Improving model controllability with input and output alterations

Emily Pitler

Google Research

Language models have some impressive capabilities

Google Research

Logical inference, no tuning, 2 exemplars

Input: Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?

Model Output: The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

PaLM: Scaling Language Modeling with Pathways, Chowdhery et al, arXiv 2022

Logical inference, no tuning, 2 exemplars

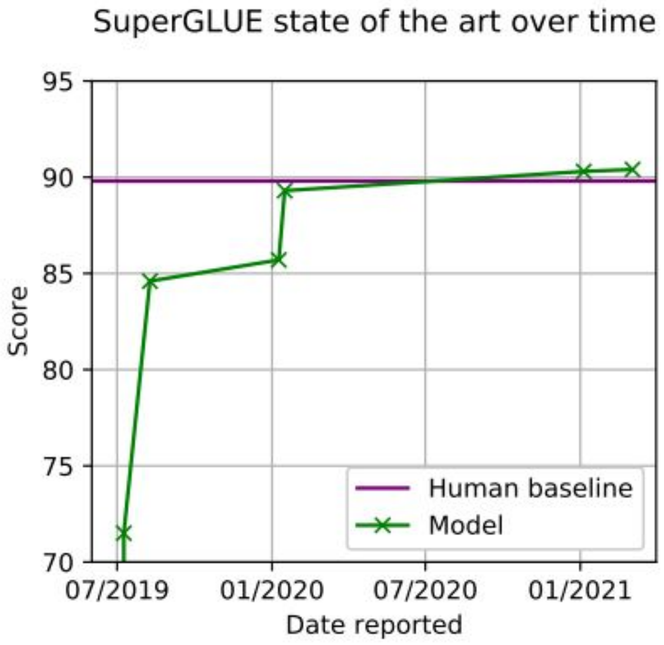
Input: Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?

Model Output: The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

Note: Still not perfectly factual

PaLM: Scaling Language Modeling with Pathways, Chowdhery et al, arXiv 2022

Even hard NLU benchmarks got saturated



[SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems]. Wang et al, NeurIPS 2019

Figure 2a from [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models] (Big-Bench), Srivastava et al, arXiv 2022

Google Research

Continued improvements from scale

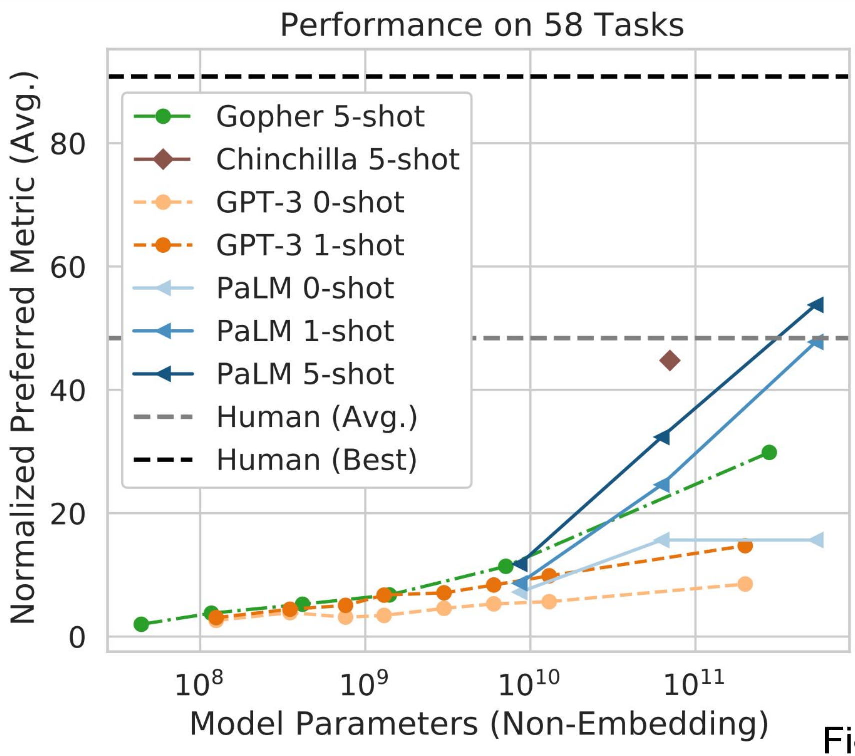


Figure 3 from PaLM

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (Big-Bench), Srivastava et al, arXiv 2022

PaLM: Scaling Language Modeling with Pathways, Chowdhery et al, arXiv 2022

Scaling language models: Methods, analysis & insights from training gopher.

Training compute-optimal large language models

Language models are few-shot learners

Google Research

Focus: Complementary challenges

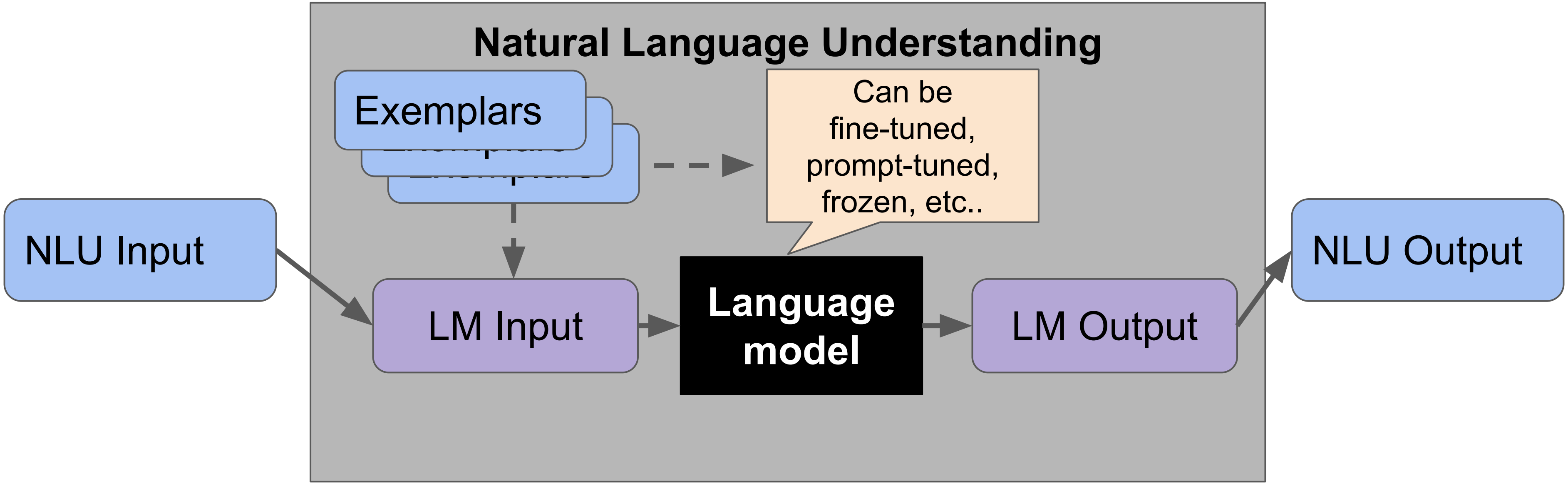
Typical NLU benchmarks: train (and/or prompt)-eval, input → output

Increasingly relevant other criteria:

- Controllability
- Adapting to
  - Changing gold standards
  - New domains
  - New information
- Interpretability

Making the LM part better is not the only way to help on these

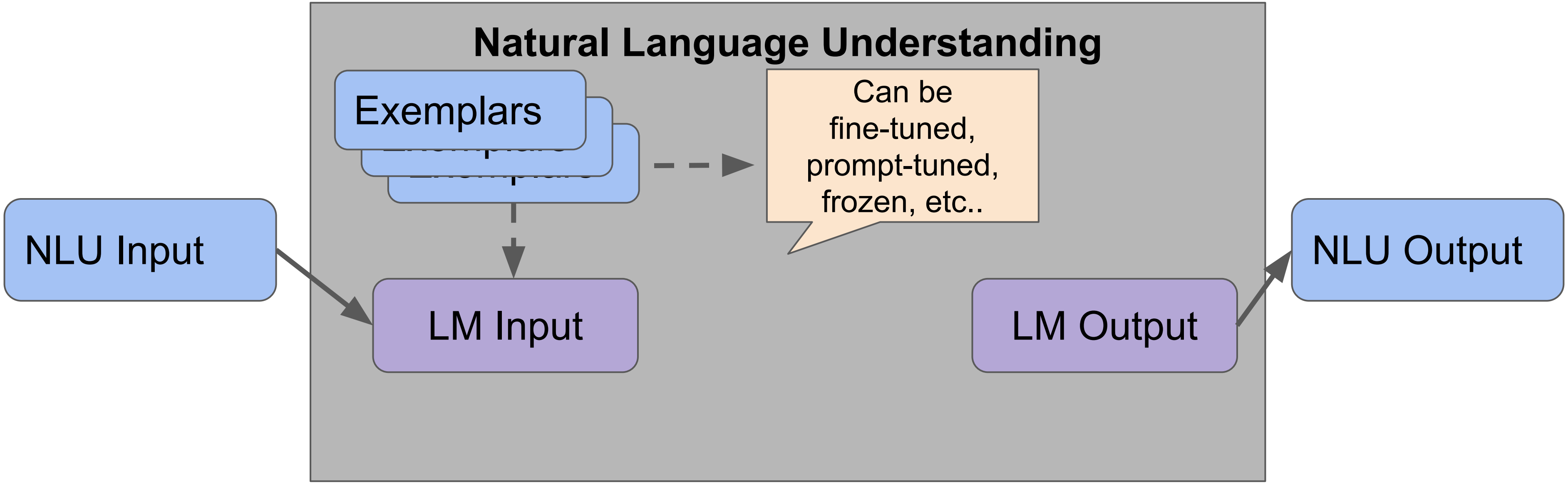
Focus: Non-LM aspects of a NLU recipe



Google Research

Focus: Non-LM aspects of a NLU recipe

What's left?

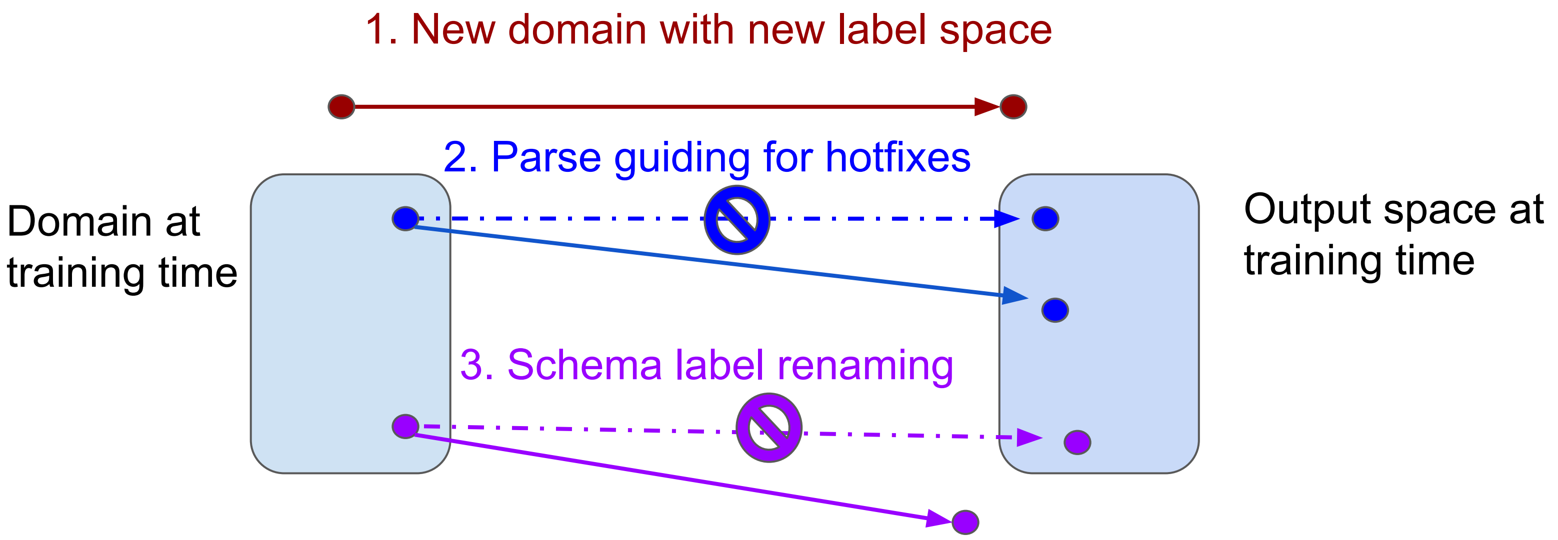


Controllable Semantic Parsing via Retrieval Augmentation

Panupong Pasupat, Yuan Zhang, Kelvin Guu

EMNLP 2021

Controllability scenarios in semantic parsing



…without retraining the parameters of the encoder-decoder model

Google Research

Reasons we may want to avoid retraining

Observe results immediately, speeding up development cycle

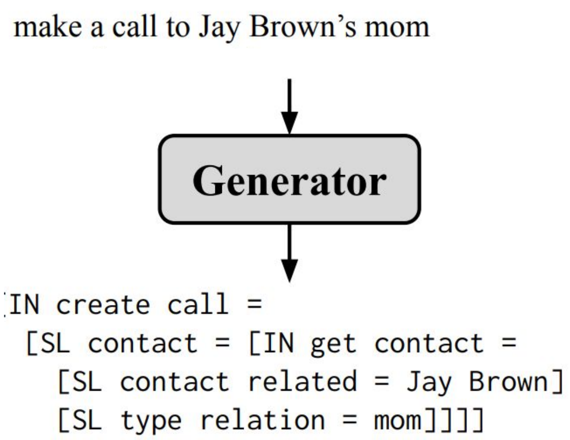
Get updates into production quickly

Avoid interacting with other clients if sharing the same base model

Training only on new might lead to catastrophic forgetting

Training on old + new computationally expensive

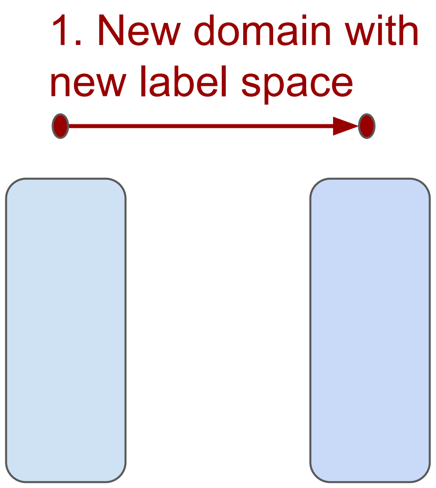
Baseline: T5-base



Experiments on English portion of MTOP dataset (Li et al, 2021)

On standard train-test, was higher than prior state-of-the-art 84.3 → 85.1

Setting 1: Domain boot-strapping



Parser trained on set of existing domains

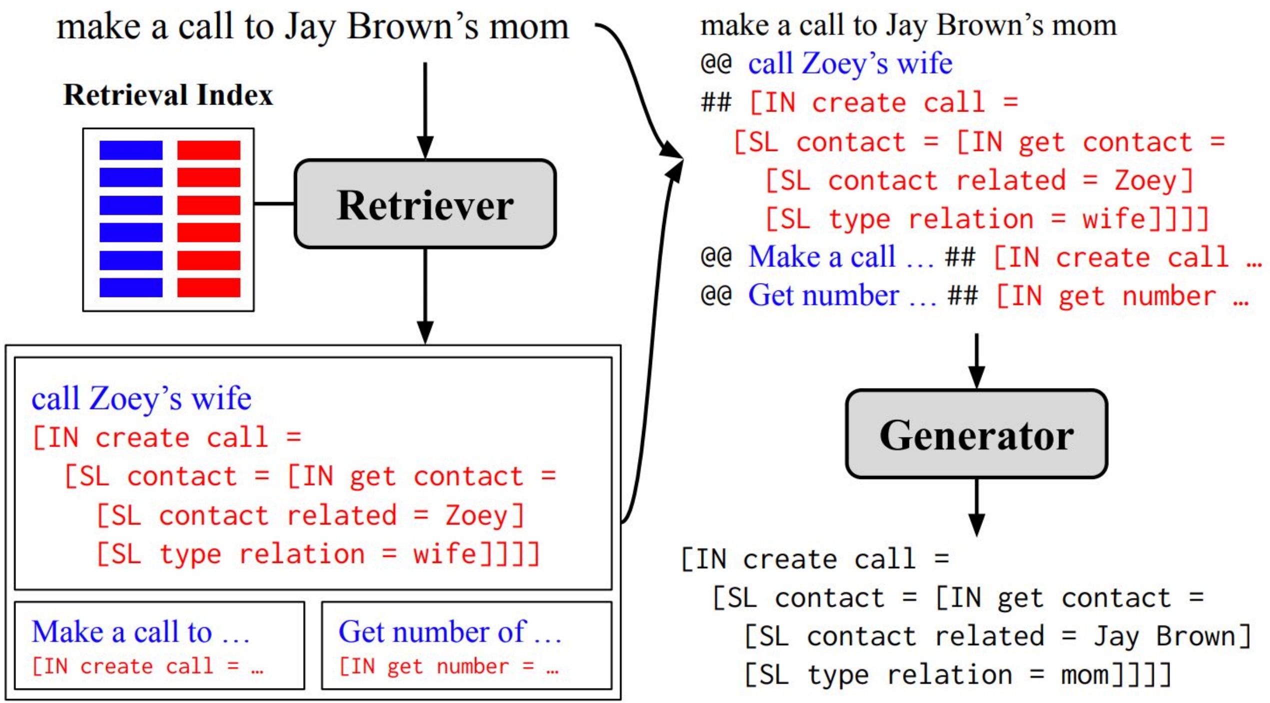
e.g., *Alarm, calling, messaging, music*

Small number of examples for a new domain available at test-time only

e.g., *event*

How can we benefit from these without any retraining?

Augment Inputs with Source Structures



Google Research

Why the model may ignore the exemplars

Adding retrieval: 5% → 39% on the new domain. Better, but why not higher?

If at training time, the exemplars are from the same set as training set:

- They don't provide any new information
- Distracting if the retriever is less accurate than the seq2seq model

Result: the model might just ignore the exemplar portion of the input

Technique for faithfulness: Anonymization

Add examples with *labels replaced with random numerical labels*

“create call” ⟶ 42, “contact” ⟶ 39, “get contact” ⟶ 53, ...

Original exemplars and target output:

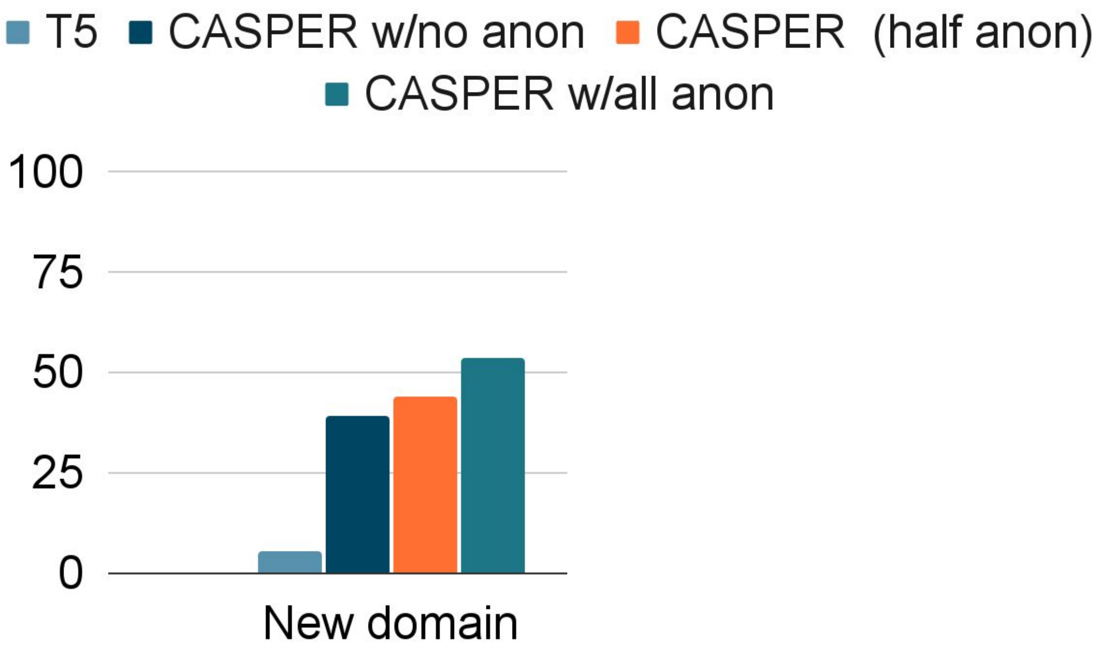
y'_1: [IN create call = [SL contact = [IN get contact = [SL contact related = Zoey] [SL relation = wife]]]]
y'_2: [IN get call = [SL contact = Jack] [SL time = today]]
y: [IN create call = [SL contact = [IN get contact = [SL relation = dad]]] [SL name app = Whatsapp]]

Anonymized:

y'_1: [IN 42 = [SL 39 = [IN 53 = [SL 6 = Zoey] [SL 94 = wife]]]]
y'_2: [IN 12 = [SL 39 = Jack] [SL 71 = today]]
y: [IN 42 = [SL 39 = [IN 53 = [SL 94 = dad]]] [SL 88 = Whatsapp]]

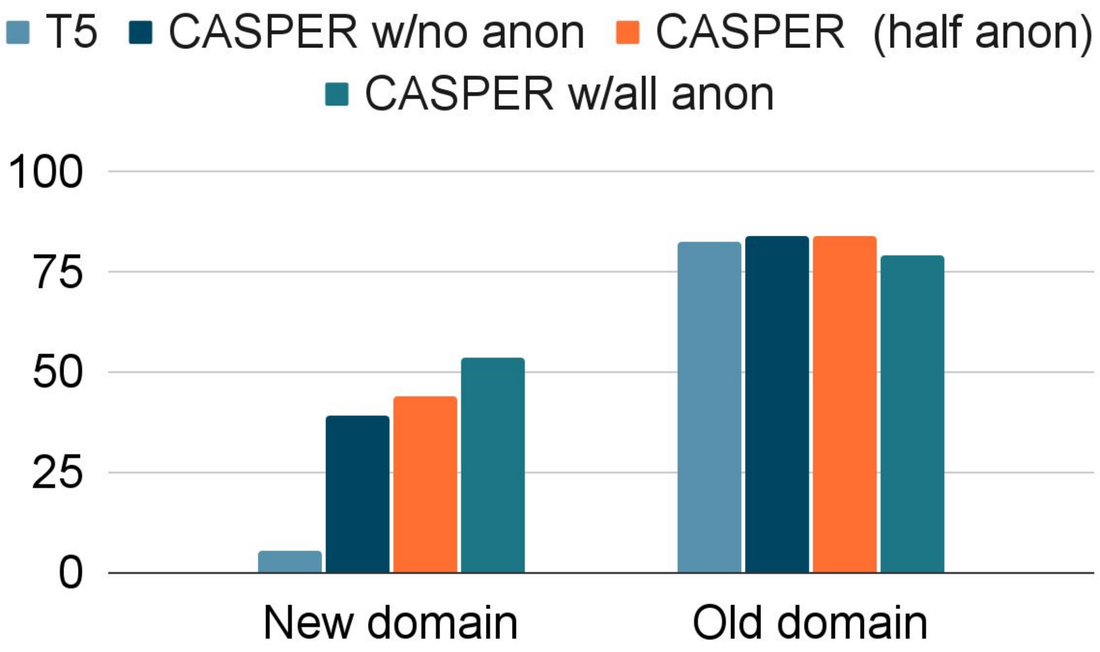
Anonymize all, none, or half of the training examples

Exemplars Improve New Domains



MTOP (Li et al, 2021), averaged over 5 choices of new domains

Exemplars Improve New Domains



MTOP (Li et al, 2021), averaged over 5 choices of new domains

Comparison with fine-tuning

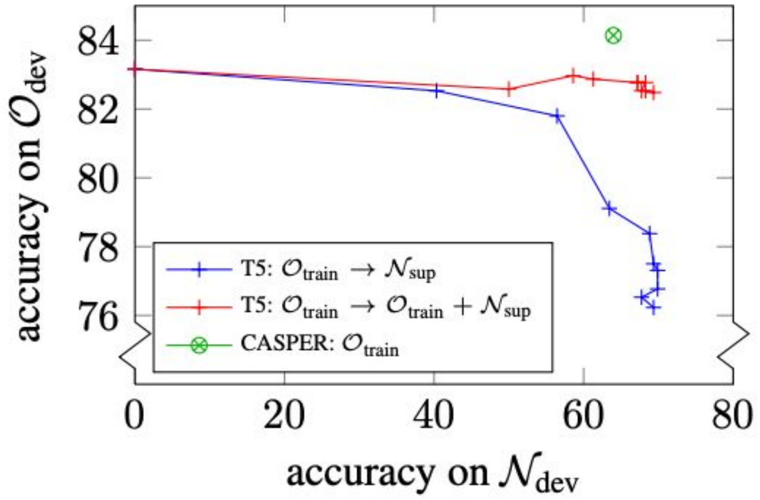
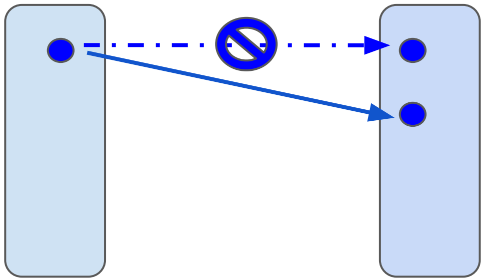


Figure 4: **Fast update for domain bootstrapping:** accuracy on $\mathcal{N}_{\text{dev}}$ and $\mathcal{O}_{\text{dev}}$ (new domain = alarm) when T5 trained on $\mathcal{O}_{\text{train}}$ is fine-tuned on either $\mathcal{N}_{\text{sup}}$ (blue) or $\mathcal{O}_{\text{train}} + \mathcal{N}_{\text{sup}}$ (red) at test time.

Google Research

Setting 2: Parser guiding

2. Parse guiding
for hotfixes



After training, we might find some problematic predictions
"Hotfixes"

How can we change selected predictions without any retraining?

And, make these generalize beyond exact matches?

Technique for Faithfulness: Guiding Tag

Add *tagged* examples with *exact match of labels+hierarchical structure*

Original input

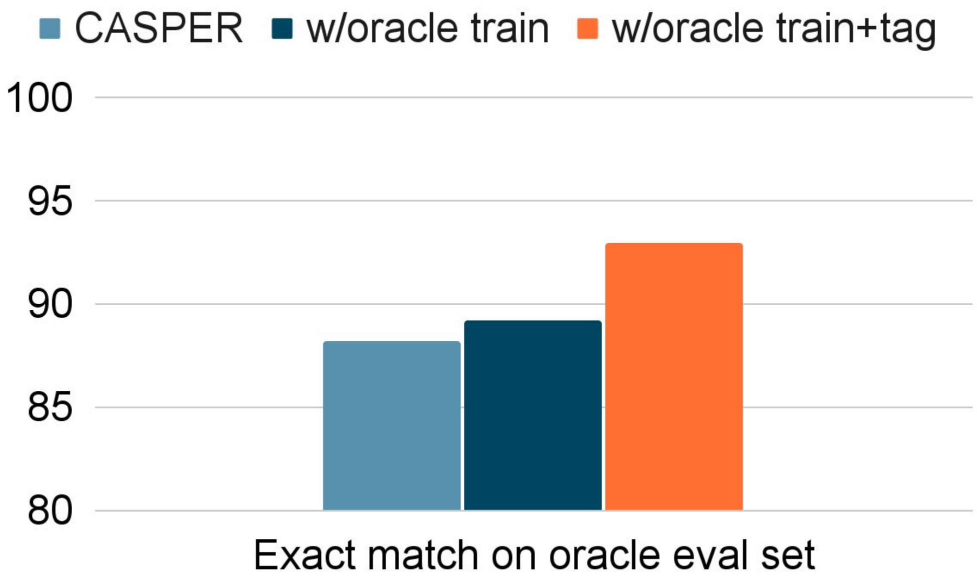
```
Make a call to Jay Brown’s mom
@@ call Zoey’s wife
## [IN create call =
  [SL contact = [IN get contact =
    [SL contact related = Zoey]
    [SL type relation = wife]]]]...
```

Input with guiding tag

```
Make a call to Jay Brown’s mom
@@ PLATINUM call Zoey’s wife
## [IN create call =
  [SL contact = [IN get contact =
    [SL contact related = Zoey]
    [SL type relation = wife]]]]...
```

Oracle experiment

Evaluate with exemplars restricted to have the same semantic template



Google Research

Adversarial examples with the guiding tag

Parser needs to balance faithfulness to the exemplar and parse quality

Gave “adversarial” guiding exemplars with very different gold standards

x: call Nicholas and Natasha **(a)**

x'_2: *PLATINUM* How do you make chicken spaghetti

y'_2: [IN get recipes =
    [SL recipes included ingredient = chicken]
    [SL recipes dish = spaghetti]]

Gold: [IN create call = [SL contact = Nicholas]
    [SL contact = Natasha]]

C_o: [IN get recipes =
    [SL recipes included ingredient = Nicholas]
    [SL recipes included ingredient = Natasha]]

Forced to predict Nicholas and Natasha are ingredients

Google Research

Setting 3: Schema refactoring

After training, the schema might change

Need new predictions for old examples

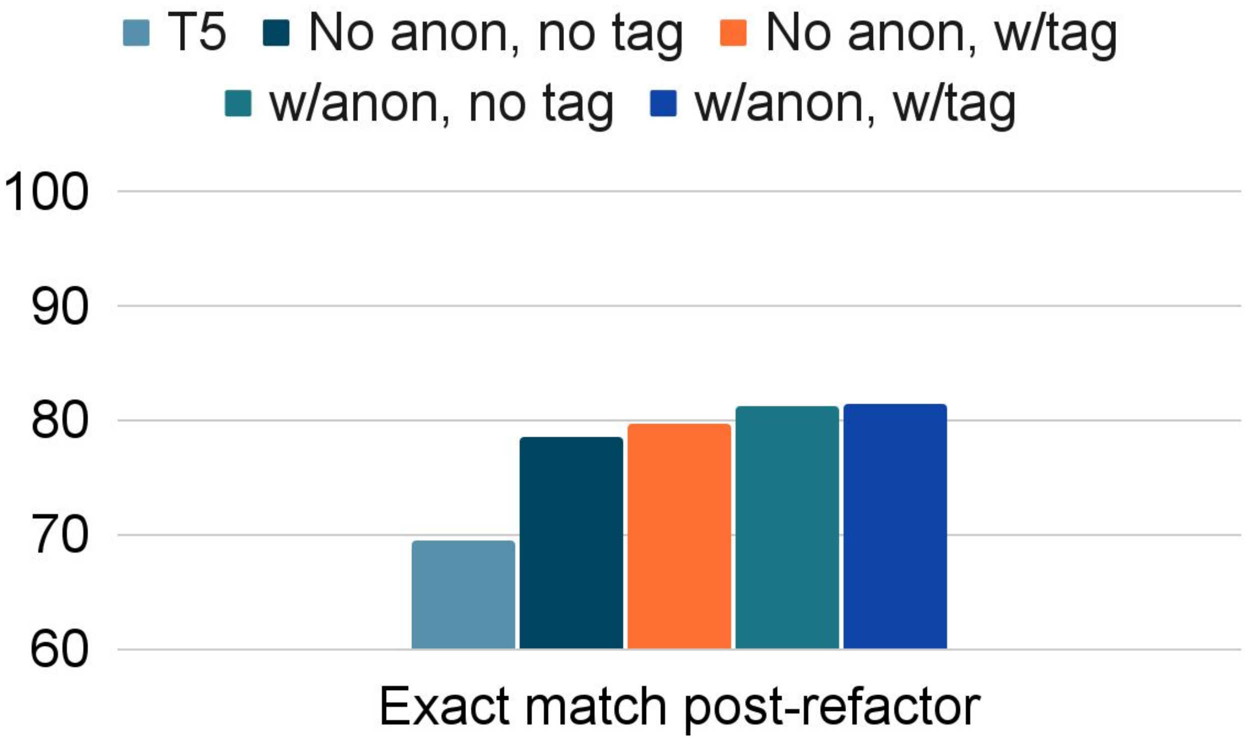
How can we better use our updated index, without retraining?

Simulate with merging labels (“pre-refactor”) then splitting out again for evaluation, following Gaddy et. al (2020)

GET_EVENT → GET_EVENT
GET_EVENT → GET_REMINDER

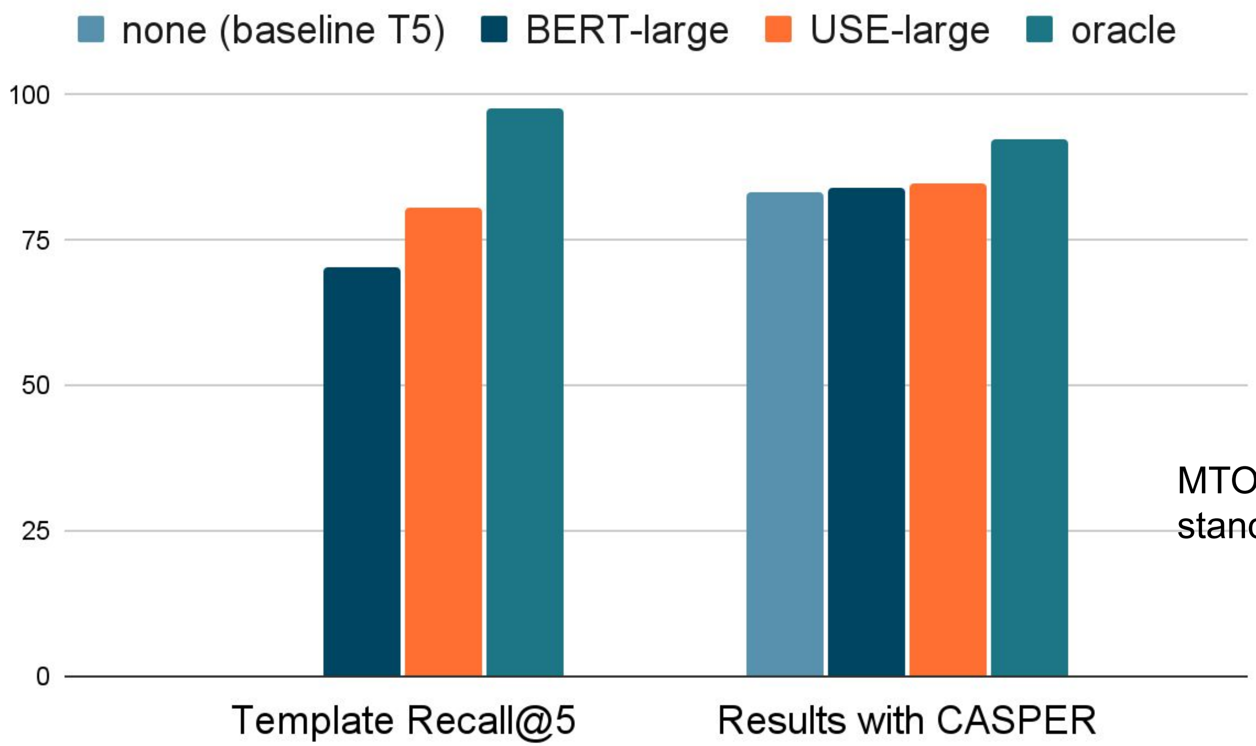
Google Research

Both anonymizing + guiding tags help



Google Research

Some analysis: Still headroom for retriever

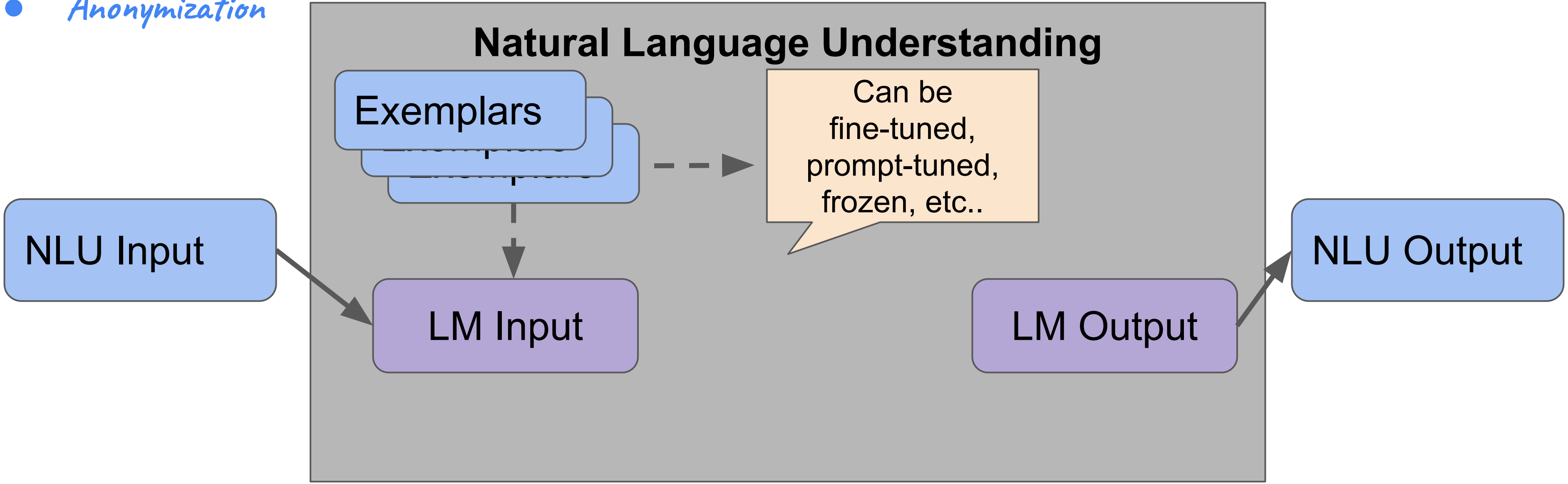


MTOP (Li et al, 2021),
standard development set

Google Research

So far: Focused on the input side. Output side?

- *Retrieval-augmentation*
- *Guiding tag*
- *Anonymization*



Google Research

Some output side interventions

Simplifying the format (removing variables)

Outputting edits (adding variables)

Adding chain-of-thought reasoning

Simplifying the output format

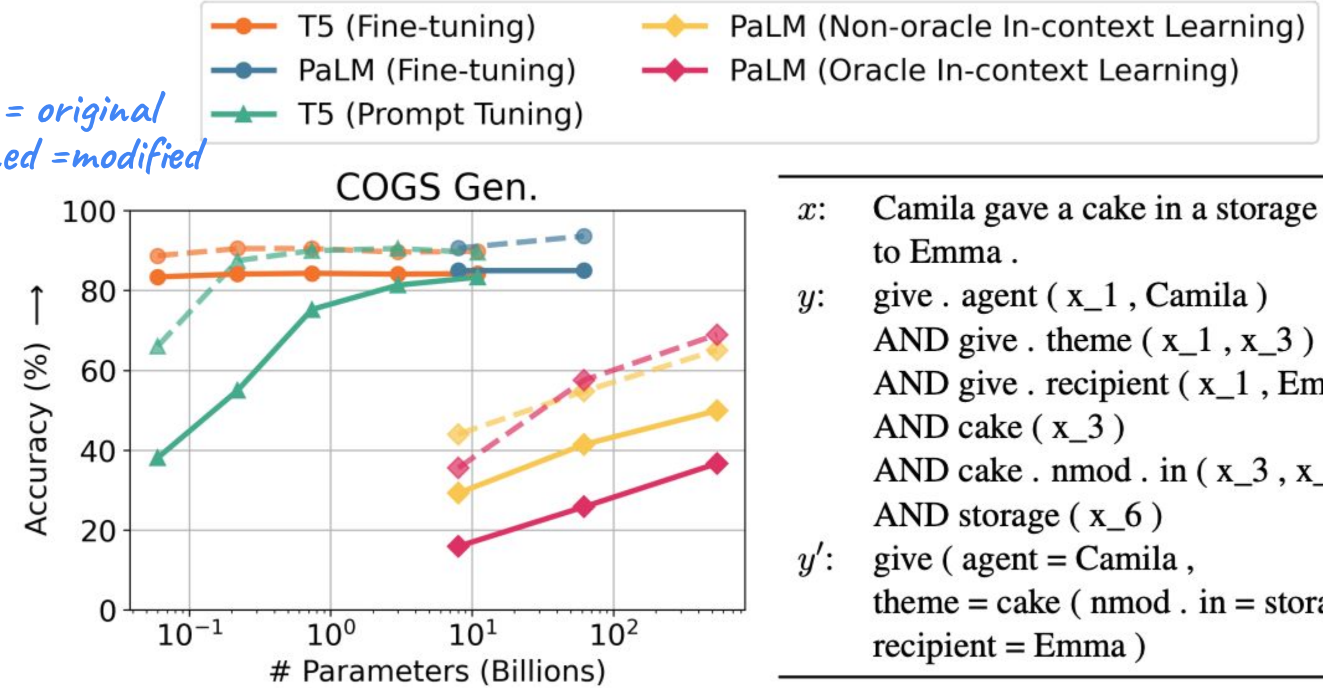
x:	Camila gave a cake in a storage to Emma .
y:	give . agent (x_1 , Camila) AND give . theme (x_1 , x_3) AND give . recipient (x_1 , Emma) AND cake (x_3) AND cake . nmod . in (x_3 , x_6) AND storage (x_6)
y':	give (agent = Camila , theme = cake (nmod . in = storage , recipient = Emma)

Rewriting the COGS: A compositional generalization challenge based on semantic interpretation (Kim and Linzen, 2020) to be variable free

Experiments in Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing, Qiu et al arXiv 2022

Improves in-context, less so for tuned models

- Solid = original
- Dashed =modified



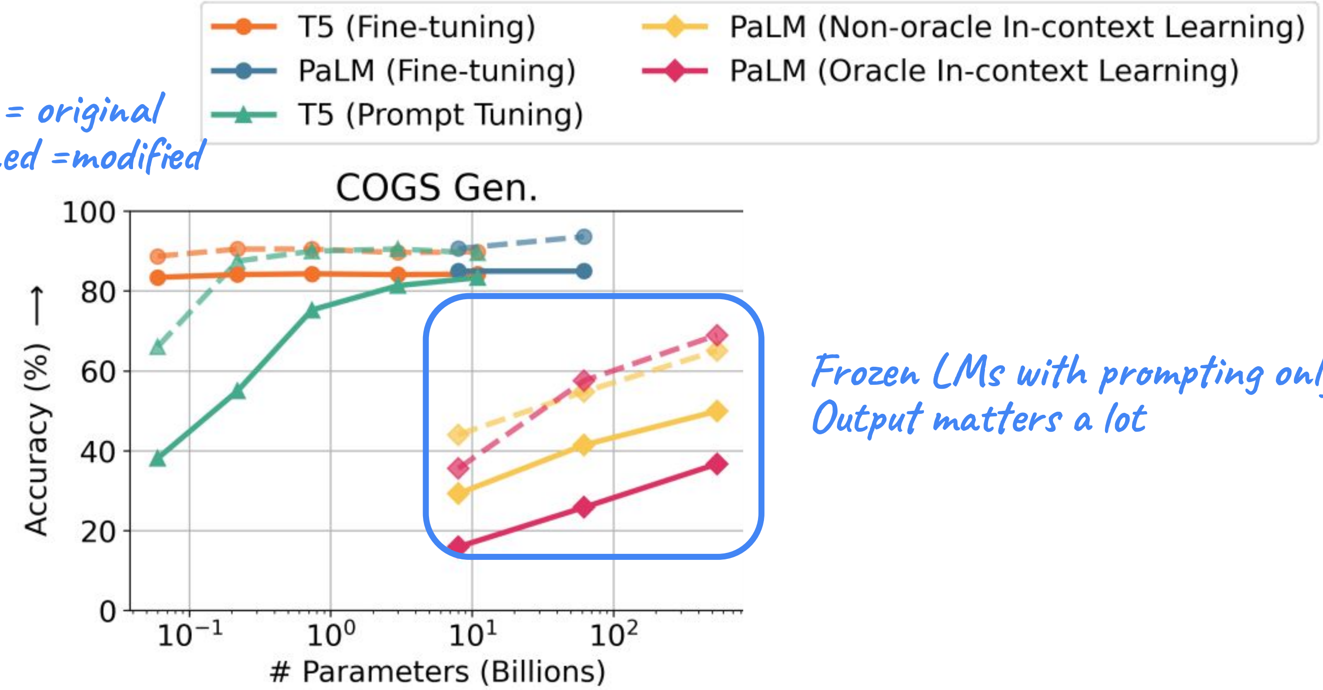
x: Camila gave a cake in a storage to Emma .

y: give . agent (x_1 , Camila)
AND give . theme (x_1 , x_3)
AND give . recipient (x_1 , Emma)
AND cake (x_3)
AND cake . nmod . in (x_3 , x_6)
AND storage (x_6)

y': give (agent = Camila ,
theme = cake (nmod . in = storage ,
recipient = Emma)

Improves in-context, less so for tuned models

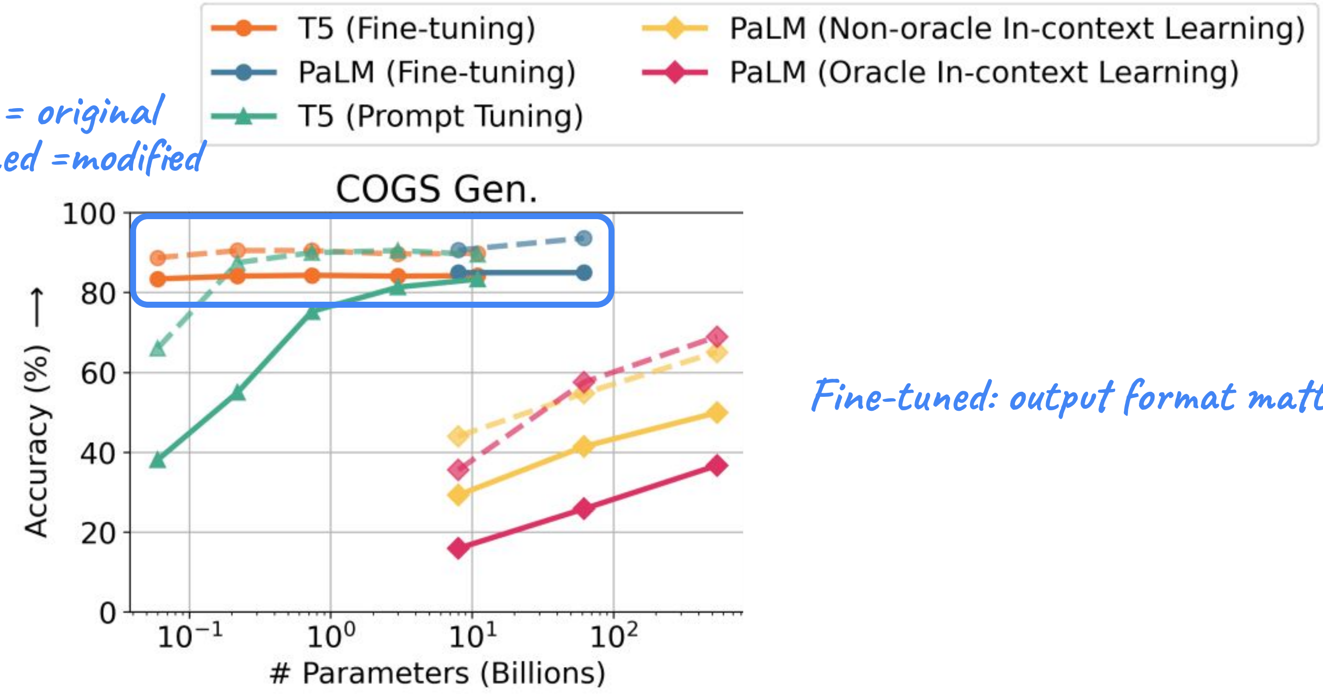
- Solid = original
- Dashed =modified



Frozen LMs with prompting only
Output matters a lot

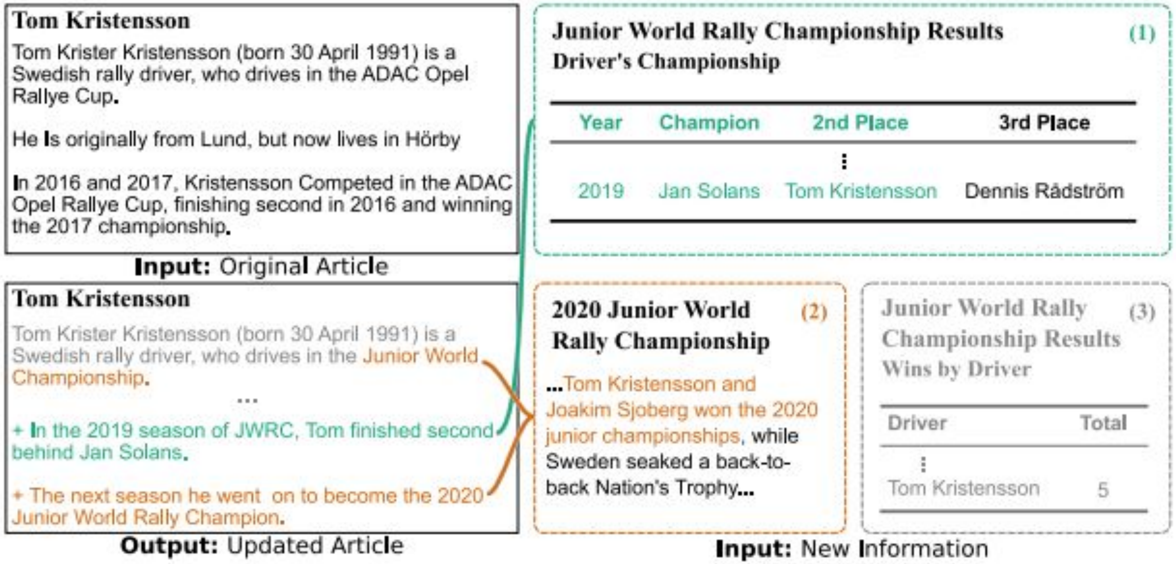
Improves in-context, less so for tuned models

- Solid = original
- Dashed =modified



Fine-tuned: output format matters much less

Task: Update articles with new information



FRUIT: Faithfully Reflecting Updated Information in Text, Logan et al, NAACL 2022 (Best New Task)

Output: new article text, or variables?

```
(2) [0] [1] [2] [3] [4] [5] [6] In the House of Representatives, she holds the
seat that was held by her father from 1979 to 1989.  (6) She is known for her
neoconservative foreign policy views, and her affiliation with the Trump campaign.
(0) (1) (2) (3) (4) Cheney is under fire for her role in the second impeachment of
Donald Trump in January 2021.
```

[Indices] refer to sentences copied.
(Indices) refer to references for supporting evidence

Figure A3: **EDIT5 Output Format.**

```
(2) Elizabeth Lynne Cheney ( ; born July 28, 1966) is an American attorney and
politician who has served as the U.S. Representative for since 2017.  She was
the Chair of the House Republican Conference, the third-highest position in the
House Republican leadership.  She is the third woman elected to that position
after Deborah Pryce and Cathy McMorris Rodgers.  She held several positions in
the U.S. State Department during the George W. Bush administration, notably as
Deputy Assistant Secretary of State for Near Eastern Affairs and Coordinator for
Broader Middle East and North Africa Initiatives.  She promoted regime change in
Iran while chairing the Iran Syria Policy and Operations Group with Elliott Abrams.
In 2009 Cheney and Bill Kristol founded Keep America Safe, a nonprofit organization
concerned with national security issues that advocated the positions of the former
Bush administration.  She was a candidate for the 2014 election to the U.S. Senate
in Wyoming, challenging three-term incumbent Mike Enzi, before withdrawing from the
race.  In the House of Representatives, she holds the seat her father held for a
decade, representing Wyoming from 1979 to 1989.  Cheney is a neoconservative.  She
later supported the second impeachment of Donald Trump for his role in the 2021
storming of the U.S. Capitol.
```

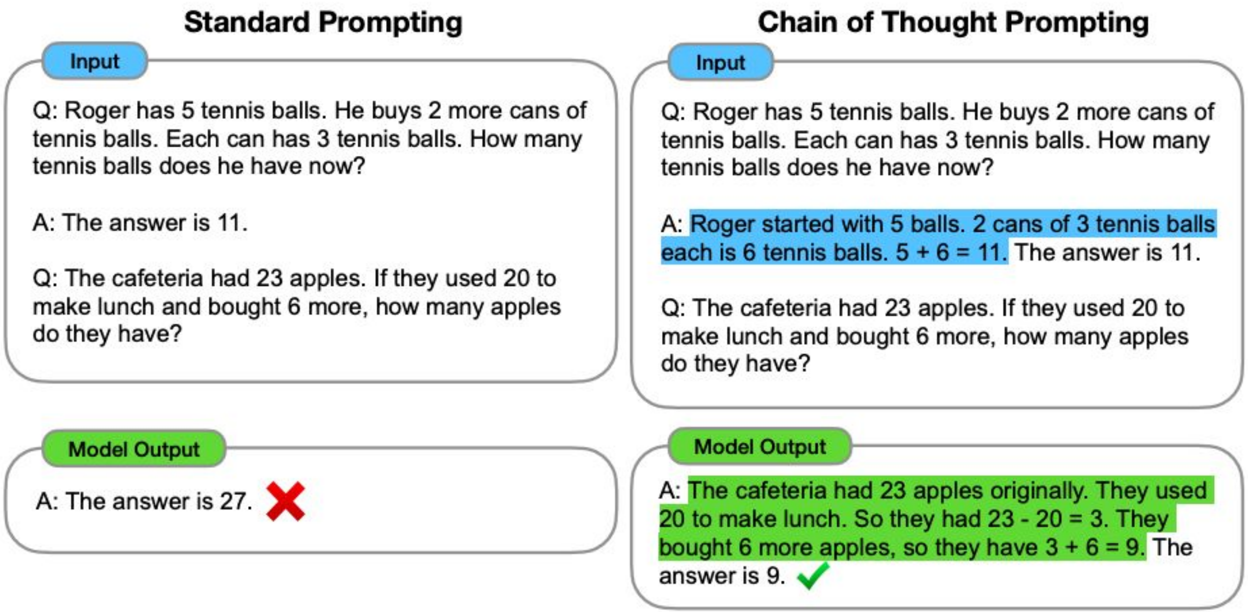
Figure A2: **T5 Output Format.**

Google Research

Output as edits improves over T5

	UpdateROUGE			Entity		Unsup.
	1	2	L	Prec.	Recall	Tokens
Copy Source	0.0	0.0	0.0	0.0	0.0	0.00
+ All Evidence	18.8	6.9	12.0	37.9	64.9*	0.00
T5-Large	31.1	18.4	24.4	52.7	44.9	2.67
+ Evidence Input	44.3	29.4	36.8	62.2	50.7	2.34
EDIT5-Small	41.2	27.3	35.3	62.4	44.9	1.71
EDIT5-Base	47.0	32.1	39.7	62.2	**54.9**	2.28
EDIT5-Large	46.3	32.4	39.6	67.2	53.1	**1.54**
EDIT5-3B	**47.4**	**34.0**	**41.1**	**69.9**	52.5	1.58

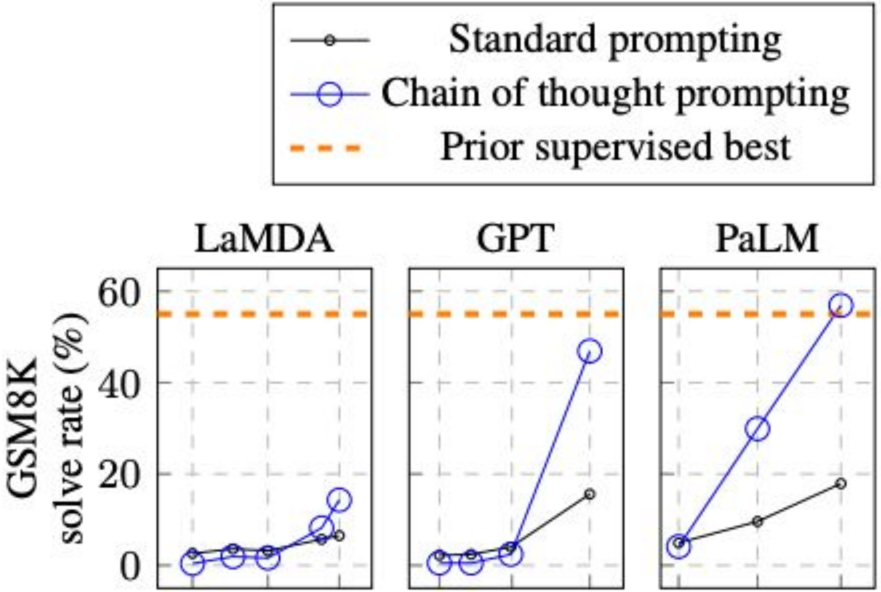
Chain-of-thought-prompting



Google Research

Chain of Thought Prompting Elicits Reasoning in Large Language Models, Wei et al, arXiv 2022

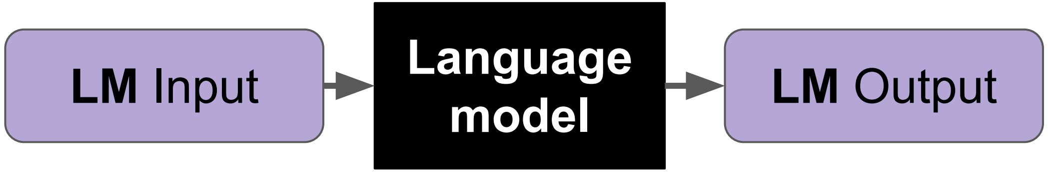
Gives large gains when prompting, math Qs



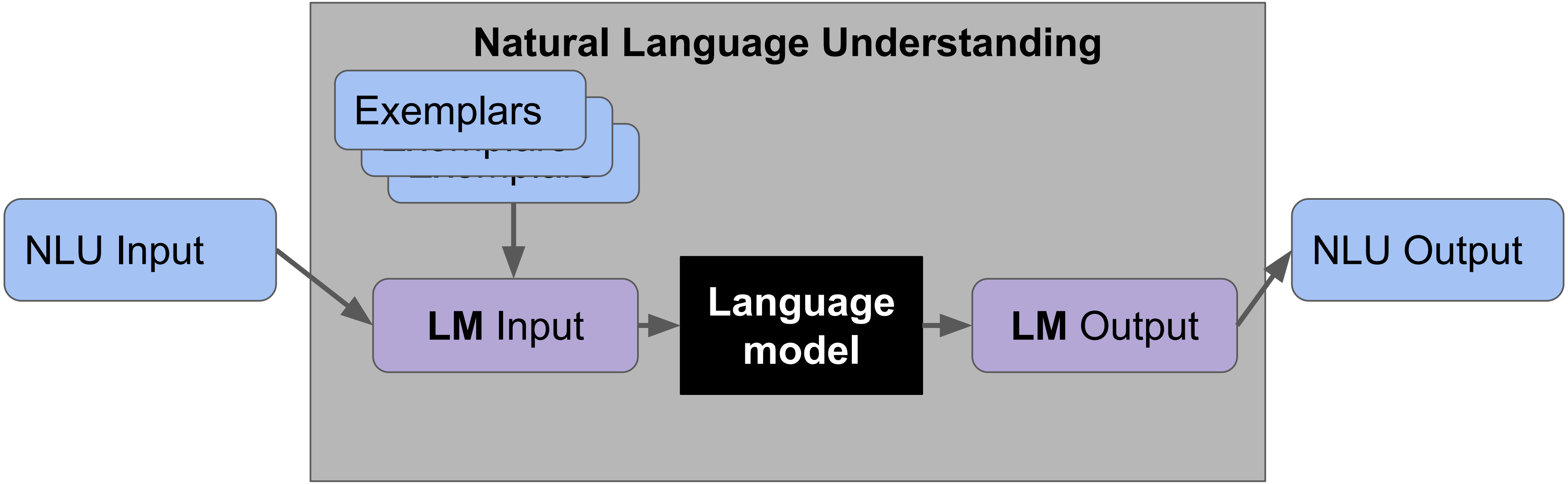
Google Research

Beyond natural language understanding

Language models are increasingly good

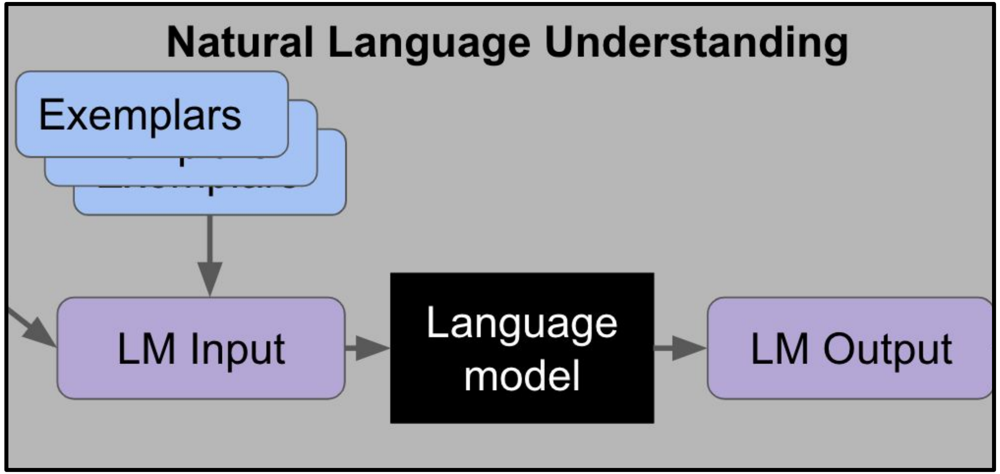


NLU based around *language models*



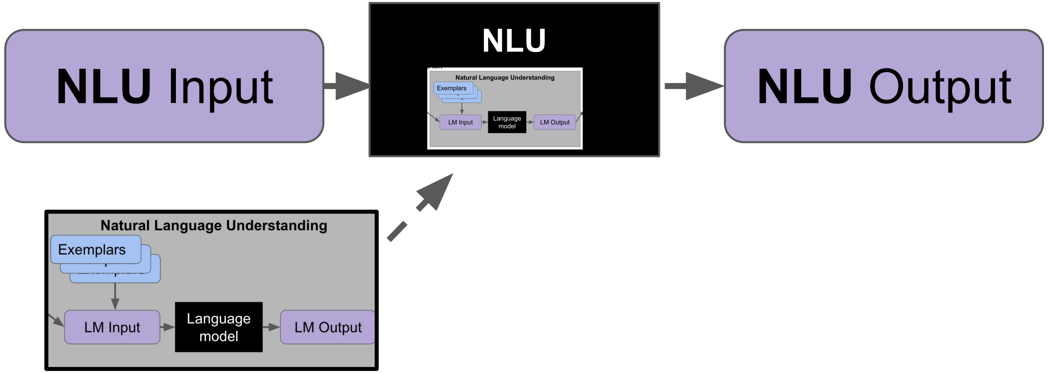
Google Research

What can we now base around good *NLU*?

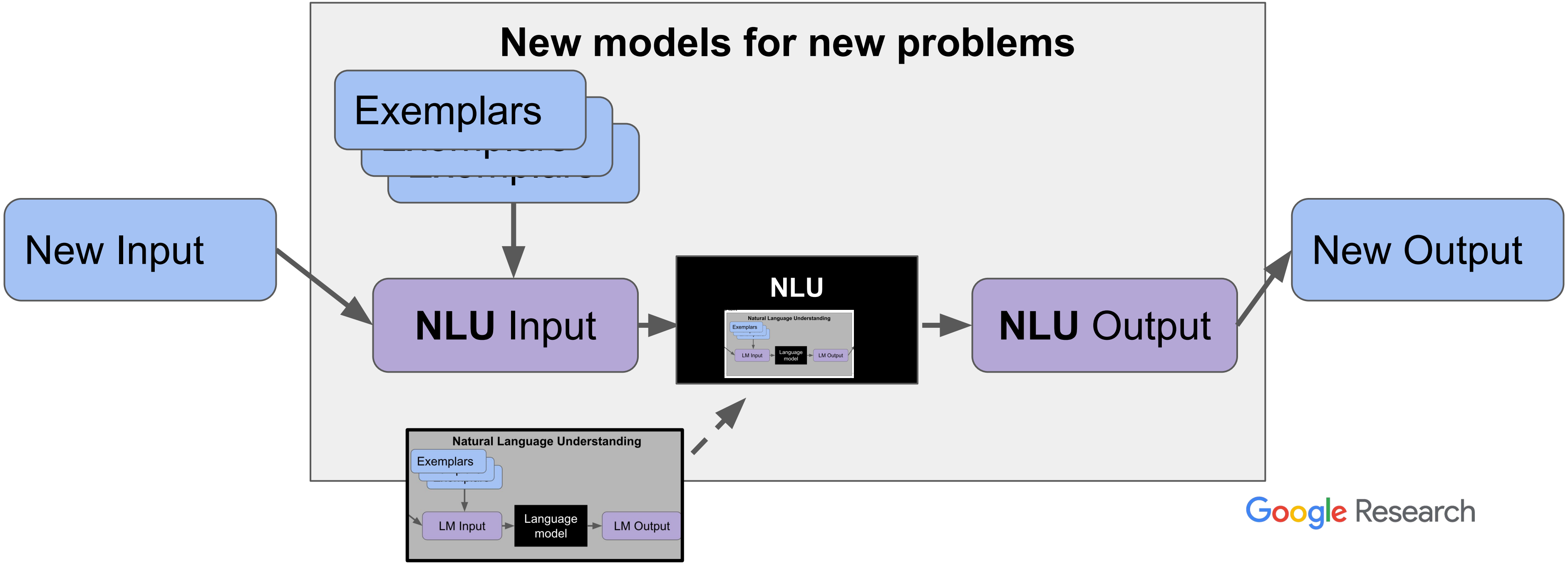


Google Research

What can we now base around good *NLU*?



What can we now base around good *NLU*?



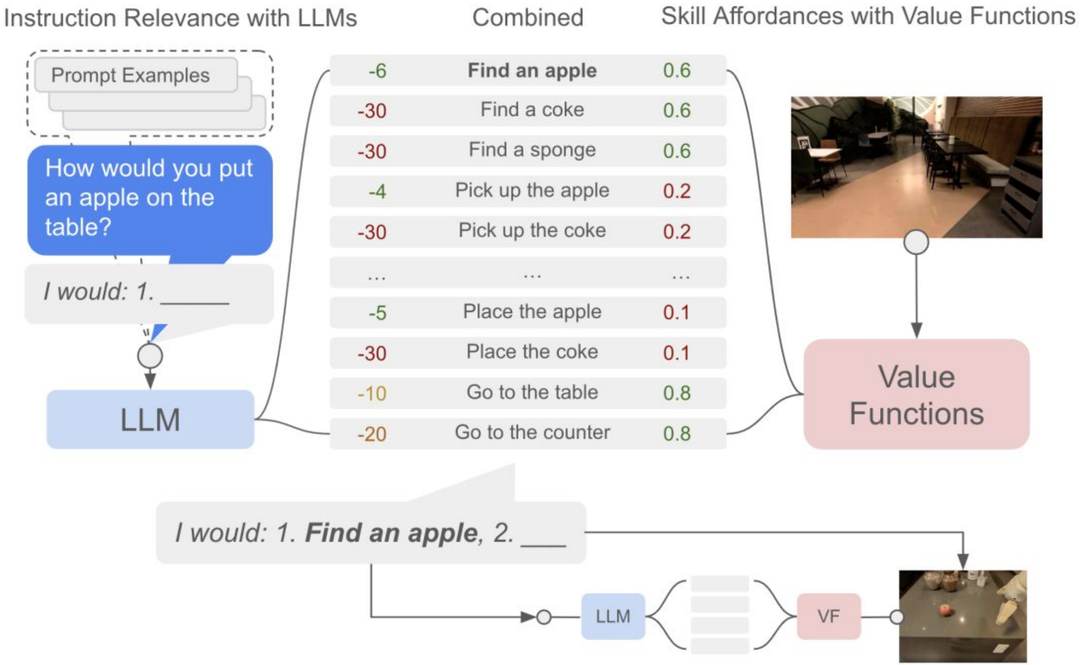
Google Research

One promising example: robotics

(a) “I just worked out, can you bring me a drink and a snack to recover?”

[Do As I Can, Not As I Say: Grounding Language in Robotic Affordances]
Ahn et al, arXiv 2022

One promising example: robotics



[Do As I Can, Not As I Say: Grounding Language in Robotic Affordances]()
Ahn et al, arXiv 2022

Conclusion

Complementary *challenges* beyond train-test accuracy for LM-based NLU

Complementary *components* for LM-based NLU

Impressive language models → advances in language understanding

↓

Impressive language understanding → advances in ______(?)

Google Research

Thank You

Emily Pitler

Google Research