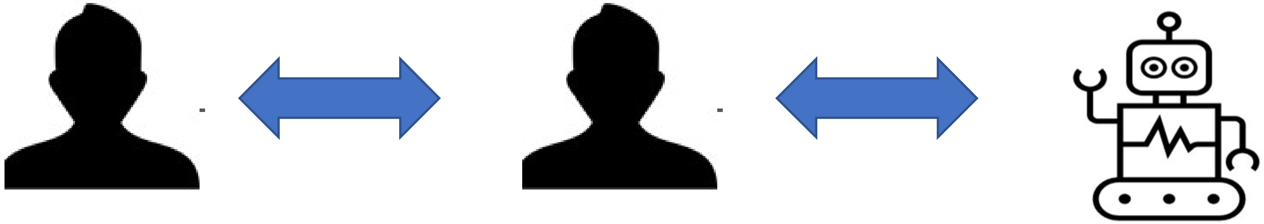
On robustifying NLP beyond scale

Seung-won Hwang
Professor
Department of CSE, Seoul National University

NLP as Ambient Intelligence for Human-Machine Cooperation

- Human/Human: Translation, Simplification / Summarization, Grammatical Error Correction
- Human/Machine: Code Generation, Task-oriented Dialogue (robot, speaker)



NLP 2.0/3.0: PLM + Finetuning (low-code)

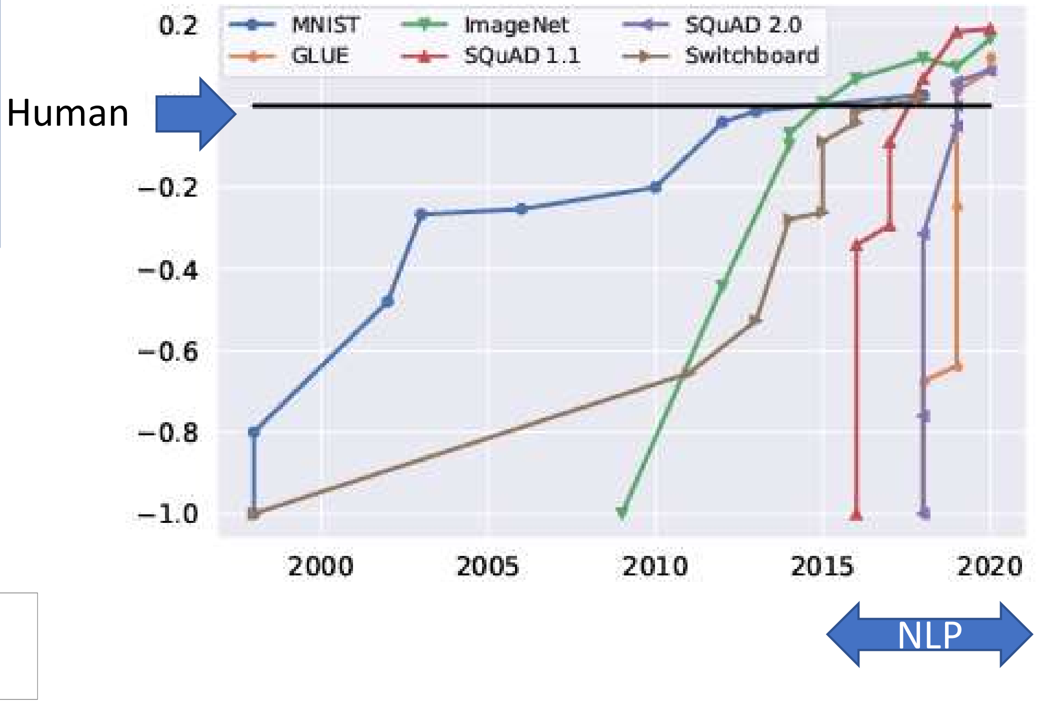
3.0: BERT, ELMo, GPT, ...

x: Capital of california?

Output (Classification)

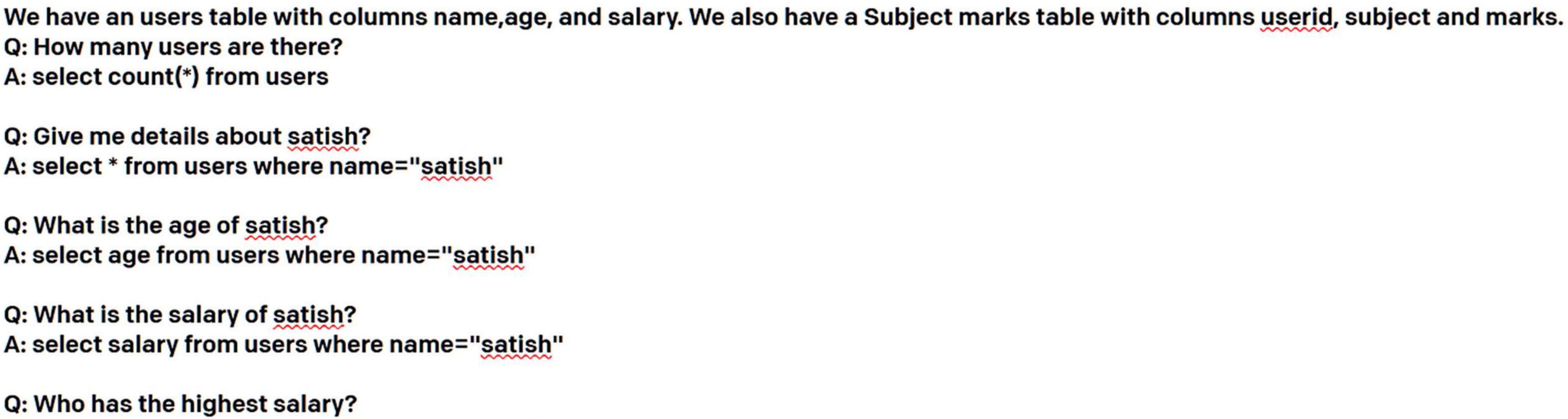
Output (Generation)

y: Select capital from
states
where state='CA'



https://arxiv.org/abs/2104.14337

NLP 3.0: Task as PLM Generation (no-code)



We have an users table with columns name,age, and salary. We also have a Subject marks table with columns userid, subject and marks.

Q: How many users are there?

A: select count(*) from users

Q: Give me details about satish?

A: select * from users where name="satish"

Q: What is the age of satish?

A: select age from users where name="satish"

Q: What is the salary of satish?

A: select salary from users where name="satish"

Q: Who has the highest salary?

mask

https://gpt3experiments.substack.com/p/a-hitchhikers-guide-to-gpt-3-part

Challenge: Robustness of NLP 3.0

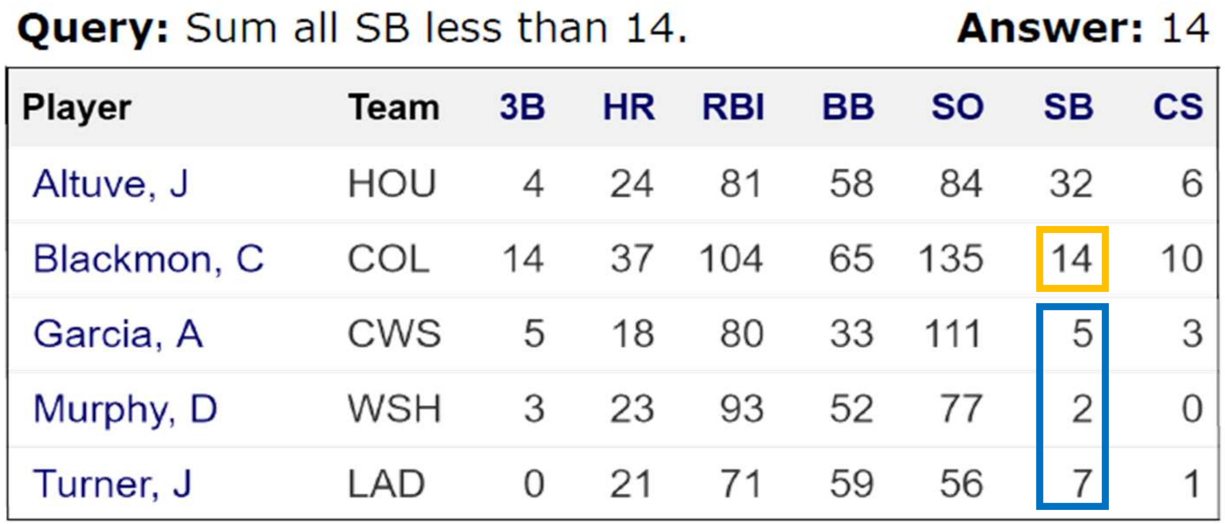
$$\frac{16}{64} = \frac{1\cancel{6}}{\cancel{6}4} = \frac{1}{4}.$$

- PLM may generate text with inconsistency, hallucination, and biases

- Task accuracy is over-estimated, with lucky guesses that cannot generalize

Motivation: On Robustifying Code Generation

- Accurate model in one dataset would not generalize in another, due to dataset artifacts (“14”)



Query: Sum all SB less than 14. **Answer:** 14

Player	Team	3B	HR	RBI	BB	SO	SB	CS
Altuve, J	HOU	4	24	81	58	84	32	6
Blackmon, C	COL	14	37	104	65	135	14	10
Garcia, A	CWS	5	18	80	33	111	5	3
Murphy, D	WSH	3	23	93	52	77	2	0
Turner, J	LAD	0	21	71	59	56	7	1

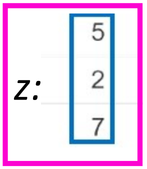
Explanatory and Actionable Debugging for Machine Learning: A TableQA Demonstration, SIGIR 2019

Proposed: Tighter Human-AI Cooperation

- Attention as rationalization supervision

$$L = L_{task}(\hat{y}, y) + \boxed{\mu \cdot L_{att}(\hat{\alpha}, \alpha)}$$

- SQL parser can be plugged in to enrich (x,y)→ *x: Sum all SB less than 14, y: 14, z:*



5

2

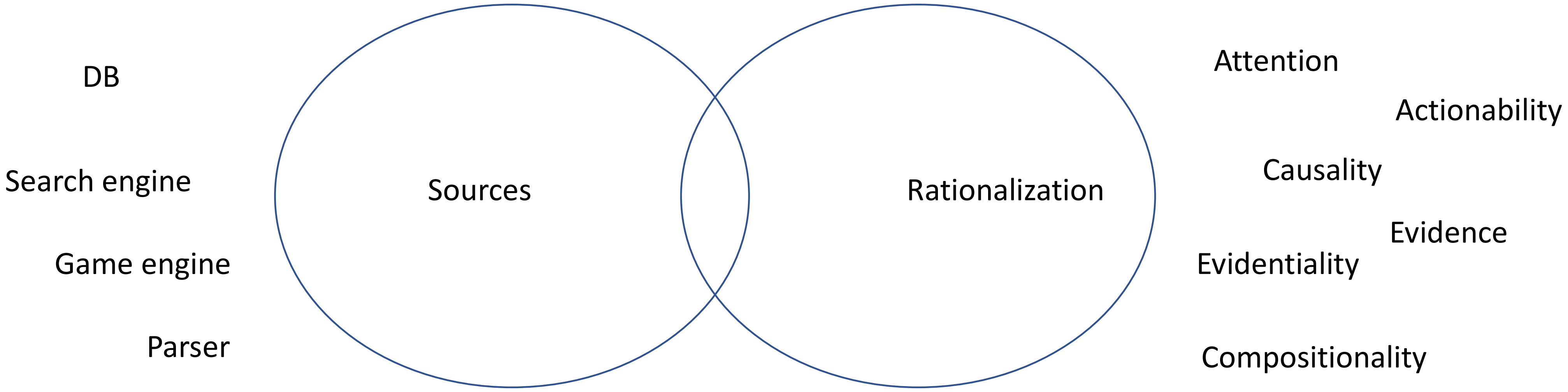
7

Model	rationalization	answer
-	60.9	59.5
With parser	**77.2** (+26.8)	**73.5** (+26.6)

Explanatory and Actionable Debugging for Machine Learning: A TableQA Demonstration, SIGIR 2019

Showcase: Human-Model Aligned Teaching

- Robustifying fine-tuning by distilling *z* from neuro/symbolic models
- Robustifying PLM generation by richer interactions



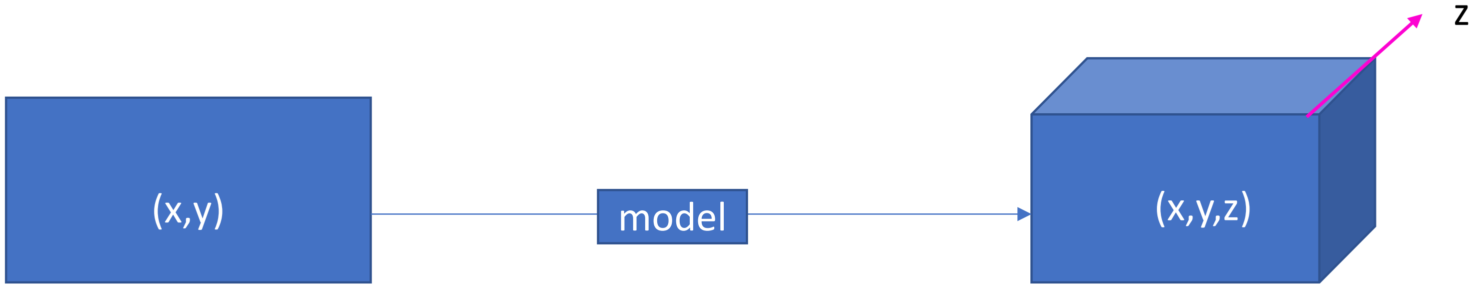
Causality as Rationalization

- Human Annotation (x,y)

- Ask for Rationalization

x: “This *Spielberg* film was wonderful” → “This *Spielberg* film was bad” (-)

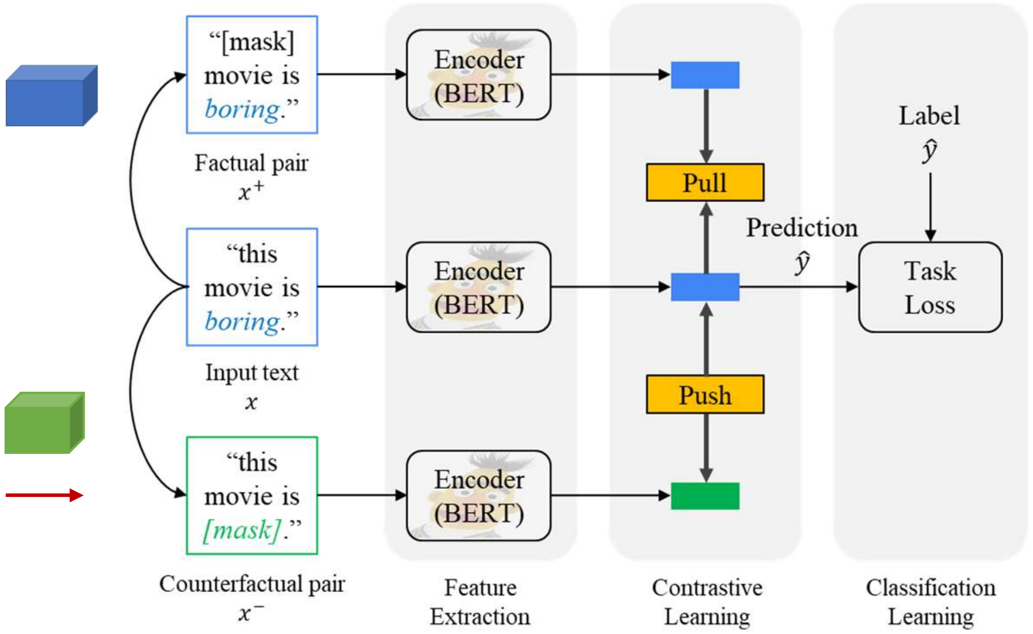
y: + → “This *Bong* film was good” (+)



C2L: Causally Contrastive Learning for Robust Text Classification, AAAI 2022

Causally Contrastive Learning for Robustness

Classifier f_ϕ

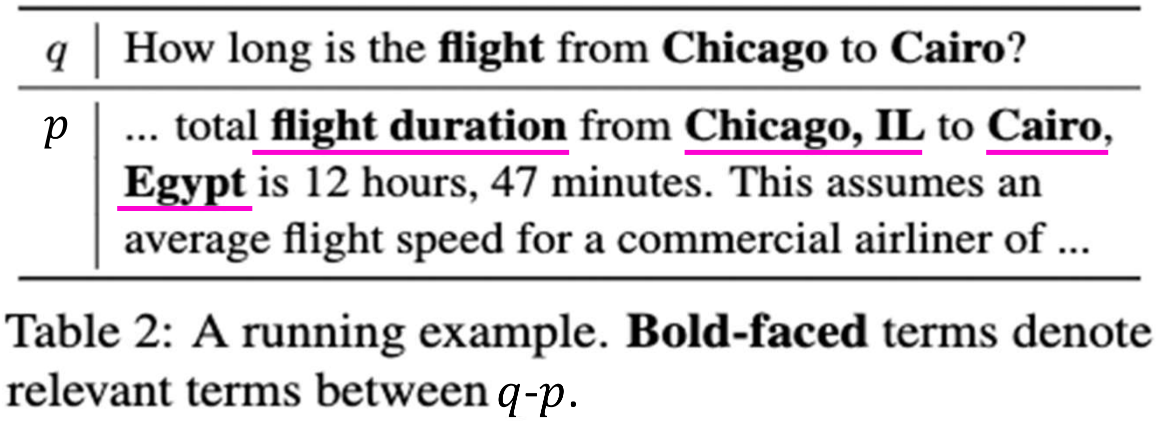


Classifier f_θ

C2L: Causally Contrastive Learning for Robust Text Classification, AAAI 2022

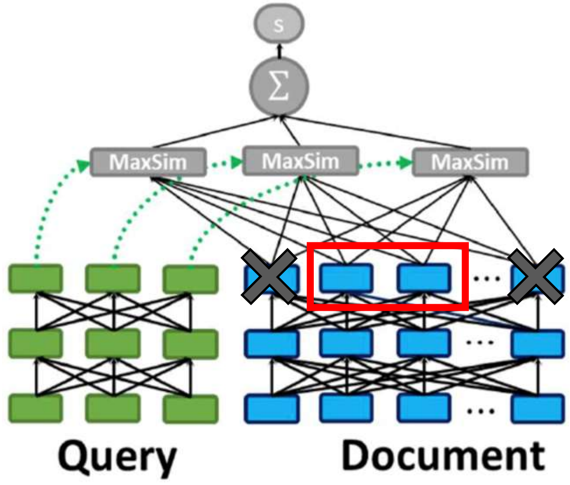
Evidence as Rationalization

- x for search engines are (query, passage) from clicks and relevance annotations
- (x,y) can be rationalized with evidential terms z, distilled from search engines
- Transferred to a commercial-scale search engine for document indexing with 10- and 3-fold decreases in latency and memory footprint



q	How long is the **flight** from **Chicago** to **Cairo**?
p	... total **flight duration** from **Chicago, IL** to **Cairo, Egypt** is 12 hours, 47 minutes. This assumes an average flight speed for a commercial airliner of ...

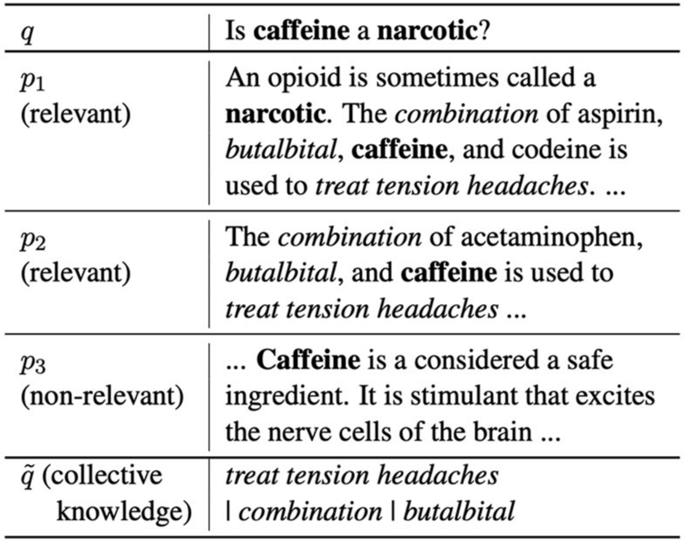
Table 2: A running example. **Bold-faced** terms denote relevant terms between q-p.



Pseudo-Relevance for Enhancing Document Representation, EMNLP 2022

Evidence for Sparse Labels

- Relevance labels (y) are inherently sparse: + is partial and – is implicit
- Pseudo-Relevance Feedback (PRF) positively labels top-k for q for inference
- Using z, PRF can be efficiently done during training for augmenting positive/negative labels



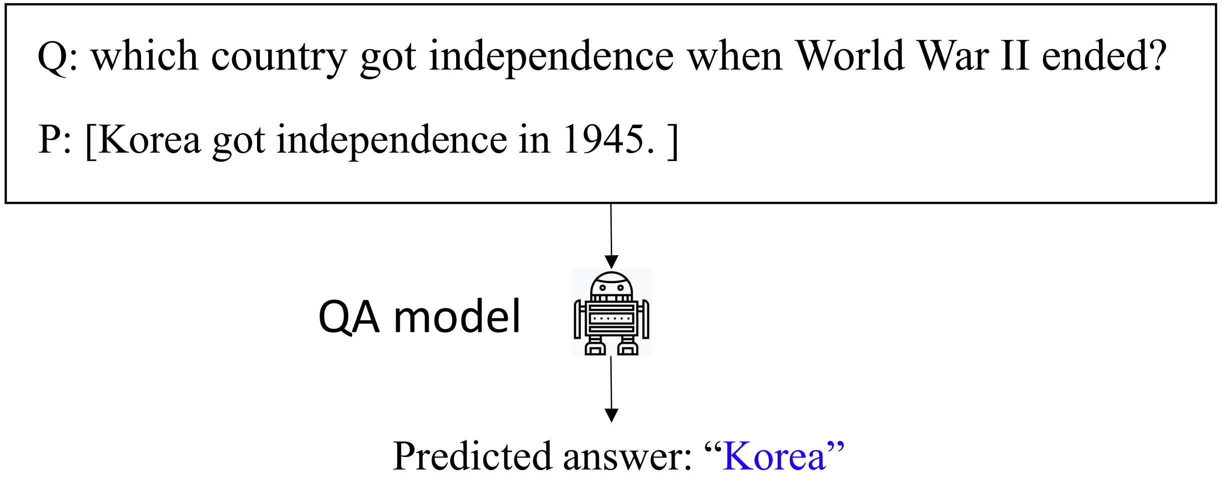
q	Is **caffeine** a **narcotic**?
p_1 (relevant)	An opioid is sometimes called a **narcotic**. The *combination* of aspirin, *butalbital*, **caffeine**, and codeine is used to *treat tension headaches*. ...
p_2 (relevant)	The *combination* of acetaminophen, *butalbital*, and **caffeine** is used to *treat tension headaches* ...
p_3 (non-relevant)	... **Caffeine** is a considered a safe ingredient. It is stimulant that excites the nerve cells of the brain ...
$\tilde{q}$ (collective knowledge)	*treat tension headaches* | *combination* | *butalbital*

Step 1. Train a neural model, to produce z for every passage.
Step 2. Gather top-k (e.g., 3 in left fig) efficiently using z
Step 3. Remove false positives using collective knowledge on z.

Collective Relevance Labeling for Passage Retrieval, NAACL 2022

Evidentiality as Rationalization

- QA model hallucinates an answer using answer type as a spurious pattern



Q: which country got independence when World War II ended?

P: [Korea got independence in 1945.]

QA model

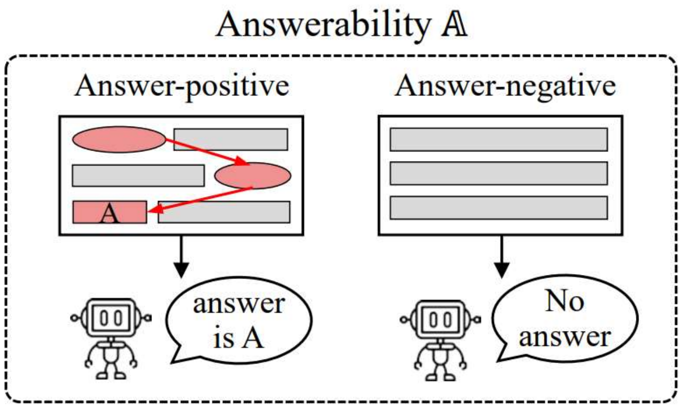
Predicted answer: "Korea"

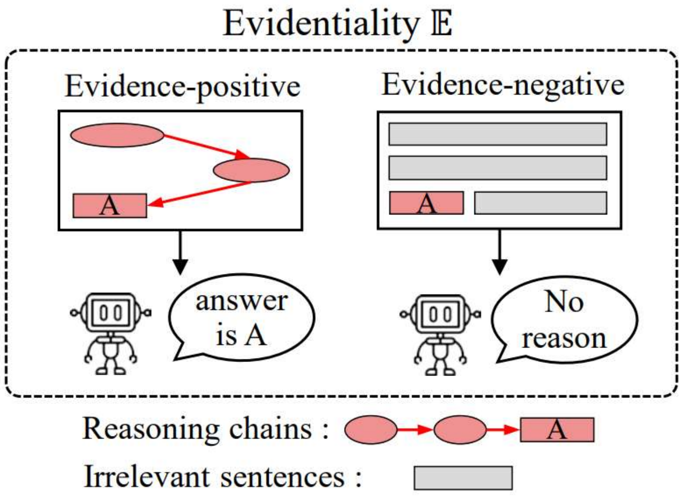
→ overconfident over missing information

Robustifying Multi-hop QA through Pseudo-Evidentiality Training, ACL 2021

Evidentiality for Robustness

- NLP model should be explicitly taught to say I don't know.





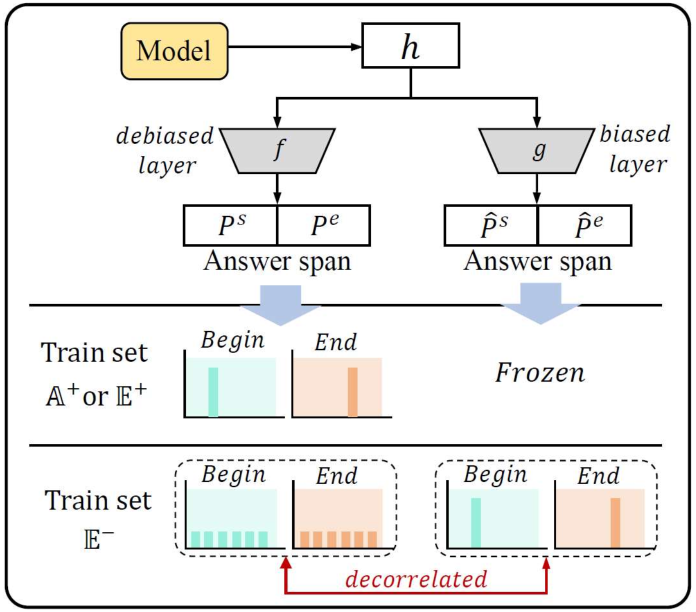
Robustifying Multi-hop QA through Pseudo-Evidentiality Training, ACL 2021

Teaching Evidentiality

- Enforcing to be not confident on the evidence-negative set

$$\mathcal{R} = \sum_{i \in \mathbb{E}^-} D_{KL}(P(\mathcal{A}_i|\mathcal{Q}_i, \mathcal{D}_i)||P_{uniform})$$

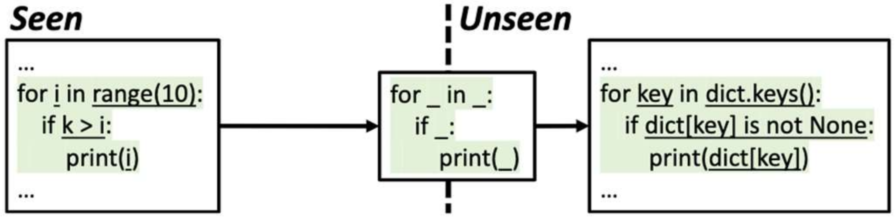
- Teaching to decorrelate answers with biased models



Robustifying Multi-hop QA through Pseudo-Evidentiality Training, ACL 2021

Compositionality as Rationalization

- NLP 3.0 fails to capture structure as evidence
- Seen structure, and its composition is considered unseen

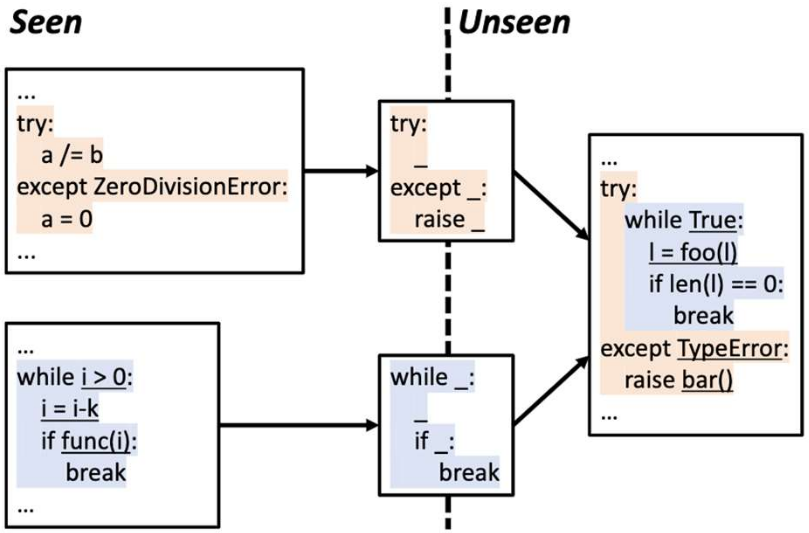


```
...
for i in range(10):
    if k > i:
        print(i)
...
```

```
for _ in _:
    if _:
        print(_)
```

```
...
for key in dict.keys():
    if dict[key] is not None:
        print(dict[key])
...
```

(a) Lexical generalization.



```
...
try:
    a /= b
except ZeroDivisionError:
    a = 0
...
```

```
...
while i > 0:
    i = i-k
    if func(i):
        break
...
```

```
try:
    _
except _:
    raise _
```

```
while _:
    _
    if _:
        break
```

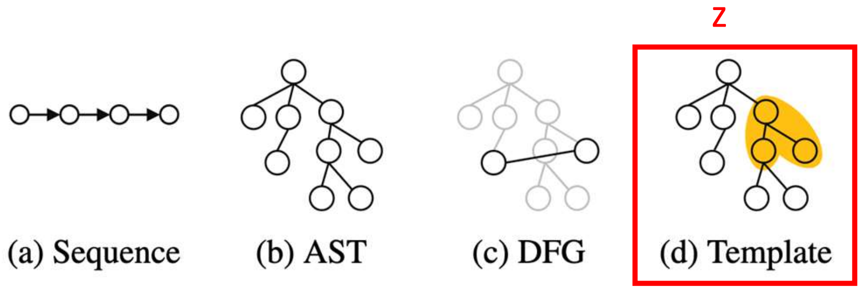
```
...
try:
    while True:
        l = foo(l)
        if len(l) == 0:
            break
except TypeError:
    raise bar()
...
```

(b) Structural generalization.

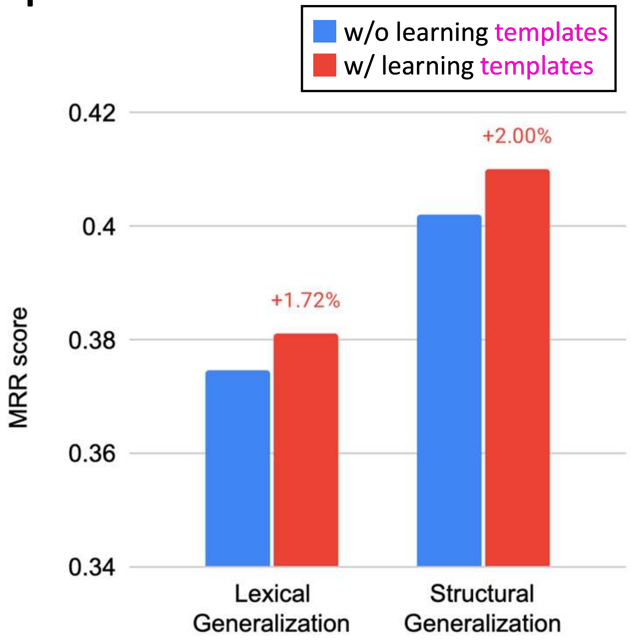
Towards Compositional Generalization in Code Search, EMNLP 2022

Teaching Compositionality

- Template repository can be extracted from parsers
- Codes can be represented as a template composition



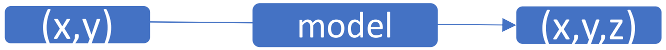
(a) Sequence (b) AST (c) DFG (d) Template

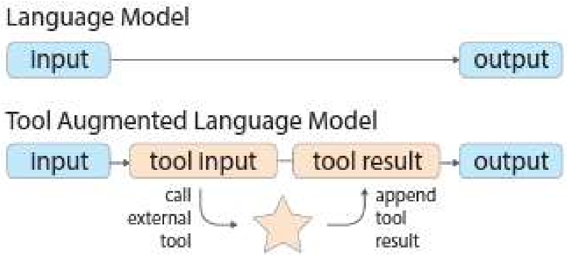


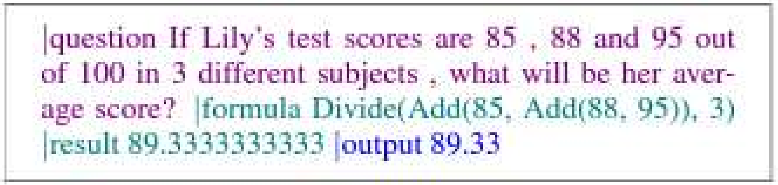
Towards Compositional Generalization in Code Search, EMNLP 2022

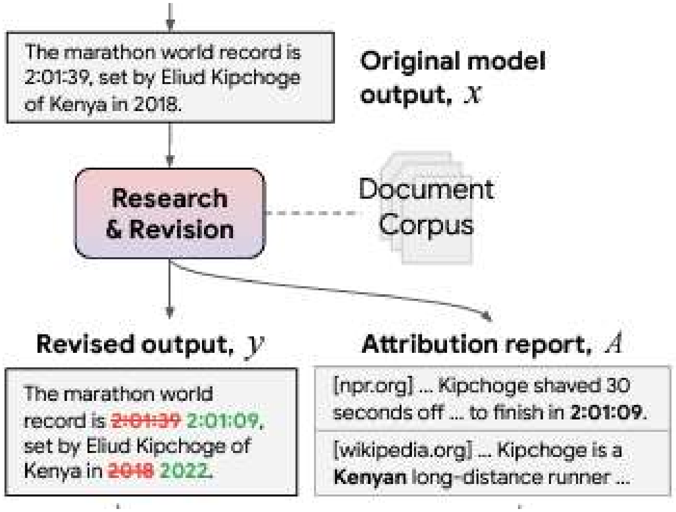
Robustness for Generation

- Model-augmented z for robustifying fine-tuning
- Generation can be similarly “model-augmented” for ensuring robustness (similar to dynamic web page with sql scripts)





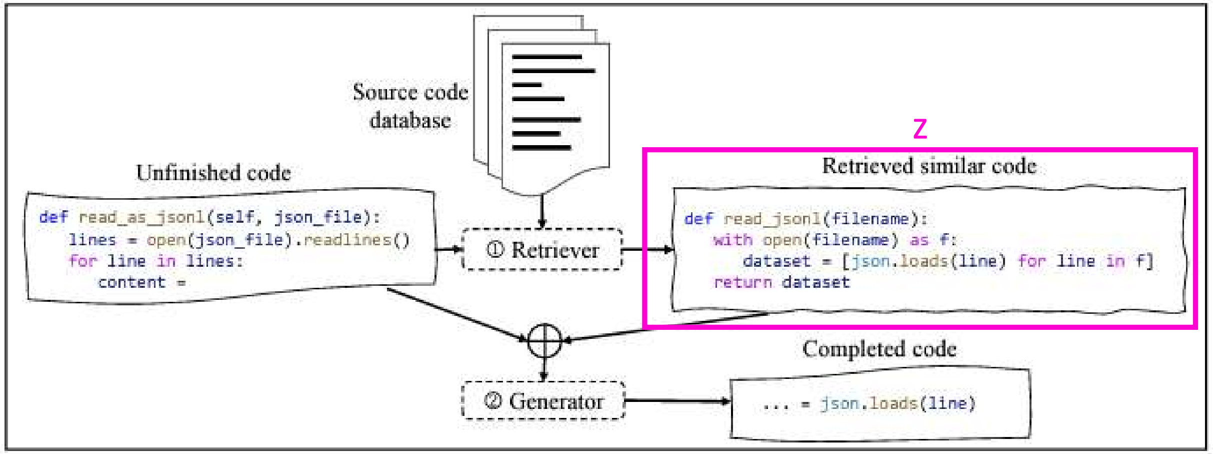

|question If Lily's test scores are 85 , 88 and 95 out of 100 in 3 different subjects , what will be her average score? |formula Divide(Add(85, Add(88, 95)), 3) |result 89.3333333333 |output 89.33


The marathon world record is 2:01:39, set by Eliud Kipchoge of Kenya in 2018.
Original model output, x
Research & Revision
Document Corpus
Revised output, y
Attribution report, A
The marathon world record is ~~2:01:39~~ 2:01:09, set by Eliud Kipchoge of Kenya in ~~2018~~ 2022.
[npr.org] ... Kipchoge shaved 30 seconds off ... to finish in 2:01:09.
[wikipedia.org] ... Kipchoge is a Kenyan long-distance runner ...

https://arxiv.org/pdf/2205.12255.pdf; https://arxiv.org/abs/2210.08726

Robustness for Generation: Evidence

- Inspired by human behaviors of copying from the related code snippets when writing code



```
def read_as_jsonl(self, json_file):
    lines = open(json_file).readlines()
    for line in lines:
        content =
```

```
def read_jsonl(filename):
    with open(filename) as f:
        dataset = [json.loads(line) for line in f]
    return dataset
```

```
... = json.loads(line)
```

Figure 1: An illustration of ReACC framework. Given an unfinished code snippet to complete, ReACC first retrieves the similar code from source code database. Then the similar code is concatenated with the unfinished code, the completed code will be generated based on them.

ReACC: ARetrieval-Augmented Code Completion Framework, ACL 2022

Robustness for Generation: Evidentiality

- Facts during LM training can change or emerge.

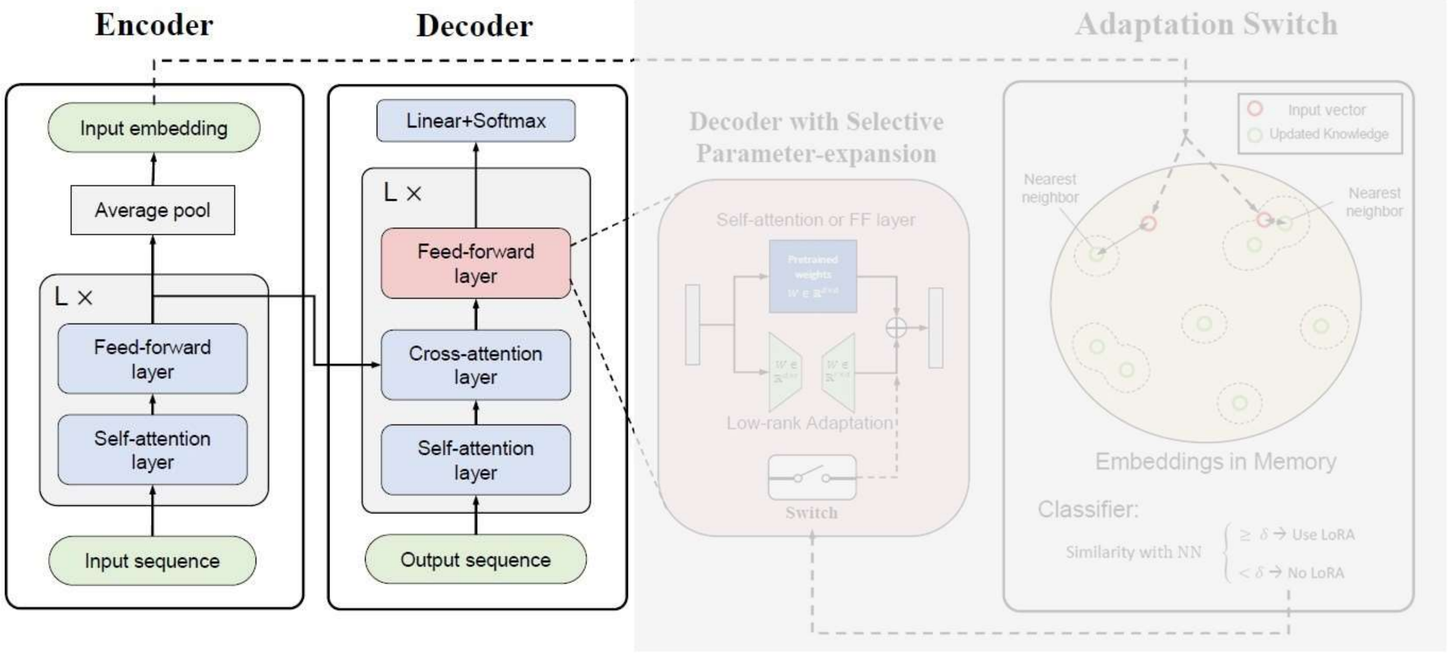


Figure 2: An overview of our proposed architecture.

Plug-and-Play Adaptation for Continuously-updated QA, EMNLP 2022

Robustness for Generation

- Parameter-efficient training for new knowledge using adapter

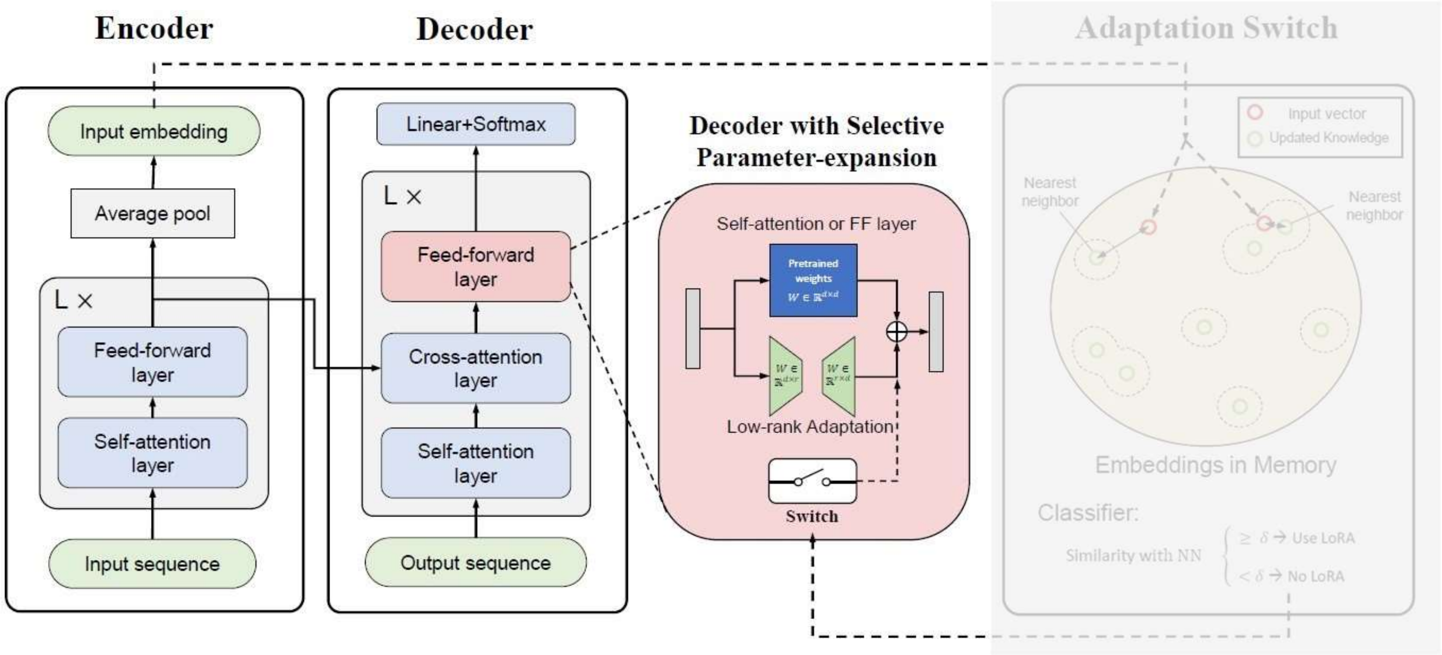


Figure 2: An overview of our proposed architecture.

Plug-and-Play Adaptation for Continuously-updated QA, EMNLP 2022

Robustness for Generation

- Evidentiality switch enables to plug-in multiple knowledge updates without forgetting old (ongoing: updates from heterogeneous sources)

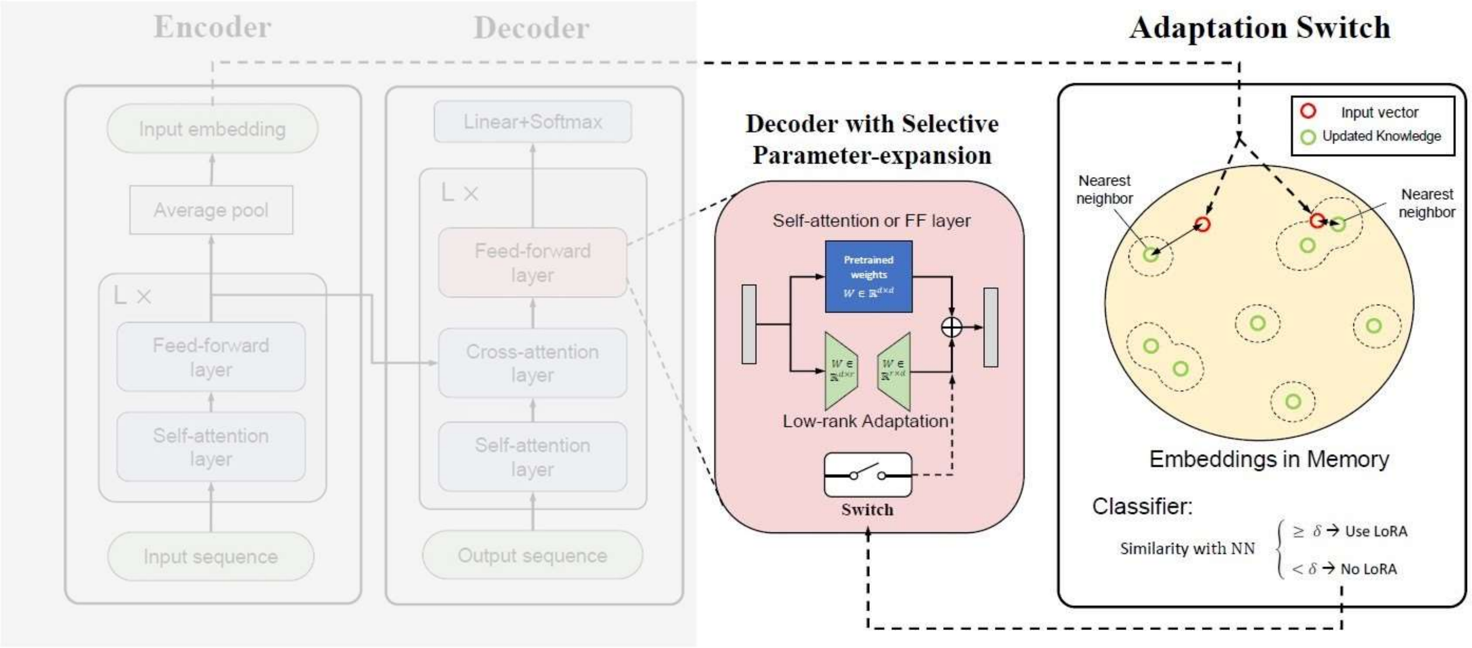


Figure 2: An overview of our proposed architecture.

Plug-and-Play Adaptation for Continuously-updated QA, EMNLP 2022

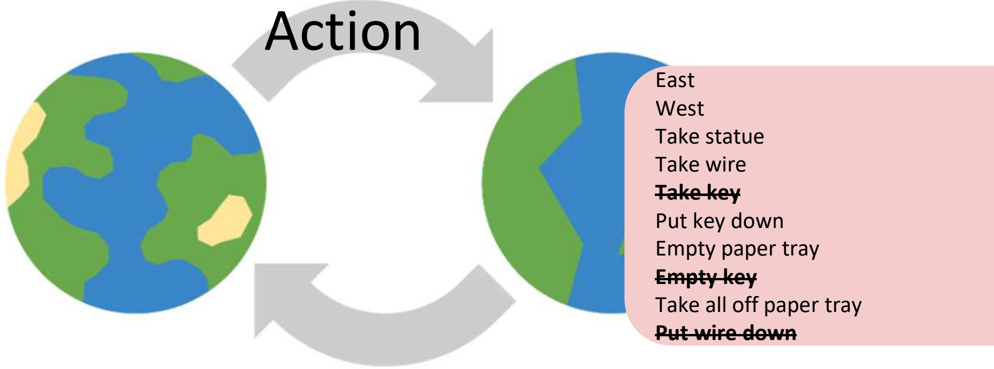
Robustness for Generation: Actionability

- Generation for robots, text-based games, code leads to actions
- Generation should keep a ***world model*** to predict how action changes the world
- Actionability-constrained PLM generation reflects should be consistent with engine and changes
- Constrained-LM (e.g., code LM) outperforms natural-LM in some language tasks

[OBS] Location: Copier Room The copier room doesn't contain any windows, and vibrates slightly with fluorescent light. A big copier sits quietly in the corner. Doors lead east and west. You can see a Dragon Statue here.

[Action] put wire down

World Model



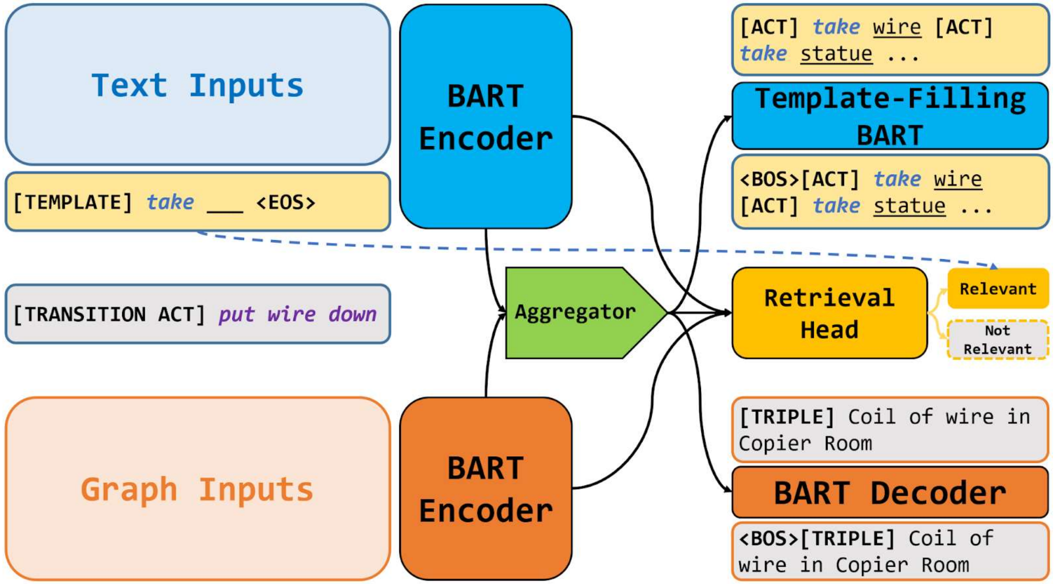
East
West
Take statue
Take wire
~~**Take key**~~
Put key down
Empty paper tray
~~**Empty key**~~
Take all off paper tray
~~**Put wire down**~~

https://arxiv.org/abs/2204.01691

Robustness for Generation: Actionability

- We built ***AWM-BART***, constraining generation as a ***template filling***
- As an actionable world model (AWM), we built a ***template retriever***, conditioned by states represented as a graph

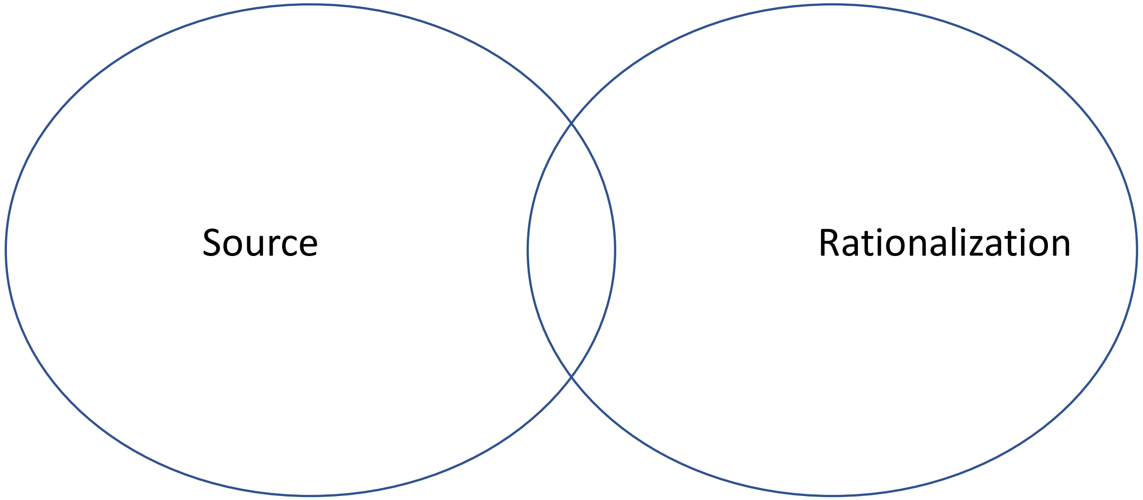
[GRAPH] <Copy Machine, in, Copier Room> <you, have, Long Key> <Sheet of Paper, in, Paper Tray> <you, have, Coil of wire> <you, have, Gun> <Paper Tray, in, Copier Room>


Text Inputs
[TEMPLATE] take ___ <EOS>
[TRANSITION ACT] put wire down
Graph Inputs
BART Encoder
Aggregator
BART Encoder
[ACT] take wire [ACT] take statue ...
Template-Filling BART
<BOS>[ACT] take wire [ACT] take statue ...
Retrieval Head
Relevant
Not Relevant
[TRIPLE] Coil of wire in Copier Room
BART Decoder
<BOS>[TRIPLE] Coil of wire in Copier Room

PLM-based World Models for Text-based Games, EMNLP 2022

Conclusion

- Deep AI models are achieving super-human accuracy on NLP tasks
- Lack of robustness explains gaps between task- and perceived-accuracy
- Richer Human-AI alignments can be distilled from models for robustifying finetuning and PLM generation



DB

Search engine

Game engine

Parser

Attention

Actionability

Causality

Evidence

Evidentiality

Compositionality

Questions?

Reference

- *PLM-based World Models for Text-based Games, EMNLP 2022*
- *Towards Compositional Generalization in Code Search, EMNLP 2022*
- *Pseudo-Relevance for Enhancing Document Representation, EMNLP 2022*
- *Collective Relevance Labeling for Passage Retrieval, NAACL 2022*
- *ReACC: A Retrieval-Augmented Code Completion Framework, ACL 2022*
- *Plug-and-Play Adaptation for Continuously-updated QA, Findings of ACL 2022*
- *C2L: Causally Contrastive Learning for Robust Text Classification, AAAI 2022*
- *Robustifying Multi-hop QA through Pseudo-Evidentiality Training, ACL 2021*
- *Explanatory and Actionable Debugging for Machine Learning: A TableQA Demonstration, SIGIR 2019 (demo)*

Visit http://seungwonh.github.io for more