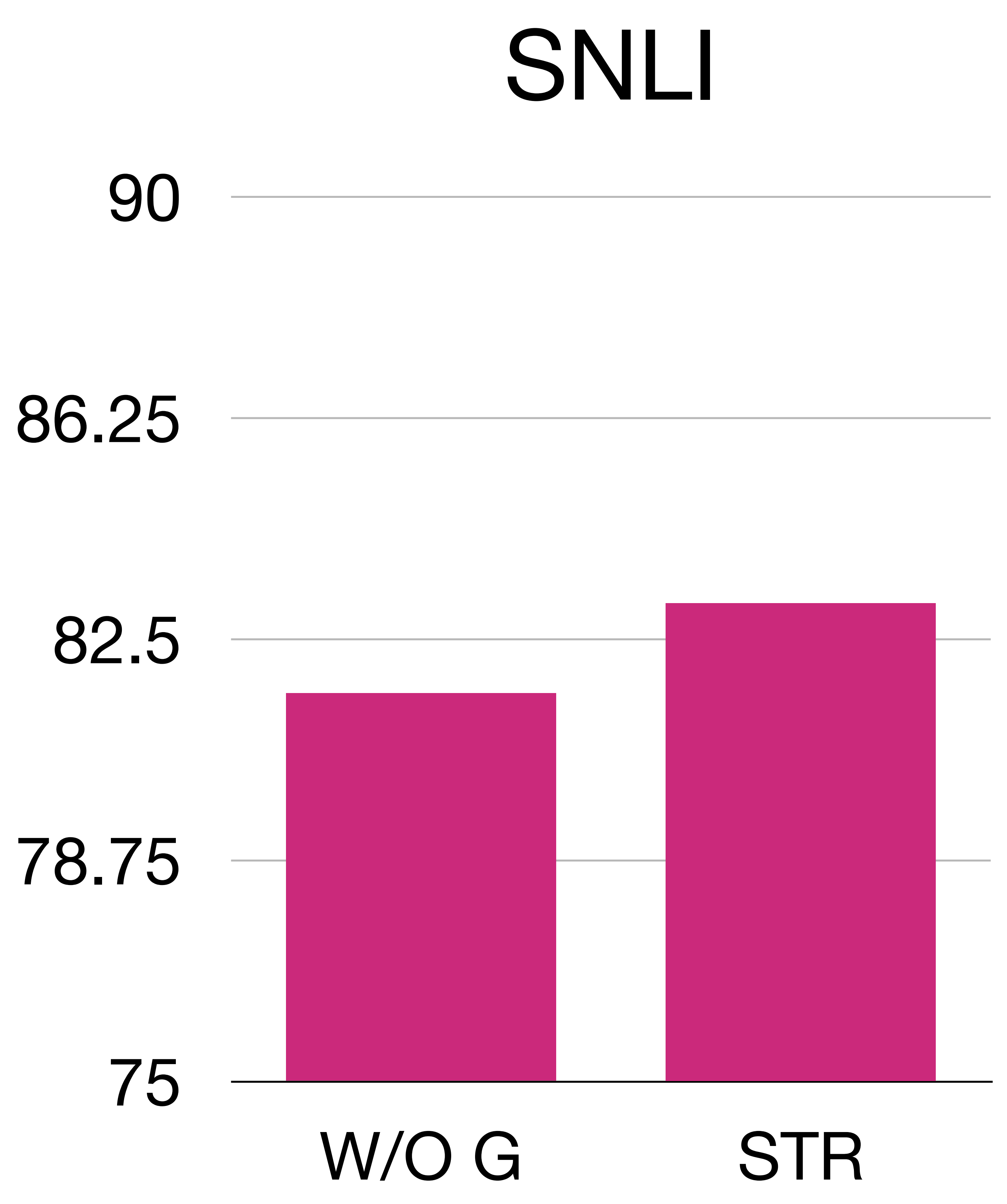
- **STR**: Append G in a string format after PHU, and fine-tune
  - Break down the graph into node - edge - node triplets
  - Append the triplets with the query as a string

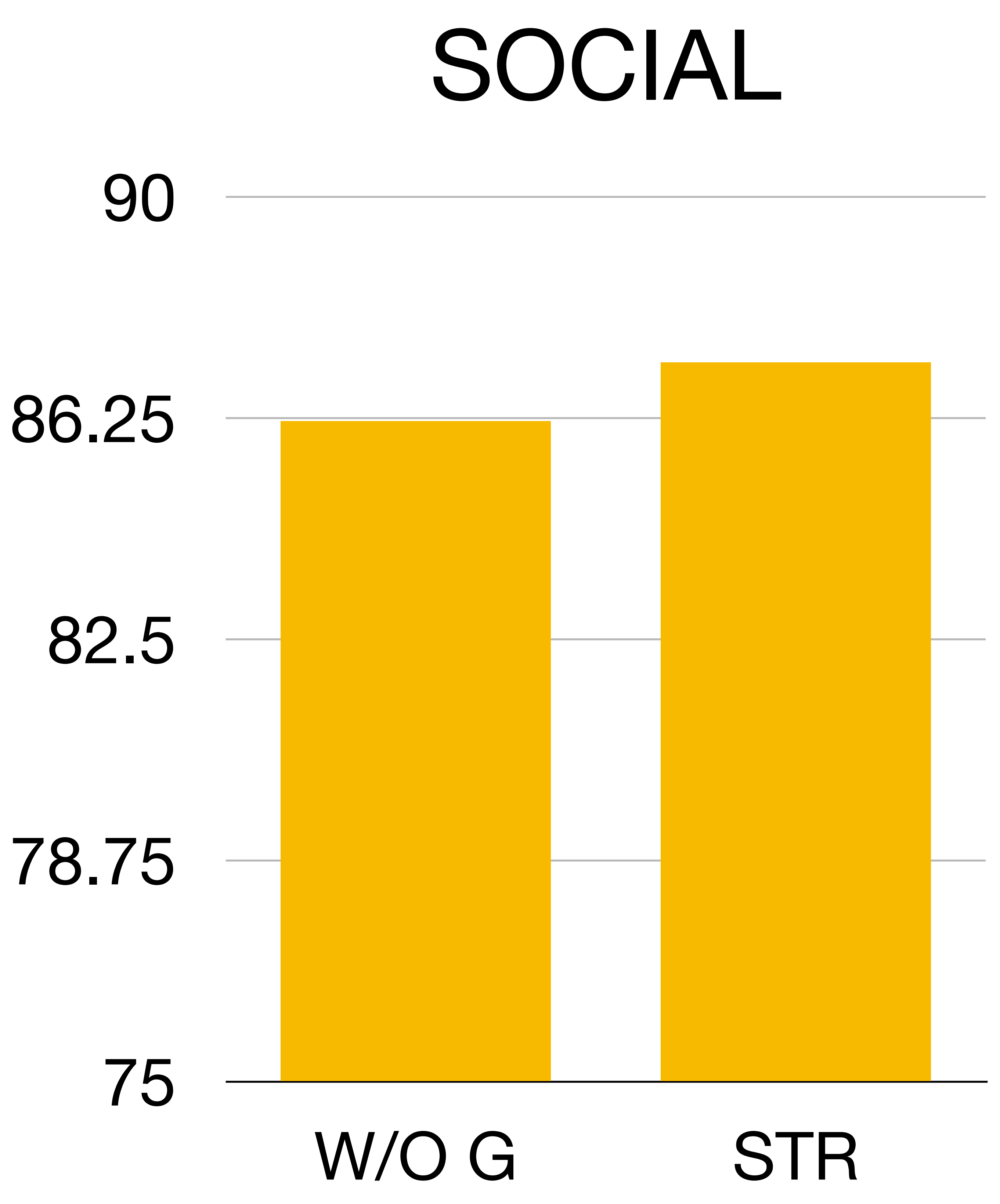
- Use RoBERTa as the encoder

Results

Baselines







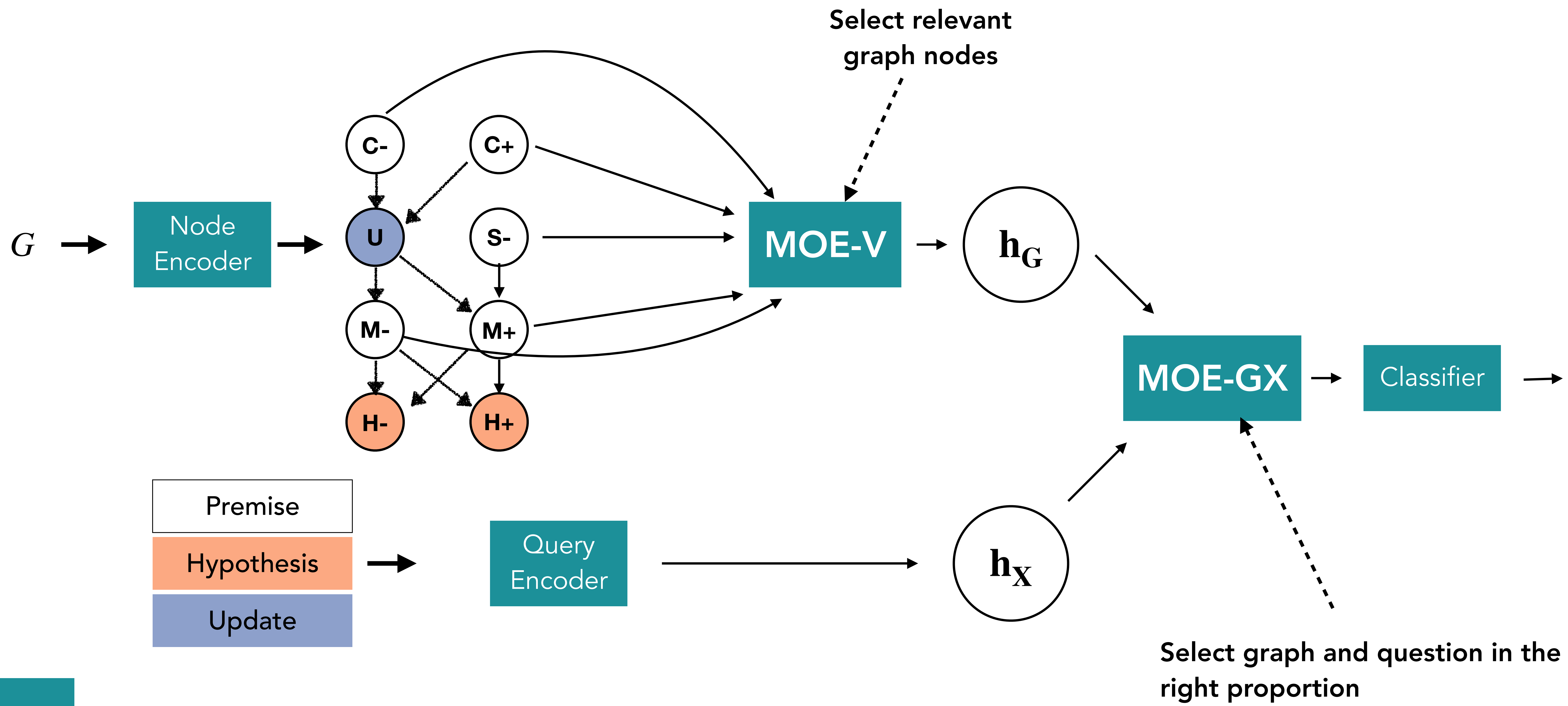
Performance: can we do better?

Explainability: can we identify which parts of the graph are more useful?

How to best use graphs to increase performance on the task?

- **STR**: Discards the semantics of various parts of G (mediator, external situation etc.)
- From human evaluation:
  - **Not every part** of the augmented graph was useful
  - The augmented graph was **not always useful**
- Thus the model needs to be able to:
  - **Selectively use** parts of the input
  - **Discard** augmentation completely

Mixture of experts for pooling graph representations



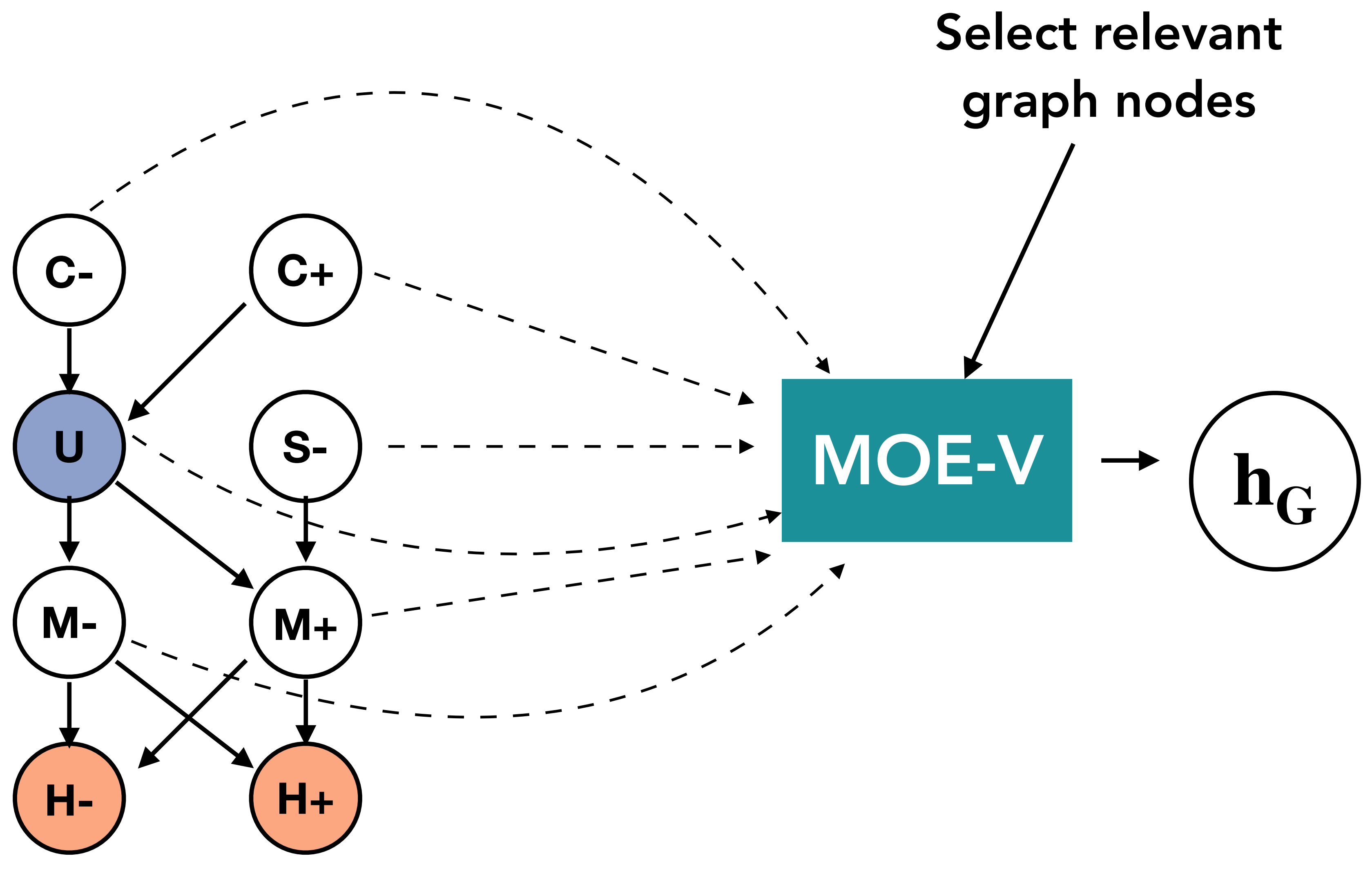
Details of the MoE model in the paper

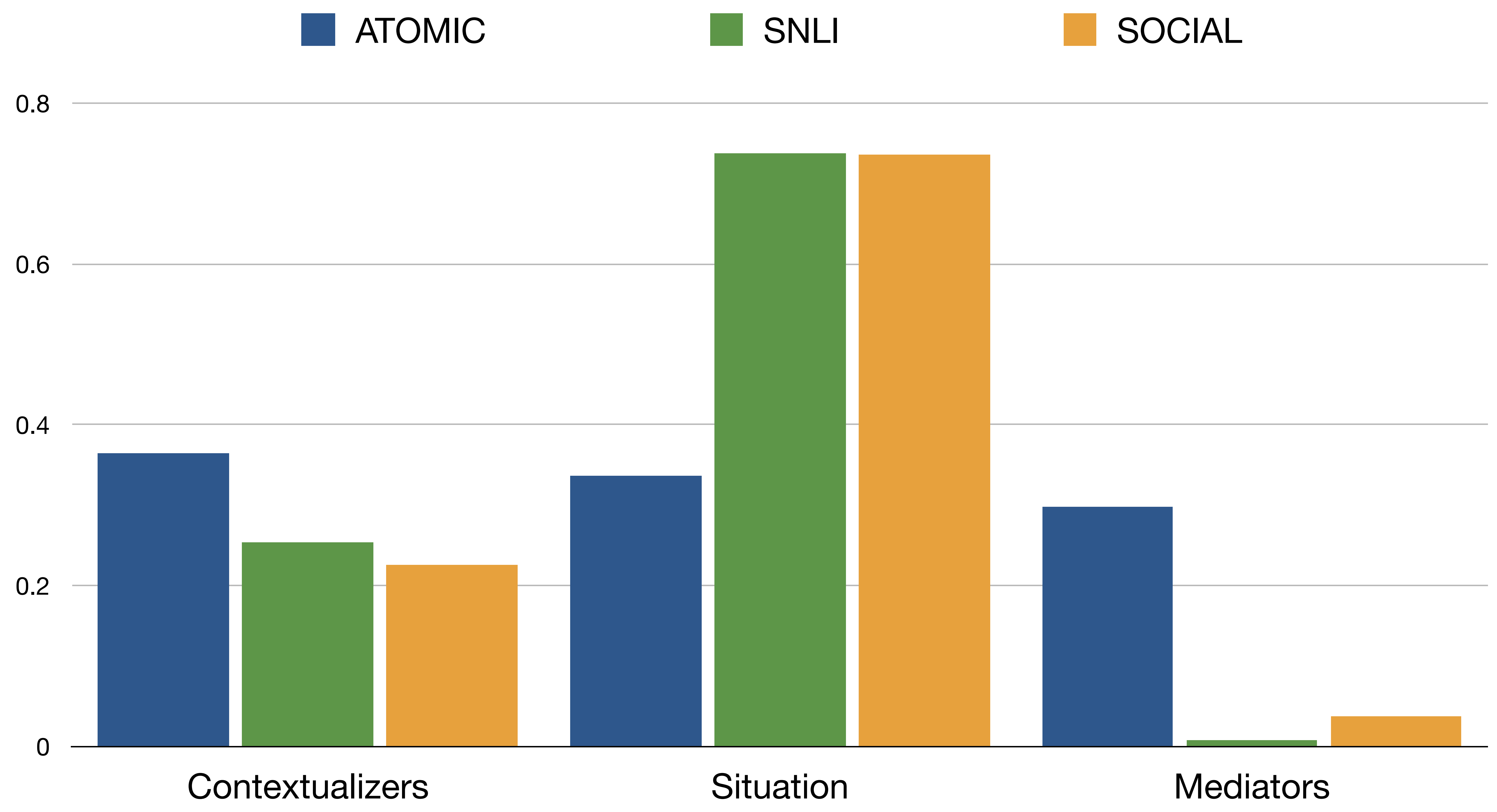
R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive mixtures of local experts", *Neural Comput.*, vol. 3, no. 1, pp. 79-87, 1991.

Results

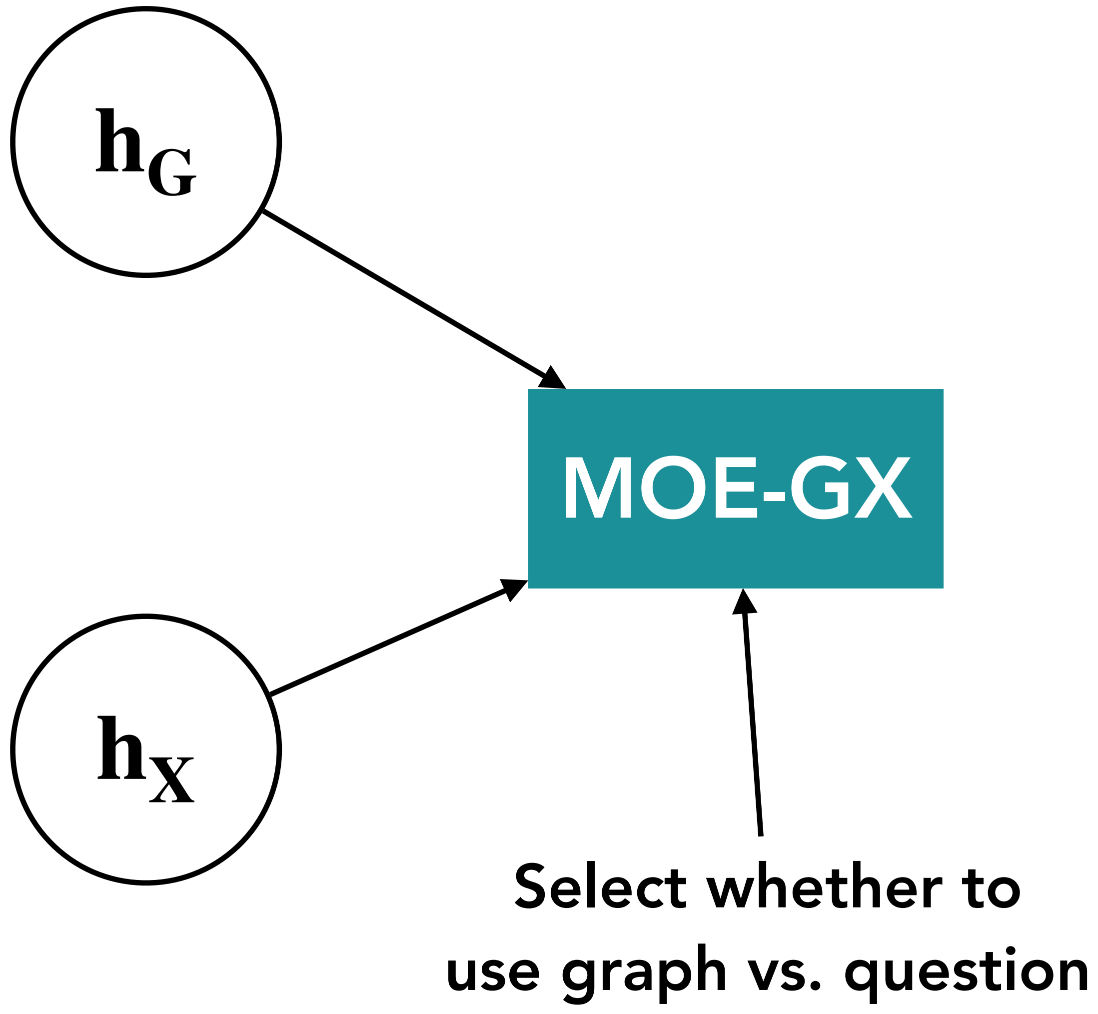

Atomic
81
80.25
79.5
78.75
78
W/O G
STR
GCN
MoE
SNLI
86
84.75
83.5
82.25
81
W/O G
STR
GCN
MoE
SOCIAL
89
88.25
87.5
86.75
86
W/O G
STR
GCN
MoE

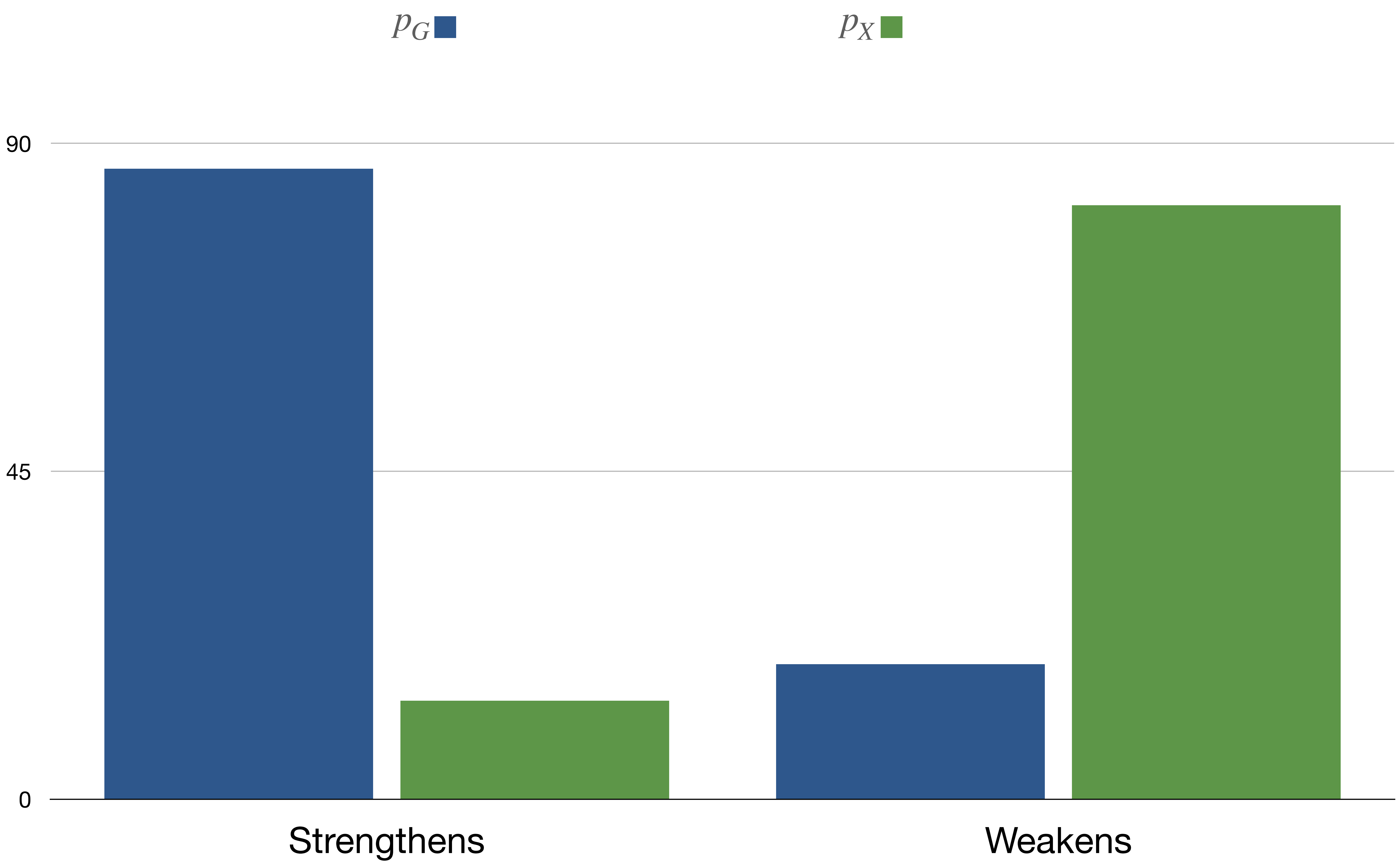
Situation nodes are more important



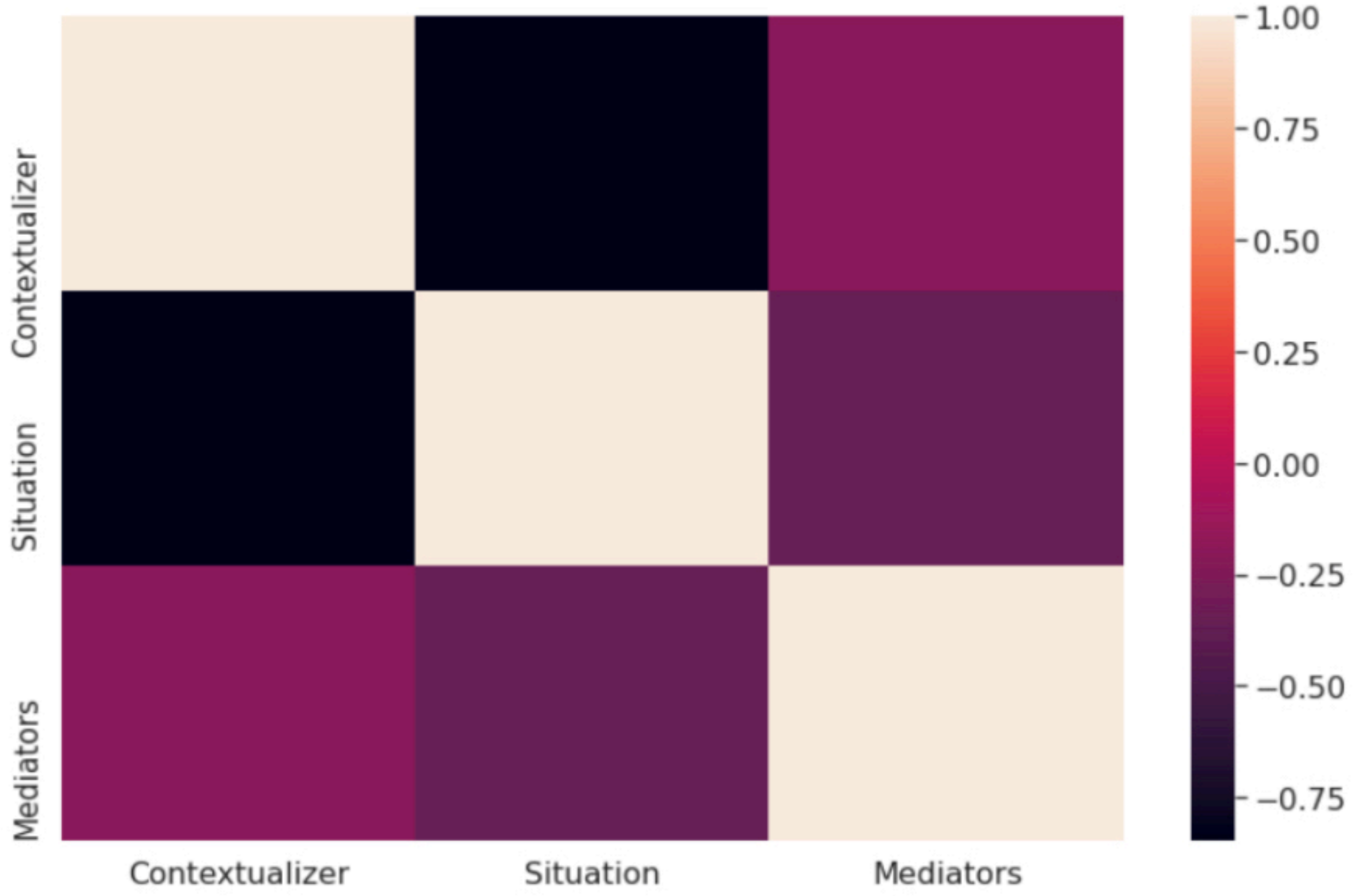


Graphs are more useful for strengthens questions





MOE-V learns the node semantics



More discussion on explainability in the paper!

Summary

- Thinking about a question scenario before modeling it helps the models

- Mixture-of-experts allows effective and explainable learning over graphs

- For KAIROS, similar strategies can be used to highlight the part of schemas that were used in matching or prediction

Code, pre-trained models, data for the EMNLP 2021 paper: https://github.com/madaan/thinkaboutit

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning ✓
- Pre-trained language models ✓
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training ✓
2. Retrieval-based augmentation ✓
3. Model-based augmentation ✓
4. Formal logic and symbolic reasoning

Formal logic and symbolic reasoning

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief

Nora Kassner[1,2], Oyvind Tafjord[1], Hinrich Schütze[2], Peter Clark[1]
[1]Allen Institute for AI, Seattle, WA
[2]Center for Information and Language Processing, LMU Munich, Germany
kassner@cis.lmu.de
{oyvindt,peterc}@allenai.org

EMNLP 2021

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief

Overview

- Language models are getting bigger to the point that even fine-tuning is intractable
- Can we add formal constraints on the model to improve its performance?
- Test of a consistent belief (e.g., “eagles are birds”)
  - Re-phrasings are Are eagles birds? Is an eagle a type of bird?
  - Consistently talk about all the downstream tasks

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief

Definitions

- **Belief**: a weighted triple (s, l, w)
  - s is a sentence (*a poodle is a dog*)
  - l is the label ∈ {true, false} (*true*)
  - w is system's strength of the belief (*0.9*)
- **Belief-bank**: a set of beliefs
- **Constraint**: a 5 tuple $(s_i . l_i \rightarrow s_j . l_j, w_i)$
  - Connects two beliefs with a weight if they are violated.
  - "X is a dog".T → "X has a tail".T, 0.8
    - Dogs usually have a tail
  - "X is a bird".T → "X is a fish".T, 1.0
    - A fish cannot be a bird
- **Consistency:**
  - $\tau = |\{ c_i \mid \neg(s_i.l_i \rightarrow s_j.l_j) \}| \ / \ |\{ c_i \mid s_i.l_i \}|$

Beliefs	12.5k
Constraints	2600

BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief

Overview

- Given:
  - A ***stream*** of sentences **Q**, each with a truth value (i.e., true or false)
  - A set of constraints **C(s)** between sentences in Q, each with a penalty w

    Constraints
  - A **Model M** that maps Q → {true, false}
  - A **SAT solver**, that can flip the truth value of sentences to incur minimum penalty

    SAT Solver
- Task:
  - Accumulate the labels for Q as predicted by M, so that they are globally consistent

Architecture

a swallow is a bird T
a swallow has fur F
A swallow has wings T

SAT Solver

Constraints

A swallow is a fish?

Model

T

a swallow is a bird T
a swallow is a fish ~~T~~ F
a swallow has fur F
A swallow has wings T

Belief bank

SAT Solver

Constraints

Relevant beliefs from Belief bank

A swallow is not a fish

+

Do swallows have gills?

Model

F

a swallow is a bird T
a swallow is a fish F
a swallow has fur F
A swallow has wings T

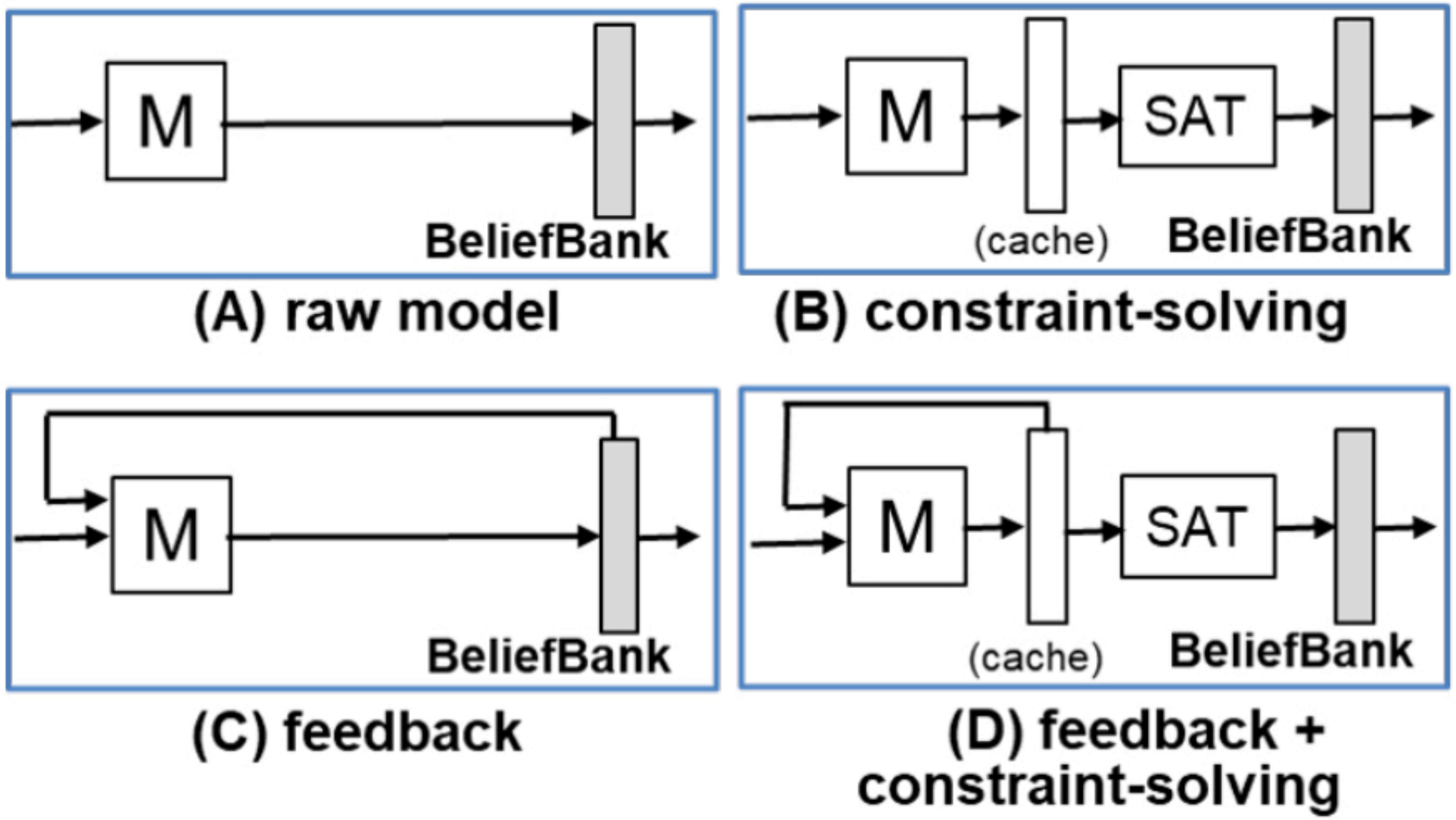
Belief bank

SAT Solver

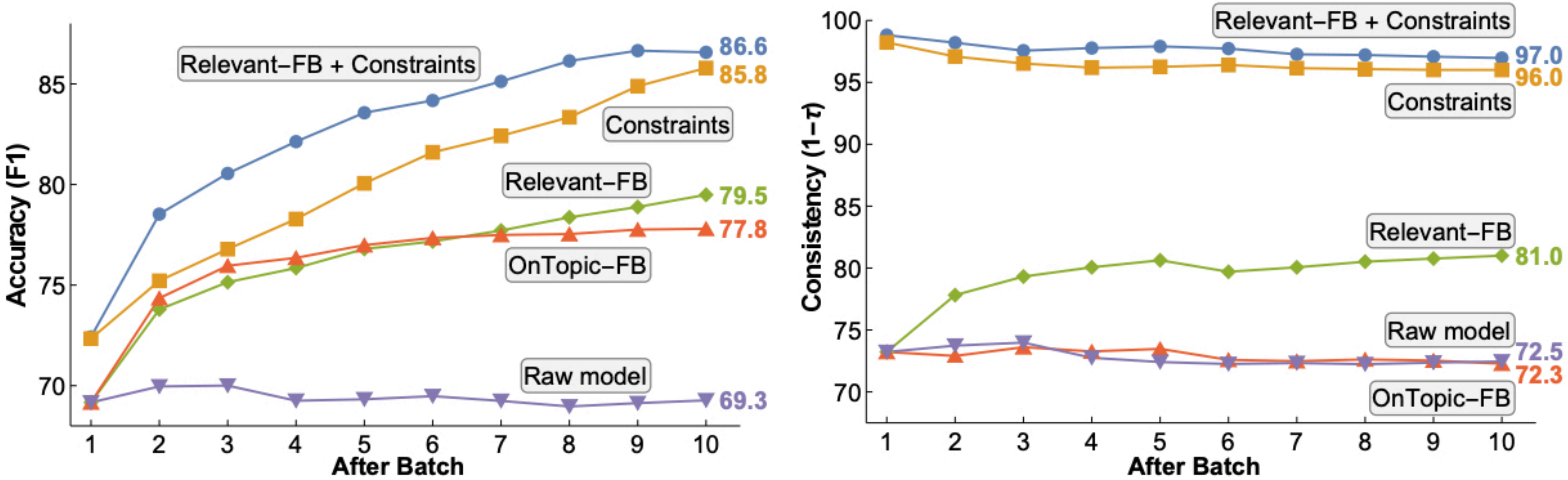
Constraints

Just asking *Do swallows have gills?* Leads to True!

Method


M
BeliefBank
(A) raw model
M
SAT
(cache)
BeliefBank
(B) constraint-solving
M
BeliefBank
(C) feedback
M
SAT
(cache)
BeliefBank
(D) feedback +
constraint-solving

Results



OnTopic-FB = using (randomly selected) **on-topic** feedback from old answers for new queries.
Relevant-FB = using most **relevant** on-topic feedback for new queries.
Constraints = running the constraint-solver after each batch.

Figure 3: Accuracy (left) and consistency (right) of the growing BeliefBank, as the system answers incrementally more questions (each batch = 10% of the queries). Relevant feedback, constraint-solving, and both, all help improve both F1 and Consistency.

Formal logic and symbolic reasoning

Additional references

- Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic RulesTom Mitchell

Arabshahi, Forough, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom Mitchell. "Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules." ***EMNLP 2021***

- Improving GPT-3 after deployment with a dynamic memory of feedback

https://openreview.net/forum?id=6DBkg64mzt6

Today: Language models + commonsense reasoning

Outline

- Commonsense reasoning
- Pre-trained language models
- The four ways of using PLTM for commonsense reasoning:

1. Pre-training
2. Retrieval-based augmentation
3. Model-based augmentation
4. Formal logic and symbolic reasoning

Do the models really have commonsense?

Clever Hans

- Giving right answer for the wrong reasons?
- Are the models *really* doing commonsense reasoning?
- Does it even matter?

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar[1,2] Hongming Zhang[3,4] Yoav Goldberg[1,2] Dan Roth[4]
[1]Bar Ilan University, [2]AI2, [3]HKUST, [4]UPenn
{yanaiela,yoav.goldberg}@gmail.com
hzhangal@cse.ust.hk, danroth@seas.upenn.edu

EMNLP 2021

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Overview

The trophy doesn't fit into *the brown suitcase* because **it** is too large.

The trophy doesn't fit into *the brown suitcase* because **it** is too small.

🏆 trophy

💼 suitcase

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Overview

- Test if the model is giving the right answer for the right reasons
- If the model *really knew*
    - It should have no trouble getting both the questions in the pair right
    - Performance on questions that do not have enough information should be random

Original		
twin-1	*The trophy* doesn't fit into *the brown suitcase* because **it** is too large.	🏆 trophy
twin-2	*The trophy* doesn't fit into *the brown suitcase* because **it** is too small.	💼 suitcase

Baselines		
no-cands	doesn't fit into because **it** is too large.	?
part-sent	because **it** is too large.	?

Results

Dataset	Setup	Single	Group
WSC	original	89.71	79.41
	no-cands	60.72	40.35
	part-sent	64.88	33.88
WSC-na	original	89.45	79.09
	no-cands	58.06	34.41
	part-sent	59.90	25.00
Winogrande	original	71.49	58.45
	no-cands	53.07	31.05
	part-sent	53.11	22.34

Additional references

Zhou, Pei, Pegah Jandaghi, Bill Yuchen Lin, Justin Cho, Jay Pujara, and Xiang Ren. "Probing Causal Common Sense in Dialogue Response Generation." *EMNLP 2021*

Wang, Peifeng, Filip Ilievski, Muhao Chen, and Xiang Ren. "Do Language Models Perform Generalizable Commonsense Inference?." *arXiv preprint arXiv:2106.11533* (2021).

What’s next?

- Exploring what exactly are these large language models learning?
- How much data do they need to generalize?
- How does that knowledge transfer to the real world?
- Interactive learning
- Multi-modal commonsense reasoning

Language models + commonsense reasoning

Summary

- Using large pre-trained language models (PTLM) for commonsense reasoning
- **The four paths to commonsense reasoning:**
    1. **Pre-training**

       Pre-train with novel objectives
    2. **Retrieval-based augmentation**

       Supplement LM with additional information
    3. **Model-based augmentation**

       Use another model to generate open-ended augmentation
    4. **Formal logic and symbolic reasoning**

       Drastically different techniques, not everything is an embedding
- **Do the models really have commonsense?**
    - Depends on the definition
    - Probably not (yet), but more investigation is needed
- **Resources:** ACL 2020 Tutorial: https://homes.cs.washington.edu/~msap/acl2020-commonsense/