Collaborative Creative Writing

Vishakh Padmakumar

NYU

Machine Learning for Language

Acknowledgements

Coauthors: He He, Tuhin Chakrabarty

Feedback and Discussions: Richard Pang, Nitish Joshi, Alicia Parrish, Angelica Chen, Nicholas Lourie, Chen Zhao, Jason Phang and the rest of the ML^2 group :)

VI What are the important and timely research directions in NLP?

VI

What are the important and timely research directions in NLP?

There are several important and timely research directions in natural language processing (NLP), some of which include:

1. Pre-training and fine-tuning: Developing techniques for pre-training large language models on large amounts of data, and then fine-tuning them on smaller, task-specific datasets. This has been very successful in a number of NLP tasks, and researchers are continuing to work on improving the performance and efficiency of these approaches.
2. Transfer learning: Developing techniques that allow NLP models to transfer knowledge from one language or domain to another, which can be particularly useful in situations where there is limited data available in a particular language or domain.
3. Dialogue systems: Developing NLP models that can effectively handle complex, open-ended conversations, such as those that might be used in virtual assistants or chatbots. This involves developing models that can understand and generate natural language, as well as models that can reason about and respond appropriately to different types of input.
4. Multilingual and cross-lingual NLP: Developing NLP models that can handle multiple languages and translate between languages, which is important for applications such as machine translation and information retrieval in a global context.
5. Sentiment analysis: Developing NLP models that can accurately identify and classify the sentiment (e.g., positive, negative, or neutral) of text, which has applications in a

VI

What are the important and timely research directions in NLP?

There are several important and timely research directions in natural language processing (NLP), some of which include:

1. Pre-training and fine-tuning: Developing techniques for pre-training large language models on large amounts of data, and then fine-tuning them on smaller, task-specific datasets. This has been very successful in a number of NLP tasks, and researchers continuing to work on improving the performance and efficiency of these approach
2. Transfer learning: Developing techniques that allow NLP models to transfer knowled from one language or domain to another, which can be particularly useful in situations where there is limited data available in a particular language or domain.
3. Dialogue systems: Developing NLP models that can effectively handle complex, open-ended conversations, such as those that might be used in virtual assistants or chatbots. This involves developing models that can understand and generate natural language, as well as models that can reason about and respond appropriately to different types of input.
4. Multilingual and cross-lingual NLP: Developing NLP models that can handle multiple languages and translate between languages, which is important for applications such as machine translation and information retrieval in a global context.
5. Sentiment analysis: Developing NLP models that can accurately identify and classify the sentiment (e.g., positive, negative, or neutral) of text, which has applications in a

Maybe a little too generic

VI

Can you write a paragraph about why we should work on human-AI collaboration

VI

Can you write a paragraph about why we should work on human-AI collaboration

There are several reasons why it is important to work on human-AI collaboration.

First, as AI continues to advance and become more prevalent in a wide range of industries, it is important to find ways to effectively integrate AI into human workflows and decision-making processes. By working on human-AI collaboration, we can help to ensure that AI is used in a way that is beneficial to both humans and AI, and that it is aligned with ethical and societal norms.

Second, human-AI collaboration has the potential to bring many benefits, including increased efficiency, productivity, and effectiveness. By combining the strengths of both humans and AI, it is possible to achieve outcomes that would not be possible with either humans or AI alone.

Third, as AI becomes more advanced and capable of handling more complex tasks, it is important to consider the potential negative impacts of AI, such as the displacement of human workers or the potential for AI to make biased or unfair decisions. By involving humans in the decision-making process, it is possible to incorporate human judgment and values into AI systems, helping to mitigate these negative impacts.

VI

Can you write a paragraph about why we should work on human-AI collaboration

There are several reasons why it is important to work on human-AI collaboration.

First, as AI continues to advance and become more prevalent in a wide range of industries, it is important to find ways to effectively integrate AI into human workflows and decision-making processes. By working on human-AI collaboration, we can help to ensure that AI is used in a way that is beneficial to both humans and AI, and that it is aligned with ethical and societal norms.

Second, human-AI collaboration has the potential to bring many benefits, including increased efficiency, productivity, and effectiveness. By combining the strengths of both humans and AI, it is possible to achieve outcomes that would not be possible with either humans or AI alone.

Third, as AI becomes more advanced and capable of handling more complex tasks, it is important to consider the potential negative impacts of AI, such as the displacement of human workers or the potential for AI to make biased or unfair decisions. By involving humans in the decision-making process, it is possible to incorporate human judgment and values into AI systems, helping to mitigate these negative impacts.

Overview

- Broad Direction: How can we assist writers at creative writing tasks?

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Content generation is exploding and there is a lot of demand for high-quality prose

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Content generation is exploding and there is a lot of demand for high-quality prose
- LLMs generate fluent text and can be finetuned for various purposes

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Content generation is exploding and there is a lot of demand for high-quality prose
- LLMs generate fluent text and can be finetuned for various purposes
- Opportune time to work on writing assistants[1,2,3]
    - Verse-by-Verse, HemingwayApp, Adept

[1] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. CoRR. abs/2107.07430
[2] Lee, Mina, Percy Liang, and Qian Yang. "CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities." arXiv preprint arXiv:2201.06796 (2022).
[3] Du, Wanyu, et al. "Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision." arXiv preprint arXiv:2204.03685 (2022).

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Many open questions **bridging NLP and HCI**:
    - How do we design the ***most effective collaboration setup*** to help human users?
    - How do we train models to generate ***helpful*** suggestions?
    - What is the ***best form of user feedback*** and how do we ***incorporate it in model training***?
    - How do we assist users in ***content planning*** for long form creative writing?
    - How do we ensure ***equitable creative writing assistance*** to all users?

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Many open questions **bridging NLP and HCI**:
    - How do we design the ***most effective collaboration setup*** to help human users?
    - How do we train models to generate ***helpful*** suggestions?
    - What is the ***best form of user feedback*** and how do we ***incorporate it in model training***?
    - How do we assist users in ***content planning*** for long form creative writing?
    - How do we ensure ***equitable creative writing assistance*** to all users?
- Our work:
    - Machine-in-the-Loop Rewriting for Creative Image Captioning
    - Collaborative Poetry writing with Instruction Tuning

Overview

- Broad Direction: How can we assist writers at creative writing tasks?
- Many open questions **bridging NLP and HCI**:
  - How do we design the ***most effective collaboration setup*** to help human users?
  - How do we train models to generate ***helpful*** suggestions?
  - What is the ***best form of user feedback*** and how do we ***incorporate it in model training***?
  - How do we assist users in ***content planning*** for long form creative writing?
  - How do we ensure ***equitable creative writing assistance*** to all users?
- Our Work:
  - **Machine-in-the-Loop Rewriting for Creative Image Captioning**
  - Collaborative Poetry writing with Instruction Tuning

Machine-in-the-Loop Rewriting for Creative Image Captioning

Vishakh Padmakumar, He He

NAACL 2022

NYU ML² Machine Learning for Language

Machine-In-The-Loop Creative Writing

- Creative writing tasks can be challenging for both humans and machines.

Machine-In-The-Loop Creative Writing

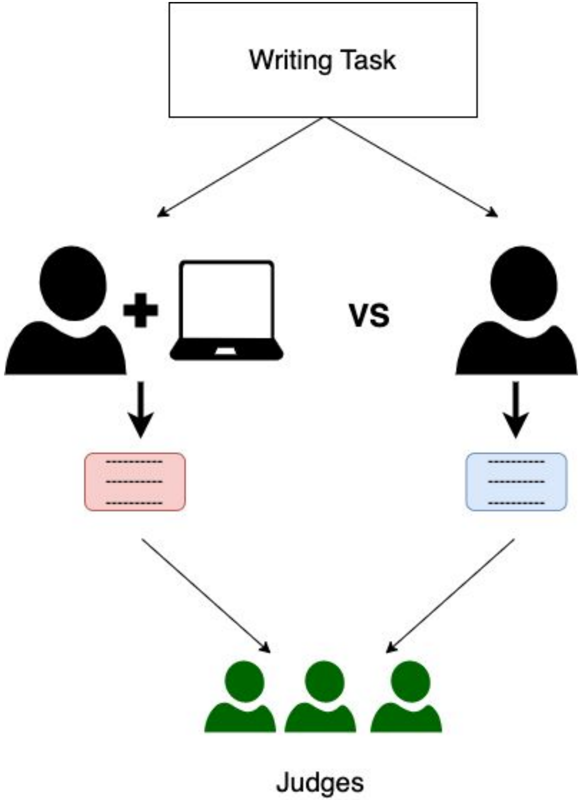
- Creative writing tasks can be challenging for both humans and machines.
    - Humans would benefit from suggestions on wording and framing their ideas[1]
    - Models are able to rewrite spans of text[4] but struggle with global coherence[2,3]

[1] Monica J Garfield. 2008. Creativity support systems. In *Handbook on Deci*sion Support Systems 2, pages 745–758. Springer
[2] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In 23rd International Conference on Intelligent User Interfaces, pages 329–340.
[3] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484
[4] Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In Association for Computational Linguistics (ACL)

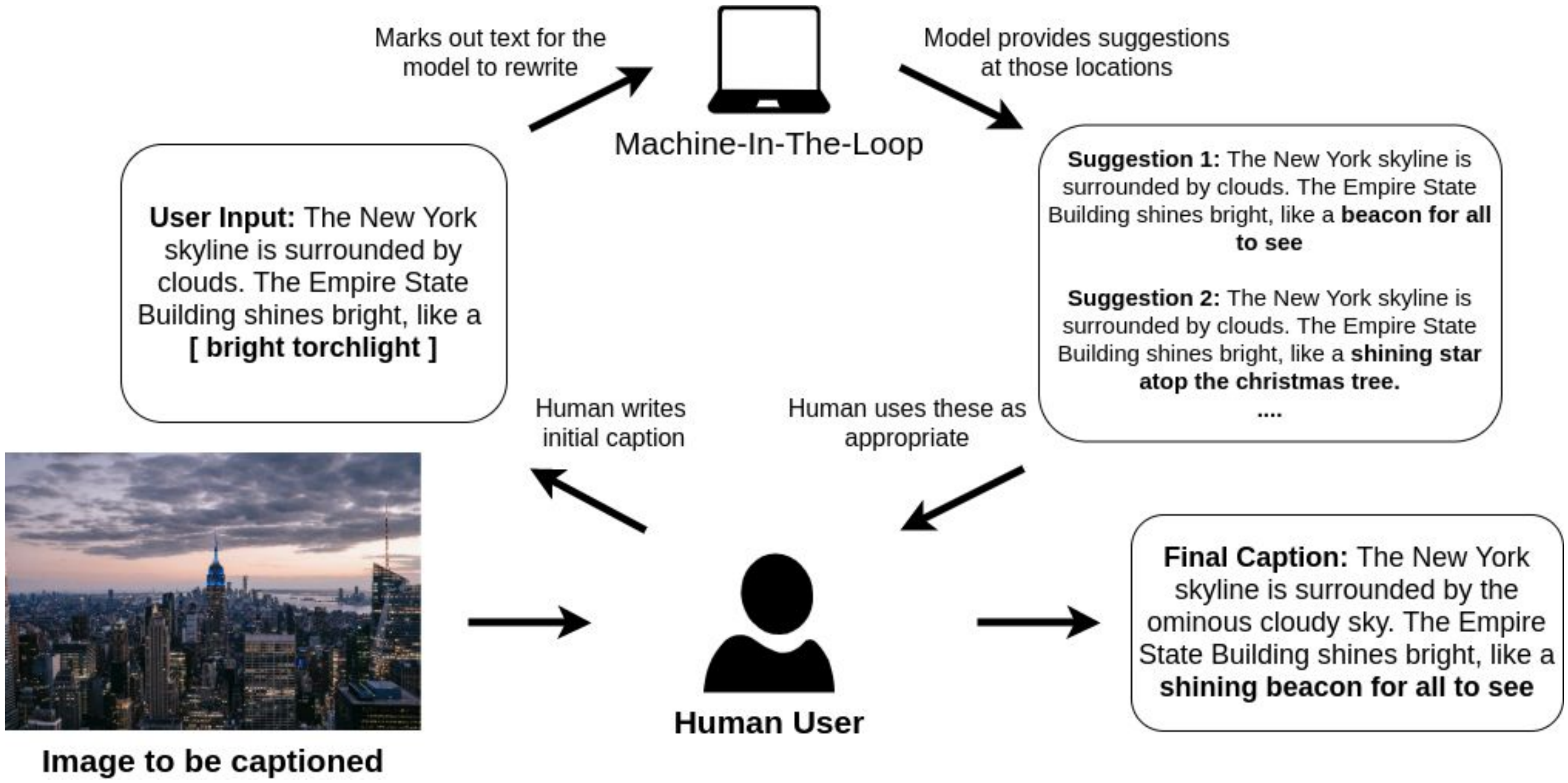
Machine-In-The-Loop Creative Writing

- Creative writing tasks can be challenging for both humans and machines.
    - Humans would benefit from suggestions on wording and framing their ideas
    - Models are able to rewrite spans of text but struggle with global coherence
- Motivates a cooperative setting: Can a model help the author improve their creative output?

Machine-In-The-Loop Creative Writing



Task Setup - Creative Image Captioning



Training the Creative Rewriting Assistant (CRA) Model

- Fine-tuning Data: We create a pseudo-parallel corpus of creative sentences (annotated for literary devices) and corresponding generic sentences
- **CRA** is a fine-tuned BART-Large model

Demo

Project Roadmap

- Do users find CRA model suggestions helpful?
- Are users more effective at creative image captioning with model help?
- How does collaboration with the model impact different users?
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Project Roadmap

- **Do users find CRA model suggestions helpful?**
- Are users more effective at creative image captioning with model help?
- How does collaboration with the model impact different users?
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Do Users Find Model Suggestions Helpful?

We compare the **CRA model** to a **baseline BART model** with an A/B user study (n=50)

Do Users Find Model Suggestions Helpful?

We compare the **CRA model** to a **baseline BART model** with an A/B user study (n=50)

- On average, users find the **CRA model** to be more helpful than **BART** by a statistically significant margin

Results from Post Completion Survey

	BART	CRA
Model Helpfulness	2.23	**3.06**

Do Users Find Model Suggestions Helpful?

We compare the **CRA model** to a **baseline BART model** with an A/B user study (n=50)

- On average, users find the **CRA model** to be more helpful than **BART** by a statistically significant margin

- Users accept larger fraction of suggestions from the **CRA model**

Results from Post Completion Survey

	BART	CRA
Model Helpfulness	2.23	**3.06**

Model Comparison via Interaction Logs

	BART	CRA
Avg # Requests	**3.02**	2.82
% Acceptance	24.5%	**31.9%**
Rouge-L Retention	0.744	**0.824**

Project Roadmap

- **Do users find CRA model suggestions helpful?**
- Are users more effective at creative image captioning with model help?
- How does collaboration with the model impact different users?
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Are users more effective at creative image captioning with model help?**
- How does collaboration with the model impact different users?
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Are Users More Effective With Model Help?

- We collected captions for 100 images from **solo writers**, from users collaborating with the **CRA model** and the **BART model**.

Are Users More Effective With Model Help?

- We collected captions for 100 images from **solo writers**, from users collaborating with the **CRA model** and the **BART model**.
- For each caption pair, we collect 3 annotations for which is better and take a majority vote

Are Users More Effective With Model Help?

- We collected captions for 100 images from **solo writers**, from users collaborating with the **CRA model** and the **BART model**.

	# Majority Vote Wins		
Human + CRA	**57**	43	**Solo Writers**
Human + CRA	**54**	46	**Human + BART**
Human + BART	**55**	45	**Solo Writers**

Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Are users more effective at creative image captioning with model help?**
- How does collaboration with the model impact different users?
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Collaborative users are more effective** at the creative writing task
- **How does collaboration with the model impact different users?**
- Can the model be adapted to learn from observed user interactions to provide better assistance?

How Does **CRA** Impact Different Users?

Effect of User Skill Level: We divide users into two groups, novice and skilled, based on their self-rated writing skill.

How Does CRA Impact Different Users?

Effect of User Skill Level: We divide users into two groups, novice and skilled, based on their self-rated writing skill.

Takeaway: Skilled users find the CRA model to be significantly more helpful

Results from Post Completion Survey

	Novice (n=22)	Skilled (n=28)
Helpfulness	2.27	**3.23**
# Requests	**3.04**	2.64
% Accepted	29.8%	**33.7%**

When is the Model Effective?

Profile of Suggestion:

- The model performs best when rewriting shorter spans of larger texts

Figurative Rewriting - Accepted Suggestion

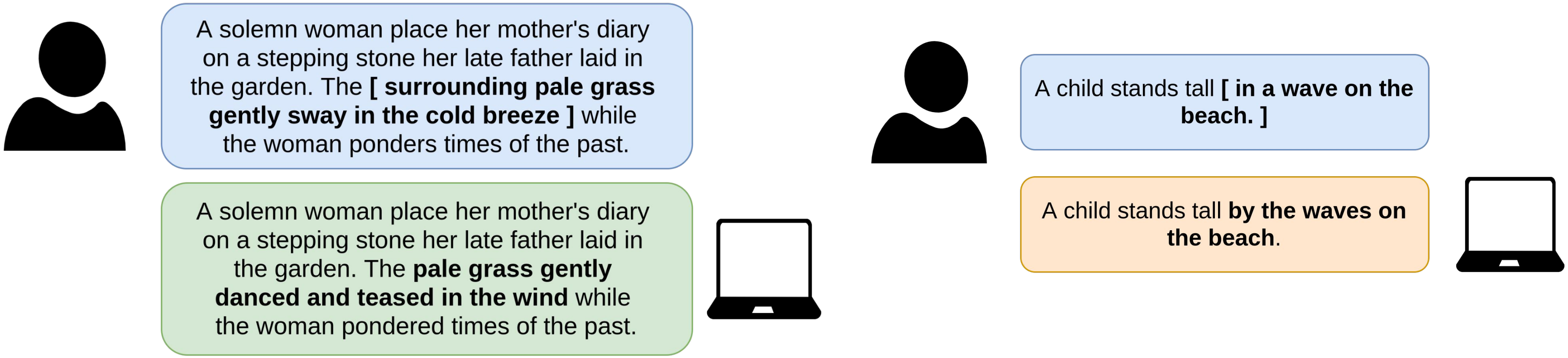
A solemn woman place her mother's diary on a stepping stone her late father laid in the garden. The **[surrounding pale grass gently sway in the cold breeze]** while the woman ponders times of the past.

A solemn woman place her mother's diary on a stepping stone her late father laid in the garden. The **pale grass gently danced and teased in the wind** while the woman pondered times of the past.

When is the Model Effective?

Profile of Suggestion:

- The model performs best when rewriting shorter spans of larger texts
- Skilled writers tend to request this profile of suggestion



Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Collaborative users are more effective** at the creative writing task
- **How does collaboration with the model impact different users?**
- Can the model be adapted to learn from observed user interactions to provide better assistance?

Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Collaborative users are more effective** at the creative writing task
- **Model helps skilled writers more** potentially widening the gap in performance
- **Can the model be adapted to learn from observed user interactions to provide better assistance?**

Can We Learn from User Feedback?

- We create a dataset from 50 sets of observed interactions.
- Sentence Pairs:
    - Original Text ↦ Accepted Suggestion
    - Rejected Suggestion ↦ Original Text
- Fine-tune the **initial CRA model** to **User-adapted CRA Model**

Can We Learn from User Feedback?

- Fine-tune the **initial CRA model** to **User-adapted CRA Model**

- Compare the two models with an A/B user study (n=50)

Can We Learn from User Feedback?

- Fine-tune the **initial CRA model** to **User-adapted CRA Model**
- Compare the two models with an A/B user study (n=50)
- On average, users find the **User-adapted CRA model** to be **more helpful** than **CRA model**, but not by a statistically significant margin

Results from Post Completion Survey

	Initial CRA	User-adapted CRA
Helpfulness	2.81	**3.05**
Satisfaction	3.67	**3.78**

Model Comparison via Interaction Logs

	CRA	User-adapted CRA
# Requests	2.88	2.76
% Acceptance	**31.9%**	31.8%

Project Roadmap

- Users find **CRA suggestions more helpful** than a baseline model
- **Collaborative users are more effective** at the creative writing task
- **Model helps skilled writers more** potentially widening the gap in performance
- **Can the model be adapted to learn from observed user interactions to provide better assistance?**

Takeaways

- Users find **CRA suggestions more helpful** than a baseline model
- **Collaborative users are more effective** at the creative writing task
- **Model helps skilled writers more** potentially widening the gap in performance
- The **model becomes more helpful after updating it** from user interactions, **but not by much**

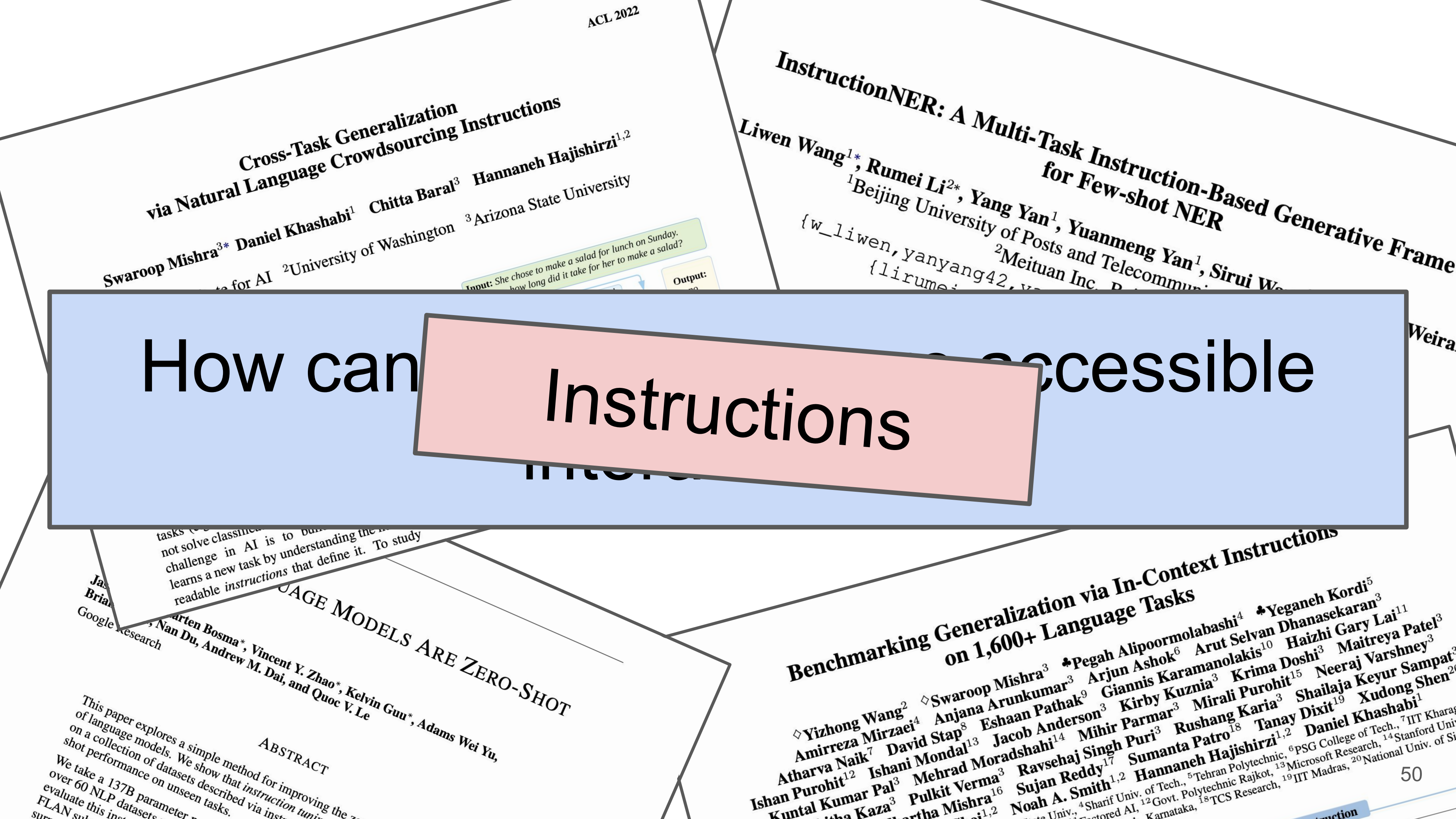
What Next?

- **Model helps skilled writers more** potentially widening the gap in performance
  - How to better assist writers who aren’t as comfortable with the language?
- The **model becomes more helpful after updating it** from user feedback, **but not by much**
  - How to learn more effectively from aggregated observed interactions?

What Next?

- **Model helps skilled writers more** potentially widening the gap in performance
  - How to better assist writers who aren’t as comfortable with the language?
- The model becomes more helpful after updating it from user feedback, but not by much
  - How to learn more effectively from aggregated observed interactions?

How can we design more accessible interactions?


ACL 2022
Cross-Task Generalization
via Natural Language Crowdsourcing Instructions
Swaroop Mishra[3*] Daniel Khashabi[1] Chitta Baral[3] Hannaneh Hajishirzi[1,2]
[2]University of Washington [3]Arizona State University
Input: She chose to make a salad for lunch on Sunday.
how long did it take for her to make a salad?
Output:
InstructionNER: A Multi-Task Instruction-Based Generative Frame
for Few-shot NER
Liwen Wang[1*], Rumei Li[2*], Yang Yan[1], Yuanmeng Yan[1], Sirui Wa
[1]Beijing University of Posts and Telecommun
[2]Meituan Inc.
{w_liwen,yanyang42,
How can
accessible
Instructions
not solve classification
challenge in AI is to
learns a new task by understanding the
readable instructions that define it. To study
UAGE MODELS ARE ZERO-SHOT
Google Research
Nan Du, Andrew M. Dai, and Quoc V. Le
Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu,
ABSTRACT
This paper explores a simple method for improving the
of language models. We show that instruction tuning
on a collection of datasets described via instr
shot performance on unseen tasks.
We take a 137B parameter
over 60 NLP datasets
Benchmarking Generalization via In-Context Instructions
on 1,600+ Language Tasks
Yizhong Wang[2] Swaroop Mishra[3] Pegah Alipoormolabashi[4] Yeganeh Kordi[5]
Amirreza Mirzaei[4] Anjana Arunkumar[3] Arjun Ashok[6] Arut Selvan Dhanasekaran[3]
Atharva Naik[7] David Stap[8] Eshaan Pathak[9] Giannis Karamanolakis[10] Haizhi Gary Lai[11]
Ishan Purohit[12] Ishani Mondal[13] Jacob Anderson[3] Kirby Kuznia[3] Krima Doshi[3] Maitreya Patel[3]
Kuntal Kumar Pal[3] Mehrad Moradshahi[14] Mihir Parmar[3] Mirali Purohit[15] Neeraj Varshney[3]
Pulkit Verma[3] Ravsehaj Singh Puri[3] Rushang Karia[3] Shailaja Keyur Sampat[3]
Sujan Reddy[17] Sumanta Patro[18] Tanay Dixit[19] Xudong Shen[20]
Noah A. Smith[1,2] Hannaneh Hajishirzi[1,2] Daniel Khashabi[1]

Collaborative Poetry Writing with Instructions

{Tuhin Chakrabarty, Vishakh Padmakumar}, He He

EMNLP 2022

NYU ML[2] Machine Learning for Language

Project Roadmap

- Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?
- Can LLMs compose instructions seen at train time in unseen combinations?
- Can we help users complete creative writing tasks (poetry writing) using natural language instructions?

Project Roadmap

- **Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?**
- Can LLMs compose instructions seen at train time in unseen combinations?
- Can we help users complete creative writing tasks (poetry writing) using natural language instructions?

Dataset of Instructions

- Staying on Subject:

Write a poetic verse that ends in a word which rhymes with 'late' → And homeward he turned, his path now **straight.**

Write a poetic sentence about 'god' and ending in 'eyes' → An all-powerful **God**, no escaping his **eyes**

Dataset of Instructions

- Control on Literary Devices:

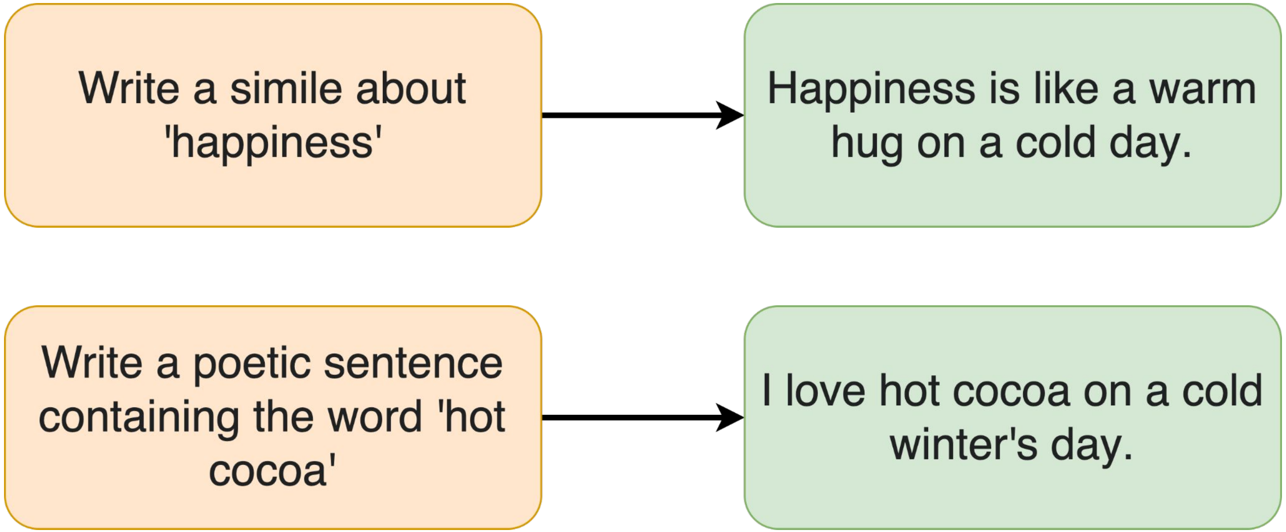
Write a simile about 'hopelessness'	Feels **like a glass that is empty all the way**
Generate a metaphor about 'literature'	**Literature is a splendid mistress**, but a bad wife.
Write a haiku that contains the word 'Earth'	**Earth**. Giver of life / Most exceptional planet / Mother to us all.

Evaluation

- Hand crafted test sets of instructions for different kinds of capabilities

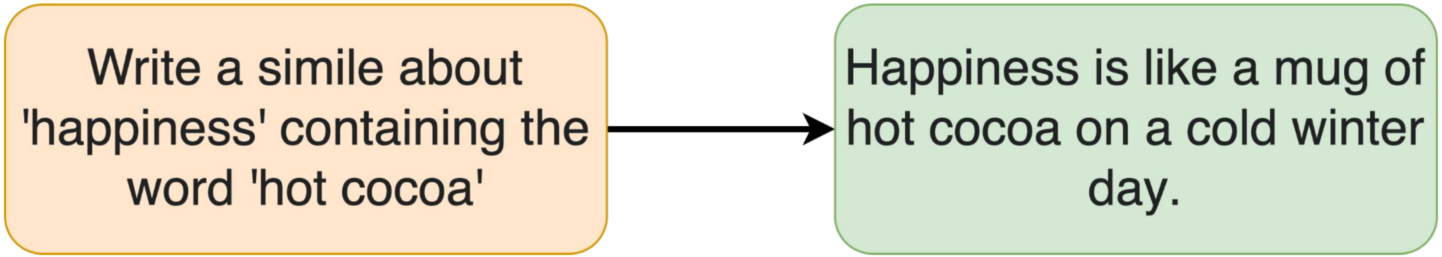
Evaluation

- Hand crafted test sets of instructions for different kinds of capabilities
    - **Known Instruction Templates**



Evaluation

- Hand crafted test sets of instructions for different kinds of capabilities
    - Known Instruction Templates
    - **Compositional Instruction Templates**



Evaluation

- Hand crafted test sets of instructions for different kinds of capabilities
  - Known Instruction Templates
  - Compositional Instruction Templates
- Baselines
  - T0 models
    - **T0 - 3B** Finetuned + **T0pp - 11B** Few-Shot
  - **InstructGPT - 175B** - Zero Shot + Few-Shot

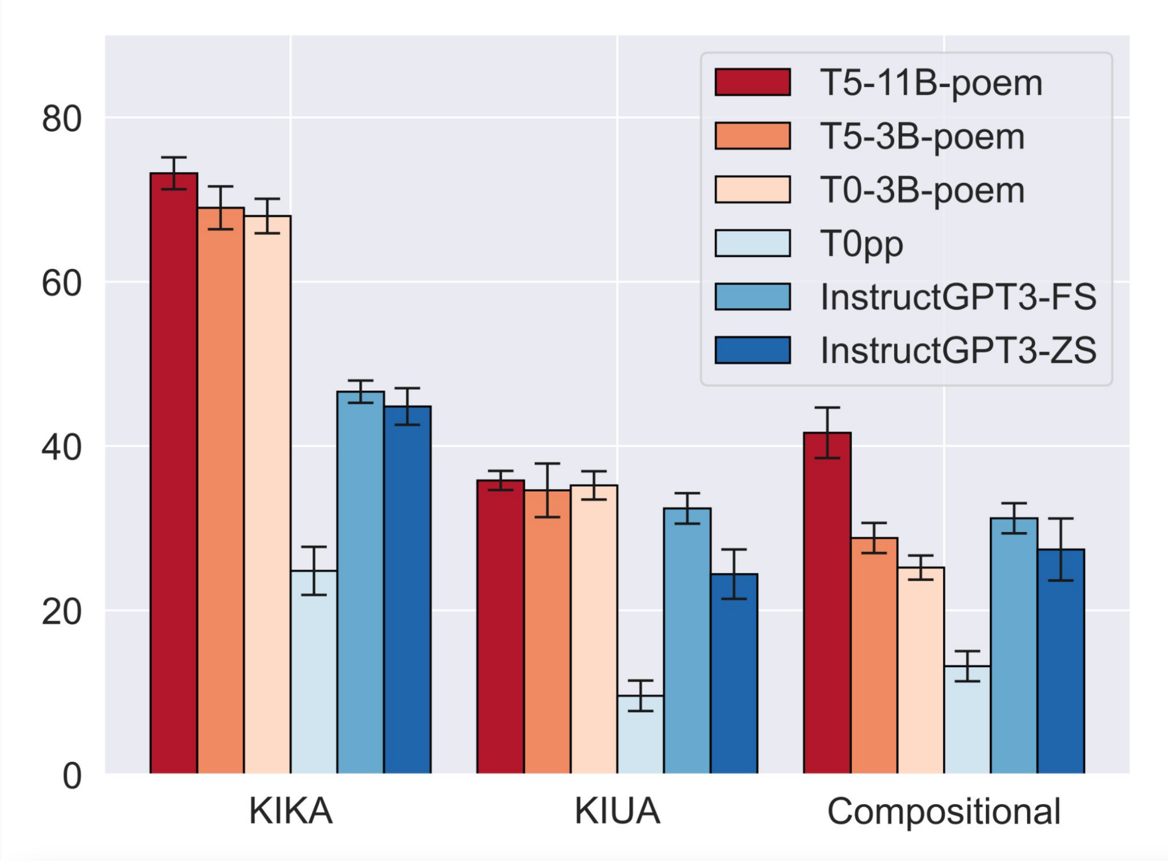
Research Questions

- Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?
- Can models compose instructions seen at train time in unseen combinations?
- Can we help users complete creative writing tasks using natural language instructions?

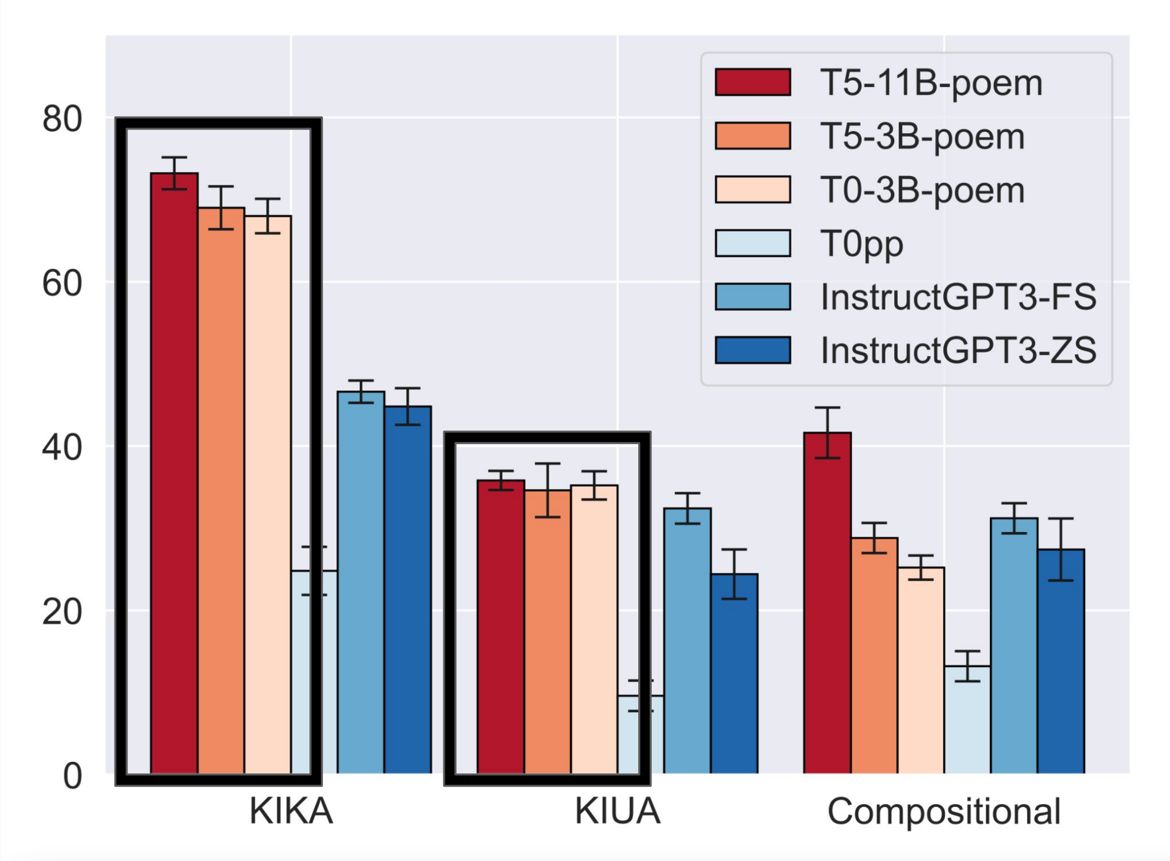
Research Questions

- **Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?**
- **Can models compose instructions seen at train time in unseen combinations?**
- Can we help users complete creative writing tasks using natural language instructions?

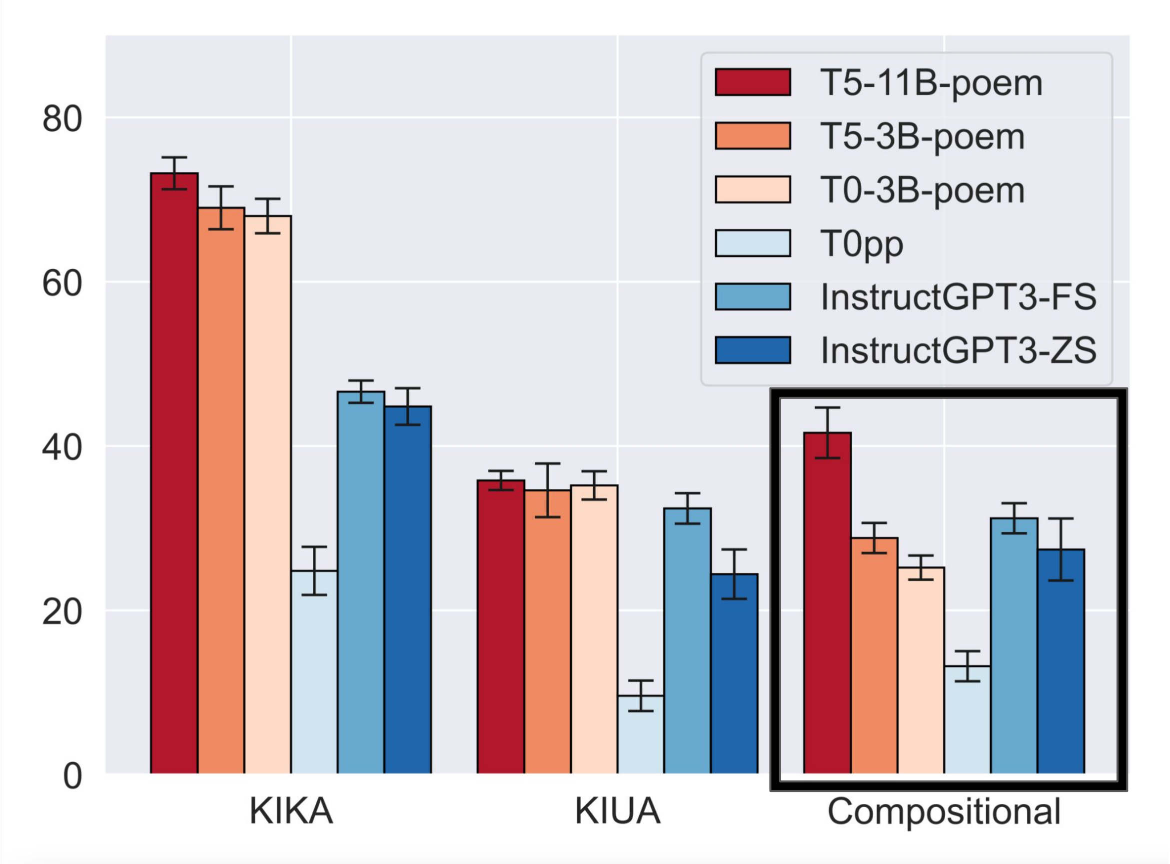
Instruction Tuning - Evaluation



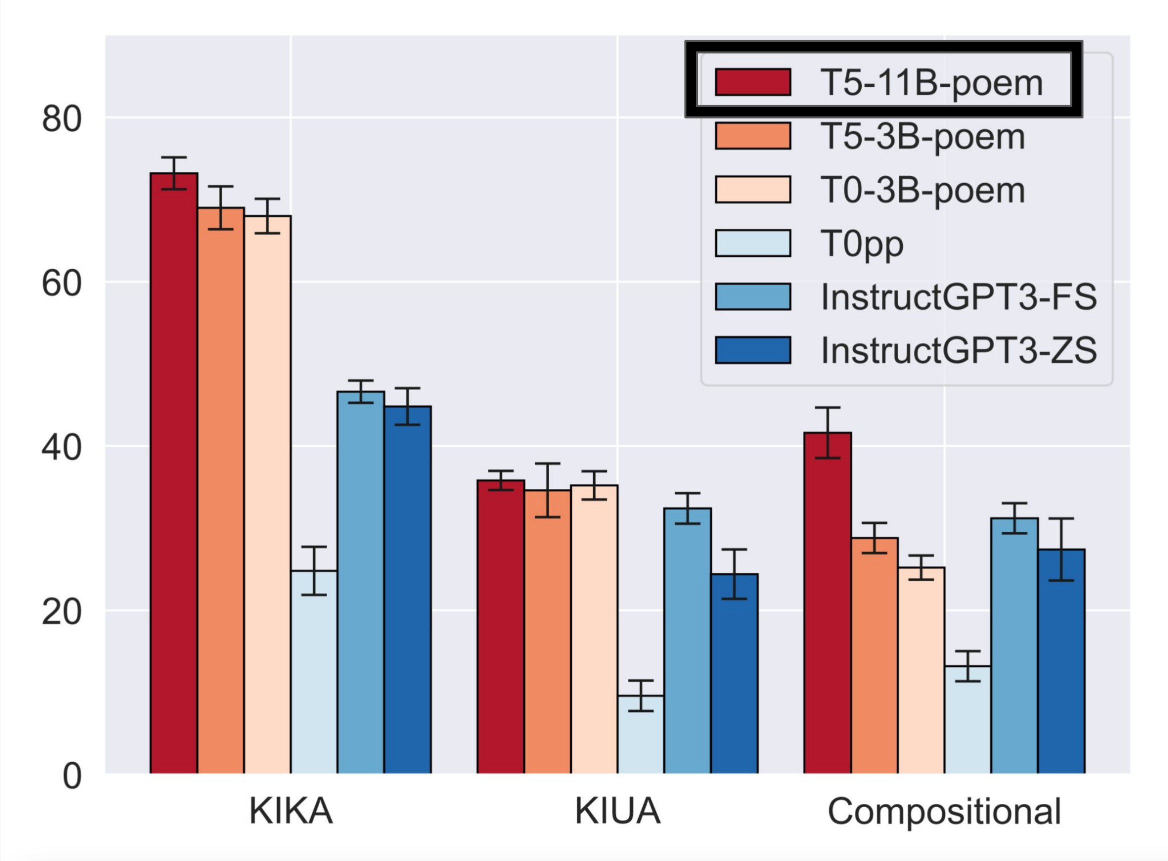
Finetuned Models Are Strong In-Domain But Drop on Out-Of-Domain Data



Larger Models Compose Instructions Better

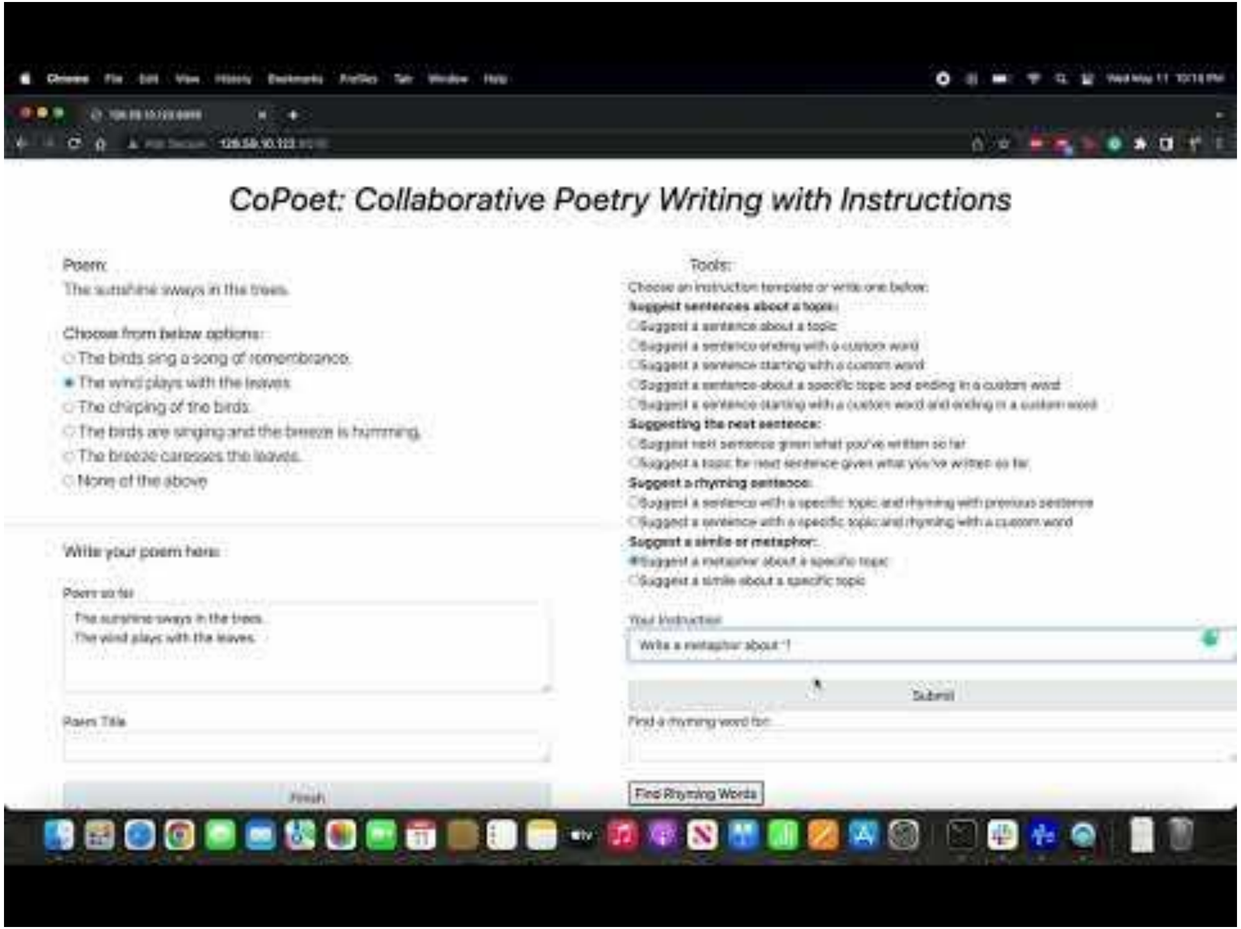


We use T5-11B for the User Study.

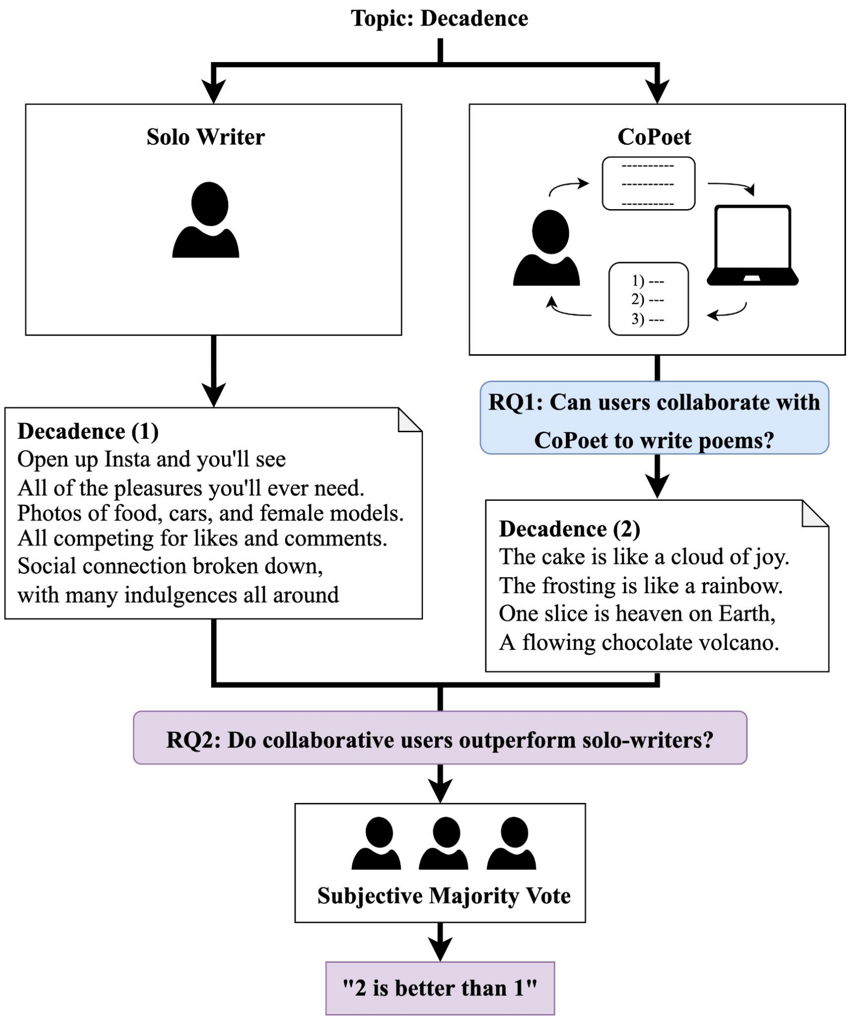


Research Questions

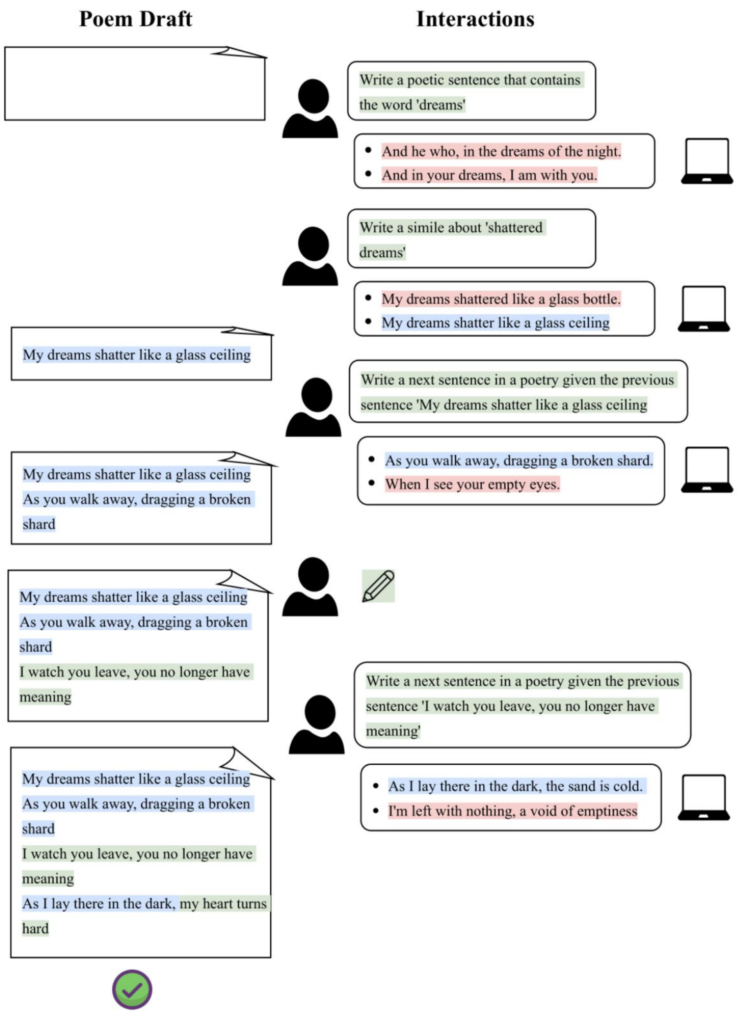
- Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?
- Can models compose instructions seen at train time in unseen combinations?
- **Can we help users complete creative writing tasks using natural language instructions?**


CoPoet: Collaborative Poetry Writing with Instructions
Poem:
The sunshine sways in the trees.
Choose from below options:
The birds sing a song of remembrance.
The wind plays with the leaves.
The chirping of the birds.
The birds are singing and the breeze is humming.
The breeze caresses the leaves.
None of the above
Write your poem here:
Poem so far
The sunshine sways in the trees.
The wind plays with the leaves.
Poem Title
Finish
Tools:
Choose an instruction template or write one below:
Suggest sentences about a topic:
Suggest a sentence about a topic
Suggest a sentence ending with a custom word
Suggest a sentence starting with a custom word
Suggesting the next sentence:
Suggest a rhyming sentence:
Suggest a simile or metaphor:
Suggest a metaphor about a specific topic
Suggest a simile about a specific topic
Your Instruction
Submit
Find Rhyming Words

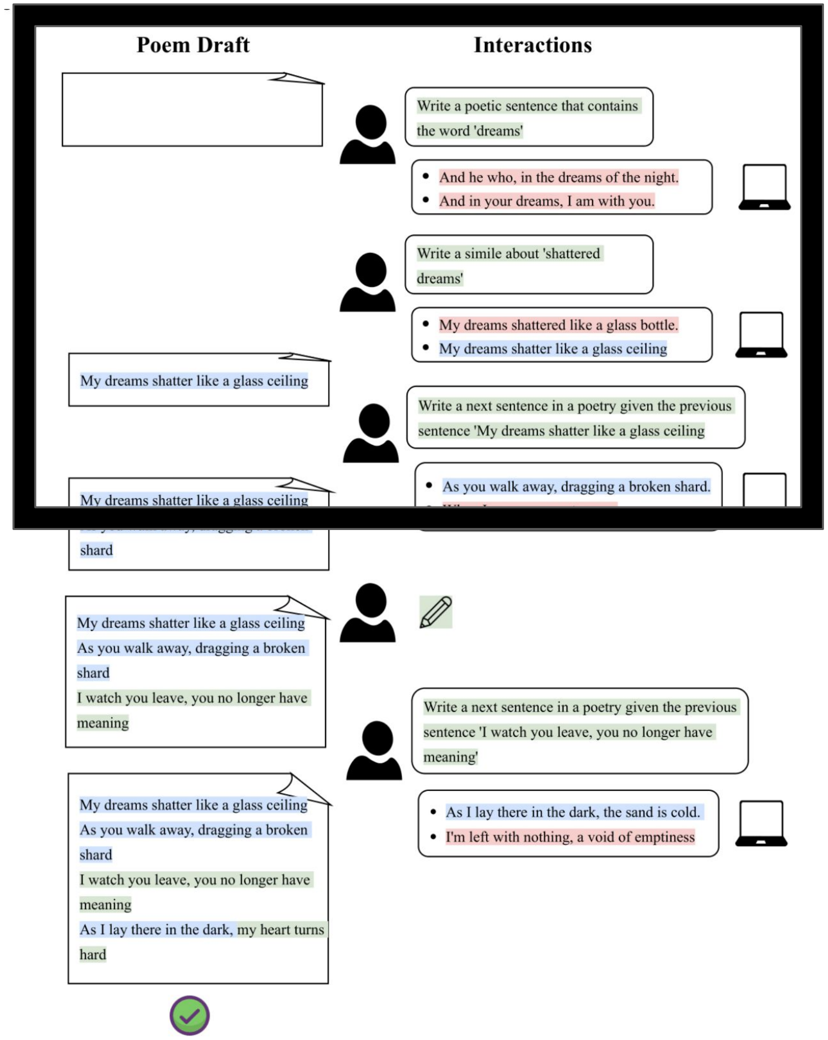
Overview of User Study



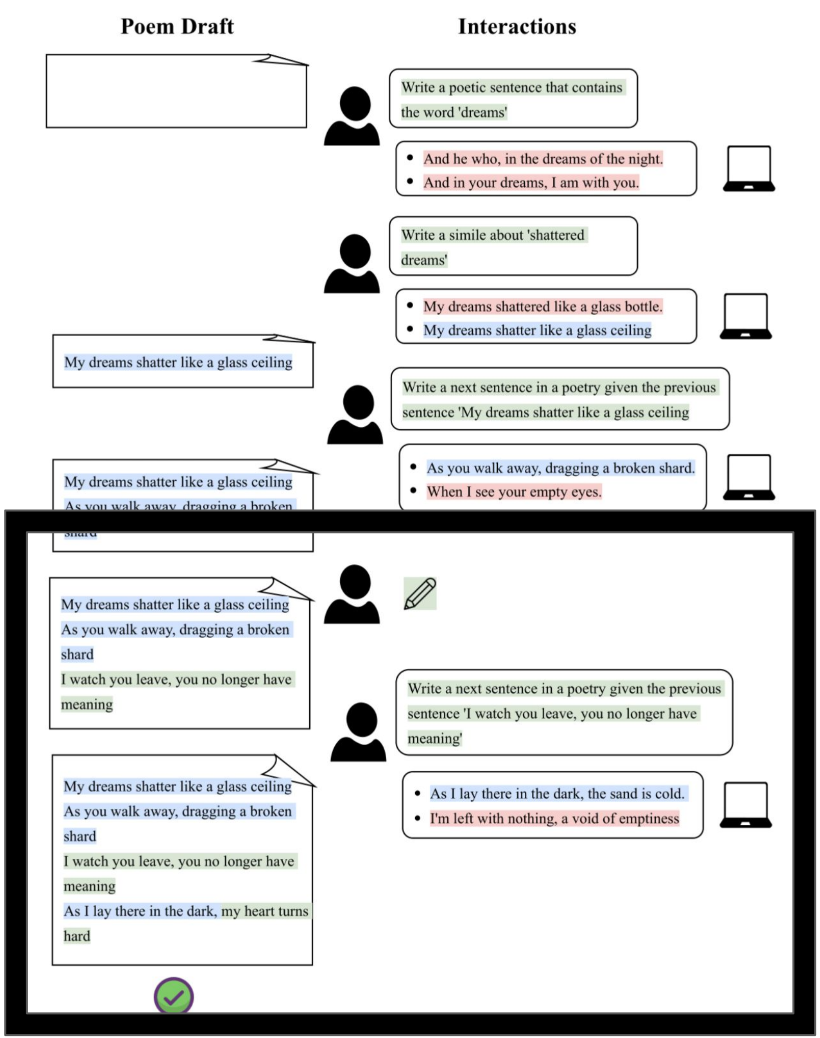
Collaborative Writing Process



Collaborative Writing Process



Collaborative Writing Process



Collaborative Poem Writing

Another Day

The world has not yet awakened.

Darkness still creeps, but the day is not far.

Oh wait! there's the sun, and thus a solitary regret.

I still can't believe I haven't been to bed yet.

Instructions:

- Write a poetic sentence that contains the word 'Morning'
- Write a simile about 'Night'
- **Write a poetic sentence that contains the word 'sun' and ending in a rhyme for 'yet'**
- Write a poetic sentence that contains the word 'Darkness' and ending in a rhyme for 'awakened'

Collaborative Poem Writing

Another Day

The world has not yet awakened.

Darkness still creeps, but the day is not far.

Oh wait! there's the sun, and thus a solitary regret.

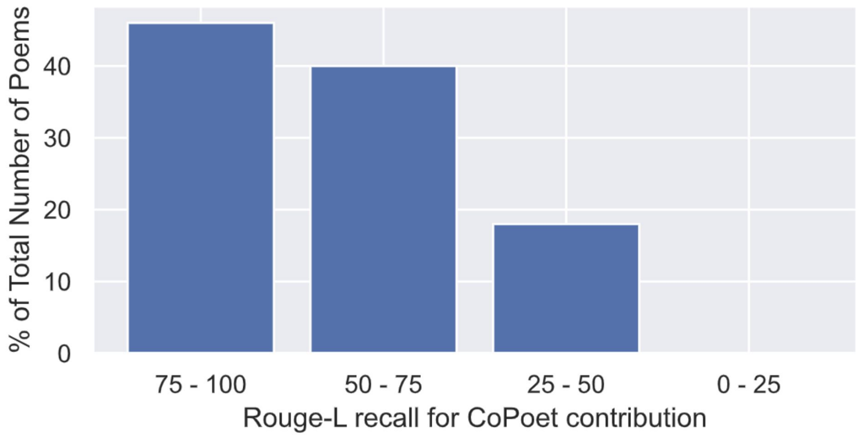
I still can't believe I haven't been to bed yet.

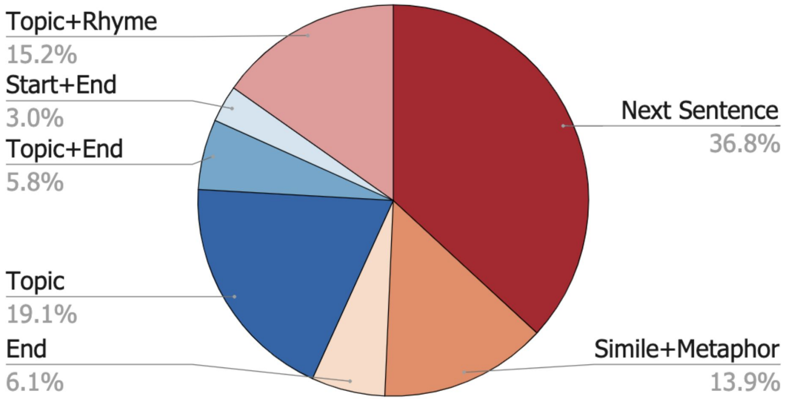
Human Written / Model Written

Instructions:

- Write a poetic sentence that contains the word 'Morning'
- Write a simile about 'Night'
- **Write a poetic sentence that contains the word 'sun' and ending in a rhyme for 'yet'**
- Write a poetic sentence that contains the word 'Darkness' and ending in a rhyme for 'awakened'

Do Users Find The Model Helpful?





Do Users Write *Better* Poems With Model Help?

	Relevant %	Preferred %
Solo	96	43
Collaborative	98	**57**

	Preferred %	Not Preferred %
Diversity	63.0	37.0
Rhyme	72.5	27.5

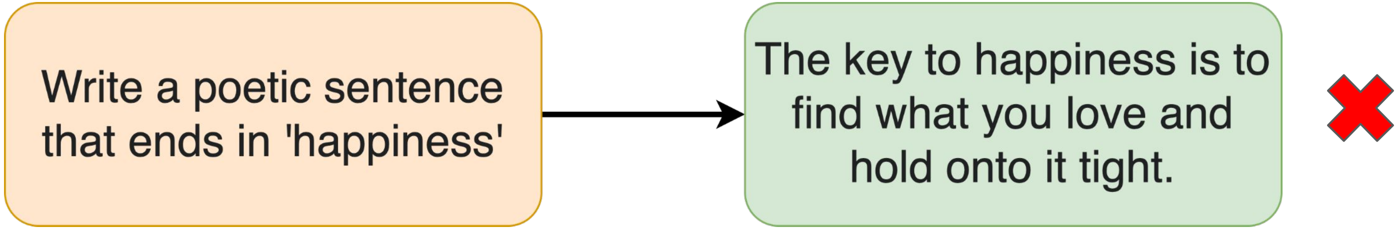
More Examples + Model Contributions

Takeaways

- **Instruction tuning** can be an **effective way to help users write poems**

Takeaways

- **Instruction tuning** can be an **effective way to help users write poems**
- InstructGPT3 is **pretty good at staying on subject** but still has **difficulty with more challenging generation instructions**



Takeaways

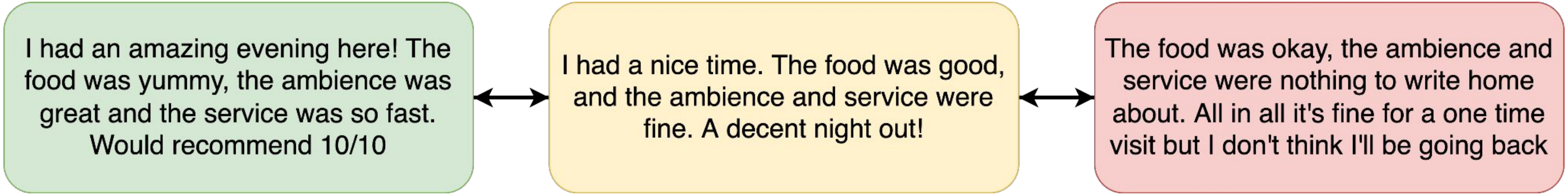
- **Instruction tuning** can be an **effective way to help users write poems**
- InstructGPT3 is **pretty good at staying on subject** but still has **difficulty with more challenging generation instructions**

Big Picture

- Many open questions bridging NLP and HCI:
  - How do we design the ***most effective collaboration setup*** to help human users?
  - How do we train models to generate ***helpful*** suggestions?
  - What is the ***best form of user feedback*** and how do we ***incorporate it in model training***?
  - How do we assist users in ***content planning*** for long form creative writing?
  - How do we ensure ***equitable creative writing assistance*** to all users?
- My Projects:
  - Machine-in-the-Loop Rewriting for Creative Image Captioning
  - Collaborative Poetry writing with Instruction Tuning

Big Picture

- How do we train models to generate ***helpful*** suggestions?
  - Controlling stylistic attributes of text such as sentiment (hopefully going to ICML)
    - Also works on non-textual sequences like proteins



Big Picture

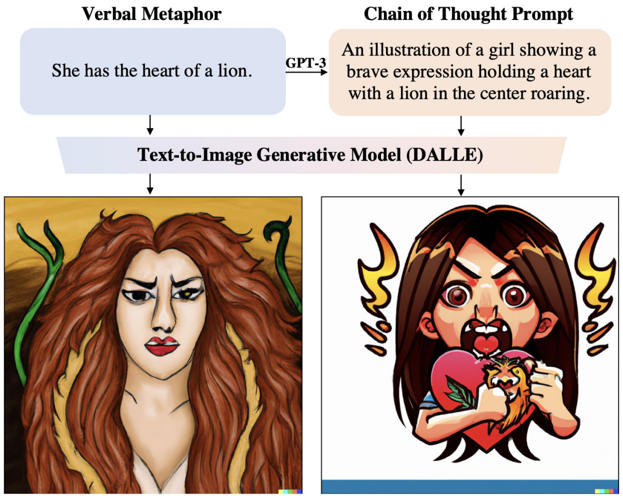
- What is the ***best form of user feedback*** and how do we ***incorporate it in model training***?
    - Few-shot LLM personalization as an alternative to aggregation of feedback
    - Machine Teaching
        - Can we train models that can perform the content selection + presentation task needed to help human students?

Big Picture

- How do we design the ***most effective collaboration setup*** to help human users?
    - Providing assistance in more specialised domains (medical texts)

Big Picture

- How do we design the ***most effective collaboration setup*** to help human users?
  - Multimodal Creativity - Check out Tuhin's work on Visual Metaphors :)



Thank You

NYU ML² Machine Learning for Language

Backup Slides

How Well Do Models Compose Instructions?

		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- ZS (175B)	InstructGPT - FS (175B)	T0pp
Subject (55)	% - Match	76.36%	60%	54.54%	72.72%	**87.87%**	65.45%
	% - Match w/ Ending (34)	**47.05%**	41.17%	38.23%	26.31%	29.41%	29.41%
Rhyme (16)	% - Match -w & Rhyme Success Rate	**43.75%**	25.00%	37.50%	26.31%	37.50%	0.00%
Simile (4)	% Subject + Comparator	25.00%	50.00%	0.00%	66.66%	**75%**	25.00%
	% Comparator	**100.00%**	**100.00%**	**100.00%**	83.33%	**100%**	25.00%
Metaphor (4)	% Subject + Comparator	**50.00%**	**50.00%**	25%	25%	25%	25.00%
Haiku (5)	% Subject + [15-19] Syllables	**60.00%**	20.00%	**60%**	0%	20%	0.00%

How Well Do Models Compose Instructions?

		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- ZS (175B)	InstructGPT - FS (175B)	T0pp
Subject (55)	% - Match	76.36%	60%	54.54%	72.72%	[illegible]	65.45%
	% - Match w/ Ending (34)	**47.05%**	41.17%	38.23%	26.31%	[illegible]	29.41%
Rhyme (16)	% - Match -w & Rhyme Success Rate	**43.75%**	25.00%	37.50%	[illegible]	[illegible]	0.00%
Simile (4)	% Subject + Comparator	25.00%	50.00%	0.00%	66.66%	[illegible]	25.00%
	% Comparator	**100.00%**	**100.00%**	**100.00%**	83.33%	[illegible]	25.00%
Metaphor (4)	% Subject + Comparator	**50.00%**	**50.00%**	25%	25%	25%	25.00%
Haiku (5)	% Subject + [15-19] Syllables	**60.00%**	20.00%	**60%**	0%	20%	0.00%

Composition improves with model size

T5 outperforms T0 again

How Well Do Models Compose Instructions?

		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- ZS (175B)	InstructGPT - FS (175B)	T0pp
	% - Match	76.36%	60%	54.54%	72.72%	**87.87%**	65.45%
Subject (55)	% - Match w/ Ending (34)	**47.05%**	41.17%	[illegible]	26.31%	29.41%	29.41%
Rhyme (16)	% - Match -w & Rhyme Success Rate	**43.75%**	[illegible]	[illegible]	[illegible]	37.50%	0.00%
	% Subject + Comparator	25.00%	50.00%	[illegible]	66.66%	**75%**	25.00%
Simile (4)	% Comparator	**100.00%**	**100.00%**	[illegible]	83.33%	**100%**	25.00%
Metaphor (4)	% Subject + Comparator	**50.00%**	**50.00%**	[illegible]	25%	25%	25.00%
Haiku (5)	% Subject + [15-19] Syllables	**60.00%**	20.00%	**60%**	0%	20%	0.00%

Comparable performance with T5 edging it on challenging instructions

Scalar Controlled Text Generation

w/ Richard Pang, He He and Ankur Parikh

NYU

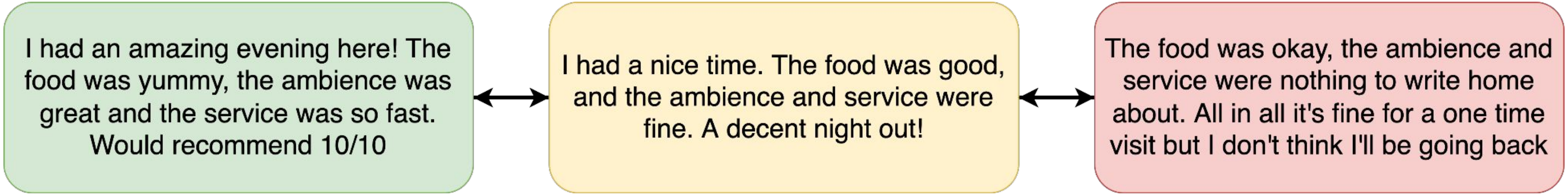
Machine Learning for Language

Motivation

- As opposed to controlling the output of text with literary devices or instructions, writers might want to control scalar attributes
  - Sentiment Control, Toxicity etc.

Motivation

- As opposed to controlling the output of text with literary devices or instructions, writers might want to control scalar attributes
- An example of sentiment control between a positive, neutral and slightly negative version of the same sentence



Problem Setup

- Assume we have an oracle scorer f_s
    - Maps from an input sequence to the range of the score
- Given an input text x and a target score s_t
- **Goal: Generate x' s.t. fs(x') = st**

Approach

- **Generate x' s.t. $f_s(x') = s_t$ iteratively**
  - First generate $x_i' = x_{i-1} + \partial f_s/\partial x$
  - Increase the number of iterations "i" in order to achieve higher/lower target scores
  - Allows generalization to OOD target scores

Approach

- Generate x' s.t. $f_s(x') = s_t$ iteratively
  - First generate $x_i' = x_{i-1} + \partial f_s / \partial x$
  - Increase the number of iterations "i" in order to achieve higher/lower target scores
  - Allows generalization to OOD target scores
  - **How do we train a model to perform this?**

Approach

- Generate x' s.t. $f_s(x') = s_t$ iteratively
  - Update $x_i' = x_{i-1} + \partial f_s/\partial x$
- **Learn to approximate ∂fs/∂x from Perturbations**
  - Create paired data using masking + infilling
  - Learn Control Tags with LLMs to edit the text and move up and down this scale

Learning from Perturbations

Example:

- **Source:** "**<dec>** The desserts come with the **easy to-believe claim that they contain just** under 200 calories.", **Score: 2.607**
- **Target:** "The desserts come with the **hard-to-believe claim that they are all** under 200 calories.", **Score: 2.132**

Example

Trained T5-base on the created dataset

X = “Top notch doctor in a top notch practise”, s(x) = 4.904, s_t = 3.0

Iteration	**Text**	**Sentiment Score**
1	<dec> Top notch doctor in a top notch practice.	4.904
2	<dec> A top notch doctor is in the practice.	4.701
3	<dec> The practice is staffed by a top notch **hygenist.**	4.003
4	<dec> The practice is managed by a top notch hygenist.	3.647
5	The practice is run by a very good hygenist.	2.955

Experiments

- For sentiment analysis, scorer model is linear regression classifier
    - Score range: [0, 4]
- Select only training data in [1, 3] range

Experiments

- For sentiment analysis, scorer model is linear regression classifier
  - Score range: [0, 4]
- Select only training data in [1, 3] range
- Evaluation:
  - Given a source sentence, alter it to two separate target scores
  - Report success rate of reaching the target
    - Very In-Domain - Source score +/- 0.3
    - In-Domain - {1.5, 2.5}
    - Out-of-Domain - {0.5, 3.5}

Results

Test set size = 1831		VID	ID	OOD
	Our Model (small, n_iter = 10, n_seq = 5)	0.965	**0.971**	0.429
	Our Model (large, n_iter = 10, n_seq = 5)	0.938	0.930	**0.623**
	Genhance (d_z = 0.15, n = 1)	0.377	0.287	0.04375
	Genhance (d_z = 0.15, n = 50)	0.9515	0.9075	0.3865
With Scorer	Genhance (d_z = 0.15, n = 100)	**0.9775**	0.9535	0.54965

Metrics of Evaluation

Test set size = 1831		VID	ID	OOD
	Ou 10, n_seq =	0.965	**0.971**	0.429
	Ou 10, n_seq =	0.938	0.930	**0.623**
	, n = 1)	0.377	0.287	0.04375
	n = 50)	0.9515	0.9075	0.3865
With Scorer	Genhance (d_z = 0.15, n = 100)	**0.9775**	0.9535	0.54965

Each value in these columns is the success rate at achieving various target attribute values
for different models

Comparison to Baseline Genhance Model

Test set size = 1831		VID	ID	OOD
	Our Model (small, n_iter = 10, n_seq = 5)	0.965	**0.971**	0.429
	Our Model (large, n_iter = 10, n_seq = 5)	0.938	0.930	**0.623**
	Genhance (d_z = 0.15, n = 1)	0.377	0.287	0.04375
	Genhance (d_z = 0.15, n = 50)	0.9515	0.9075	0.3865
With Scorer	Genhance (d_z = 0.15, n = 100)	**0.9775**	0.9535	0.54965

We outperform a controlled generation baseline on ID and OOD

Trade-off based on Perturbation Size

When we make large perturbations, we perform better on OOD But sentences differ more from the source

1831		**VID**	**ID**	**OOD**
	Our Model (small, n_iter = 10, n_seq = 5)	0.965	**0.971**	0.429
	Our Model (large, n_iter = 10, n_seq = 5)	0.938	0.930	**0.623**
	Genhance (d_z = 0.15, n = 1)	0.377	0.287	0.04375
	Genhance (d_z = 0.15, n = 50)	0.9515	0.9075	0.3865
With Scorer	Genhance (d_z = 0.15, n = 100)	**0.9775**	0.9535	0.54965

Fine-Grained Comparison

The success rate is a coarse metric. We want to examine the average case.

Given all the source examples, how much are we able to change the score of each?

Fine-Grained Comparison

The success rate is a coarse metric. We want to examine the average case.

Given all the source examples, how much are we able to change the score of each?

Our Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	0	3	12	29	40	105	12	0
	1.5-2	1	0	18	32	71	174	12	0
	2-2.5	0	0	5	21	64	194	18	0
	2.5-3	0	0	1	6	41	130	11	0
	3-3.5	0	0	0	0	0	0	0	0

Genhance Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	1	23	53	49	39	25	11	0
	1.5-2	0	12	91	81	76	33	15	0
	2-2.5	0	0	18	87	131	49	17	0
	2.5-3	0	0	0	13	114	52	10	0
	3-3.5	0	0	0	0	0	0	0	0

Fine-Grained Comparison

Given all the source examples, how much are we able to change the score of each?

We compare our approach to the same baseline model

Our Model									
Source Score	0-0.5	0	0						0
	0.5-1	0	0						0
	1-1.5	0	3	12	29	40	105	12	0
	1.5-2	1	0	18	32	71	174	12	0
	2-2.5	0	0	5	21	64	194	18	0
	2.5-3	0	0	1	6	41	130	11	0
	3-3.5	0	0	0	0	0	0	0	0

Genhance Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	1	23	53	49	39	25	11	0
	1.5-2	0	12	91	81	76	33	15	0
	2-2.5	0	0	18	87	131	49	17	0
	2.5-3	0	0	0	13	114	52	10	0
	3-3.5	0	0	0	0	0	0	0	0

Fine-Grained Comparison

Given all the source examples, how much are we able to change the score of each?

Our Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	-0.5	0	0	0	0	0	0	0	0
	.5-1	0	0	0	0	0	0	0	0
	-1.5	0	3	12	29	40	105	12	0
	.5-2	1	0	18	32	71	174	12	0
	-2.5	0	0	5	21	64	194	18	0
	.5-3	0	0	1	6	41	130	11	0
	-3.5	0	0	0	0	0	0	0	0

Genhance Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
	0-0.5								0
	0.5-1								0
									0
Score	1.5-2								0
	2-2.5								0
	2.5-3								0
	3-3.5	0	0	0	0	0	0	0	0

For each example, we examine the source score and the corresponding target score achieved. Here we attempt to increase the score

Fine-Grained Comparison

Given all the source examples, how much are we able to change the score of each?

Our Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	0	3	12	29	40	105	[illegible]	[illegible]
	1.5-2	1	0	18	32	71	174	12	0
	2-2.5	0	0	5	21	64	194	18	0
	2.5-3	0	0	1	6	41	130	11	0
	3-3.5	0	0	0	0	0	0	0	0

Genhance Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25	11	0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	33	15	0
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	49	17	0
	2.5-3	0	0	0	13	114	52	10	0
	3-3.5	0	0	0	0	0	0	0	0

This cell means that there were **105 examples** with **source score in 1-1.5** that achieved **target score between 3.5 and 4**

Fine-Grained Comparison

Given all the source examples, how much are we able to change the score of each?

Our Model		Target Score Achieved							
		1-1.5	1.5-2	2-2	2.5-3	3-3.5	3.5-4	4-4.5	.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	0	3	12	29	40	105	12	0
	1.5-2	1	0	18	32	71	174	12	0
	2-2.5	0	0	5	21	64	194	18	0
	2.5-3	0	0	1	6	41	130	11	0
	3-3.5	0	0	0	0	0	0	0	0

Our model clusters most examples in 3.5 to 4 range, which outperforms the baseline where it is mostly 3-3.5

Target Score Achieved						
1.5-2	2-2.	2.5-3	3-3.5	3.5-4	4-4.5	.5-5
0	0	0	0	0	0	0
0	0	0	0	0	0	0
2[illegible]	53	49	39	25	11	0
[illegible]	91	81	76	33	15	0
0	18	87	131	49	17	0
0	0	13	114	52	10	0
0	0	0	0	0	0	0

Fine-Grained Comparison

Our model is able to better shift the distribution of scores of examples in the desired direction.

Our Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	0	3	12	29	40	105	12	0
	1.5-2	1	0	18	32	71	174	12	0
	2-2.5	0	0	5	21	64	194	18	0
	2.5-3	0	0	1	6	41	130	11	0
	3-3.5	0	0	0	0	0	0	0	0

Genhance Model		Target Score Achieved							
		1-1.5	1.5-2	2-2.5	2.5-3	3-3.5	3.5-4	4-4.5	4.5-5
Source Score	0-0.5	0	0	0	0	0	0	0	0
	0.5-1	0	0	0	0	0	0	0	0
	1-1.5	1	23	53	49	39	25	11	0
	1.5-2	0	12	91	81	76	33	15	0
	2-2.5	0	0	18	87	131	49	17	0
	2.5-3	0	0	0	13	114	52	10	0
	3-3.5	0	0	0	0	0	0	0	0

Takeaways

- We're able to achieve better controlled generation than baselines, particularly towards OOD target attribute values
- Our approach learns the distribution of the data and trains a model to move up and down the scale of an attribute along the data distribution.

Next Steps

- Nothing that we do is specific to text!
- We're trying to generate protein sequences where we control the attribute of the stability of the molecule.
    - Not possible in other approaches that require a differentiable scoring function

Collaborative Poem Writing

Sad Reality

No one prepares you for growing older

Aging is a symptom of your dreams foreclosure

You're not a musician, but you smoke like them

Hurry up, your lunch is over in 10

Instructions:

- Write a poetic sentence that contains the word 'Growing up'
- **Write a metaphor about 'Aging'**
- Write a next sentence in a poetry given the previous sentence 'Aging is a symptom of your dreams fading'
- Write a next sentence in a poetry given the previous sentence 'You're not a musician, but you smoke like them'

User Interface

Describe the image below

Enter your text

The majestic lion lies poised and ready to strike, an [awesome sight to see].

Suggest Finish

Char count: 77

Select the suggestion that you like best: 1 ○ 2 ○ 3 ○ Original Text ○

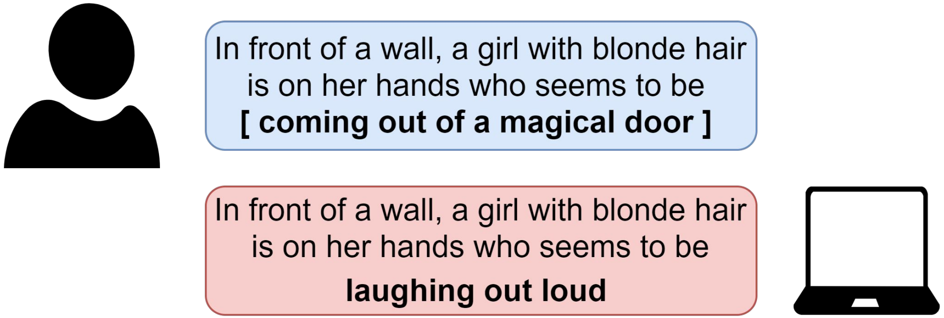
Suggestion 1
The majestic lion lies poised and ready to strike, an **alert** predator eyeing his **prey**

Suggestion 2
The majestic lion lies poised and ready to strike, an **avenger.**

Suggestion 3
The majestic lion lies poised and ready to strike, an **abyss waiting.**

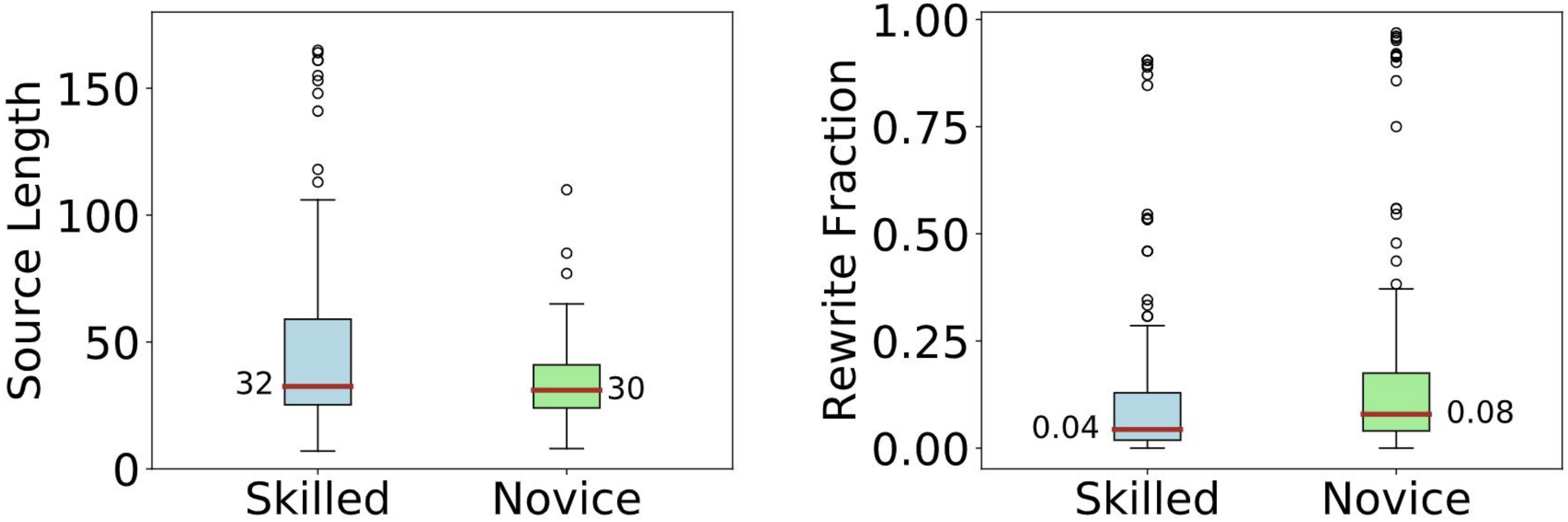
Error Analysis

- Most common error case is content drift
    - Model changes the meaning of the sentence when rewriting
- Copying of the source text verbatim
- Repetition in generated text



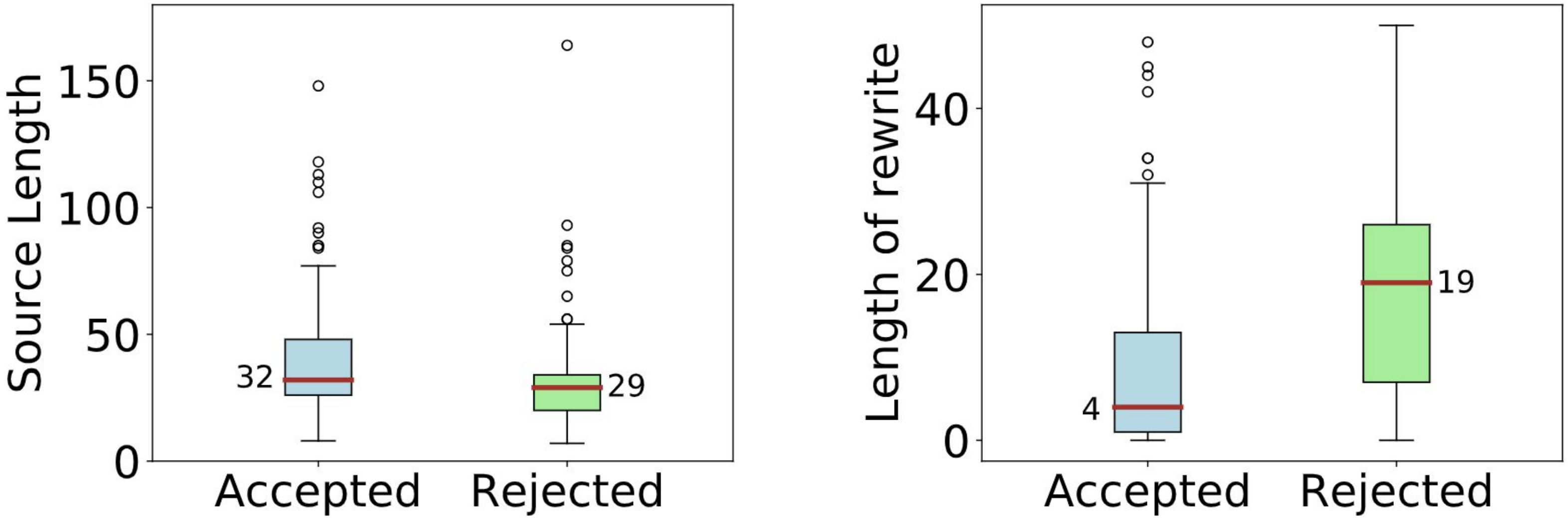
When is the Model Effective?

Skilled writers tend to write longer sentences and request shorter fractions to be rewritten



When is the Model Effective?

Shorter rewrites in longer sentences tend to be accepted more



Automatic Evaluation - Unknown Entities

		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- ZS (175B)	InstructGPT - FS (175B)	T0pp
Subject (31)	% - Match	**80.64%**	74.19%	77.41%	69.23%	74.19%	51.61%
	% - Match w/ Ending (2)	**100%**	**100%**	**100%**	0%	0.00%	0.00%
Rhyme (11)	Success Rate	**36.36%**	36.36%	9.09%	9.09%	18.18%	9.09%
Simile (9)	% - Subject + Comparator	33.33%	44.44%	33.33%	11.11%	**55.55%**	11.11%
	% - Comparator	**100.00%**	**100.00%**	**100.00%**	66.66%	88.88%	22.22%
Metaphor (7)	% - Subject + Comparator	**100.00%**	**100.00%**	**100.00%**	85.71%	**100.00%**	28.57%
Haiku (7)	% Subject + (15-19) Syllables	**71.42%**	**71.42%**	57.14%	57.14%	42.85%	0.00%

Automatic Evaluation - Unknown Entities

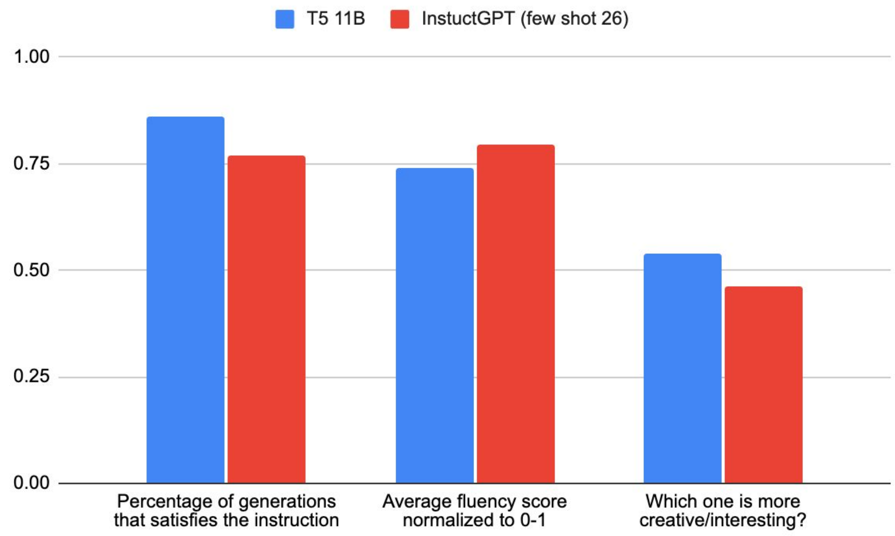
		T5 - 11B	**T5 - 3B**	**T0 - 3B**	**InstructGPT- ZS (175B)**	**InstructGPT - FS (175B)**	**T0pp**
Subject (31)	% - Match	**80.64%**	74.19%	77.41%	69.23%	74.19%	51.61%
	% - Match w/ Ending (2)	**100%**	**100%**	[illegible]	[illegible]	0.00%	0.00%
Rhyme (11)	Success Rate	**36.36%**	36.36%	[illegible]	[illegible]	18.18%	9.09%
Simile (9)	% - Subject + Comparator	33.33%	[illegible]	[illegible]	[illegible]	**55.55%**	11.11%
	% - Comparator	**100.00%**	[illegible]	[illegible]	[illegible]	88.88%	22.22%
Metaphor (7)	% - Subject + Comparator	**100.00%**	**100.00%**	[illegible]	[illegible]	**100.00%**	28.57%
Haiku (7)	% Subject + (15-19) Syllables	**71.42%**	**71.42%**	57.14%	57.14%	42.85%	0.00%

Similar trends as with known entities with lower overall performance

Human Evaluation

Known Entities

	T5 11B	**InstructGPT**
Percentage that satisfies the instruction	**0.862**	0.769
Fluency, on a scale of 1-5	3.697	**3.969**
Which one is more creative/interesting?	**0.538**	0.462



Human Evaluation

Known Entities

	T5 11B	InstructGPT
Percentage that satisfies the instruction	**0.862**	0.769
Fluency, on a scale of 1-5	3.697	**3.969**
Which one is more creative/interesting?	**0.538**	0.462

Unknown Entities

	T5 11B	InstructGPT
P[illegible]tisfies [illegible]n	**0.925**	0.865
F[illegible] of 1-5	3.865	**3.905**
[illegible]ore creative/interesting?	**0.567**	0.433

InstructGPT is more fluent

Human Evaluation - Compositional Instructions

Compositional Test Set

	T5 11B	**InstructGPT**
Percentage that satisfies the instruction	**0.776**	0.552
Fluency, on a scale of 1-5	3.483	**3.756**
Which one is more creative/interesting?	0.477	**0.523**

Collaborative Poem Writing

The harshness of time.

Time is a very harsh mistress all bitter and cold.

Time is never ending everlasting and very bold.

Time will elapse you in moments that matter.

Time will pass and I'll get older and fatter.

Instructions:

- Write a poetic sentence that contains the word 'Time'
- Write a next sentence in a poetry given the previous sentence 'Time is my horse that stays always with me.'
- Write a poetic sentence that contains the word 'time' and ending in a rhyme for 'me'
- Write a poetic sentence that contains the word 'flow' and ending in a rhyme for 'me'
- **Write a poetic sentence that contains the word 'Time' and ending in a rhyme for 'matter'**
- Write a poetic sentence that contains the word 'time' and ending in a rhyme for 'cold'

How Well Do Models Follow Instructions?

Known Templates		**T5 - 11B**	**T5 - 3B**	**T0 - 3B**	**InstructGPT- Zero Shot (175B)**	**InstructGPT - Few Shot (175B)**	**T0pp (11B)**
Subject (51)	% - Match	92.15%	92.15%	86.27%	86.27%	**96.07%**	84.31%
	% - Match w/ Ending (22)	**95.45%**	**95.45%**	86.36%	13.63%	18.18%	40.90%
Rhyme (14)	Success Rate	78.57%	**85.71%**	**85.71%**	57.14%	71.42%	0.00%
Simile (6)	% Subject + Comparator	66.66%	**83.33%**	66.66%	83.33%	66.66%	16.66%
	% Comparator	**100.00%**	**100%**	**100.00%**	100%	83.33%	16.66%
Metaphor (5)	% Subject + Comparator	**100.00%**	**100.00%**	**100.00%**	80%	**100%**	0.00%
Haiku (5)	% Subject + (15-19) Syllables	20.00%	40.00%	0.00%	**80%**	40%	0.00%

How Well Do Models Follow Instructions?

Known Templates		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- Zero Shot (175B)	InstructGPT - Few Shot (175B)	T0pp (11B)
Subject (51)	**% - Match**	92.15%	9[illegible]	[illegible]	36.27%	**96.07%**	84.31%
	% - Match w/ Ending (22)	**95.45%**	9[illegible]	[illegible]	13.63%	18.18%	40.90%
Rhyme (14)	**Success Rate**	78.57%	8[illegible]	[illegible]	57.14%	71.42%	0.00%
Simile (6)	**% Subject + Comparator**	66.6[illegible]%	8[illegible]	[illegible]	33.33%	66.66%	16.66%
	% Comparator	**100.00%**	1[illegible]	[illegible]	100%	83.33%	16.66%
Metaphor (5)	% Subject + Comparator	**100.00%**	10[illegible]	[illegible]	80%	**100%**	0.00%
Haiku (5)	**% Subject + (15-19) Syllables**	20.00%	40.00%	0.00%	**80%**	40%	0.00%

Hand Crafted Metrics for each kind of instruction

These are **soft metrics**!

How Well Do Models Follow Instructions?

Known Templates		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- Zero Shot (175B)	InstructGPT - Few Shot (175B)	T0pp (11B)
Subject (51)	% - Match	92.15%	92.15%	86.27%	86.27%	[illegible]	[illegible]1%
	% - Match w/ Ending (22)	**95.45%**	**95.45%**	86.36%	13.63%	[illegible]	[illegible]0%
Rhyme (14)	Success Rate	78.57%	**85.71%**	**85.71%**	57.14%	[illegible]	[illegible]%
Simile (6)	% Subject + Comparator	66.66%	**83.33%**	66.66%	8[illegible].33%	[illegible]	[illegible]6%
	% Comparator	**100.00%**	**100%**	**100.00%**	100%	[illegible]	[illegible]6%
Metaphor (5)	% Subject + Comparator	**100.00%**	**100.00%**	**100.00%**	80%	[illegible]	[illegible]%
Haiku (5)	% Subject + (15-19) Syllables	20.00%	40.00%	0.00%	**80%**	40%	0.00%

T5 largely outperforms T0

↓

Transfer of Instruction Tuning to new/unrelated tasks

How Well Do Models Follow Instructions?

Known Templates		T5 - 11B	T5 - 3B	T0 - 3B	InstructGPT- Zero Shot (175B)	InstructGPT - Few Shot (175B)	T0pp (11B)
			15%	86.27%	86.27%	**96.07%**	84.31%
Subject (51			**45%**	86.36%	13.63%	18.18%	40.90%
Rhyme (14)			**71%**	**85.71%**	57.14%	71.42%	0.00%
			33%	66.66%	83.33%	66.66%	16.66%
Simile (6)			**0%**	**100.00%**	100%	83.33%	16.66%
Metaphor (5)			**00%**	**100.00%**	80%	**100%**	0.00%
Haiku (5)	% Subject + (15-19) Syllables	20.00%	40.00%	0.00%	**80%**	40%	0.00%

InstructGPT performance improves in few-shot setting.

Both outperform **T0pp**

How Well Do Models Follow Instructions?

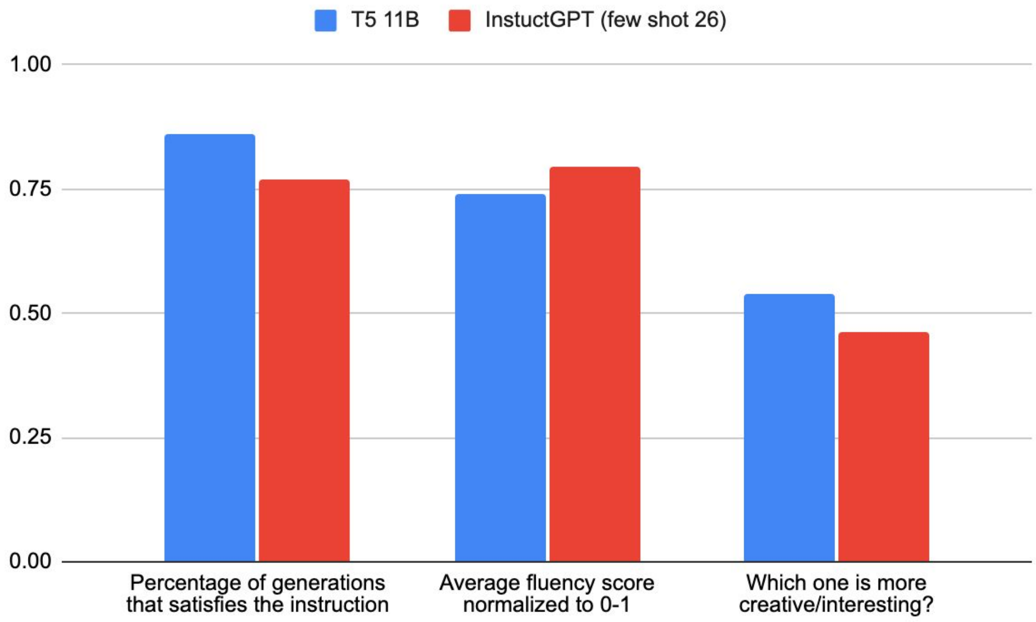
Known Templates		T5 - 11B	T5 - 3B		tGPT- t (175B)	InstructGPT - Few Shot (175B)	T0pp (11B)
Subject (51)	% - Match	92.15%	92.15%		7%	**96.07%**	84.31%
	% - Match w/ Ending (22)	**95.45%**	**95.45%**		3%	18.18%	40.90%
Rhyme (14)	Success Rate	78.57%				71.42%	0.00%
Simile (6)	% Subject + Comparator	66.66%	**83.33%**		3%	66.66%	16.66%
	% Comparator	**100.00%**	**100%**		%	83.33%	16.66%
Metaphor (5)	% Subject + Comparator	**100.00%**	**100.00%**		%	**100%**	0.00%
Haiku (5)	% Subject + (15-19) Syllables	20.00%	40.00%	0.00%	**80%**	40%	0.00%

InstructGPT is good at staying on subject.

But performance drops on harder instructions

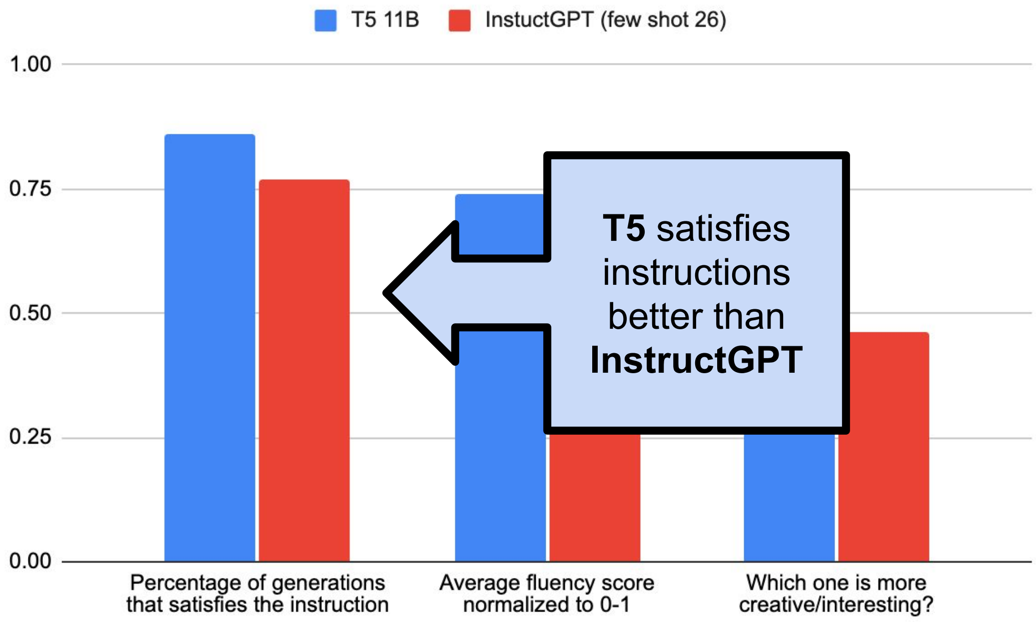
How Well Do Models Follow Instructions?

- Sample model generations for each instruction
- Majority vote from 3 human annotators on each of the following axes



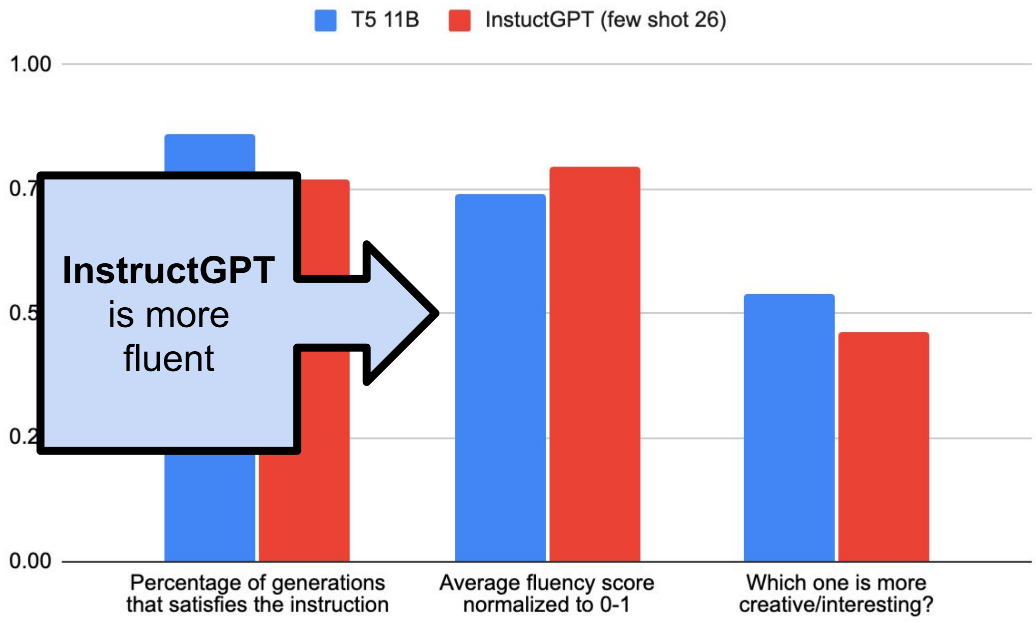
How Well Do Models Follow Instructions?

- Sample model generations for each instruction
- Majority vote from 3 human annotators on each of the following axes



How Well Do Models Follow Instructions?

- Sample model generations for each instruction
- Majority vote from 3 human annotators on each of the following axes



Research Questions

- Can we train LLMs to satisfy creative writing instructions for poetry writing tasks?
- **Can models compose instructions seen at train time in unseen combinations?**
- Can we help users complete creative writing tasks using natural language instructions?

How Well Do Models Compose Instructions?

Compositional Test Set