Improving model controllability with input and output alterations
Language models have some impressive capabilities
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".

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*
Even hard NLU benchmarks got saturated


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

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

Figure 3 from PaLM
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

NLU Input -> Exemplars -> Language model

Language model output: Can be fine-tuned, prompt-tuned, frozen, etc.

NLU Output
Focus: Non-LM aspects of a NLU recipe

What’s left?

Natural Language Understanding

Exemplars

Can be fine-tuned, prompt-tuned, frozen, etc..

NLU Input

LM Input

LM Output

NLU Output
Controllable Semantic Parsing via
Retrieval Augmentation

Panupong Pasupat, Yuan Zhang,
Kelvin Guu
EMNLP 2021
Controllability scenarios in semantic parsing

1. New domain with new label space
2. Parse guiding for hotfixes
3. Schema label renaming

...without retraining the parameters of the encoder-decoder model
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

1. New domain with new label space

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

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]]]
@@ Make a call ... ## [IN create call ... 
@@ Get number ... ## [IN get number ...

[IN create call =
  [SL contact = [IN get contact =
    [SL contact related = Jay Brown]
    [SL type relation = mom]]]
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:

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

Anonymized:

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

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 $N_{dev}$ and $O_{dev}$ (new domain = alarm) when T5 trained on $O_{train}$ is fine-tuned on either $N_{sup}$ (blue) or $O_{train} + N_{sup}$ (red) at test time.
Setting 2: Parser 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
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

\[
x: \text{call Nicholas and Natasha} \\
x'_{2}: \text{PLATINUM How do you make chicken spaghetti} \\
y'_{2}: \text{[IN get recipes =} \\
\text{[SL recipes included ingredient = chicken]} \\
\text{[SL recipes dish = spaghetti]}\]

Gold: \text{[IN create call = [SL contact = Nicholas]} \\
\text{[SL contact = Natasha}]

C_{0}: \text{[IN get recipes =} \\
\text{[SL recipes included ingredient = Nicholas]} \\
\text{[SL recipes included ingredient = Natasha]}\]
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_REMINDER
Both anonymizing + guiding tags help
Some analysis: Still headroom for retriever

MTOP (Li et al, 2021), standard development set
So far: Focused on the input side. Output side?

- Retrieval-augmentation
- Guiding tag
- Anonymization

**Natural Language Understanding**

- Exemplars
- Can be fine-tuned, prompt-tuned, frozen, etc..

**Input**
- NLU Input
- LM Input

**Output**
- NLU Output
- LM Output
Some output side interventions

Simplifying the format (removing variables)

Outputting edits (adding variables)

Adding chain-of-thought reasoning
Simplifying the output format

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

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

\[ x: \text{ Camila gave a cake in a storage to Emma.} \]
\[ y: \text{ give . agent ( } x_1 , \text{ Camila )}
\]  
\[ \text{ AND give . theme ( } x_1 , x_3 )
\]  
\[ \text{ AND give . recipient ( } x_1 , \text{ Emma )}
\]  
\[ \text{ AND cake ( } x_3 )
\]  
\[ \text{ AND cake . nmod . in ( } x_3 , x_6 )
\]  
\[ \text{ AND storage ( } x_6 )
\]
\[ y': \text{ give ( agent = Camila ,}
\]  
\[ \text{ theme = cake ( nmod . in = storage ,}
\]  
\[ \text{ recipient = Emma )}
\]
Improves in-context, less so for tuned models

- **Solid** = original
- **Dashed** = modified

Frozen LMs with prompting only

Output matters a lot

Research
Improves in-context, less so for tuned models

- **Solid = original**
- **Dashed = modified**

---

**Fine-tuned: output format matters much less**

- T5 (Fine-tuning)
- PaLM (Non-oracle In-context Learning)
- PaLM (Fine-tuning)
- PaLM (Oracle In-context Learning)
- T5 (Prompt Tuning)
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.

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.
Output as edits improves over T5

<table>
<thead>
<tr>
<th></th>
<th>UpdateROUGE</th>
<th>Entity</th>
<th>Unsup. Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>L</td>
</tr>
<tr>
<td>Copy Source + All Evidence</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>T5-Large + Evidence Input</td>
<td>31.1</td>
<td>18.4</td>
<td>24.4</td>
</tr>
<tr>
<td>EDIT5-Small</td>
<td>41.2</td>
<td>27.3</td>
<td>35.3</td>
</tr>
<tr>
<td>EDIT5-Base</td>
<td>47.0</td>
<td>32.1</td>
<td>39.7</td>
</tr>
<tr>
<td>EDIT5-Large</td>
<td>46.3</td>
<td>32.4</td>
<td>39.6</td>
</tr>
<tr>
<td>EDIT5-3B</td>
<td><strong>47.4</strong></td>
<td><strong>34.0</strong></td>
<td><strong>41.1</strong></td>
</tr>
</tbody>
</table>
Chain-of-thought-prompting

**Standard Prompting**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

**Chain of Thought Prompting**

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. \(5 + 6 = 11\). The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓
Gives large gains when prompting, math Qs
Beyond natural language understanding
Language models are increasingly good

LM Input → Language model → LM Output
NLU based around *language models*
What can we now base around good NLU?
What can we now base around good NLU?
What can we now base around good NLU?

New models for new problems

New Input → NLU Input → NLU → NLU Output → New Output

Exemplars

Natural Language Understanding

Exemplars

LM Input

Language model

LM Output

Google Research
One promising example: robotics

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

(a) “I just worked out, can you bring me a drink and a snack to recover?”
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 $\rightarrow$ advances in language understanding

$\downarrow$

Impressive language understanding $\rightarrow$ advances in _____(?)
Thank You

Emily Pitler