Why natural language is the right vehicle for complex reasoning

TEXAS
The University of Texas at Austin

TAUR LAB

Greg Durrett

NYU

April 14, 2022

Reasoning about text: entailment

premise | hypothesis

A dog is chasing a cat | *Two animals are running* → Transformer → **P entails H**

A dog is chasing a cat | *Two animals are sitting* → Transformer → **P contradicts H**

There's **some** reasoning happening, in a purely latent way:

X chasing ⇒ X is running

X chasing ⇒ X is not sitting

A dog is chasing a cat | *Three animals are running* → Transformer → **P entails H**

…but when the model's latent reasoning is flawed, it's hard to diagnose

Example: multi-hop QA where systems only do single-hop reasoning

Jifan Chen and GD, NAACL19; Sewon Min et al., ACL19; Yichen Jiang and Bansal ACL19

Where latent reasoning breaks down

Can we just improve latent reasoning models? Better data, debiasing, contrastive learning, …

18?

Source: Natural Questions
(Tom Kwiatkowski et al., 2019)

Transformer

What age do you need to be to buy a bb gun?

*From Wikipedia: In **Manitoba, Saskatchewan, Ontario, British Columbia, and Quebec**, the minimum age to purchase an **airsoft gun** is 18.*

Applies to **Canada** only (new dataset with this context: SituatedQA; Michael J.Q. Zhang and Eunsol Choi, 2021)

Are airsoft and BB guns the same? It’s complicated!

End-to-end models don't model these nuances well.

We need justified reasoning in addition to answers.

Contrast: Theorem Provers

A dog is chasing a cat

$\exists d.\exists c.\exists e.\ dog(d) \wedge cat(c) \wedge chase(e) \wedge agent(e, d) \wedge patient(e, c)$

Two animals are running

$\exists a.\exists b.\ animal(a) \wedge animal(b) \wedge running(a) \wedge running(b)$

Theorem Prover

$\forall x.\forall e.\ chase(e) \wedge agent(e, x) \Rightarrow running(x)$

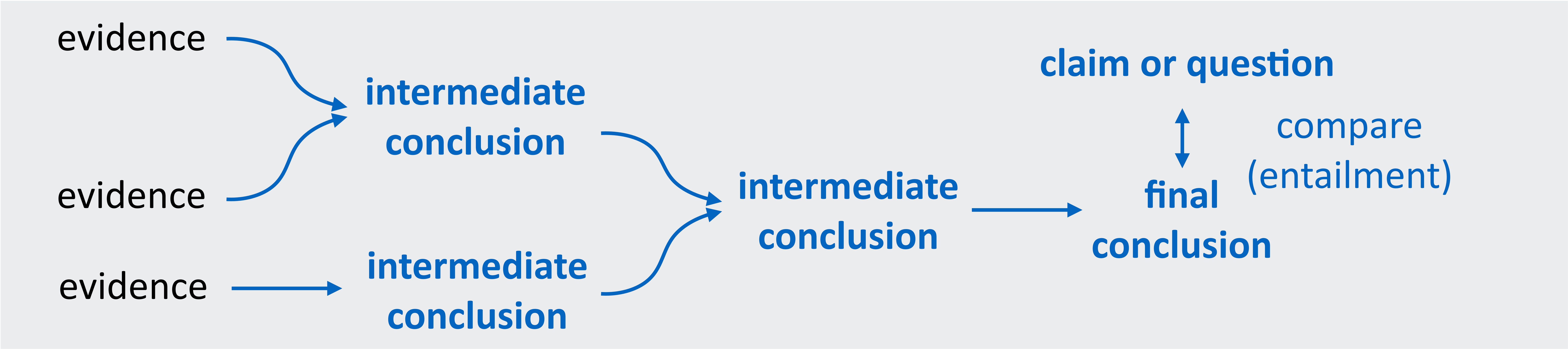
$\forall x.\forall e.\ chase(e) \wedge patient(e, x) \Rightarrow running(x)$

$\forall x.\ dog(x) \Rightarrow animal(x)$

$\forall x.\ cat(x) \Rightarrow animal(x)$

- Advantage: Articulates explicit intermediate reasoning states
- Disadvantage: requires high-coverage semantic formalism, parser into that formalism, and background knowledge hard to learn from data

Our vision



Use pre-trained models to do reasoning **directly in natural language**

- Combine **logical inference** (modus ponens, …) and **lexical inference** (paraphrasing, …)

Natural logic, theorem provers

This approach

End-to-end models, chain-of-thought

(Bill MacCartney and Manning, 2007; Hai Hu et al. “MonaLog”, 2020, …)

(Maxwell Nye et al., 2021; Jason Wei et al., 2022)

Why natural language?

A dog is chasing a cat → Two animals are chasing each other

Chasing involves running

→ Two animals are running

- **Expressive**. Text ***is already*** a broad-coverage semantic representation
- **Flexible.** Approaches operating over text can synthesize pre-trained models, Wikipedia, commonsense knowledge bases, …
- **Interpretable** reasoning chains
- But: we need the **right data** and need to ensure our models are doing sound reasoning

Outline

Entailment to verify QA

Jifan Chen, Eunsol Choi, GD. EMNLP-Findings21
Can NLI Models Verify QA Systems' Predictions?

evidence → **standalone statement** ←→ **Q+answer**
NLI

Logically manipulating statements

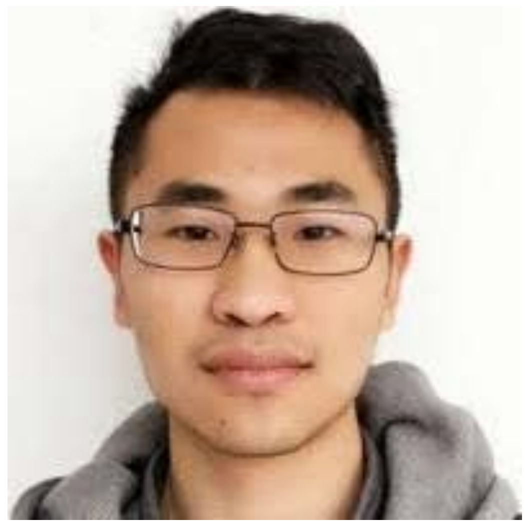
Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, GD. EMNLP21
Flexible Generation of Natural Language Deductions

statement
statement
→ **deduction**

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, GD. In submission.
Natural Language Deduction through Search over Statement Compositions

Improving diverse generation (if time)

Jiacheng Xu, GD. NAACL22.
Massive-scale Decoding for Text Generation using Lattices







Verifying Reading Comprehension

Ted Danson

RoBERTA QA model

Who plays the bad guy in the Good Place?

The first season of the fantasy comedy television series The Good Place […] *The series focuses on Eleanor Shellstrop (Kristen Bell) , a woman who wakes up in the afterlife and is introduced by Michael (Ted Danson) to a Heaven-like utopia* […]

Assume a base QA system with a latent reasoning process. Can we check the answer?

- Can better determine if question is unanswerable
- Can improve confidence, "selective QA setting" (Amita Kamath et al. 2020)
- Can validate presuppositions in the question (Najoung Kim et al., 2021)

Our Method

The series focuses on Eleanor Shellstrop (Kristen Bell) , a woman who wakes up in the afterlife and is introduced by Michael (Ted Danson) [...]

Decontextualization

Eunsol Choi et al. (2021)

standalone statement (premise)

The series ***The Good Place*** *focuses on Eleanor Shellstrop (Kristen Bell) , a woman who wakes up in the afterlife and is introduced by Michael (Ted Danson) [...]*

NLI model → **false**

Who plays the bad guy in the Good Place?

Ted Danson

Question-to-statement conversion

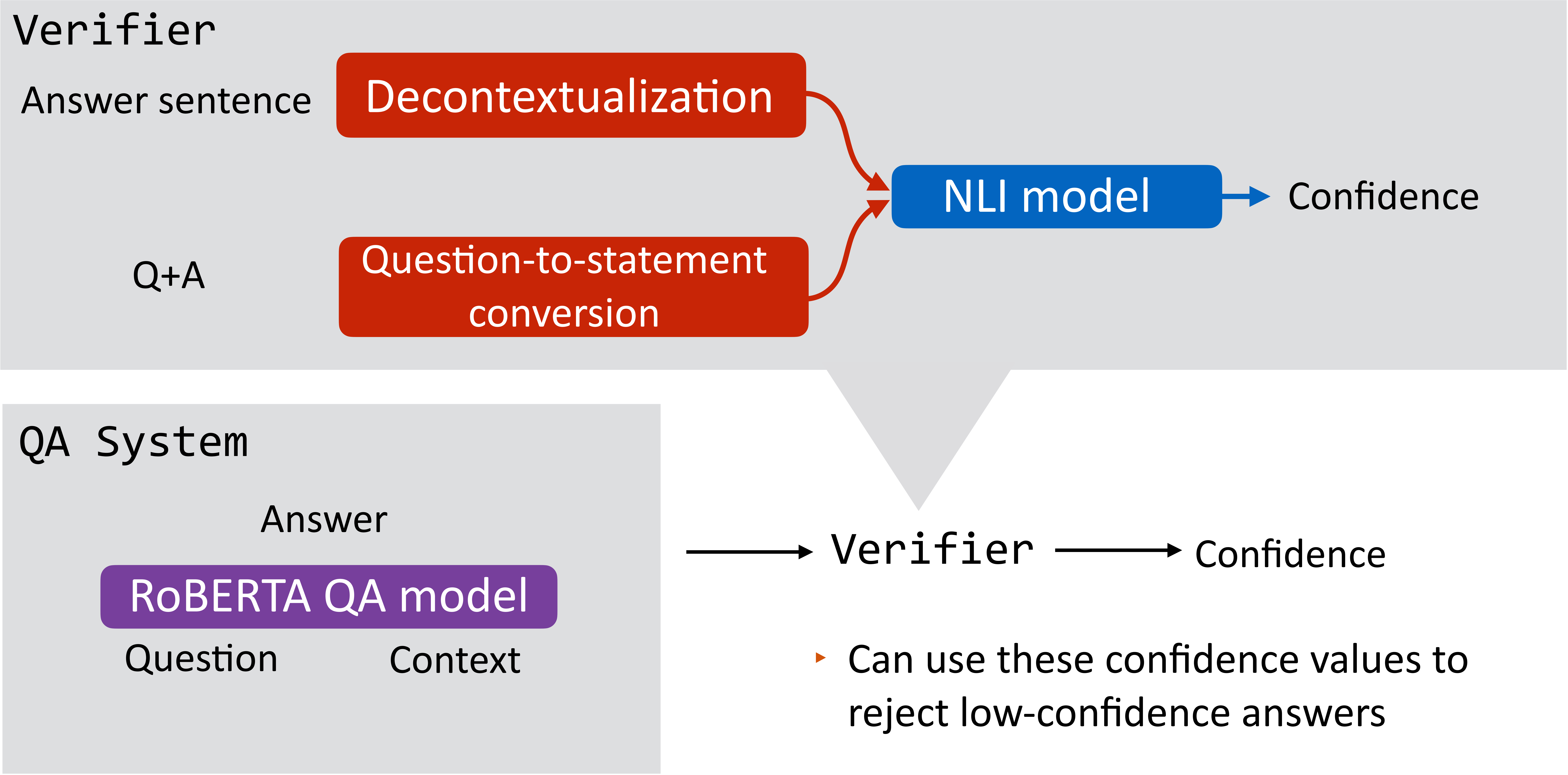
Dorottya Demszky et al. (2018)
(upgraded to use T5-3B)

Ted Danson *plays the bad guy in the Good Place.*

hypothesis

(in this case: right for the wrong reasons)

System



- Can use these confidence values to reject low-confidence answers

Results: Unanswerable Questions

- Train QA system on (En) SQuAD 1.1 (answers every question), run on SQuAD 2.0 (contains unanswerable questions), use the verifier to reject bad answers

`QA System` → `Verifier` → keep or reject answer

- A RoBERTa MNLI model can reject **78.5%** of unanswerables, accept **82.5%** of answerables. Good zero-shot performance (NLI model is not trained on SQuAD + pipeline is not optimized end-to-end)

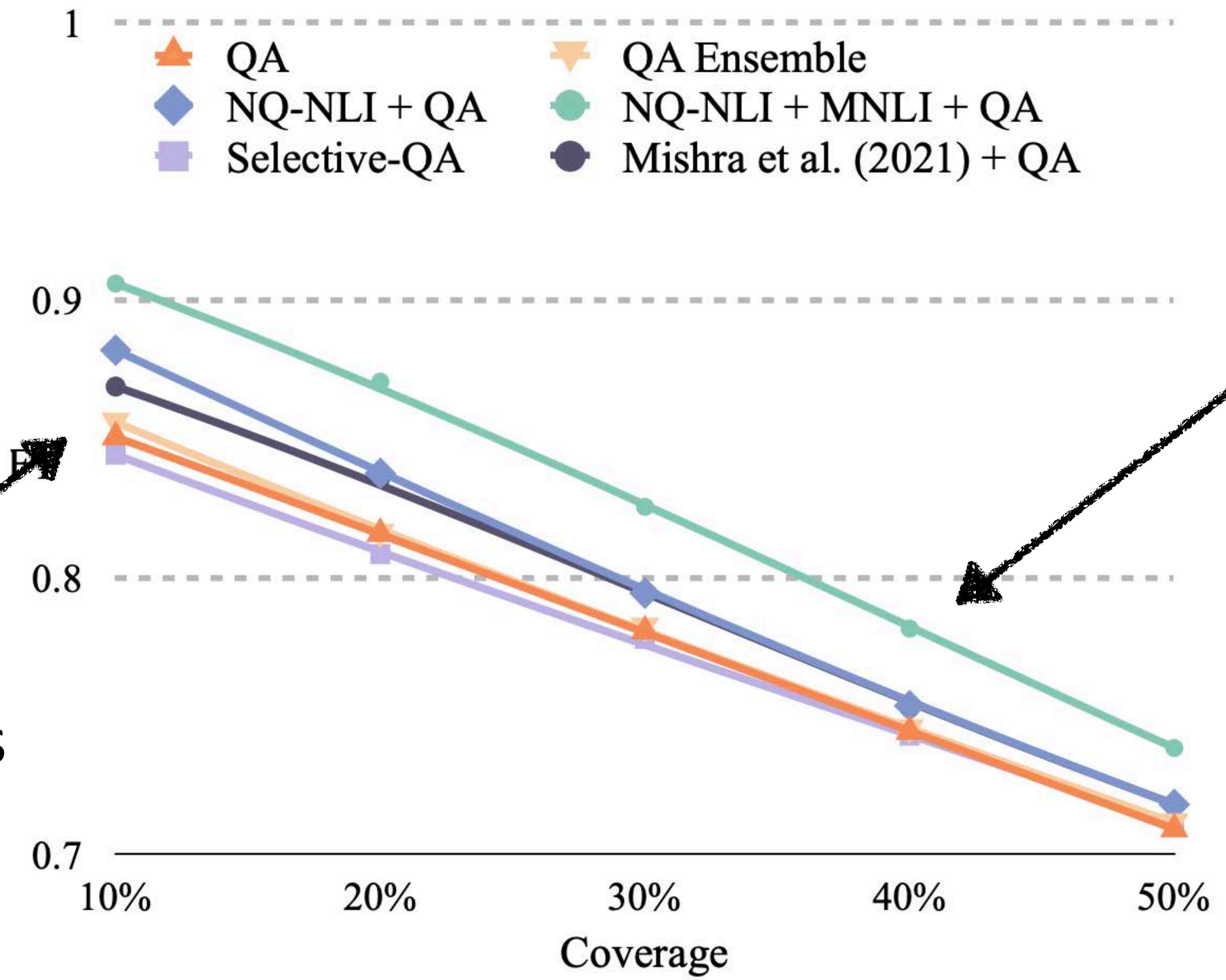
- Can use MNLI off-the-shelf here because we have single-sentence premises

(unlike Anshuman Mishra et al., 2021; Wenpeng Yin et al., DocNLI, 2021)

Results: Selective QA (5 datasets)

- Base QA: BERT-Large on (En) NaturalQuestions. Target: NQ + 4 out-of-domain (En) sets.

QA = use QA posteriors as confidence



Our best model: combine QA posteriors with verifier trained on MNLI + NQNLI

NQNLI: Natural Questions converted to NLI with our framework

NLI works well as an answer verifier

Errors

- **QA conversion / decontextualization errors are rare.** NL manipulation is great with big models like T5-3B!
- **Entailment errors** are more common. But sometimes, the entailment model disagrees with the QA dataset **and is right**

Reformulated Q+A: ***John von Neumann*** *developed the central processing unit (CPU).*

Context: *On June 30, 1945, before ENIAC was made, mathematician* ***John von Neumann*** *distributed the paper entitled First Draft of a Report on the EDVAC. It was the outline of a stored-program computer that would eventually be completed in August 1949.*

- John von Neumann is marked as the gold answer (debatable), NLI model disagrees
- Manipulation of these examples makes it easy to evaluate reasoning. **Is this evidence really sufficient to validate the answer?**

Takeaways

- We can manipulate question-answer pairs and evidence sentences **in natural language** and use NLI to check QA answers
- Manipulation was highly reliable, and the two operations we had were sufficient to allow us to employ a pre-existing model (NLI)
- NLI can improve calibration for QA and lets us audit both our models and datasets

Outline

Entailment to verify QA

Jifan Chen, Eunsol Choi, GD. EMNLP-Findings21
Can NLI Models Verify QA Systems' Predictions?

evidence → **standalone statement** ↔ **Q+answer**
NLI

Logically manipulating statements

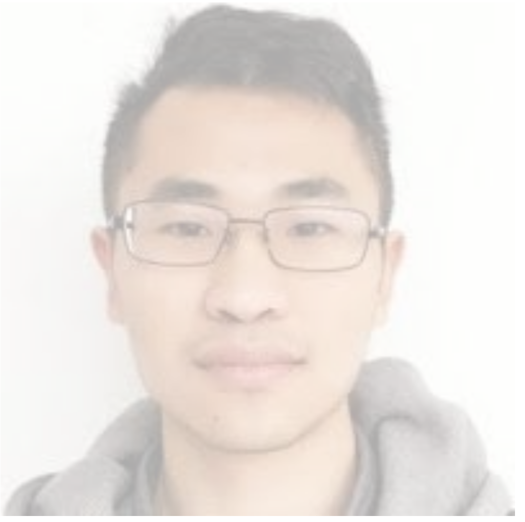
Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, GD. EMNLP21
Flexible Generation of Natural Language Deductions

statement
statement
→ **deduction**

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, GD. In submission.
Natural Language Deduction through Search over Statement Compositions

Improving diverse generation

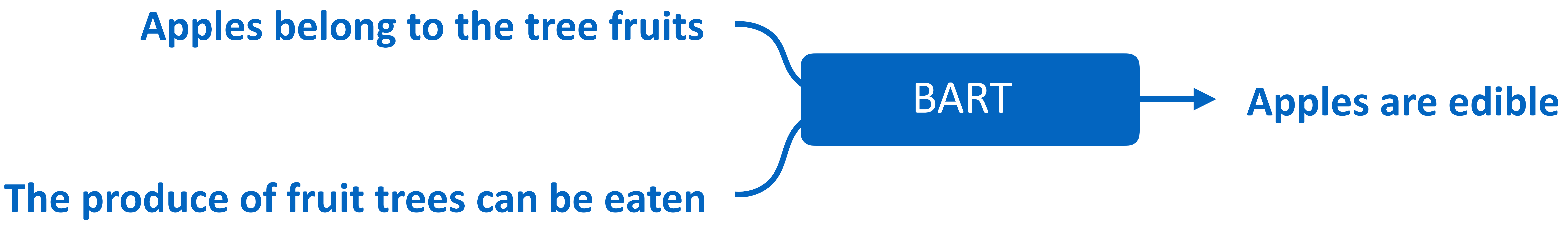
Jiacheng Xu, GD. NAACL22.
Massive-scale Decoding for Text Generation using Lattices







Natural Language Deduction



Natural language deduction: place a distribution over the set of valid (and useful) conclusions

1. Can we automate collecting this kind of data at scale?
2. Can we chain these inferences together into multiple steps?

Our Approach

We characterize inference as a blend of two processes:

Logical inference

Fruits are edible. Apples are a fruit.
→ Apples are edible.

Invariant w.r.t. lexical content
Easy to describe with a concise set of rules
Hard to learn distributionally

↓

Automatic template-based data generation

Lexical inference

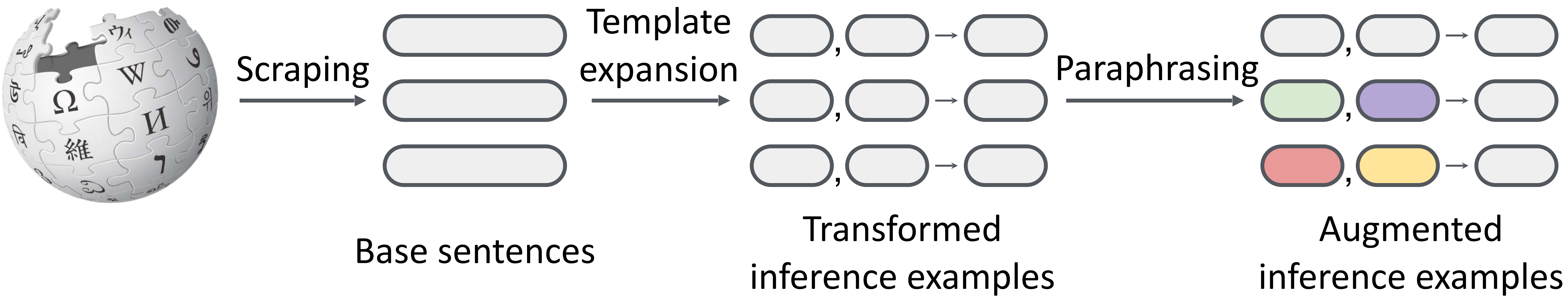
edible $\longleftrightarrow$ can be eaten

Idiosyncratic, depends on words
Hard to describe with a concise set of rules
Can be learned distributionally

↓

Transfer learning from a pre-trained LM
Data augmentation with paraphrasing

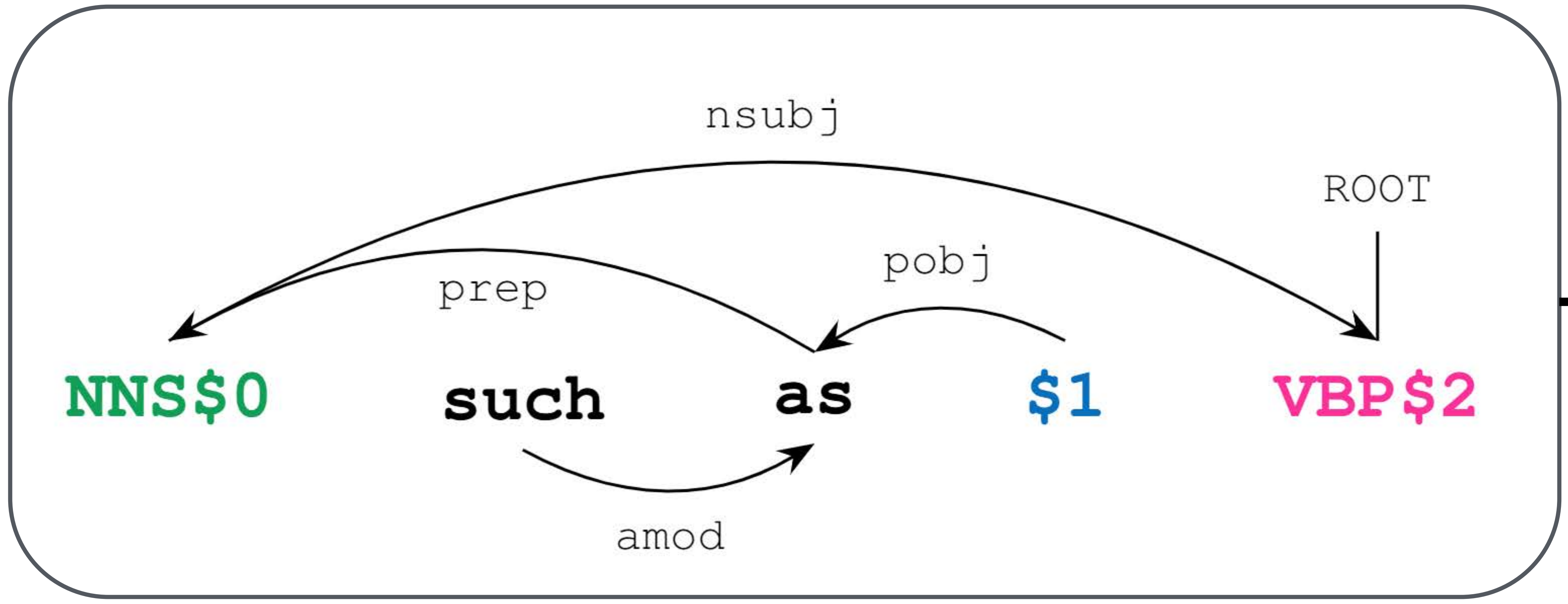
Data Generation

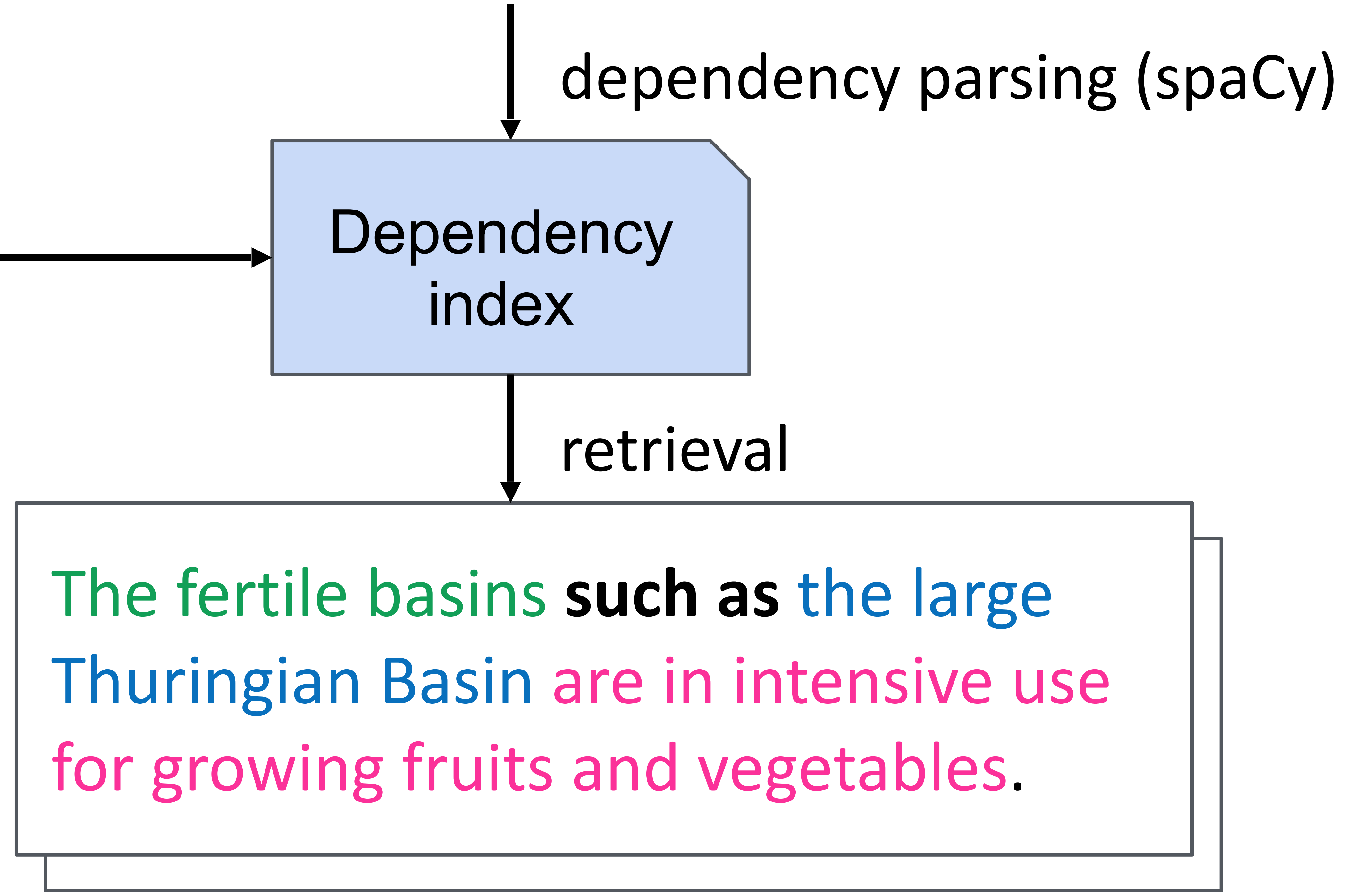


Bostrom et al. EMNLP21

Data Generation

1. Source sentence scraping using Hearst patterns (6 patterns for this type of *substitution* reasoning)





Bostrom et al. EMNLP21

Data Generation

2. Template expansion (1 template per Hearst pattern)

The fertile basins **such as** the large Thuringian Basin are in intensive use for growing fruits and vegetables.

→

X VP
Y is a X
↓
Y VP

↓ slot filling + reinflection

The fertile basins are in intensive use for growing fruits and vegetables.
The large Thuringian Basin is a fertile basin.

→ The large Thuringian Basin is in intensive use for growing fruits and vegetables.

Bostrom et al. EMNLP21

Data Generation

3. Paraphrasing

The fertile basins are in intensive use for growing fruits and vegetables.
The large Thuringian Basin is a fertile basin.

→ The large Thuringian Basin is in intensive use for growing fruits and vegetables.

Automatic paraphrasing model (PEGASUS)

Growing fruits and vegetables requires a lot of fertile basins.
The Thuringian Basin is fertile.

→ The large Thuringian Basin is in intensive use for growing fruits and vegetables.

Paraphrasing adds noise, but only to the input. We find the model **still does sound** reasoning and can handle more lexical variation

Bostrom et al. EMNLP21

Examples

- BART trained on 126k examples we automatically collect from Wikipedia

Doctors use medical tests to **diagnose** diseases and guide treatment.

Some of the most common **diagnostic procedures** include blood counts and other multi-factor panels.

No paraphrasing → [repeats first premise]

Ours → Doctors use blood counts and other multi-factor panels to diagnose diseases and guide treatment.

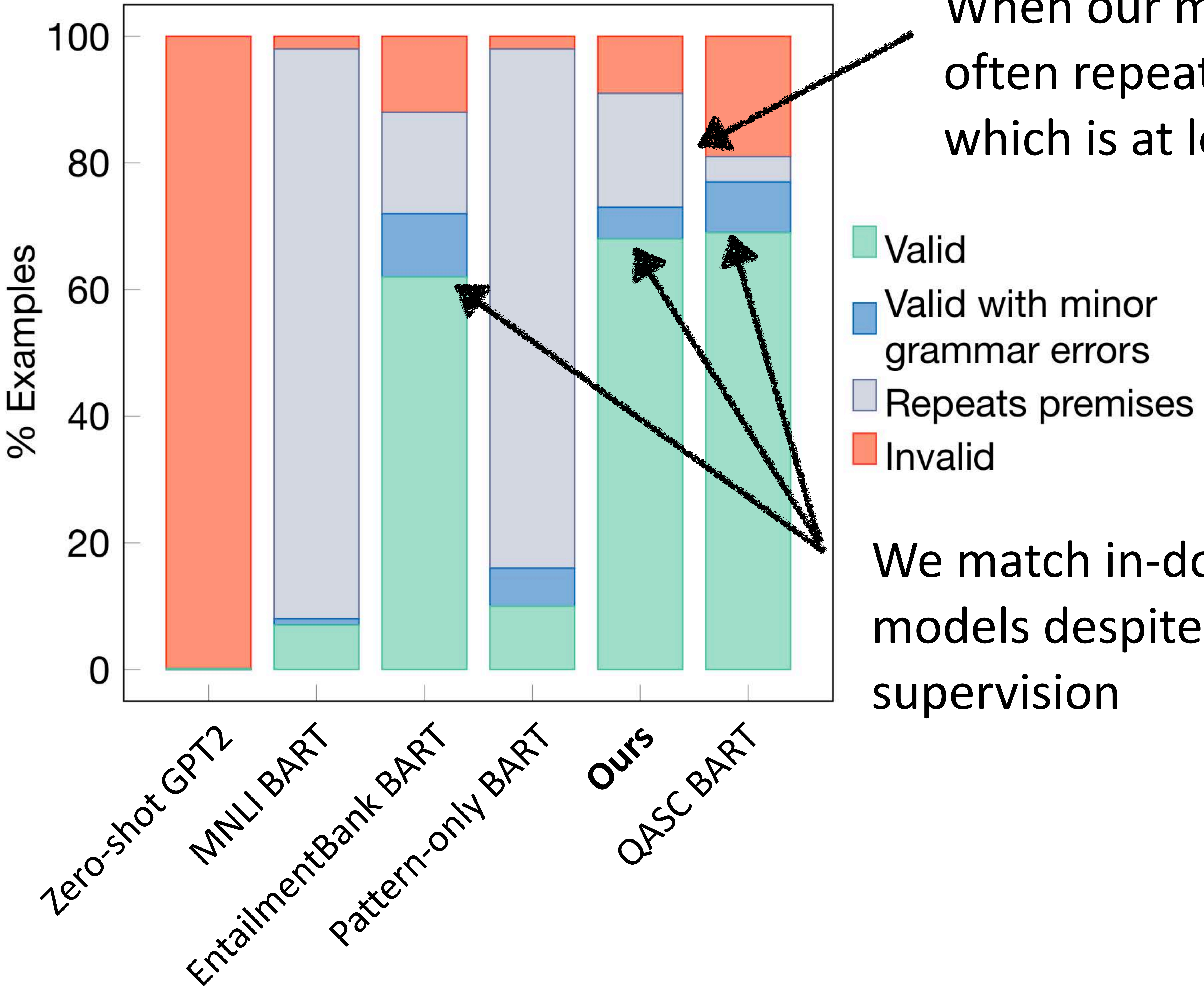
Humans **changing an ecosystem** usually has a negative impact on an ecosystem / organisms living in an ecosystem.

Humans building homes in an ecosystem **causes that ecosystem to change.**

Ours → Humans building homes in an ecosystem usually has a negative impact on an ecosystem / organisms living in an ecosystem.

Bostrom et al. EMNLP21

Results: QASC Human Eval



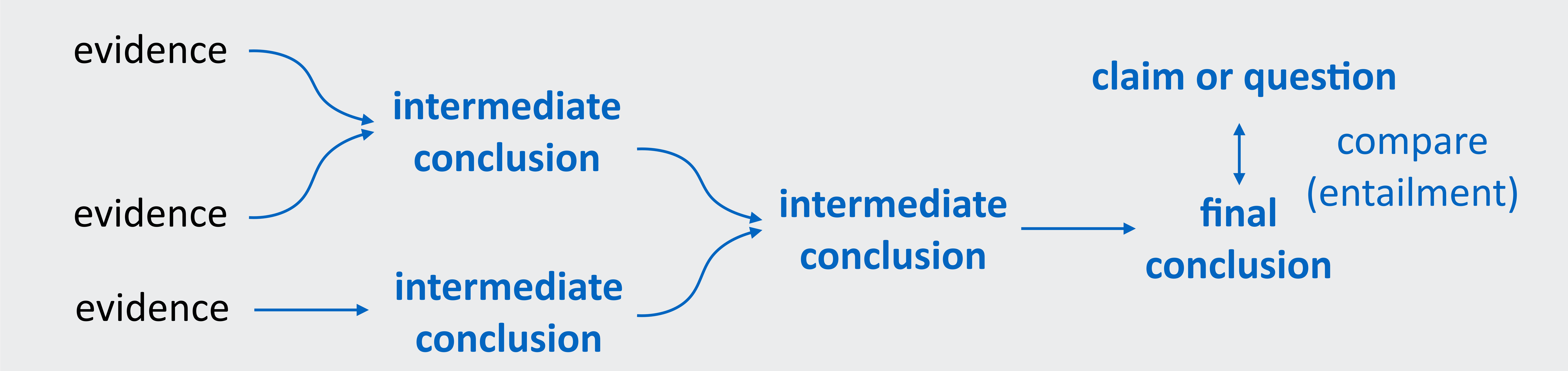
When our model fails, it often repeats premises, which is at least not wrong!

We match in-domain supervised models despite using synthetic supervision

Tushar Khot et al., 2020

Multi-step Deduction

- We showed something that works well for single steps. Let's return to our goal...



- Goal: dynamically apply operations (including deduction and decontextualization) to give the conclusions we need
- This is a hard search problem: intermediate states are potentially all natural language sentences that can be reasonably generated from the evidence

Bostrom et al. arXiv22

Multi-step Deduction: Setup

evidence

s_1	Paper is recyclable.
s_2	Recyclable means old material can be converted into new material
s_3	Cardstock is a type of paper

hypothesis

Old cardstock can be turned into new cardstock.

- Collection of evidence (science domain, taken from EntailmentBank) and hypothesis
- Can we prove the hypothesis deductively using our generative step model? (not just throwing everything into a discriminative model)

Bostrom et al. arXiv22

Search Heuristics

evidence

s_1 Paper is recyclable.

s_2 Recyclable means old material can be converted into new material

s_3 Cardstock is a type of paper

✓

✓

no meaningful inference

hypothesis

Old cardstock can be turned into new cardstock.

- Search frontier of pairs (s_i, s_j) of sentences we can combine — how to prioritize?
- Goal-conditioned heuristic: learn a model $g(s_i, s_j, h)$ — how likely will combining s_i and s_j eventually lead to h? Requires training on EntailmentBank
- **Search and deduction are decoupled.** Search conditions on the hypothesis, but the deduction itself uses only the premises

Bostrom et al. arXiv22

Multi-Step Deduction

evidence

s_1 Paper is recyclable.

s_2 Recyclable means old material can be converted into new material

s_3 Cardstock is a type of paper

Step 1 Cardstock is recyclable.

Step 2 Old cardstock can be converted into new cardstock.

NLI model

hypothesis

Old cardstock can be turned into new cardstock.

- Repeatedly prove statements and expand the search space, then check if each entails the claim with an NLI model (fine-tuned for this domain)

Bostrom et al. arXiv22

Evaluation: EntailmentBank

- Input: 25 English premises and a potentially true hypothesis.
  Goal: classify hypothesis as true/false based on premises

premises (+ 22 distractors)

A planet rotating causes cycles of day and night on that planet.

Earth rotating on its tilted axis occurs once per day.

Earth is a kind of planet.

true goal statement

Goal: The earth rotating on its tilted axis causes the cycles of day and night on earth. ✓

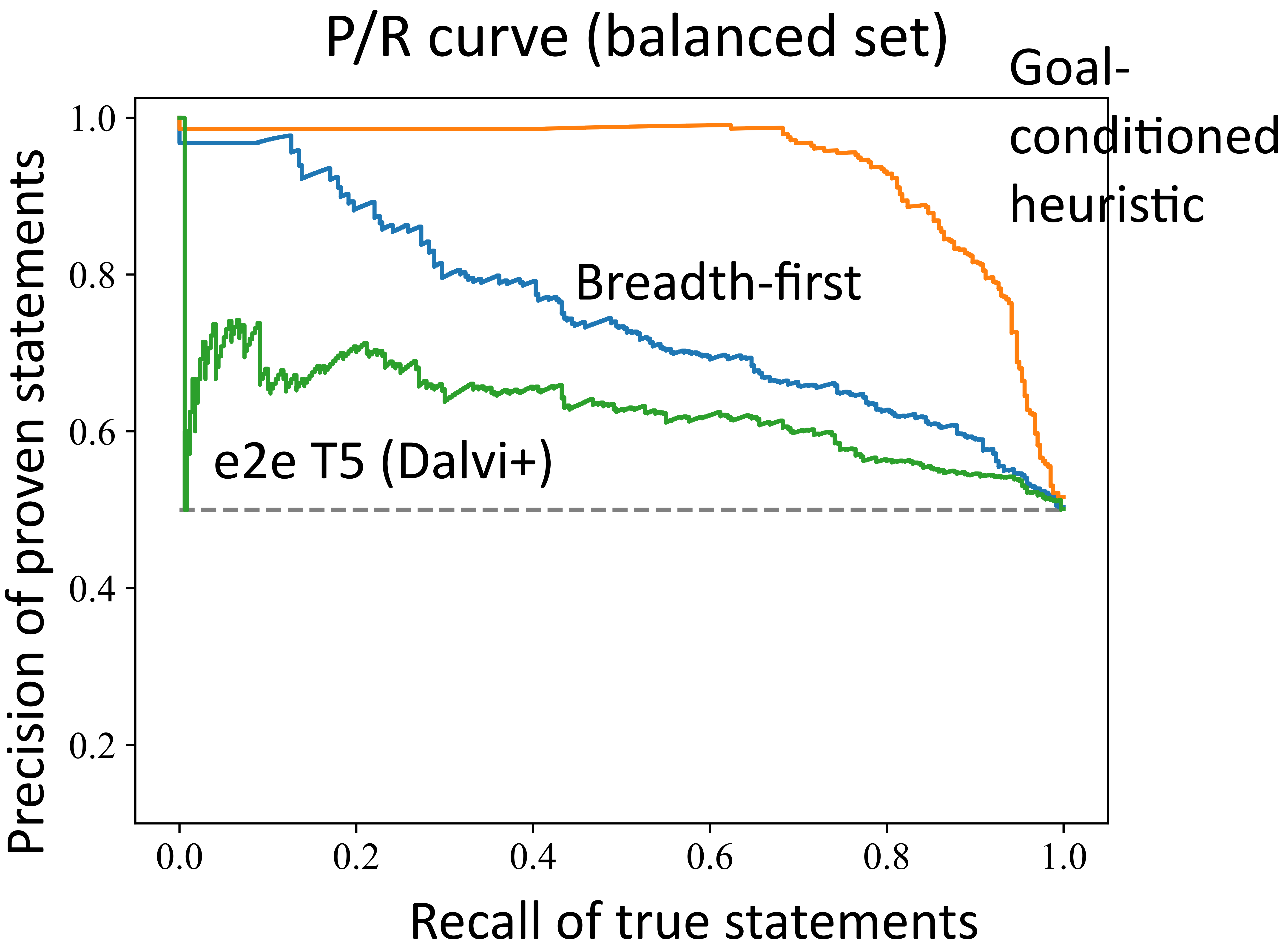
random distractor goal

Goal: Looking at the moon has less of a negative impact on the eyes. ✗

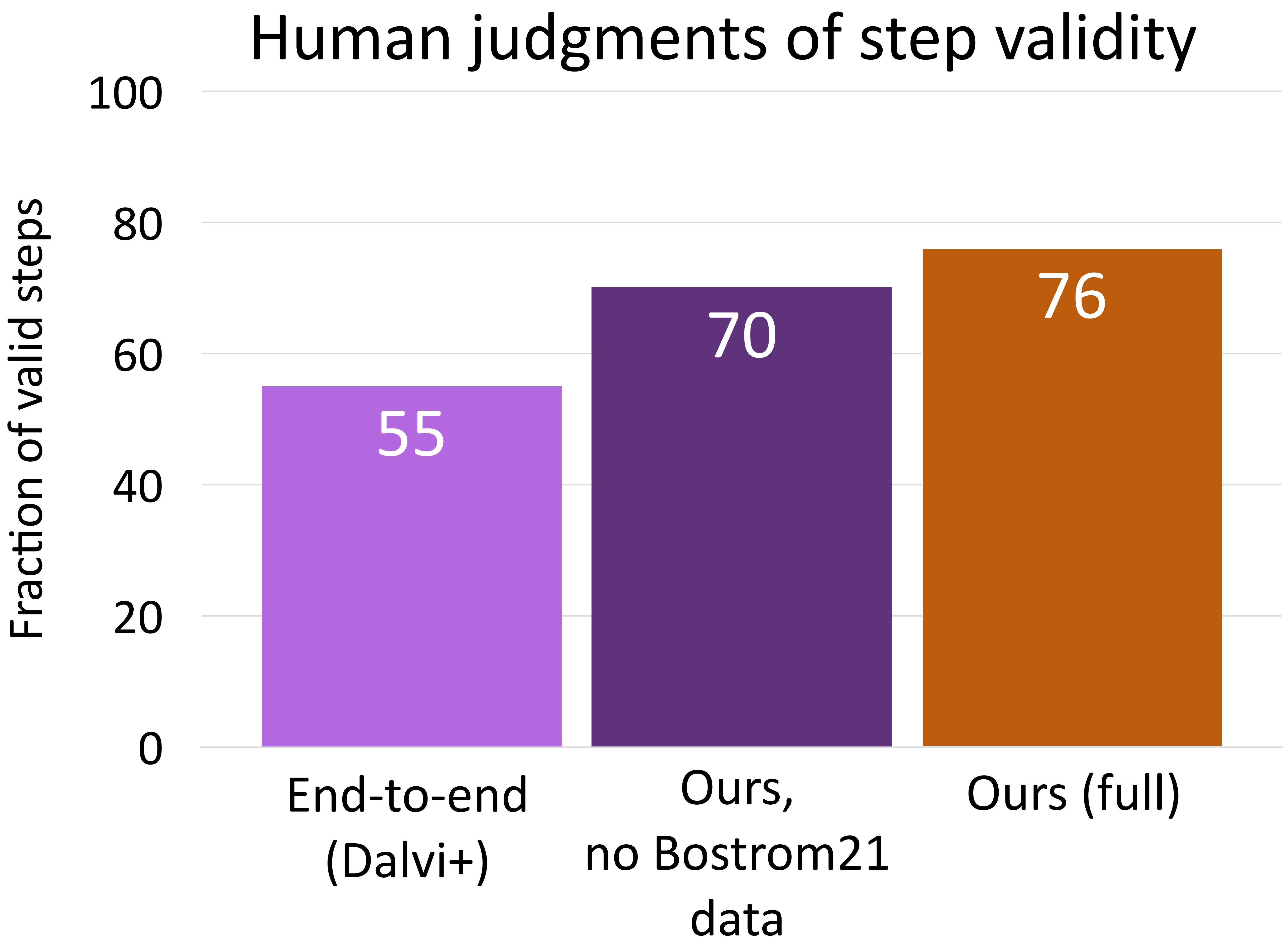
Bostrom et al. arXiv22

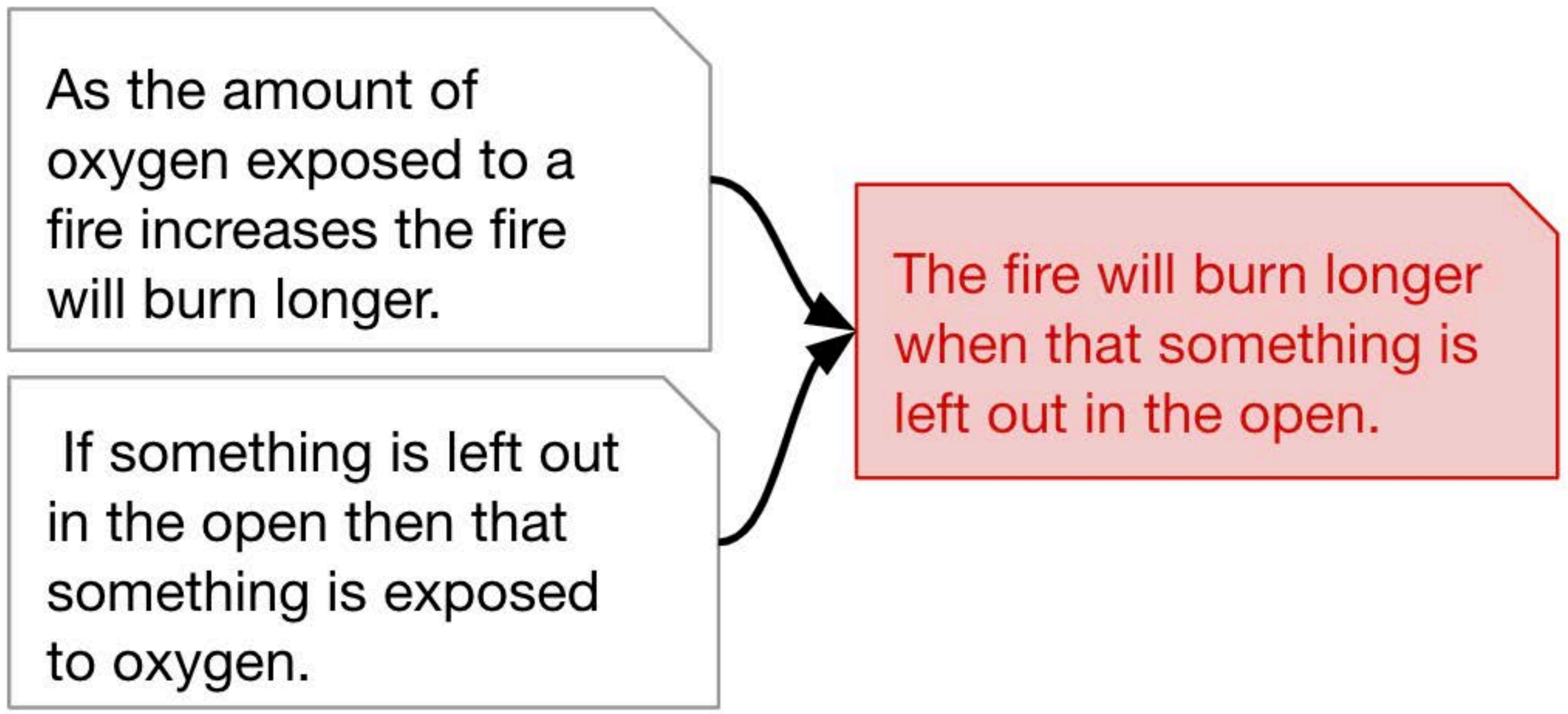
Results

- Input: 25 premises and a potentially true hypothesis. Goal: classify hypothesis as T/F
- Baseline: a pure end-to-end T5 model (Dalvi et al., 2021). Rank outputs by generation probability and apply a threshold to classify
- Separating concerns of search and deduction is important, and a good heuristic is important



Results: Individual Steps





- Our model is substantially better than end-to-end T5
- Some gray area about what's an error or not

Takeaways

- Our deduction models can capture broad-domain reasoning patterns **with little human training signal**, no logical forms
- Our models are **expressive** (can represent statements across several datasets) and are **flexible**
- A multi-step reasoning system founded on our deduction principles outperforms a pure end-to-end approach. **Structuring the reasoning this way helps!**
- Ongoing work: learning a backward model to do abductive inference, be able to hypothesize missing premises
- Ongoing work: take a step towards symbolic components in the model

Outline

Entailment to verify QA

Jifan Chen, Eunsol Choi, GD. EMNLP-Findings21
Can NLI Models Verify QA Systems' Predictions?

evidence → **standalone statement** ←NLI→ **Q+answer**

Logically manipulating statements

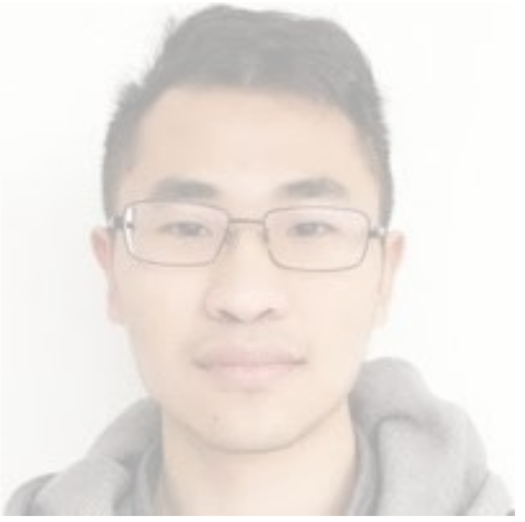
Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, GD. EMNLP21
Flexible Generation of Natural Language Deductions

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, GD. In submission.
Natural Language Deduction through Search over Statement Compositions

statement, statement → **deduction**

Improving diverse generation

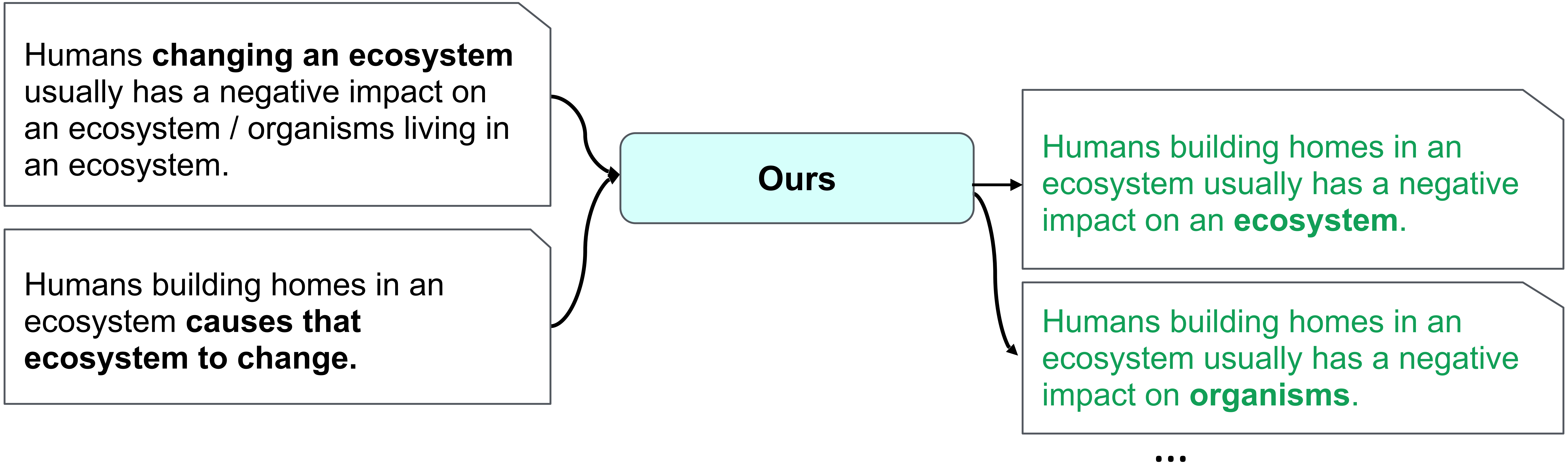
Jiacheng Xu, GD. NAACL22.
Massive-scale Decoding for Text Generation using Lattices







Advancements in Generation



- Multiple correct generations — we're not sure which ones might be useful
- How can we access **as many generation candidates as possible?**

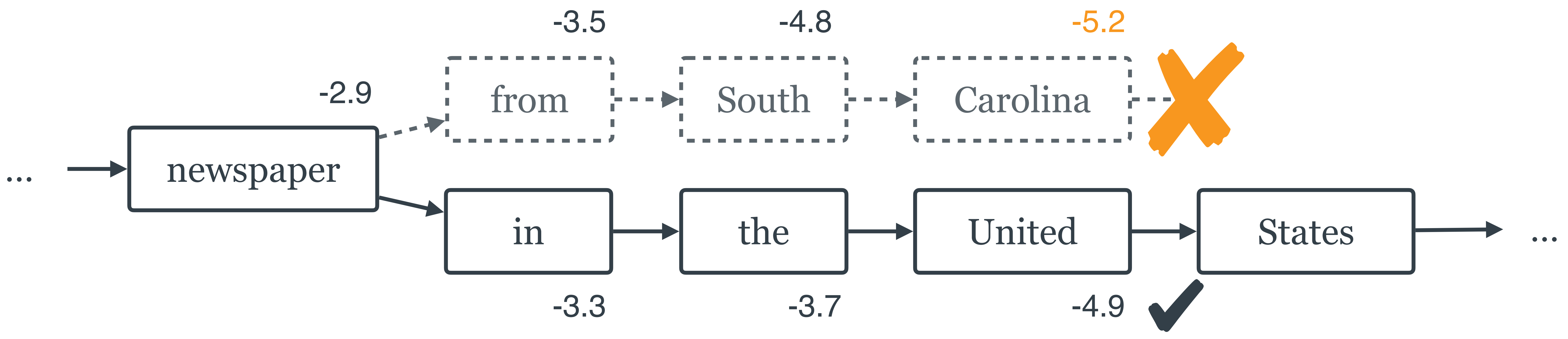
Getting Diverse Summaries

Input Doc → Model →

Generated Summaries with beam search

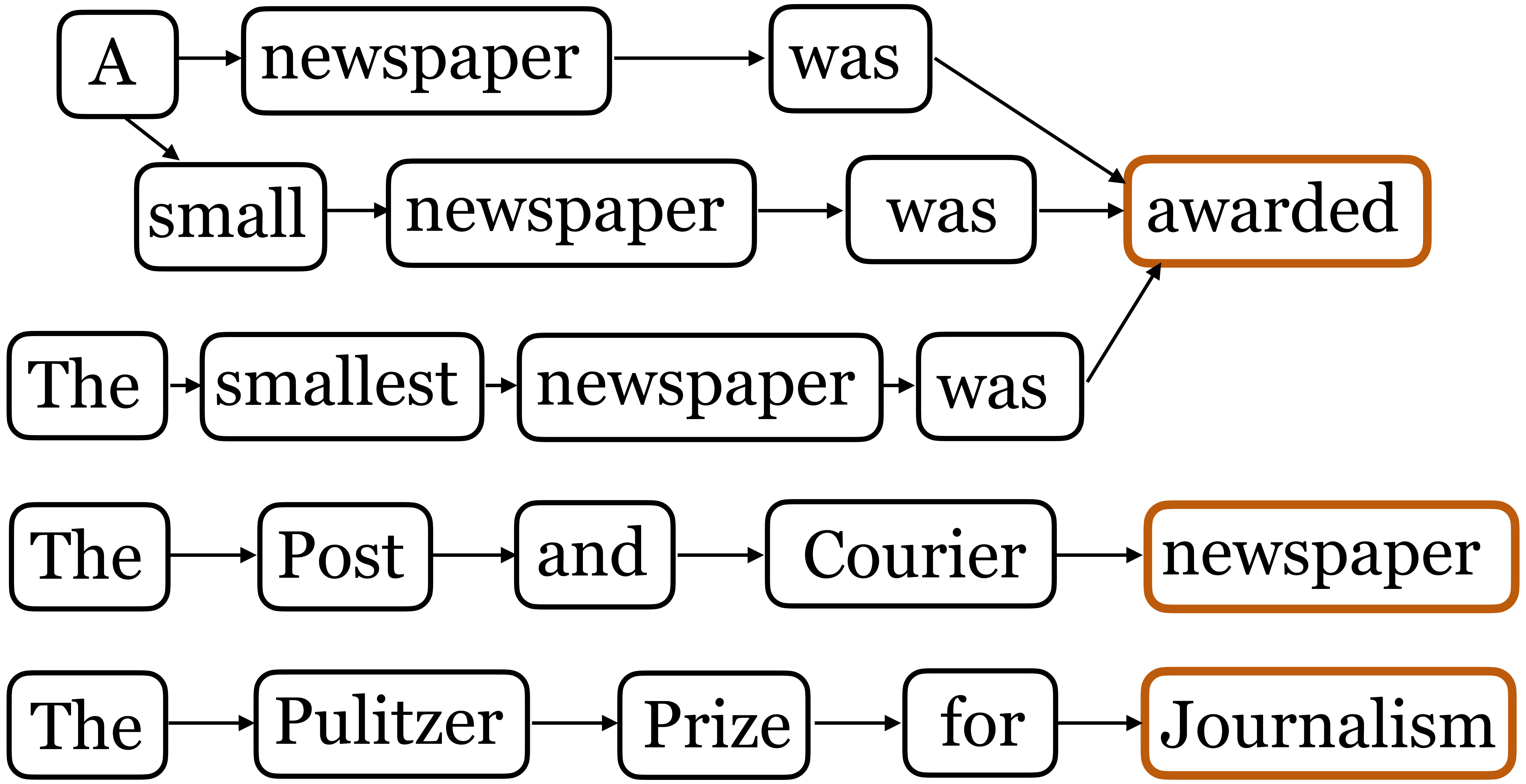
1. The smallest newspaper in the United States has won the Pulitzer Prize for Journalism for the second year in a row.
2. The smallest newspaper in the United States has won the Pulitzer Prize for Journalism.
3. The smallest newspaper in the United States has won the Pulitzer Prize for Journalism for the first time.
4. The smallest newspaper in the United States was awarded the gold medal for public service.

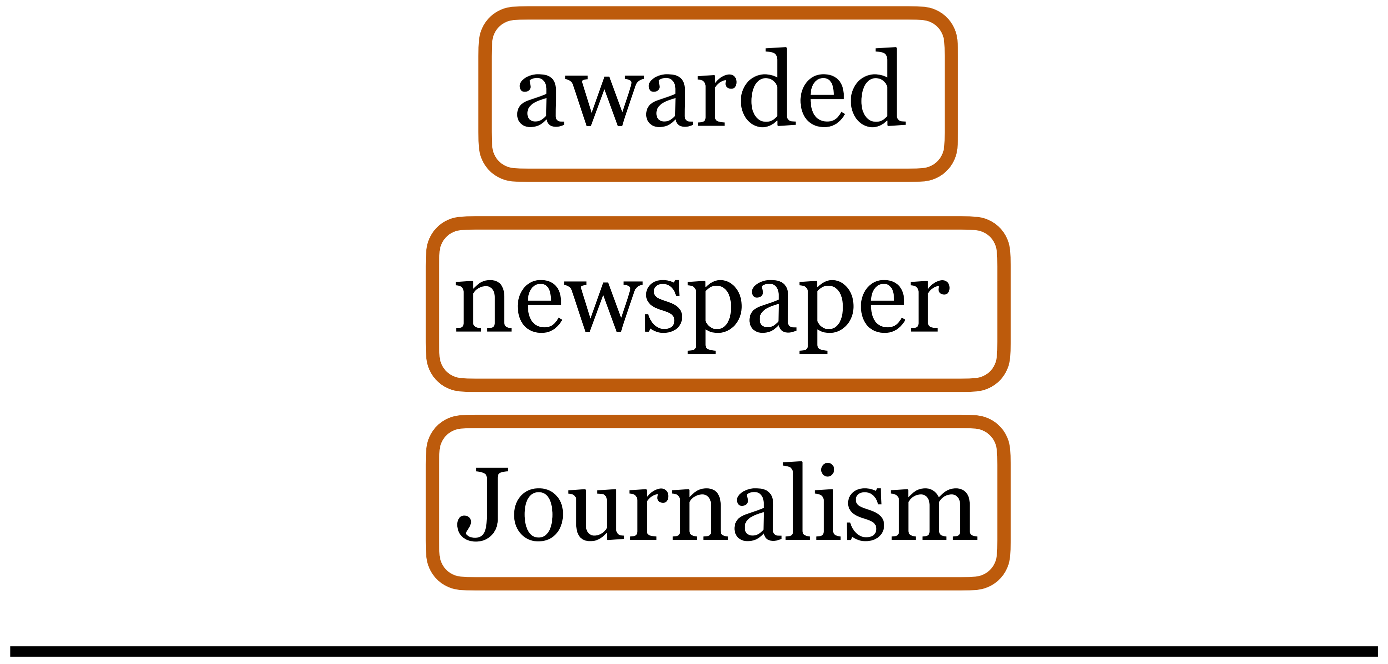
- Top summaries are similar and wrong! (not the smallest). **Too much redundancy**
- Other useful states were explored (info about location) but **pruning eliminated them**



- We're going to fix **two problems** with beam search to improve diversity

Reducing Redundancy with Recombination





Beam state

- Hypotheses are stored in a lattice. Beam search now operates over nodes in this lattice
- Expanding a node continues all of the hypotheses ending in that node

Hypothesis Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Prefix A: A small $\boxed{\text{newspaper was awarded}}$

Prefix B: A $\boxed{\text{newspaper was awarded}}$

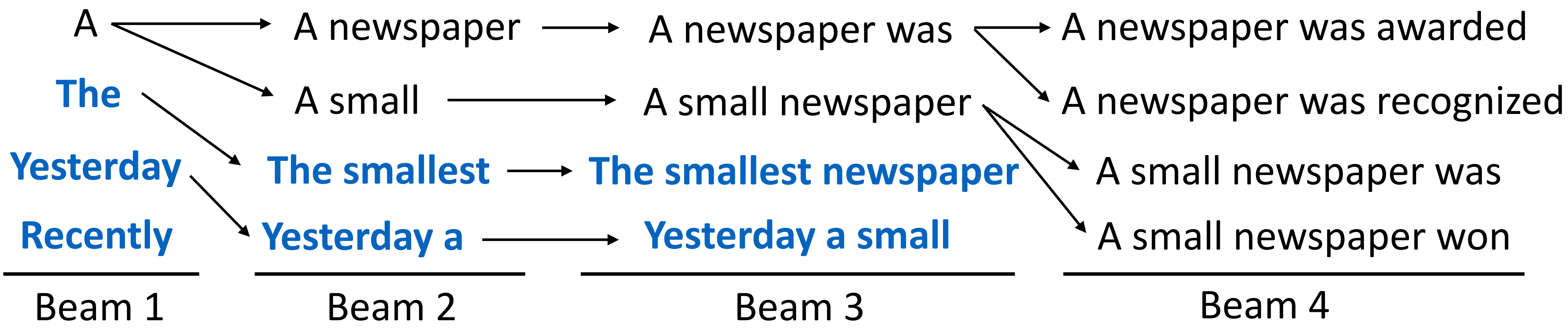
→ Assumption: if these criteria are met, the rest of the summary will be similar: P(**y** | document, **A**) ≈ P(**y** | document, **B**)

- For summarization: we find that when this heuristic applies, **~70%** of time the greedy completion of the summary is exactly the same.
- When these distributions match, merging states in the lattice is completely okay!

Zhisong Zhang et al., EMNLP 2018

Reducing Pruned States

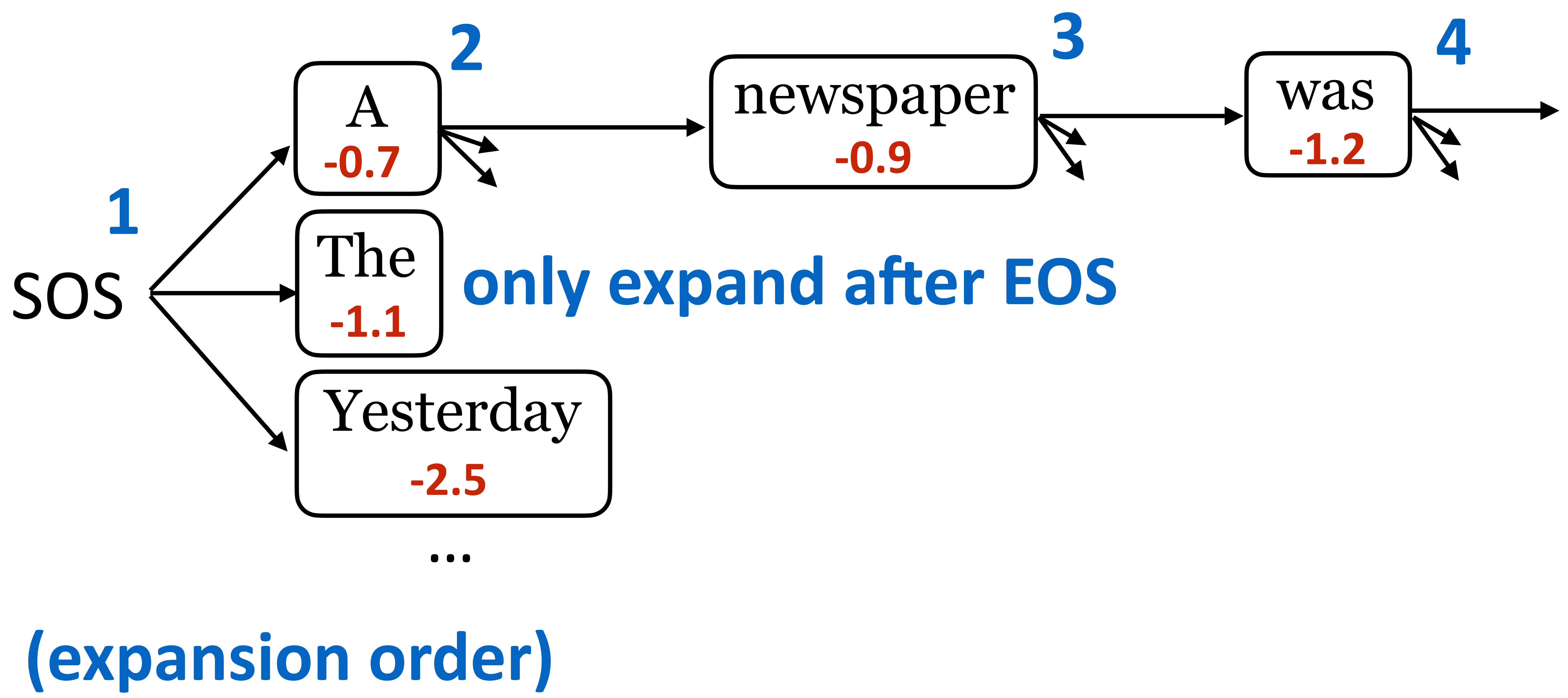
- **Beam search wastes time** — most expanded hypotheses are eventually discarded. (Makes sense if you want the one-best, but not to get diverse options)



- Every **blue step** ultimately got pruned, even though these could be good summaries

Reduce Pruning with BFS/DFS

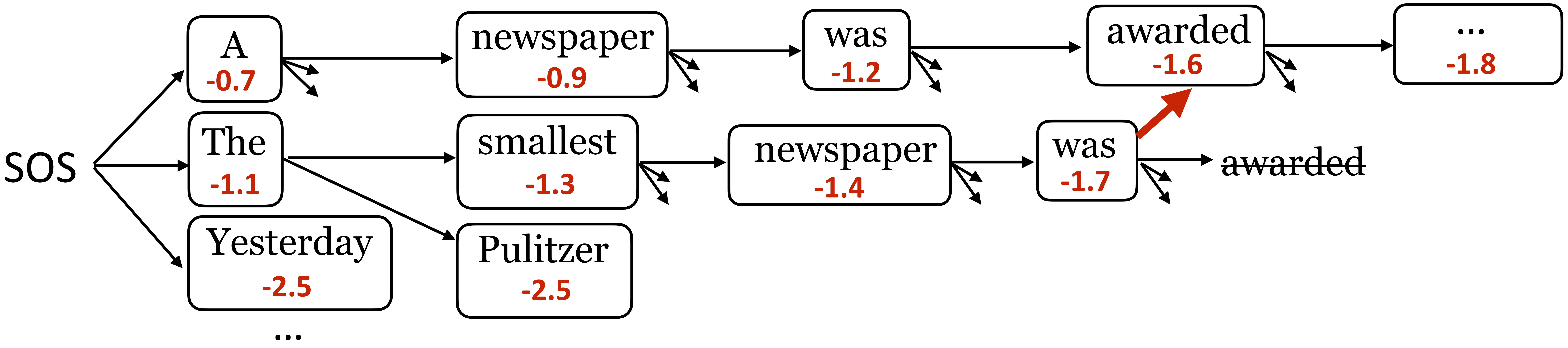
- We want a search algorithm where **every explored state** is on some finished path
- Use a modified best-first search with a depth-first stage: greedily expand each node until an EOS token is reached
- Expand by model score, shown in **red** below



...but for our depth-first stage, we continue expanding until we reach EOS. This greedy path is typically high-scoring.

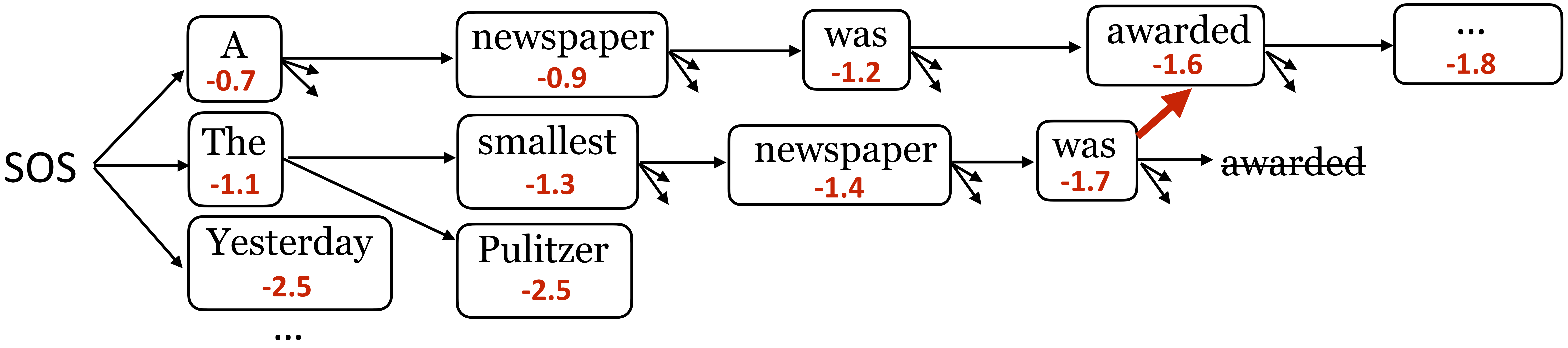
Reduce Pruning with BFS/DFS

- Effective when combined with recombination: subsequent paths that are explored may overlap with earlier ones

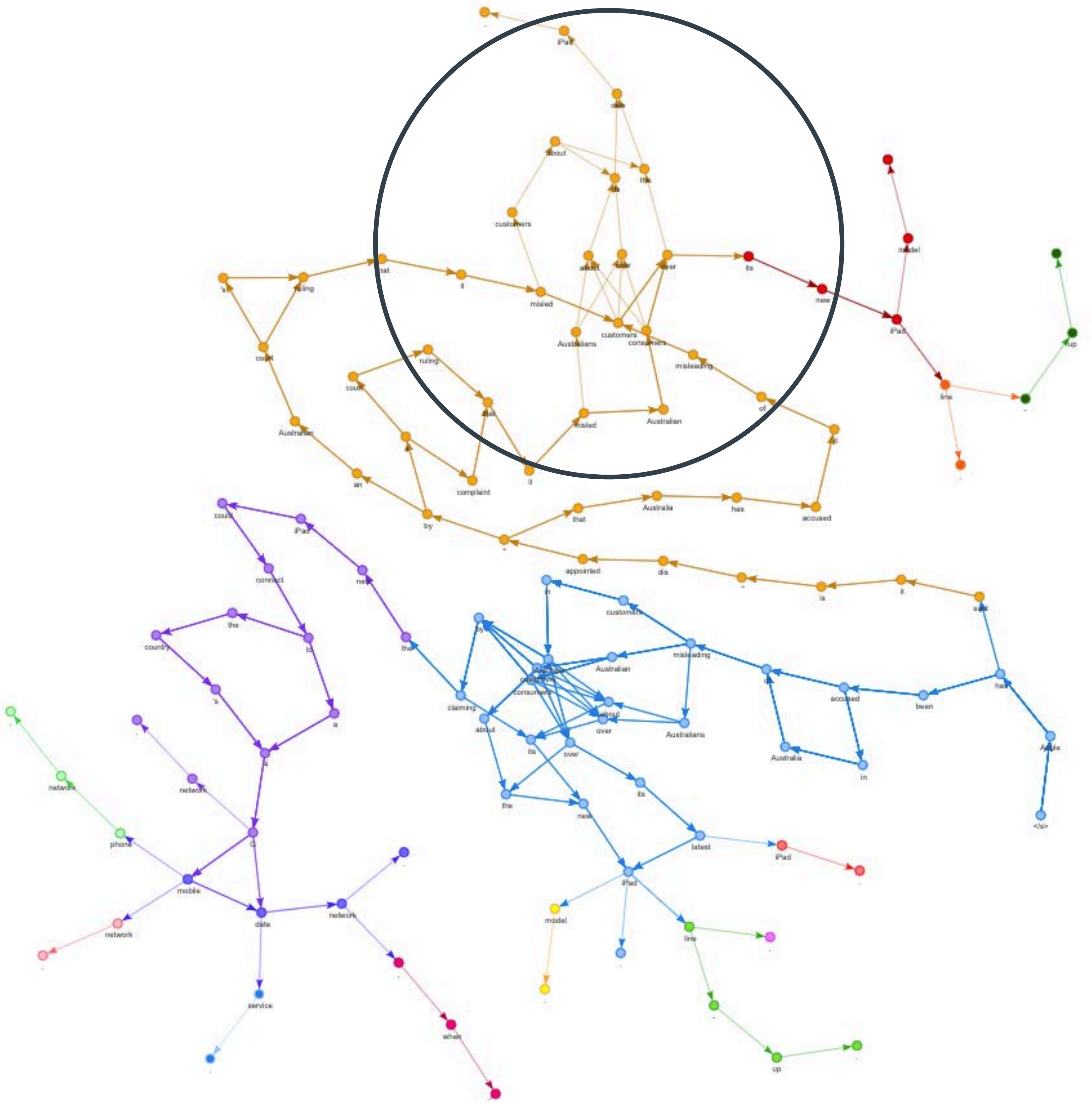


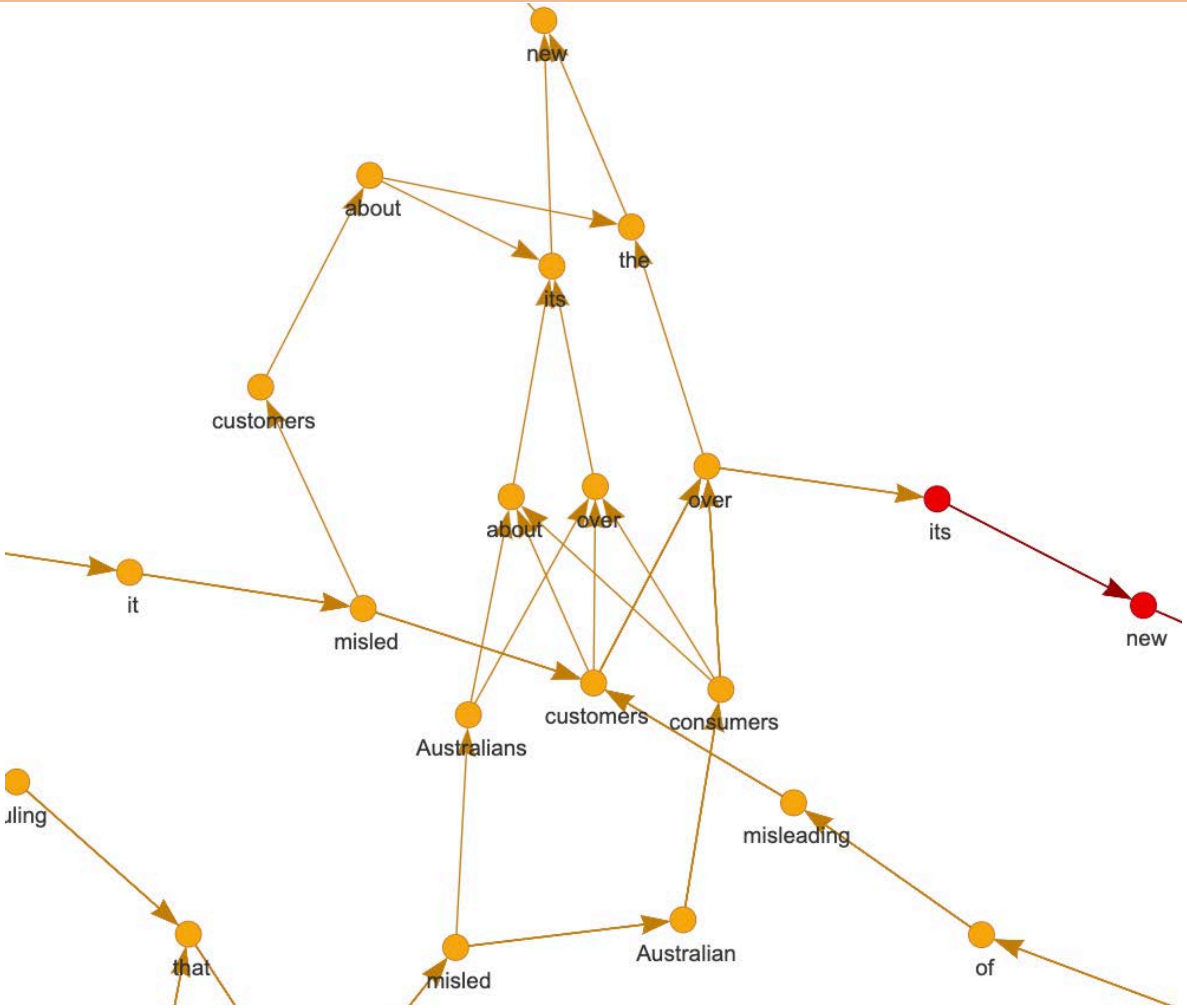
Overall Algorithm

- Explore paths with best-first/depth-first search
- Merge states when redundancy is identified
- Construct a **lattice** of many possible options



Final Output: Summary Lattice

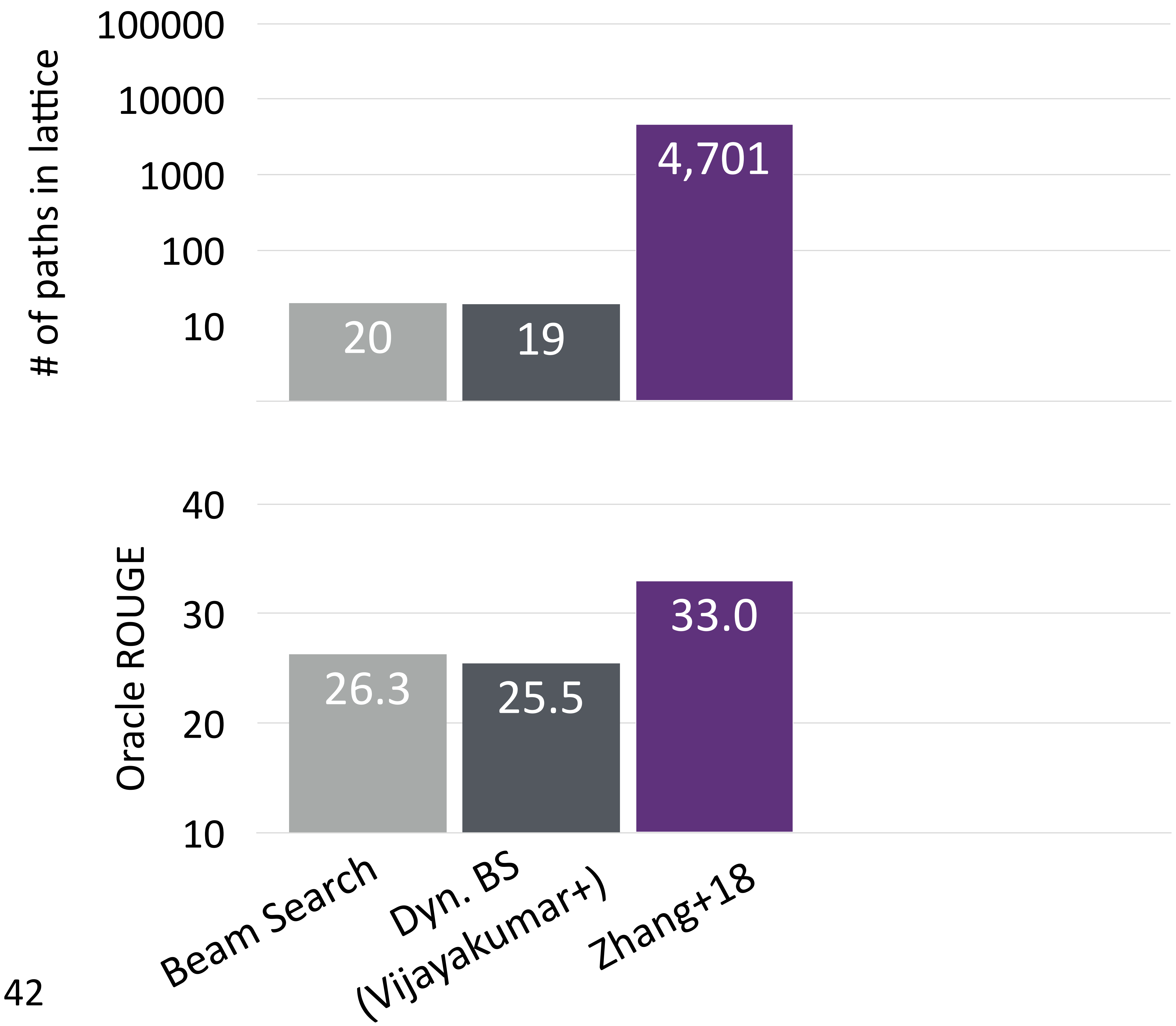




- Encode exponentially many summaries in a compact space

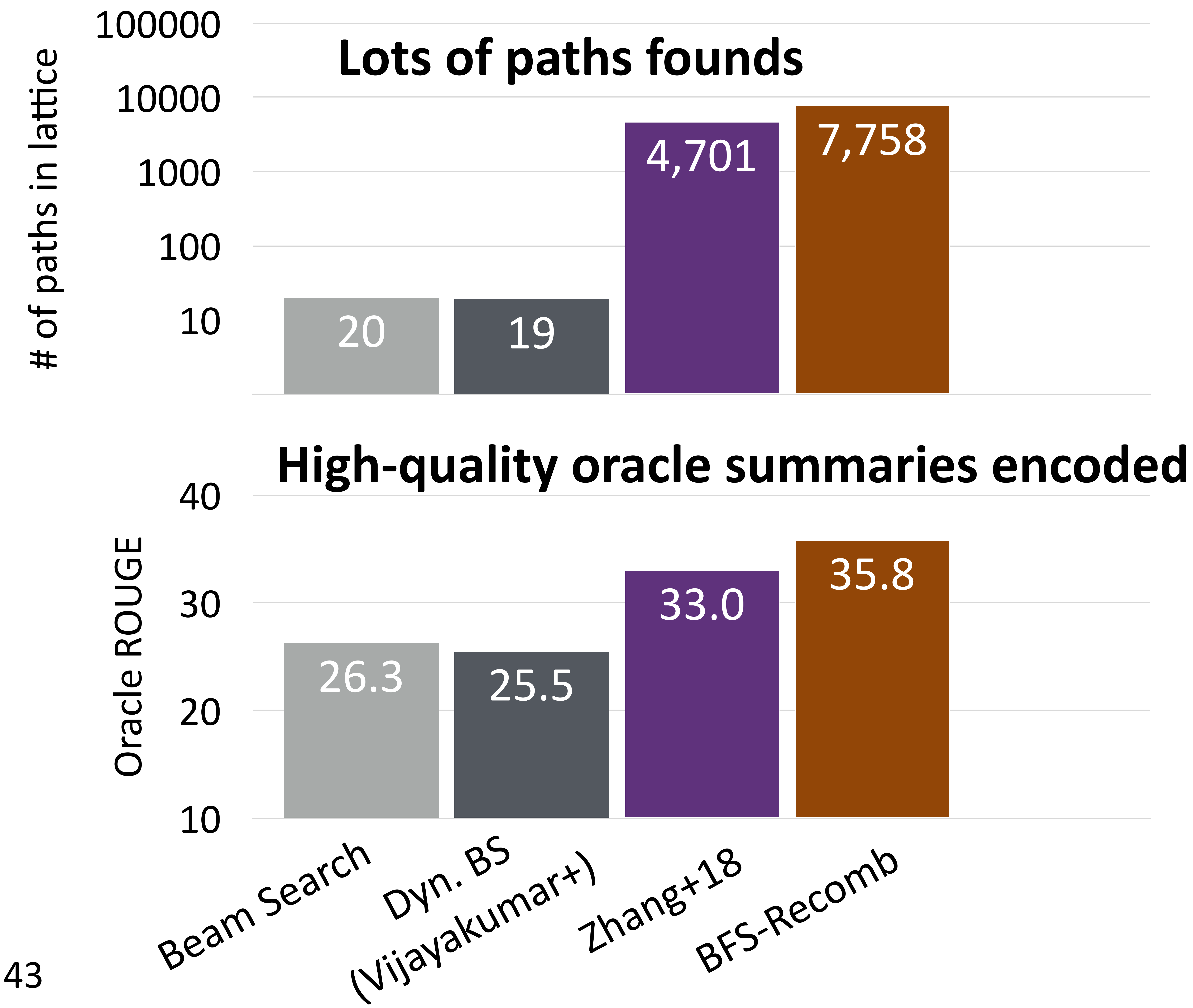
(customers | consumers) (about | over) (its | the) new iPad…

Results



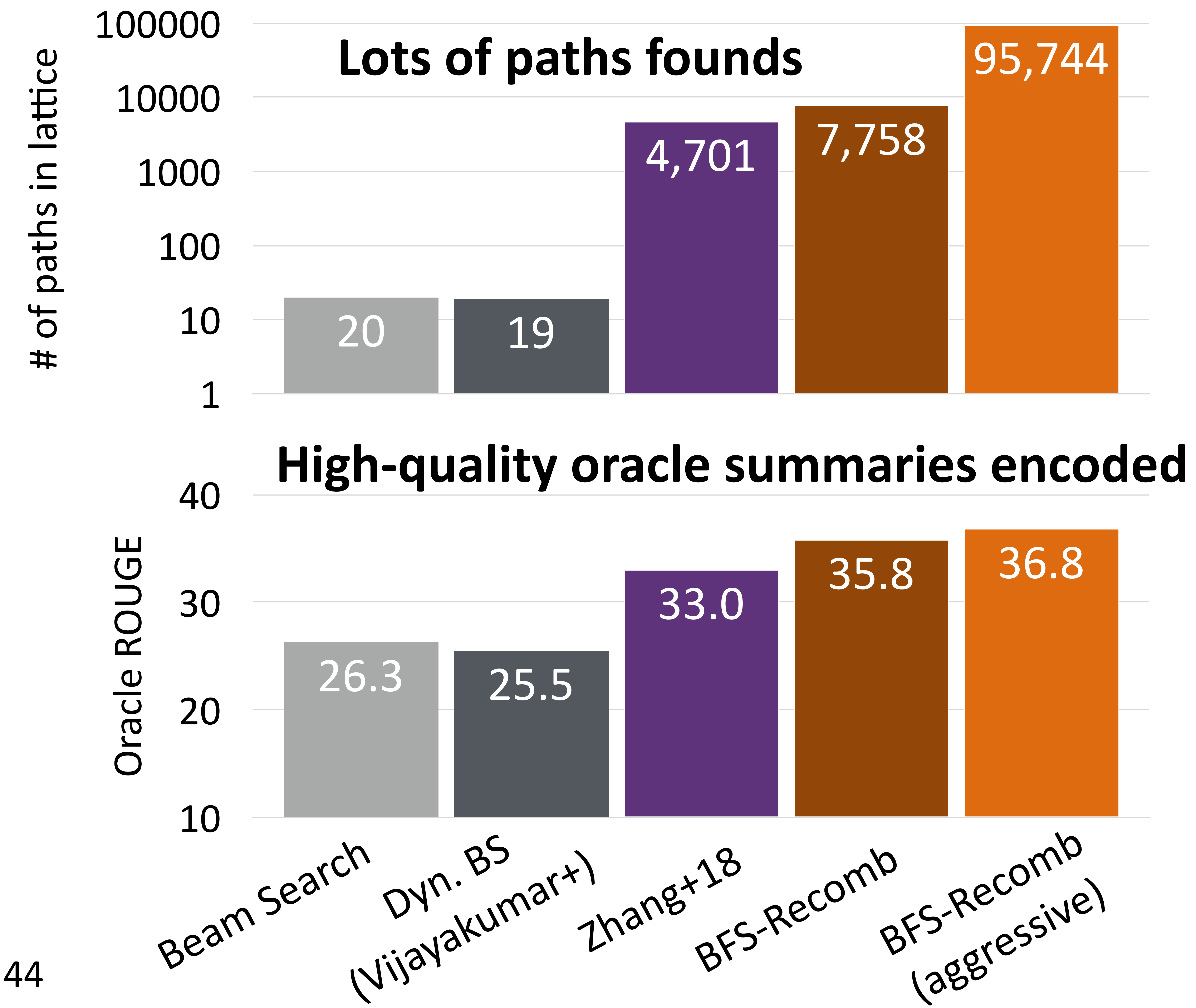
# of paths in lattice
100000
10000
1000
100
10
20
19
4,701
Oracle ROUGE
40
30
20
10
26.3
25.5
33.0
Beam Search
Dyn. BS (Vijayakumar+)
Zhang+18

Results



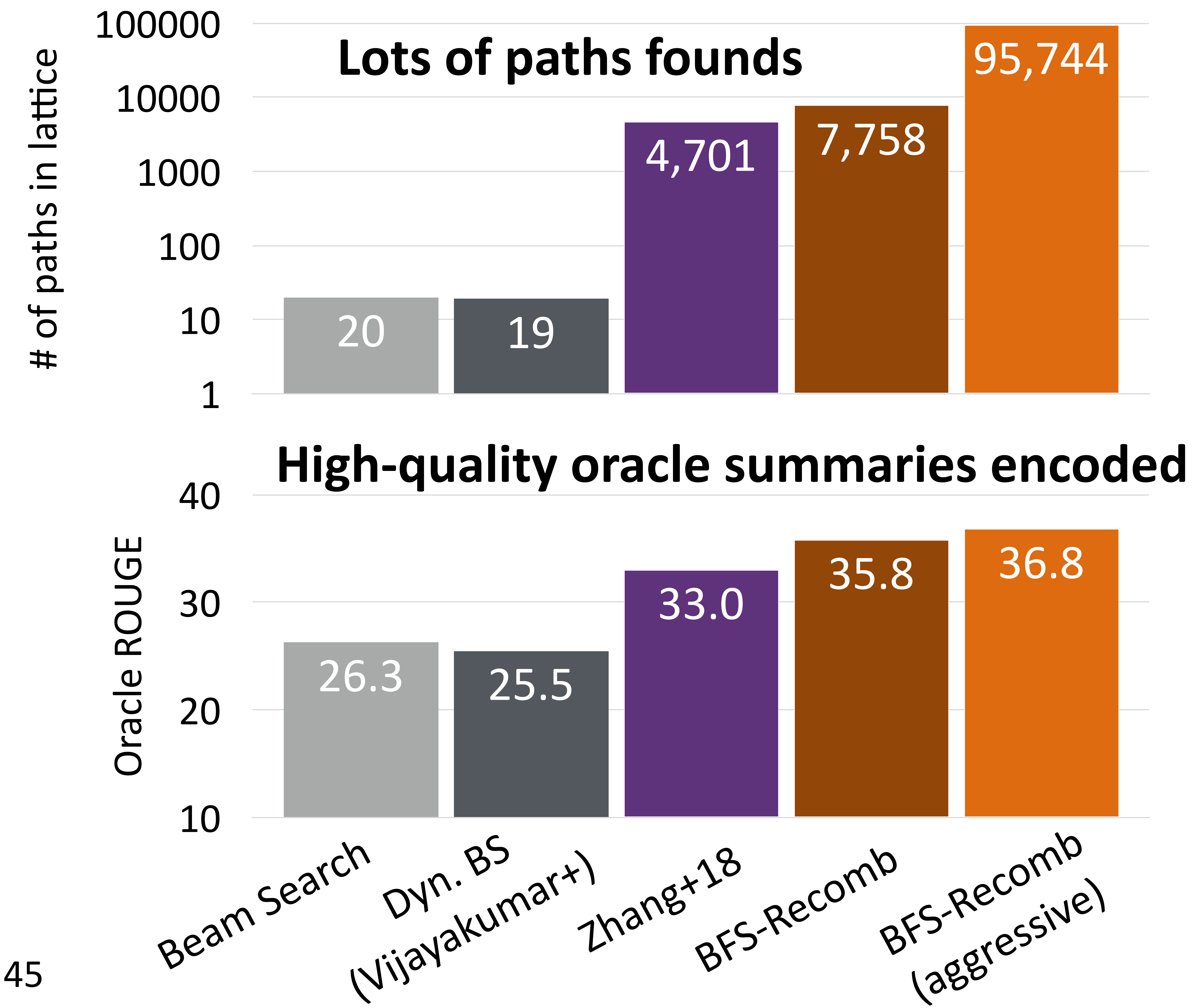
Lots of paths founds
# of paths in lattice
100000
10000
1000
100
10
20
19
4,701
7,758
High-quality oracle summaries encoded
Oracle ROUGE
40
30
20
10
26.3
25.5
33.0
35.8
Beam Search
Dyn. BS (Vijayakumar+)
Zhang+18
BFS-Recomb

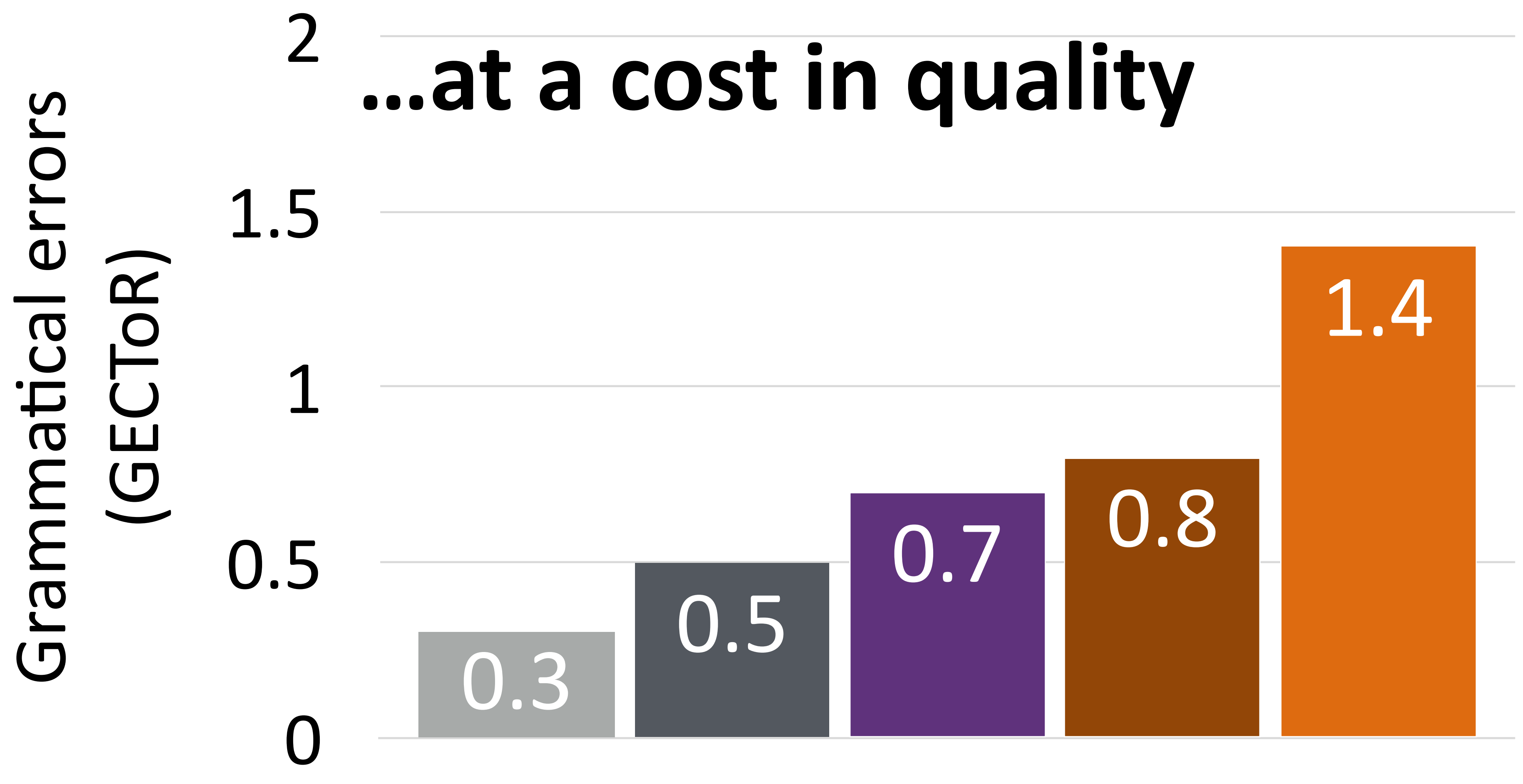
Results



Lots of paths founds
# of paths in lattice
100000
10000
1000
100
10
1
20
19
4,701
7,758
95,744
High-quality oracle summaries encoded
Oracle ROUGE
40
30
20
10
26.3
25.5
33.0
35.8
36.8
Beam Search
Dyn. BS (Vijayakumar+)
Zhang+18
BFS-Recomb
BFS-Recomb (aggressive)

Results





- Our most aggressive merging does introduce some grammatical errors. Better merging heuristics could help with this.

Goals for Lattices

Our generation systems can already encode lots of good options. We just need to be able to efficiently find them and encode them!

- Can we rerank our generated deductions and pick out the good ones?
- Can users control + correct the system on-the-fly, with the system learning those corrections?

Applications: factuality, controllable dialogue, diverse paraphrasing, and more!

Enumerating all valid deductions + a strong proof engine = effective search over natural language proofs using present-day models?

Outline

Entailment to verify QA

Jifan Chen, Eunsol Choi, GD. EMNLP-Findings21
Can NLI Models Verify QA Systems' Predictions?

evidence → **standalone statement** ←**NLI**→ **Q+answer**

Logically manipulating statements

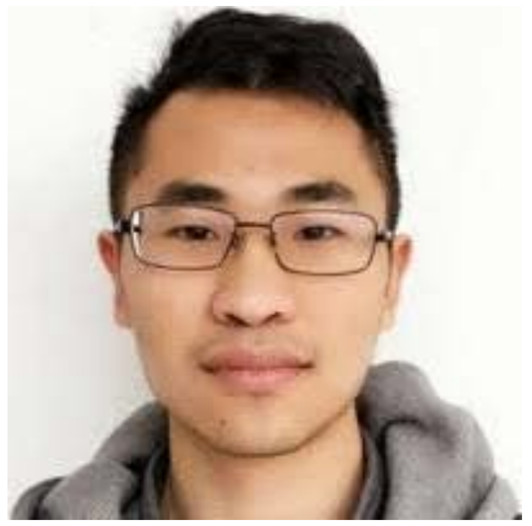
Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, GD. EMNLP21
Flexible Generation of Natural Language Deductions

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, GD. In submission.
Natural Language Deduction through Search over Statement Compositions

statement, statement → **deduction**

Improving diverse generation

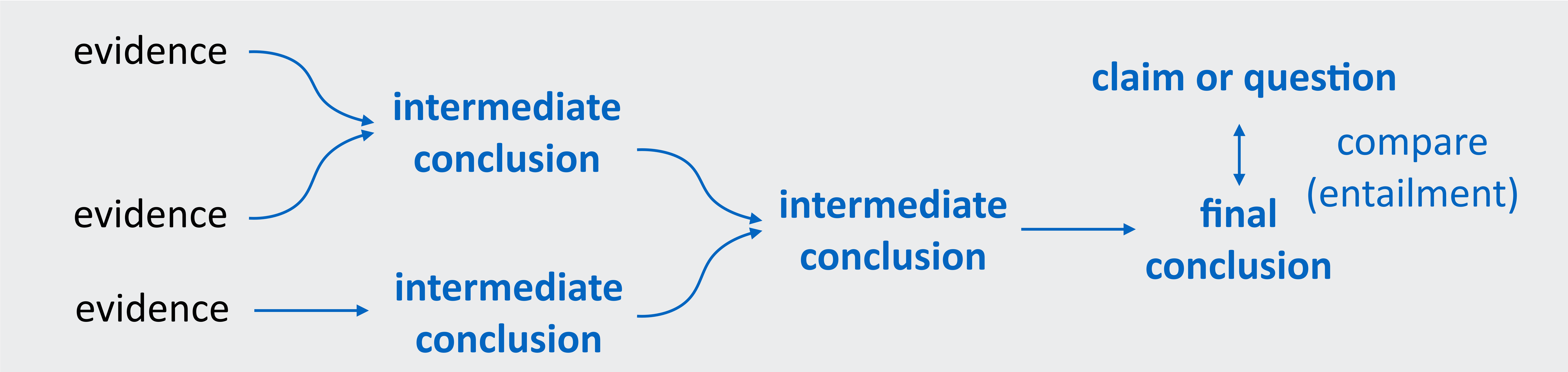
Jiacheng Xu, GD. NAACL22.
Massive-scale Decoding for Text Generation using Lattices







Path to Multi-step Reasoning



With better generation techniques and advances in pre-trained models, these steps will become easier and easier!

Goals

- Can we use it to do better, more explainable fact-checking?

Claim: Social Security was basically invented at the University of Wisconsin-Madison; that's where Franklin Roosevelt got the idea.

Decompose

Who invented social security?

Political Scientists at the University of Wisconsin-Madison, including Edwin Witte, known as the "Father of Social Security," Arthur J. Altmeyer, and Wilbur Cohen developed the 1934 proposal for a federally funded pension plan.

When did Franklin Roosevelt implement social security?

The Social Security Act was enacted August 14, 1935.

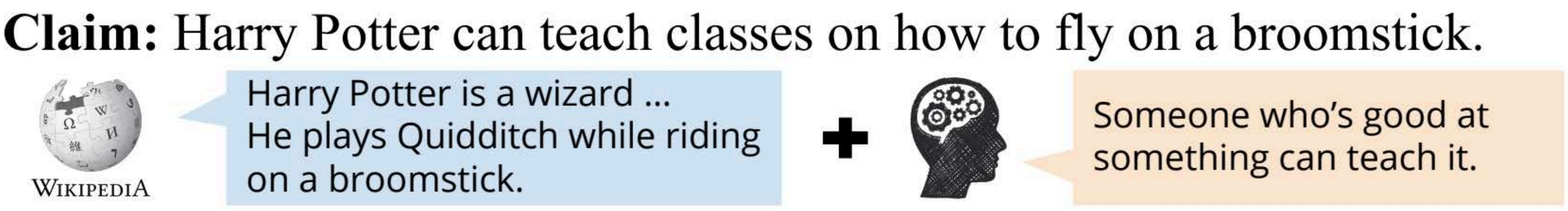
intermediate conclusion → **verify/contradict claim**

Angela Fan et al. *Generating Fact-Checking Briefs*, EMNLP 2020

Jifan Chen, Aniruddh Sriram, Eunsol Choi, GD. In preparation.

Goals

- Can we materialize reasoning about entities?



Large annotated dataset (13k total claims about entities). Can textual reasoning help materialize an explanation?



Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, GD. *NeurIPS Datasets* 2021
CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge

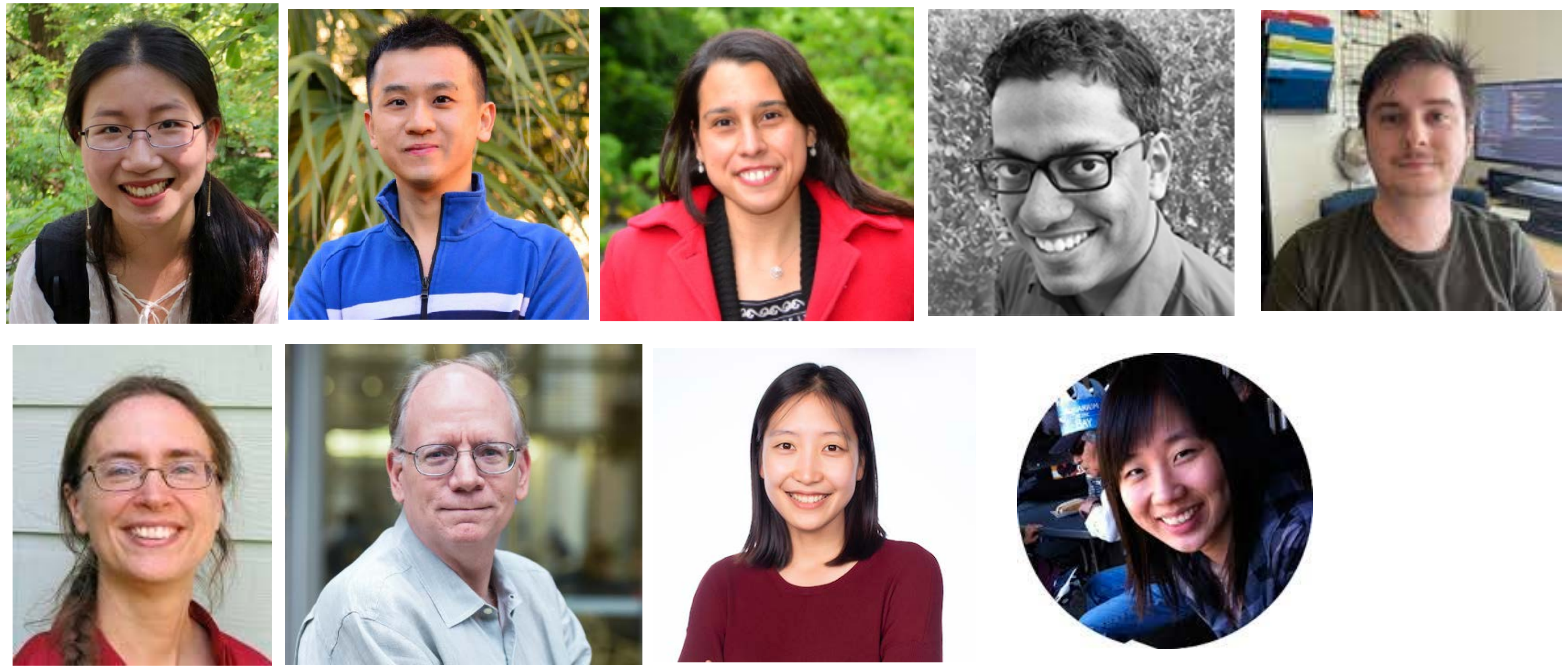
Conclusion

- Complex reasoning problems can be tackled using natural language to represent intermediate conclusions

- Natural language is an **expressive**, **flexible**, and **interpretable** vehicle for reasoning

- New datasets and better models are dramatically improving our ability to manipulate text (PaLM). **Making logical inferences in text is increasingly becoming viable.**

Acknowledgments



Thanks!

Results: EntailmentBank Human Eval

Similar performance on EntailmentBank using just one of our reasoning patterns

Caveat: we can't handle every reasoning pattern…yet :)

Bhavana Dalvi, Peter Jansen, et al., 2021