Does Writing With Language Models (LLMs) Reduce Content Diversity?

Vishakh Padmakumar, He He

ICLR 2024

NYU ML² Machine Learning for Language

Background: Collaborative Writing

- **Broad Direction:**
  - How can we assist writers at various writing tasks?
  - What is the impact of model assistance on the writing process?

Background: Collaborative Writing

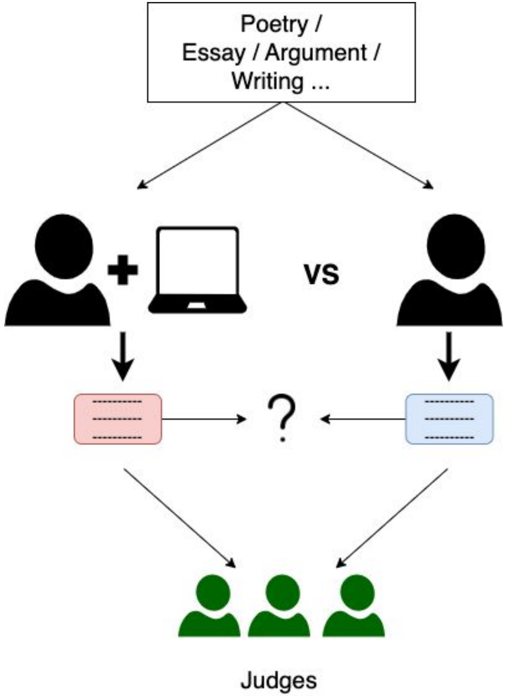
- **Broad Direction:**
  - How can we assist writers at various writing tasks?
  - What is the impact of model assistance on the writing process?

- Predates LLMs and contemporary NLP methods to work on retrieval and dictionary based systems [1, 2]

[1] Roemmele, Melissa, and Andrew S. Gordon. "Creative help: A story writing assistant." *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30-December 4, 2015, Proceedings 8*. Springer International Publishing, 2015.
[2] Kim, Joy, et al. "Mechanical novel: Crowdsourcing complex work through reflection and revision." *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 2017.

Background: Collaborative Writing

- **Broad Direction:** What is the impact of model assistance on the writing process?



Motivation

Motivation

- Evidence that LLMs can influence the views of users during co-writing [1,2]

[1] Jakesch, Maurice, et al. "Co-writing with opinionated language models affects users' views." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
[2] Bhat, Advait, et al. "Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing." *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023.

Motivation

- Evidence that LLMs can influence the views of users during co-writing [1,2]
- As different users rely on the same model for suggestions, this creates an algorithmic monoculture[3]

[1] Jakesch, Maurice, et al. "Co-writing with opinionated language models affects users' views." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
[2] Bhat, Advait, et al. "Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing." *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023.
[3] Kleinberg, Jon, and Manish Raghavan. "Algorithmic monoculture and social welfare." *Proceedings of the National Academy of Sciences* 118.22 (2021): e2018340118.

Motivation

- Evidence that LLMs can influence the views of users during co-writing [1,2]
- As different users rely on the same model for suggestions, this creates an algorithmic monoculture[3]

Does collaborative with LLMs result in different users writing more similar text, reducing the overall diversity of content produced?

[1] Jakesch, Maurice, et al. "Co-writing with opinionated language models affects users' views." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
[2] Bhat, Advait, et al. "Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing." *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023.
[3] Kleinberg, Jon, and Manish Raghavan. "Algorithmic monoculture and social welfare." *Proceedings of the National Academy of Sciences* 118.22 (2021): e2018340118.

User Study Format

User Study Format

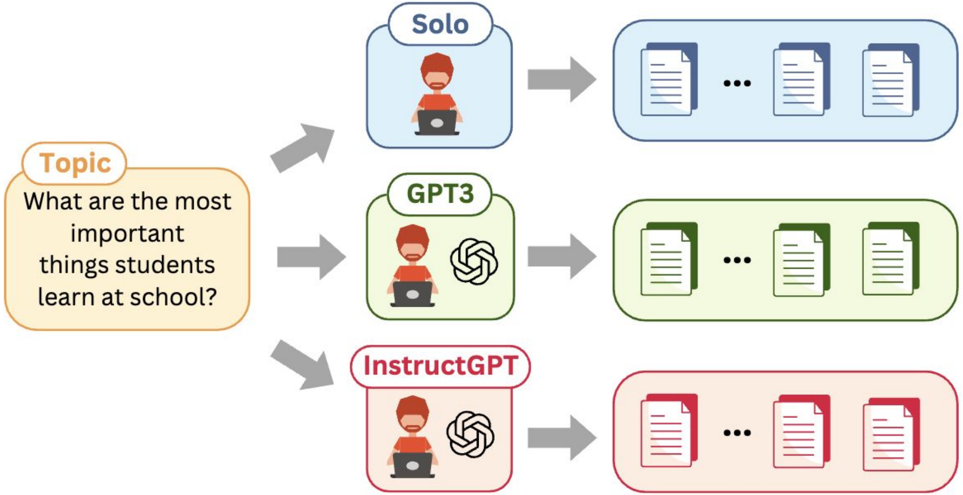
- **Task:** Argumentative Essay writing (~300-500 words) on a set of 10 open ended questions as collected by NYT
    - Example: What are the most important things students should learn at school?

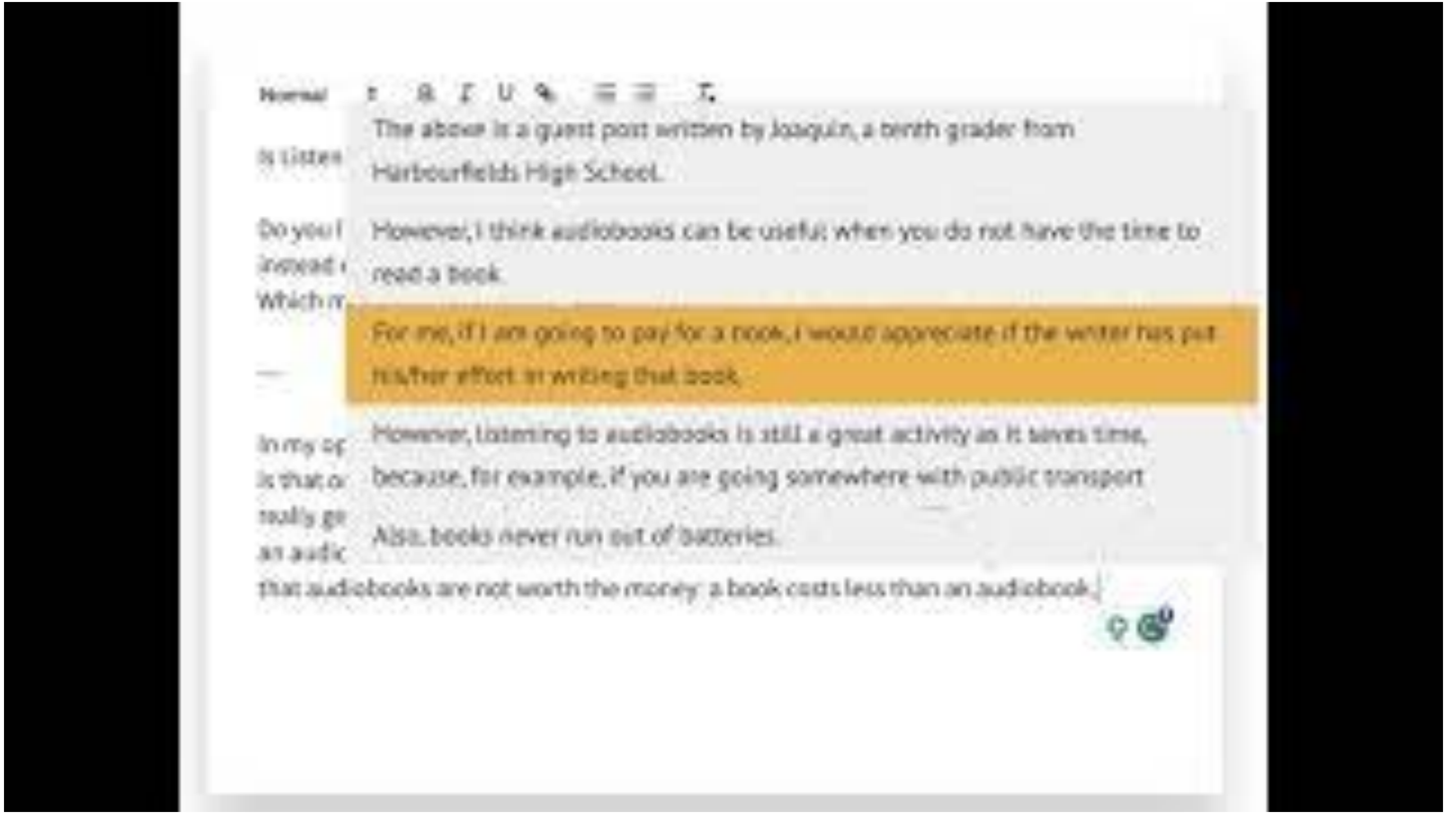
User Study Format

- **Task:** Argumentative Essay writing (~300-500 words) on a set of 10 open ended questions as collected by NYT
- (Semi) Professional writers from Upwork

User Study Format

- **Task:** Argumentative Essay writing (~300-500 words) on a set of 10 open ended questions as collected by NYT
- (Semi) Professional writers from Upwork writing with and without model help


Topic
What are the most important things students learn at school?
Solo
...
GPT3
...
InstructGPT
...



The above is a guest post written by Joaquin, a tenth grader from Harbourfields High School.

However, I think audiobooks can be useful when you do not have the time to read a book.

For me, if I am going to pay for a book, I would appreciate if the writer has put his/her effort in writing that book.

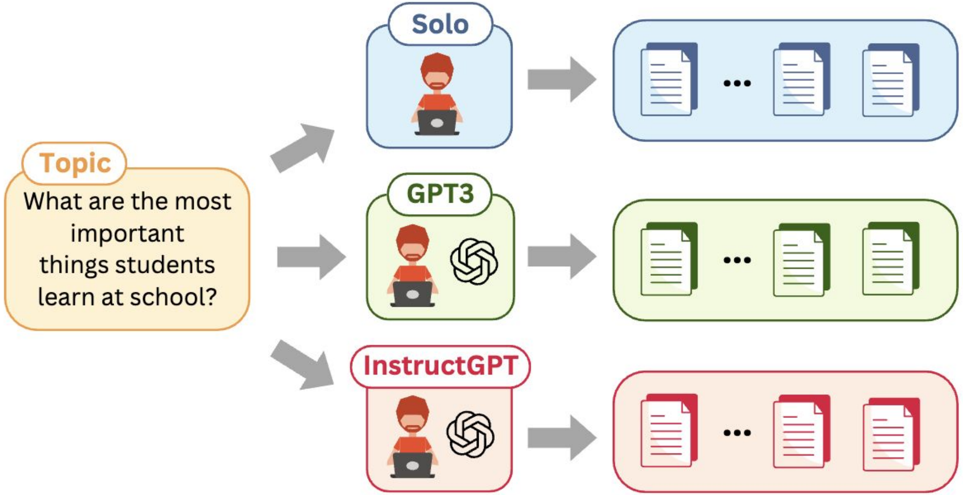
However, listening to audiobooks is still a great activity as it saves time, because, for example, if you are going somewhere with public transport

Also, books never run out of batteries.

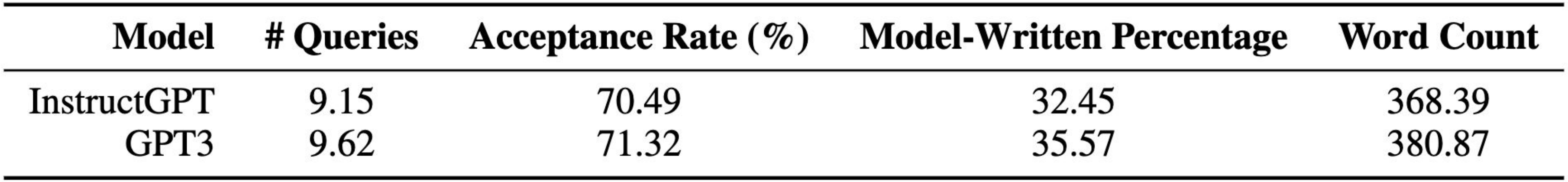
that audiobooks are not worth the money: a book costs less than an audiobook.

User Study Format

- **Task:** Argumentative essay writing (~300-500 words) on a set of 10 open ended questions as collected by NYT
- (Semi) Professional writers from Upwork writing with and without model help
- 10 topics x 10 responses = 100 essays from each setup to compare


Topic
What are the most important things students learn at school?
Solo
GPT3
InstructGPT
...
...
...

Users Find Both Models Equally Helpful for Collaborative Writing



Model	# Queries	Acceptance Rate (%)	Model-Written Percentage	Word Count
InstructGPT	9.15	70.49	32.45	368.39
GPT3	9.62	71.32	35.57	380.87

How Can We Compare the Content of Essays?

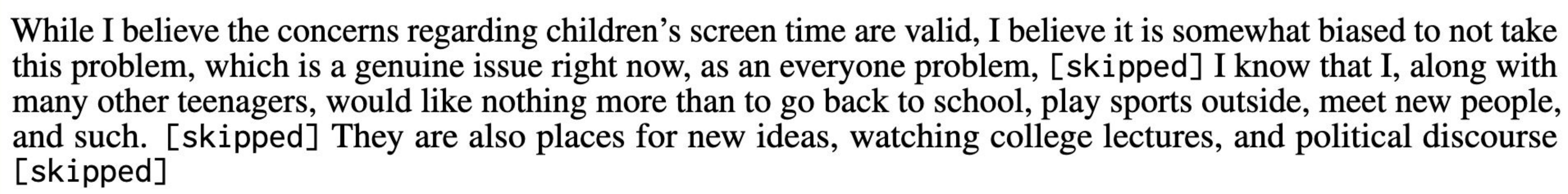
How Can We Compare the Content of Essays?

Example Essay:
Topic: How Worried Should We Be About Screen Time During the Pandemic?

How Can We Compare the Content of Essays?

Example Essay:

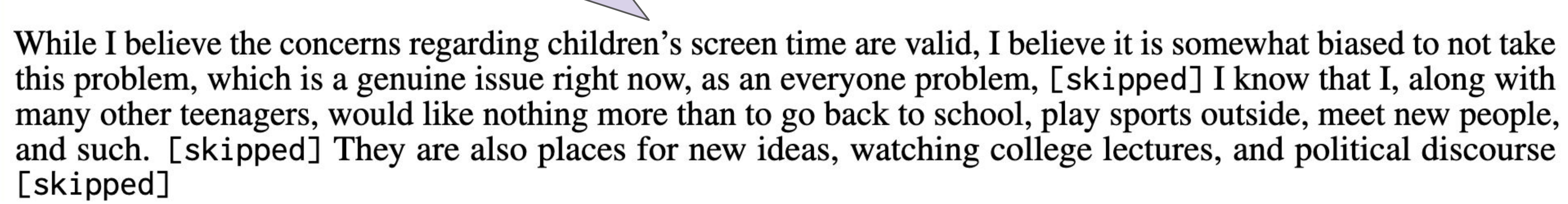
Topic: How Worried Should We Be About Screen Time During the Pandemic?



While I believe the concerns regarding children's screen time are valid, I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem, `[skipped]` I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. `[skipped]` They are also places for new ideas, watching college lectures, and political discourse `[skipped]`

How Can We Compare the Content of Essays?

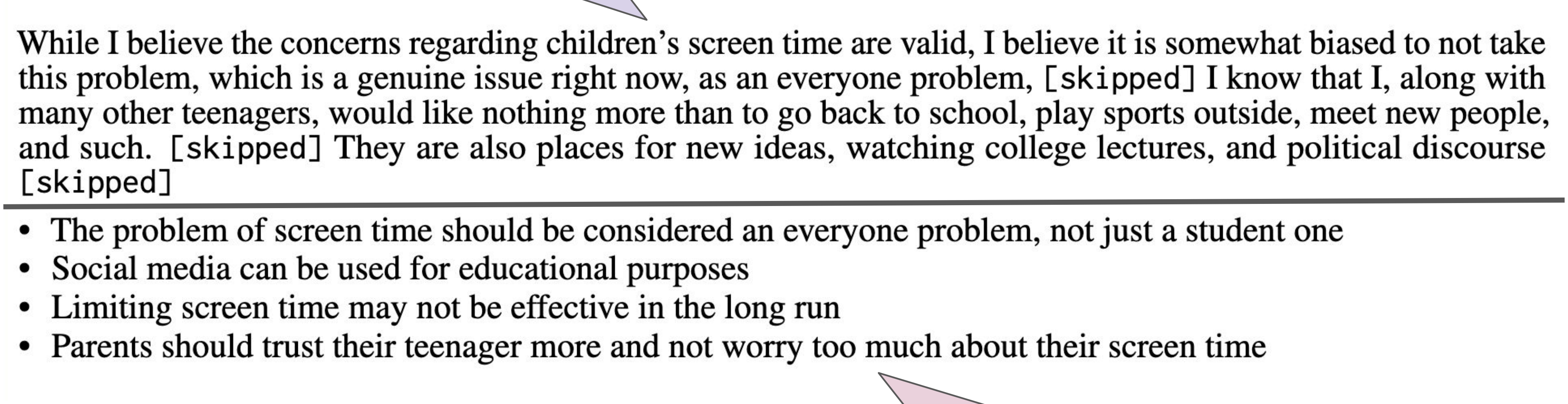
Raw Text Level



While I believe the concerns regarding children’s screen time are valid, I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem, [skipped] I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. [skipped] They are also places for new ideas, watching college lectures, and political discourse [skipped]

Comparing the Content of Essays via Summarization

Raw Text Level



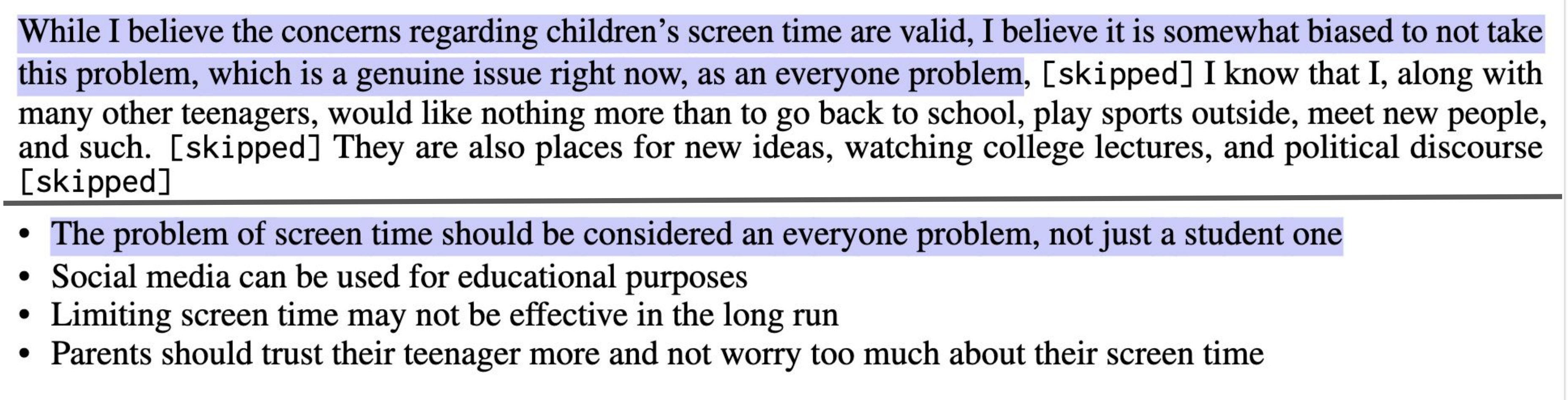
While I believe the concerns regarding children’s screen time are valid, I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem, `[skipped]` I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. `[skipped]` They are also places for new ideas, watching college lectures, and political discourse `[skipped]`

- The problem of screen time should be considered an everyone problem, not just a student one
- Social media can be used for educational purposes
- Limiting screen time may not be effective in the long run
- Parents should trust their teenager more and not worry too much about their screen time

Key Point Level

Does the Model Contribute to These Key Points?

Step 1: Aligning key points to sentences via Rouge-L

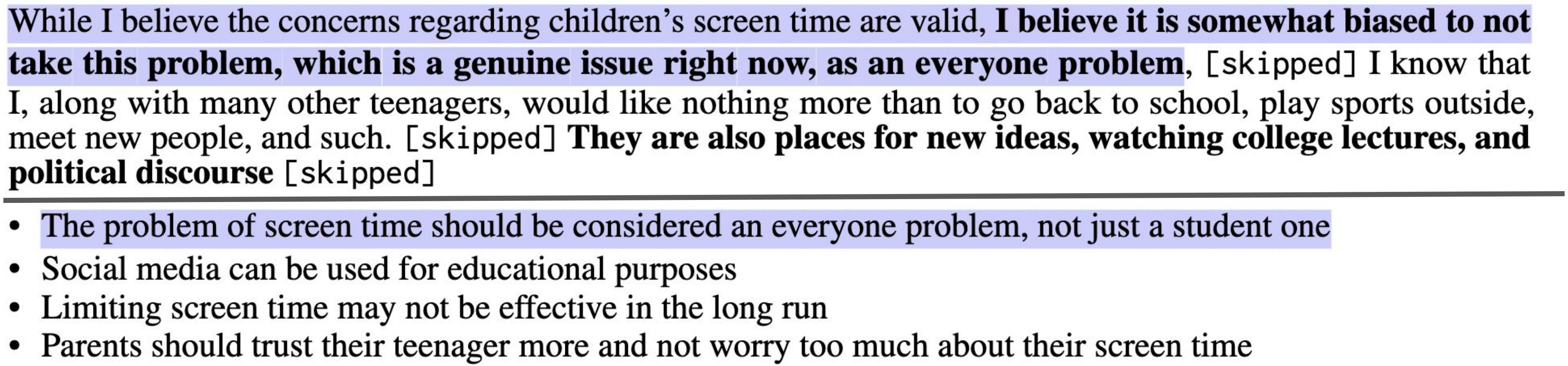


While I believe the concerns regarding children’s screen time are valid, I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem, [skipped] I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. [skipped] They are also places for new ideas, watching college lectures, and political discourse [skipped]

- The problem of screen time should be considered an everyone problem, not just a student one
- Social media can be used for educational purposes
- Limiting screen time may not be effective in the long run
- Parents should trust their teenager more and not worry too much about their screen time

Does the Model Contribute to These Key Points?

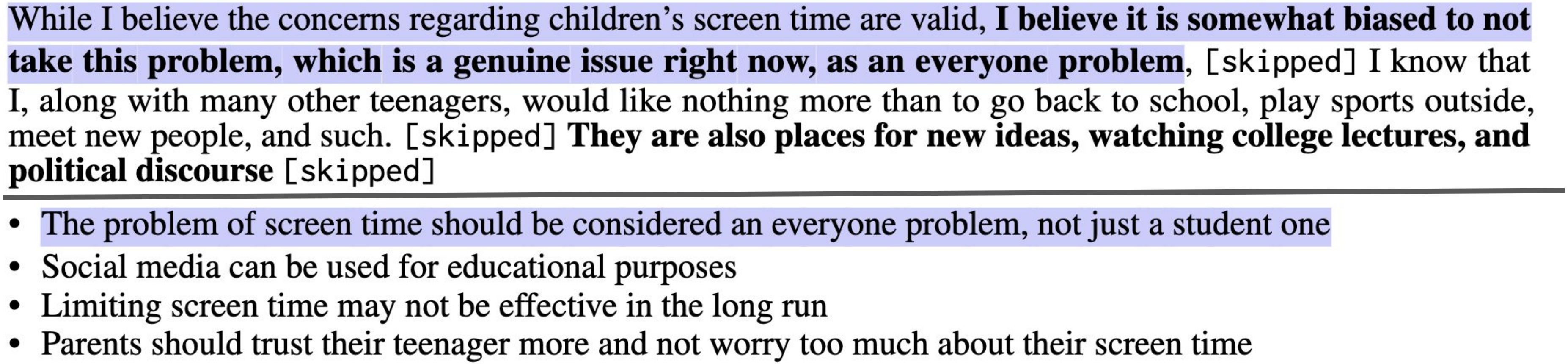
Step 1: Aligning key points to sentences via Rouge-L
Step 2: Attributing authorship of key points as User vs Model



While I believe the concerns regarding children’s screen time are valid, **I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem**, [skipped] I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. [skipped] **They are also places for new ideas, watching college lectures, and political discourse** [skipped]

- The problem of screen time should be considered an everyone problem, not just a student one
- Social media can be used for educational purposes
- Limiting screen time may not be effective in the long run
- Parents should trust their teenager more and not worry too much about their screen time

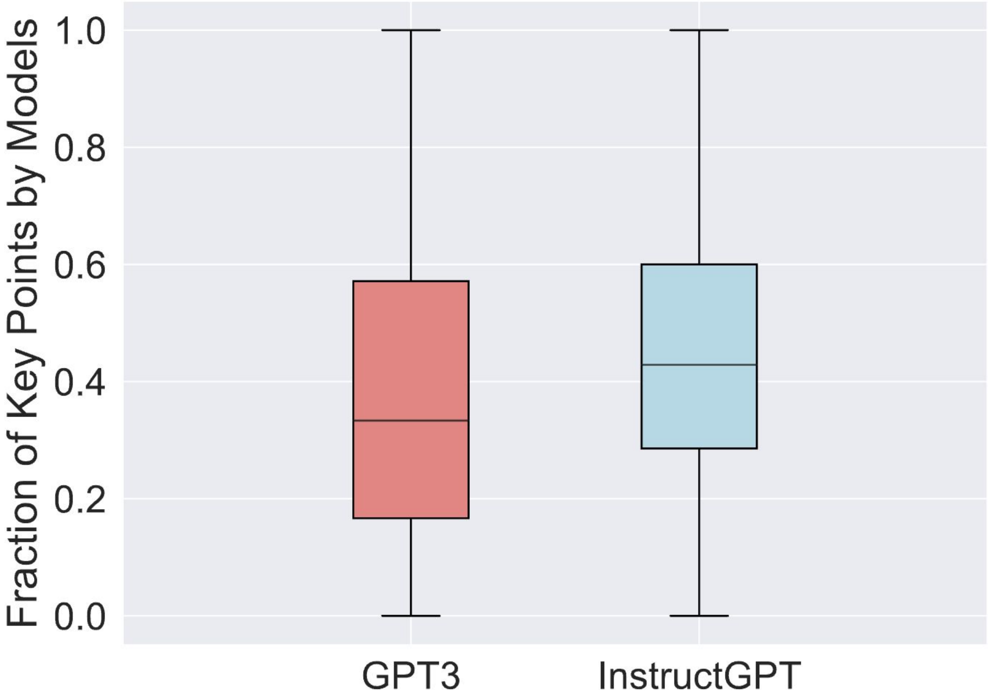
Allows for Analysis of Model Contribution to Keypoints



While I believe the concerns regarding children’s screen time are valid, **I believe it is somewhat biased to not take this problem, which is a genuine issue right now, as an everyone problem**, [skipped] I know that I, along with many other teenagers, would like nothing more than to go back to school, play sports outside, meet new people, and such. [skipped] **They are also places for new ideas, watching college lectures, and political discourse** [skipped]

- The problem of screen time should be considered an everyone problem, not just a student one
- Social media can be used for educational purposes
- Limiting screen time may not be effective in the long run
- Parents should trust their teenager more and not worry too much about their screen time

Users Use Model Suggestions to Write Key Points in the Essays



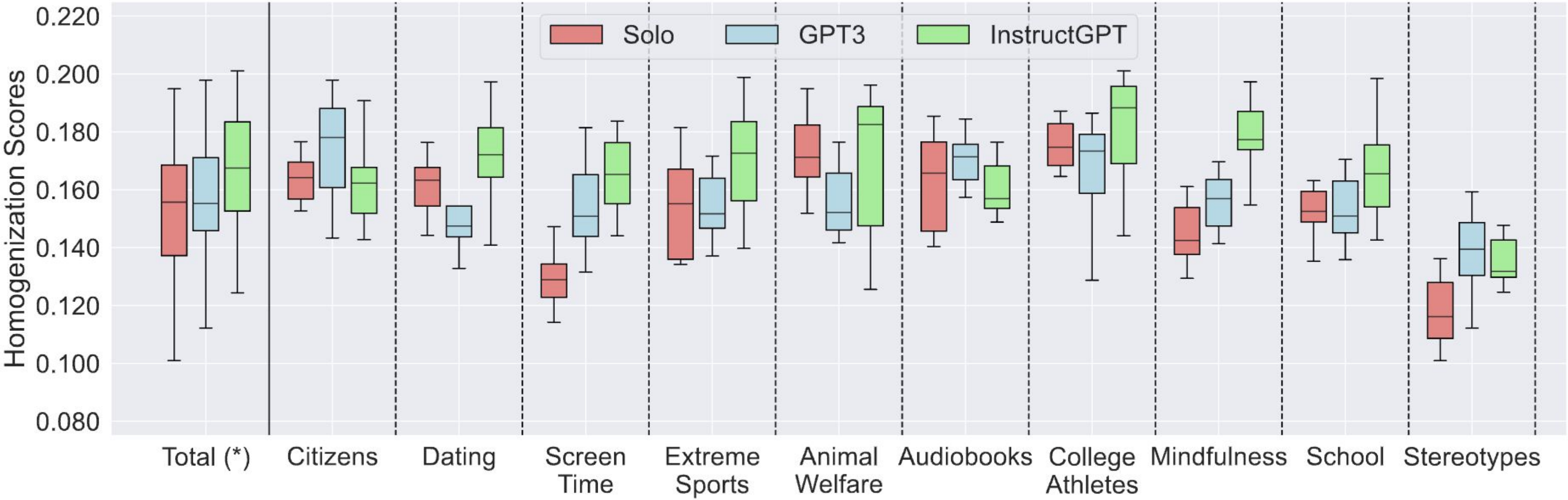
RQ1: Does Writing With LLMs Result in More Similar Essays?

Formalize Homogenization Using Pairwise Similarity

We calculate the homogenization of an essay ‘d’ written on topic ‘t’ as the average pairwise similarity to other documents (D_t) on that topic

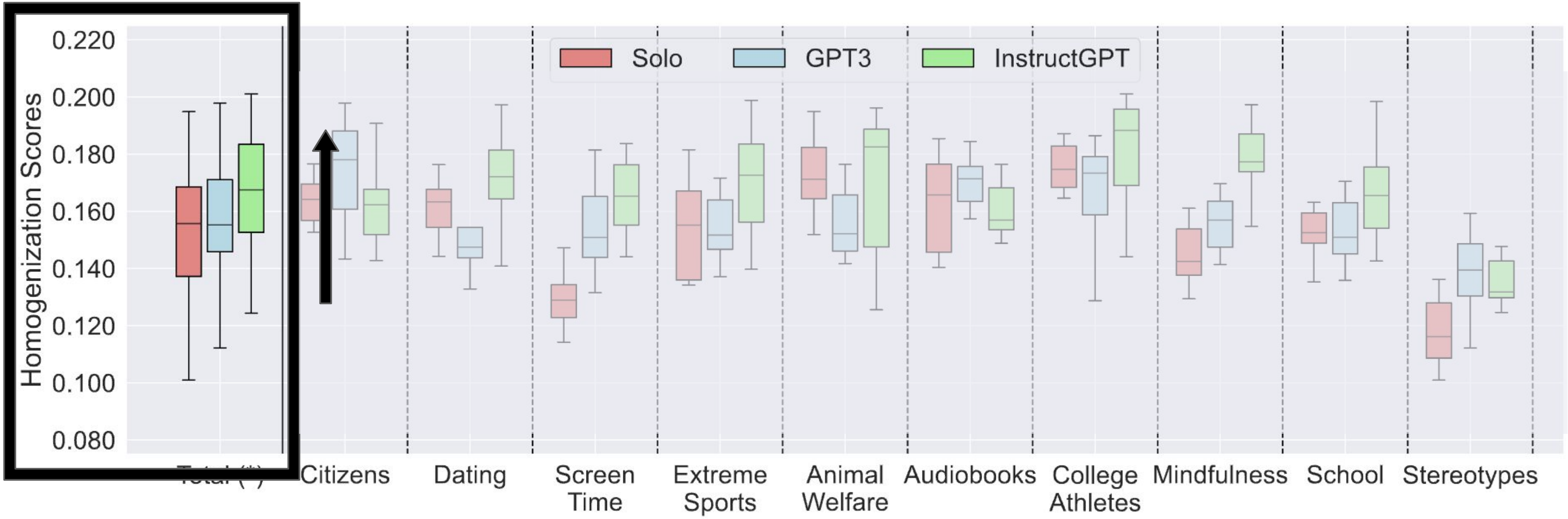
$$\mathrm{hom}(d \mid t) = \frac{1}{|D_t| - 1} \sum_{d' \in D_t \setminus d} \mathrm{sim}(d, d')$$

Results



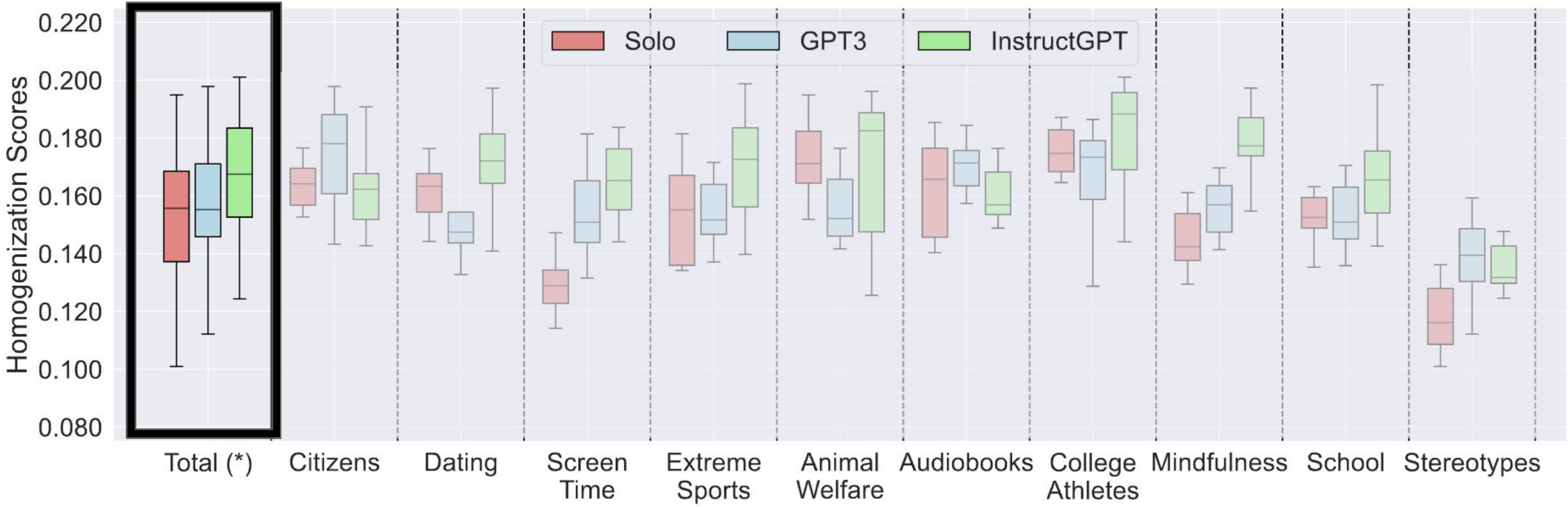
Homogenization at the key point level via Rouge-L

Higher homogenization implies more similar essays



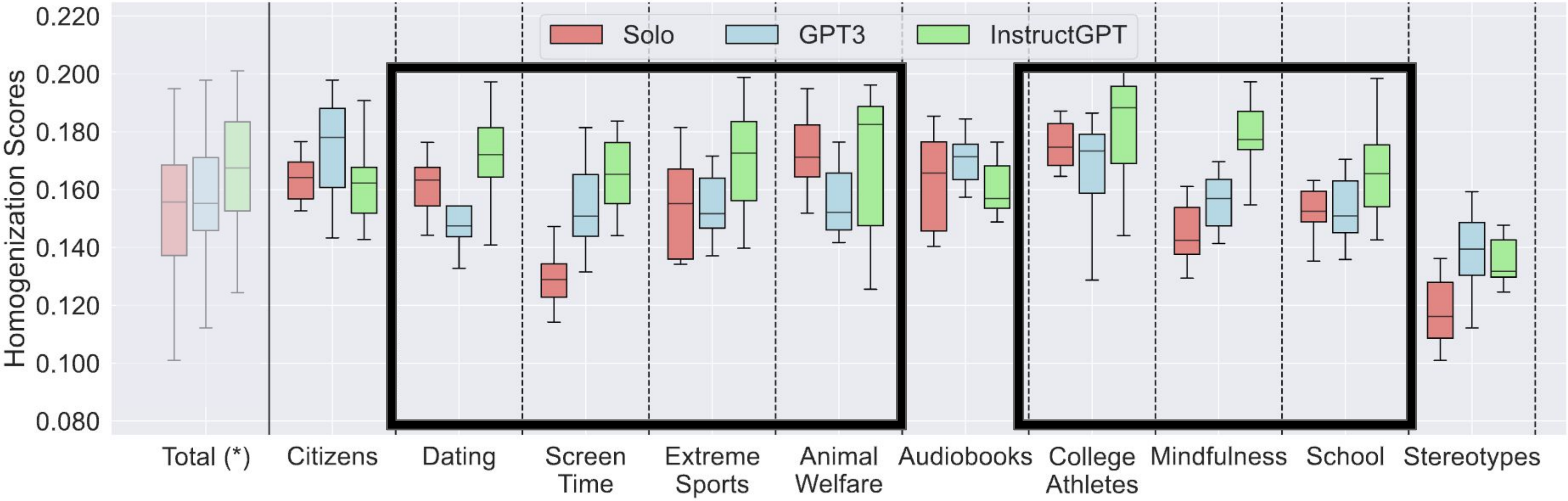
Homogenization at the key point level via Rouge-L

Writing with InstructGPT results in the highest average homogenization or most similar essays



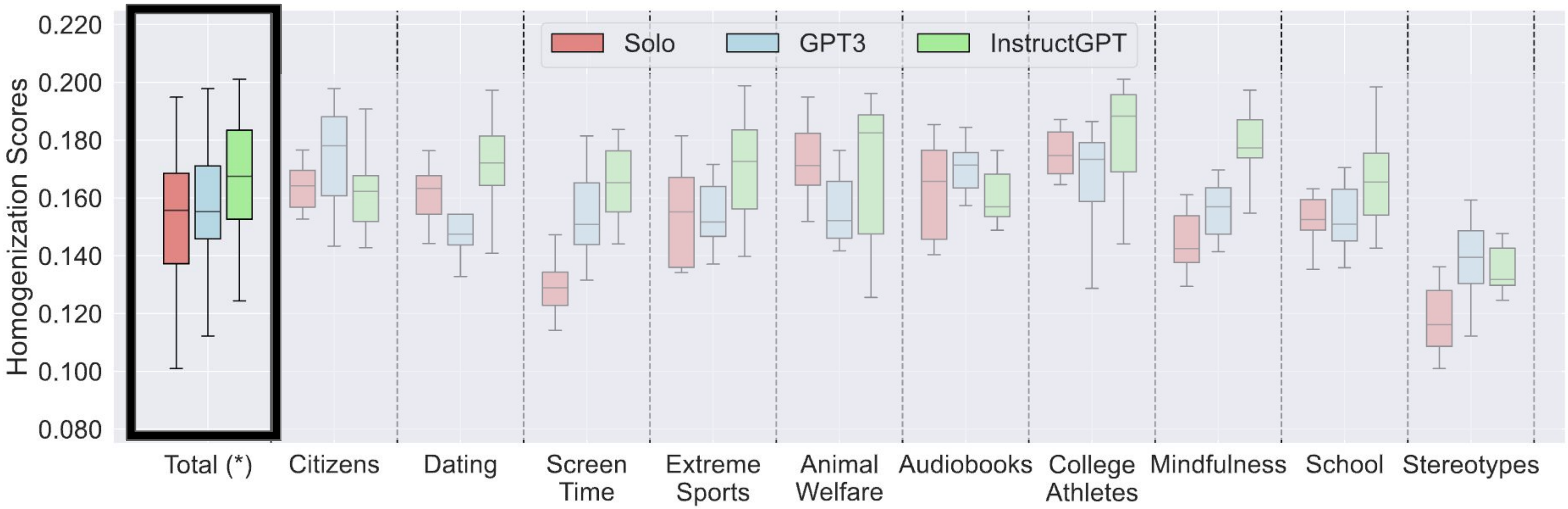
Homogenization at the key point level via Rouge-L

InstructGPT has the highest median homogenization in 7 out of 10 topics



Homogenization at the key point level via Rouge-L

Writing with GPT3 does not change the average homogenization from Solo Writers



Homogenization at the key point level via Rouge-L

RQ2: Does Writing With LLMs Reduce Overall Diversity?

Formalize Diversity Using Unique Information

We calculate the diversity of a set of essays D as the total amount of unique information in them

Formalize Diversity Using Unique Information

We calculate the diversity of a set of essays D as the total amount of unique information in them

Content Diversity:

Information Unit: Key Points

Diversity Measure: Fraction of Unique Clusters of Key Points

Example of Clustering of Key Points

Essay 1:

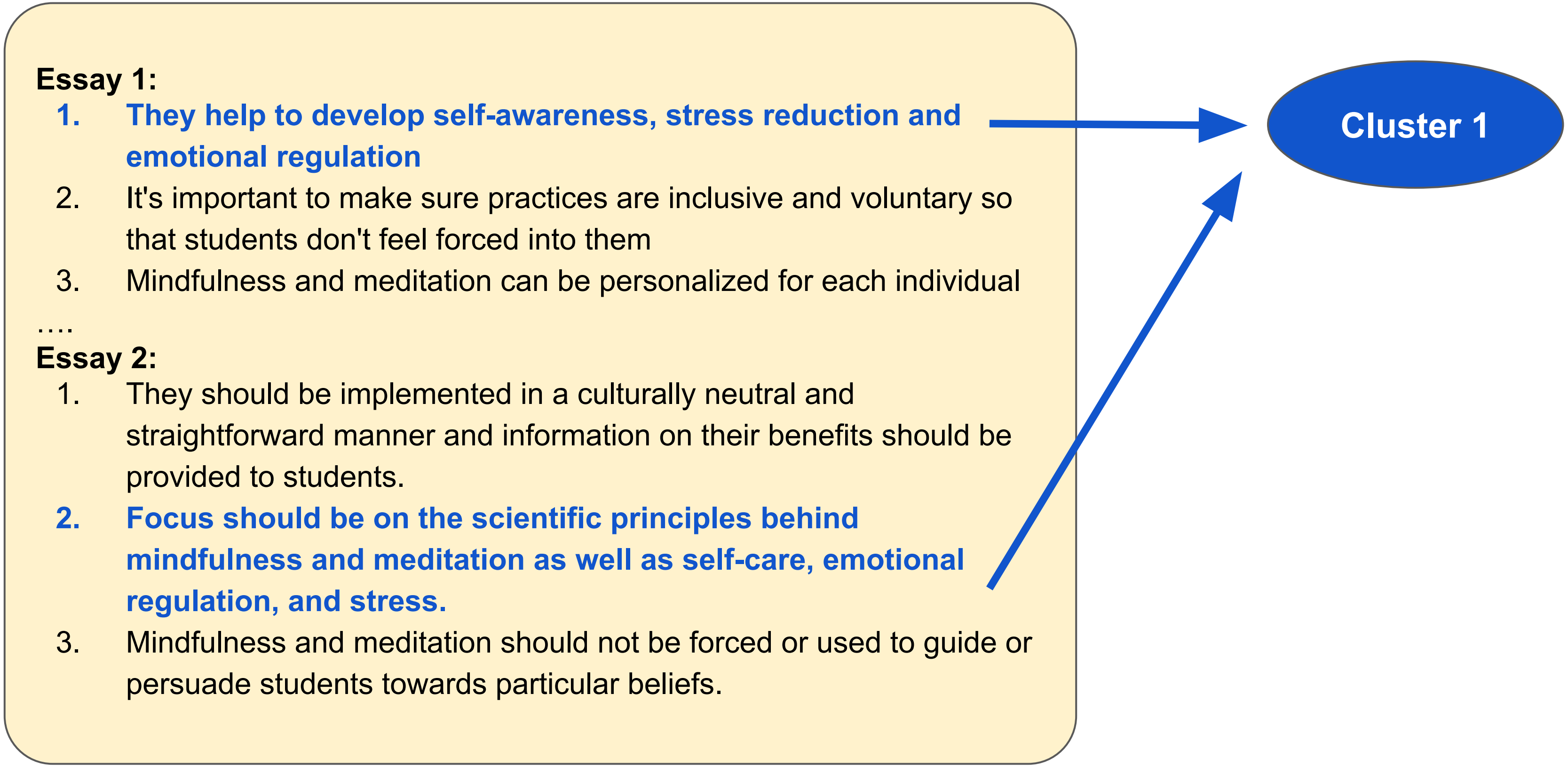
1. They help to develop self-awareness, stress reduction and emotional regulation
2. It's important to make sure practices are inclusive and voluntary so that students don't feel forced into them
3. Mindfulness and meditation can be personalized for each individual

….

Essay 2:

1. They should be implemented in a culturally neutral and straightforward manner and information on their benefits should be provided to students.
2. Focus should be on the scientific principles behind mindfulness and meditation as well as self-care, emotional regulation, and stress.
3. Mindfulness and meditation should not be forced or used to guide or persuade students towards particular beliefs.

Example of Clustering of Key Points



Essay 1:

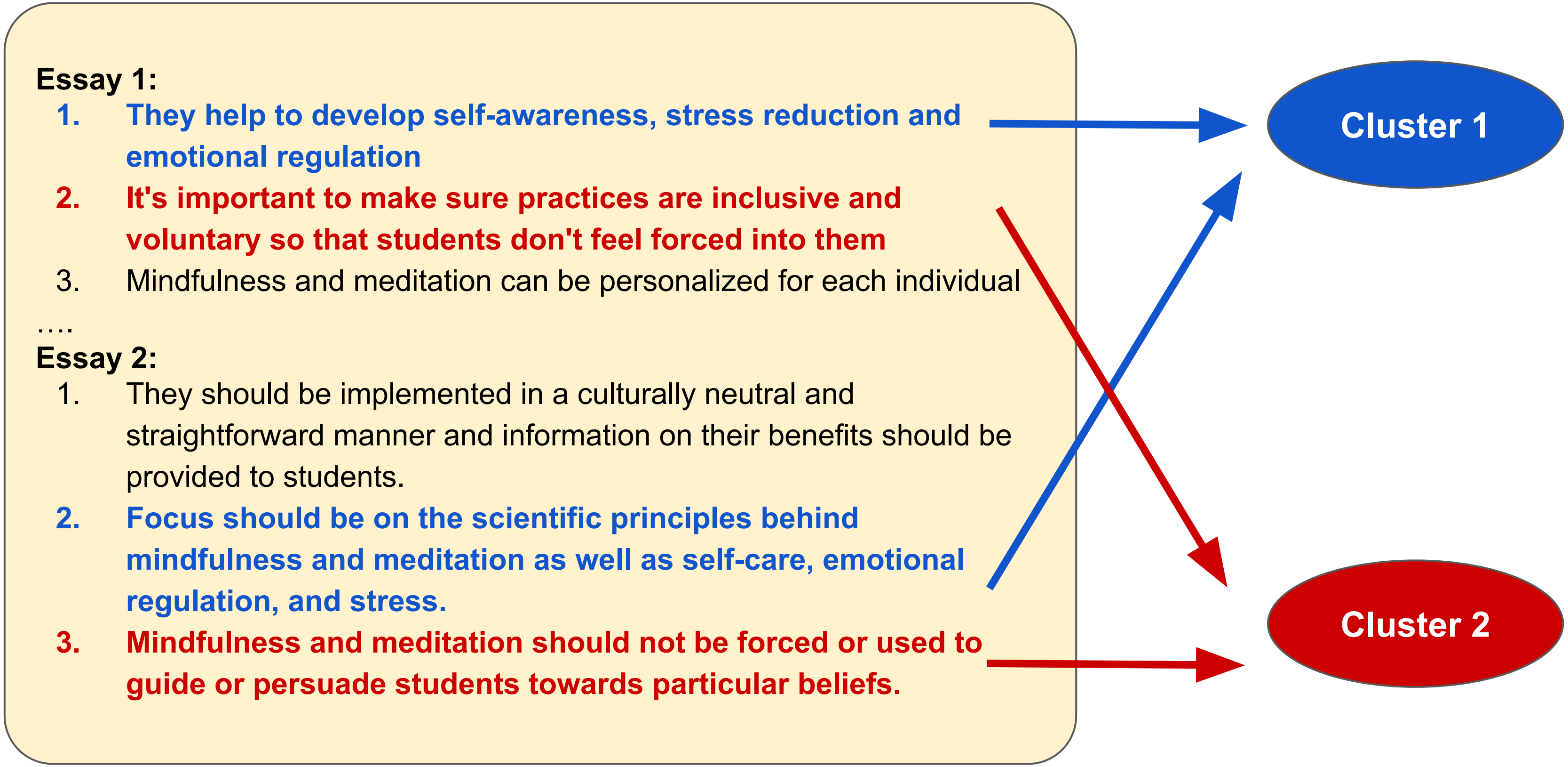
1. **They help to develop self-awareness, stress reduction and emotional regulation**
2. It's important to make sure practices are inclusive and voluntary so that students don't feel forced into them
3. Mindfulness and meditation can be personalized for each individual

….

Essay 2:

1. They should be implemented in a culturally neutral and straightforward manner and information on their benefits should be provided to students.
2. **Focus should be on the scientific principles behind mindfulness and meditation as well as self-care, emotional regulation, and stress.**
3. Mindfulness and meditation should not be forced or used to guide or persuade students towards particular beliefs.

Example of Clustering of Key Points



Essay 1:

1. **They help to develop self-awareness, stress reduction and emotional regulation**
2. **It's important to make sure practices are inclusive and voluntary so that students don't feel forced into them**
3. Mindfulness and meditation can be personalized for each individual

….

Essay 2:

1. They should be implemented in a culturally neutral and straightforward manner and information on their benefits should be provided to students.
2. **Focus should be on the scientific principles behind mindfulness and meditation as well as self-care, emotional regulation, and stress.**
3. **Mindfulness and meditation should not be forced or used to guide or persuade students towards particular beliefs.**

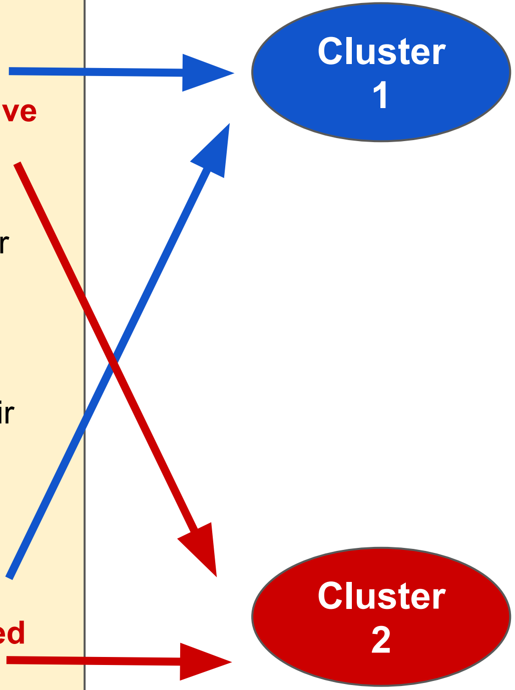
Example of Clustering of Key Points

Essay 1:

1. **They help to develop self-awareness, stress reduction and emotional regulation**
2. **It's important to make sure practices are inclusive and voluntary so that students don't feel forced into them**
3. Mindfulness and meditation can be personalized for each individual

Essay 2:

1. They should be implemented in a culturally neutral and straightforward manner and information on their benefits should be provided to students.
2. **Focus should be on the scientific principles behind mindfulness and meditation as well as self-care, emotional regulation, and stress.**
3. **Mindfulness and meditation should not be forced or used to guide or persuade students towards particular beliefs.**



Cluster 1

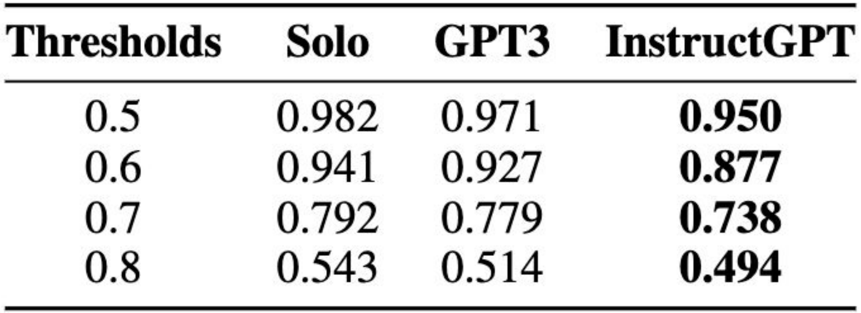
Cluster 2

Diversity =

Fraction of Unique Key Points =

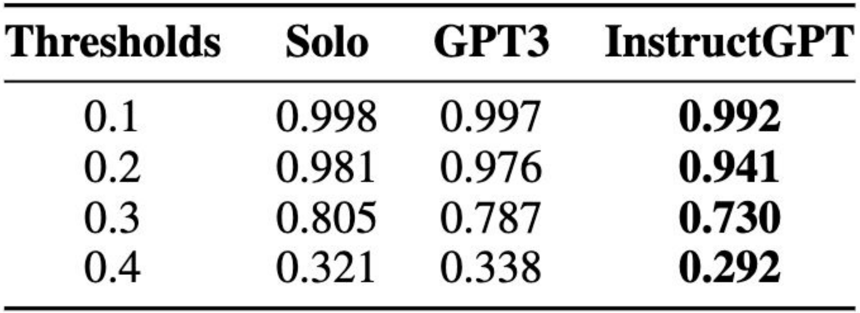
4 / 6 = 0.66

Results



Thresholds	Solo	GPT3	InstructGPT
0.5	0.982	0.971	**0.950**
0.6	0.941	0.927	**0.877**
0.7	0.792	0.779	**0.738**
0.8	0.543	0.514	**0.494**

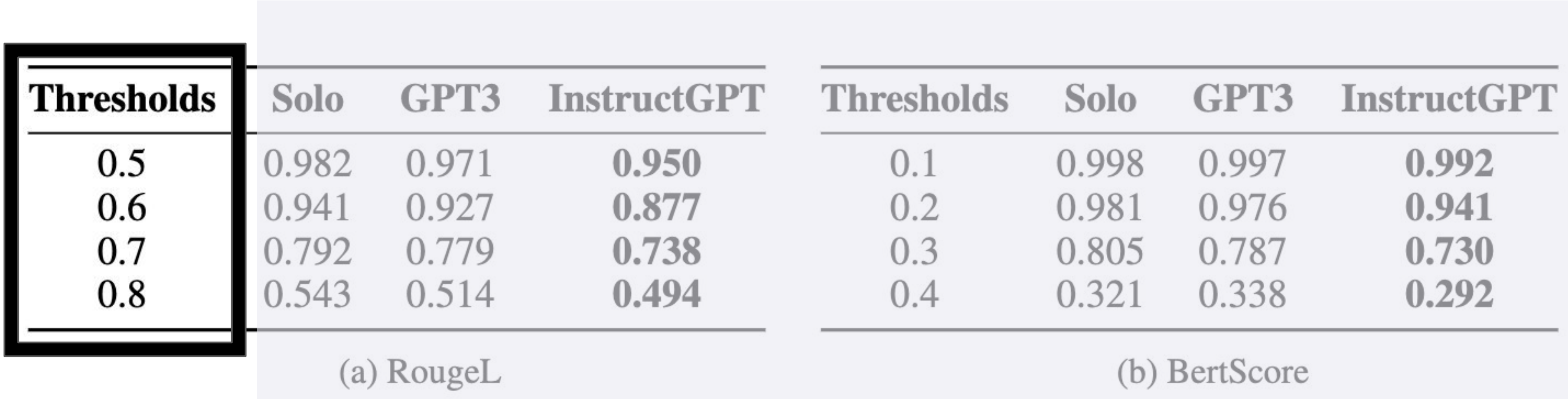
(a) RougeL



Thresholds	Solo	GPT3	InstructGPT
0.1	0.998	0.997	**0.992**
0.2	0.981	0.976	**0.941**
0.3	0.805	0.787	**0.730**
0.4	0.321	0.338	**0.292**

(b) BertScore

Effect of Thresholds on Clustering



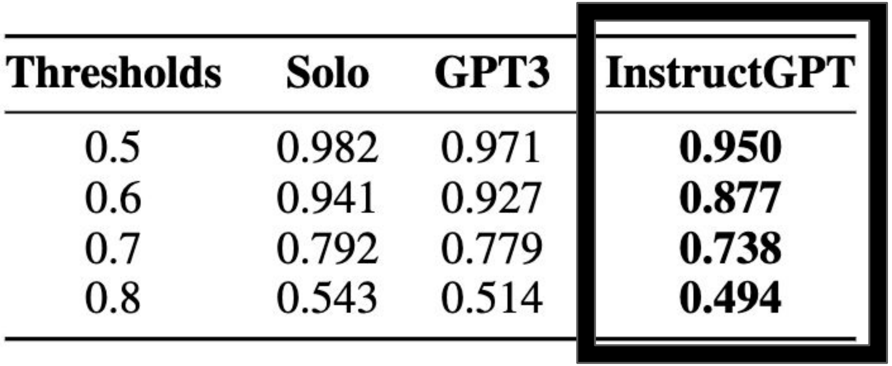
Thresholds	Solo	GPT3	InstructGPT
0.5	0.982	0.971	**0.950**
0.6	0.941	0.927	**0.877**
0.7	0.792	0.779	**0.738**
0.8	0.543	0.514	**0.494**

(a) RougeL

Thresholds	Solo	GPT3	InstructGPT
0.1	0.998	0.997	**0.992**
0.2	0.981	0.976	**0.941**
0.3	0.805	0.787	**0.730**
0.4	0.321	0.338	**0.292**

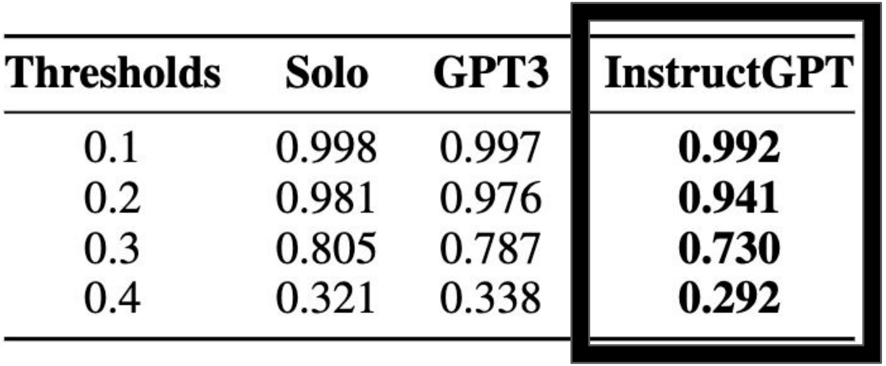
(b) BertScore

Writing with InstructGPT reduces key point diversity across both metrics and across all thresholds



Thresholds	Solo	GPT3	InstructGPT
0.5	0.982	0.971	**0.950**
0.6	0.941	0.927	**0.877**
0.7	0.792	0.779	**0.738**
0.8	0.543	0.514	**0.494**

(a) RougeL

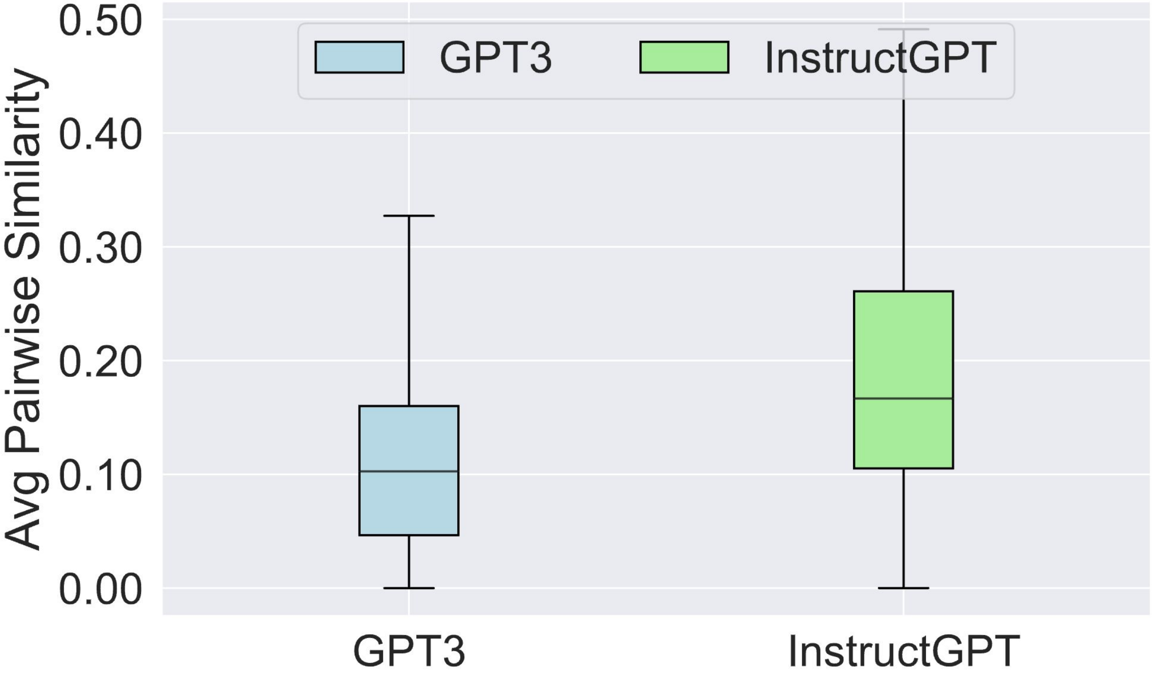


Thresholds	Solo	GPT3	InstructGPT
0.1	0.998	0.997	**0.992**
0.2	0.981	0.976	**0.941**
0.3	0.805	0.787	**0.730**
0.4	0.321	0.338	**0.292**

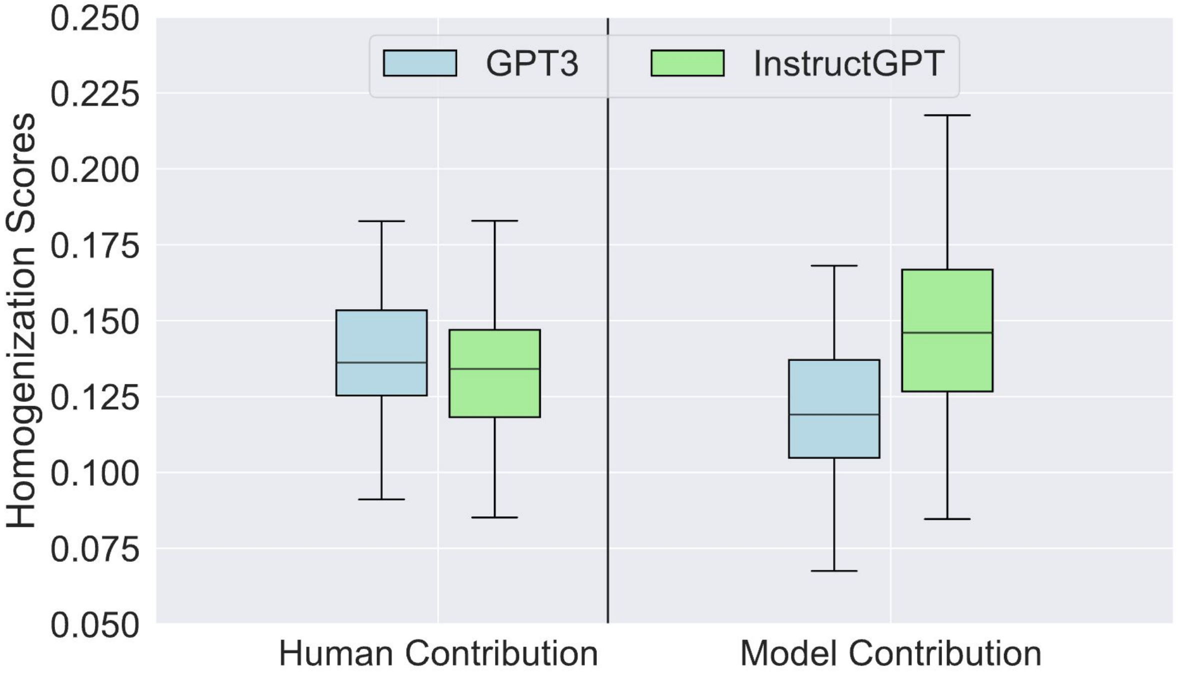
(b) BertScore

Why does InstructGPT have a stronger impact on diversity than GPT3?

InstructGPT presents users with more similar suggestions



The key points attributed to InstructGPT are more homogeneous than GPT3, user behavior is the same



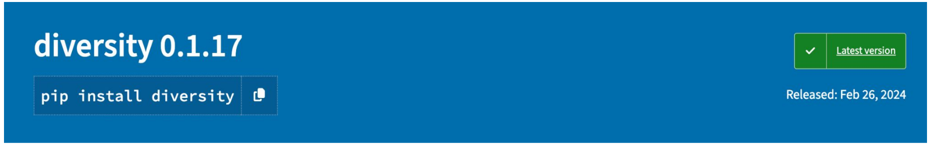
Takeaways

- Collaboration with InstructGPT makes users write more similar essays, reducing the overall diversity as well

Takeaways

- Collaboration with InstructGPT makes users write more similar essays, reducing the overall diversity as well
- This effect is not observed with GPT3 highlighting that the bump in performance from tuning the model on human feedback comes at the cost of more homogeneous content

Contemporary/Follow-up works support our findings :)

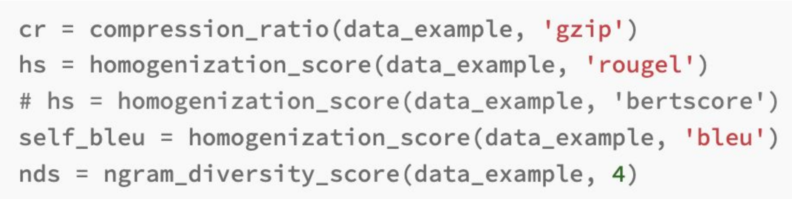


diversity 0.1.17

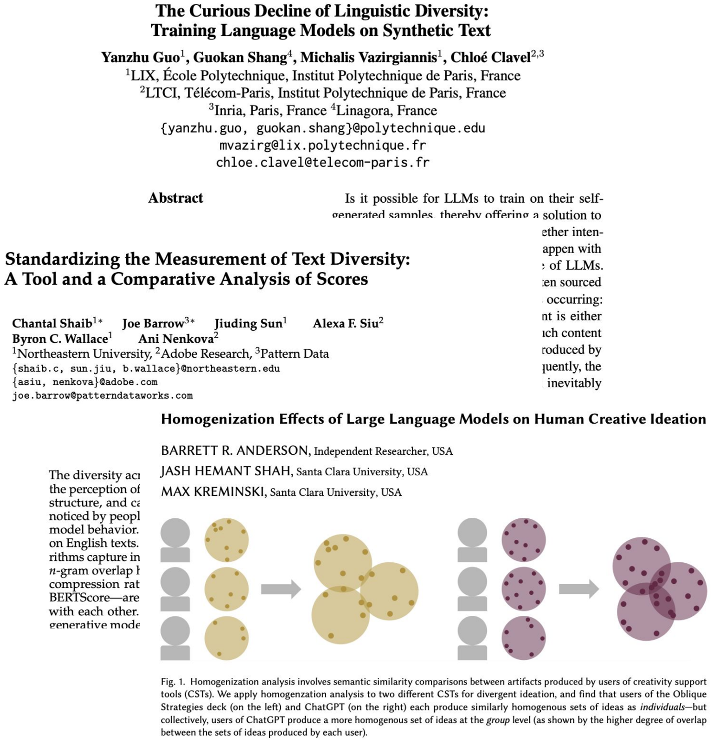
```
pip install diversity
```

✓ Latest version

Released: Feb 26, 2024



```
cr = compression_ratio(data_example, 'gzip')
hs = homogenization_score(data_example, 'rougel')
# hs = homogenization_score(data_example, 'bertscore')
self_bleu = homogenization_score(data_example, 'bleu')
nds = ngram_diversity_score(data_example, 4)
```



The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text

Yanzhu Guo[1], Guokan Shang[4], Michalis Vazirgiannis[1], Chloé Clavel[2,3]
[1]LIX, École Polytechnique, Institut Polytechnique de Paris, France
[2]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
[3]Inria, Paris, France [4]Linagora, France
{yanzhu.guo, guokan.shang}@polytechnique.edu
mvazirg@lix.polytechnique.fr
chloe.clavel@telecom-paris.fr

Abstract

Is it possible for LLMs to train on their self-generated samples, thereby offering a solution to
ether inten-
appen with
: of LLMs.
ten sourced
occurring:
nt is either
ich content
roduced by
quently, the
inevitably

Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores

Chantal Shaib[1*] **Joe Barrow**[3*] **Jiuding Sun**[1] **Alexa F. Siu**[2]
Byron C. Wallace[1] **Ani Nenkova**[2]
[1]Northeastern University, [2]Adobe Research, [3]Pattern Data
{shaib.c, sun.jiu, b.wallace}@northeastern.edu
{asiu, nenkova}@adobe.com
joe.barrow@patterndataworks.com

The diversity acı
the perception of
structure, and ca
noticed by peopl
model behavior.
on English texts.
rithms capture in
n-gram overlap h
compression rat
BERTScore—are
with each other.
generative mode

Homogenization Effects of Large Language Models on Human Creative Ideation

BARRETT R. ANDERSON, Independent Researcher, USA
JASH HEMANT SHAH, Santa Clara University, USA
MAX KREMINSKI, Santa Clara University, USA

Fig. 1. Homogenization analysis involves semantic similarity comparisons between artifacts produced by users of creativity support tools (CSTs). We apply homogenzation analysis to two different CSTs for divergent ideation, and find that users of the Oblique Strategies deck (on the left) and ChatGPT (on the right) each produce similarly homogenous sets of ideas as *individuals*—but collectively, users of ChatGPT produce a more homogenous set of ideas at the *group* level (as shown by the higher degree of overlap between the sets of ideas produced by each user).

Limitations

- Single interactions with users so unclear how long term behavior changes
- Ablations into the kind of interactions with users
- The writing-setting is still not natural i.e. we hire folks to perform tasks for us

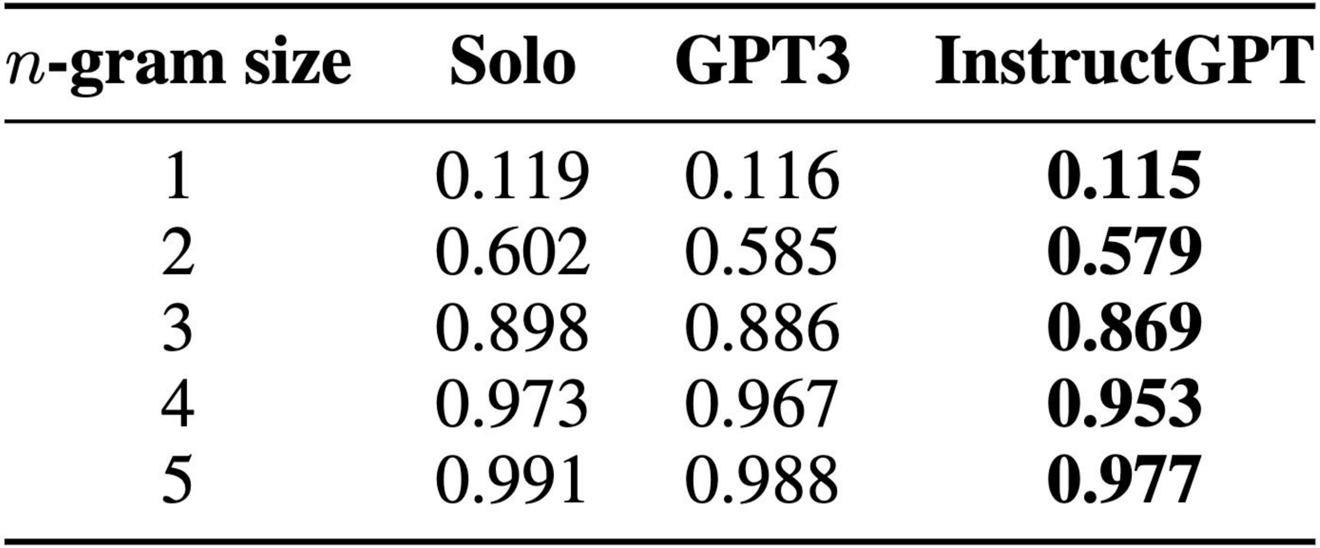
Questions and Comments?

Limitations

- Single interactions with users so unclear how long term behavior changes
- Ablations into the kind of interactions with users
- The writing-setting is still not natural i.e. we hire folks to perform tasks for us

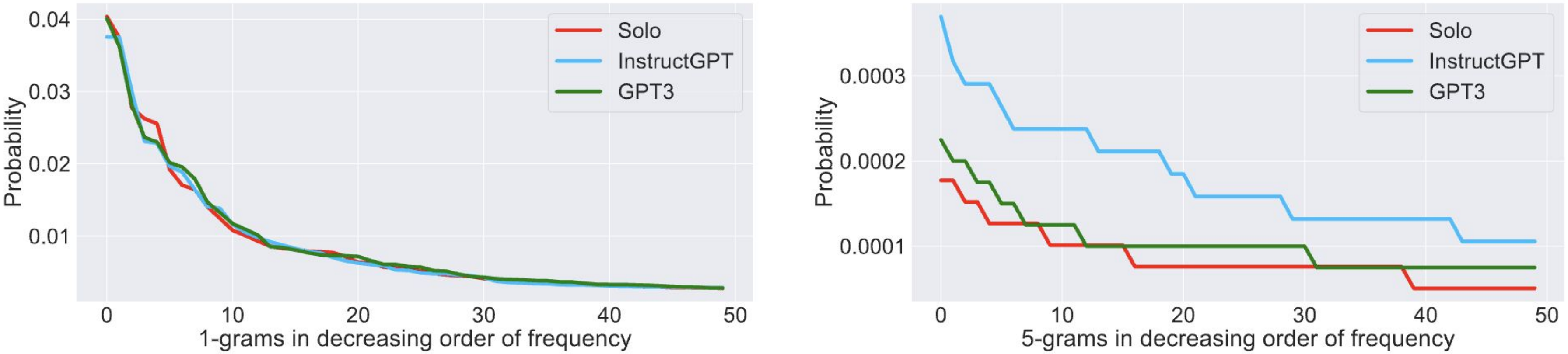
Backup Slides

Writing with InstructGPT reduces lexical diversity



n-gram size	Solo	GPT3	InstructGPT
1	0.119	0.116	**0.115**
2	0.602	0.585	**0.579**
3	0.898	0.886	**0.869**
4	0.973	0.967	**0.953**
5	0.991	0.988	**0.977**

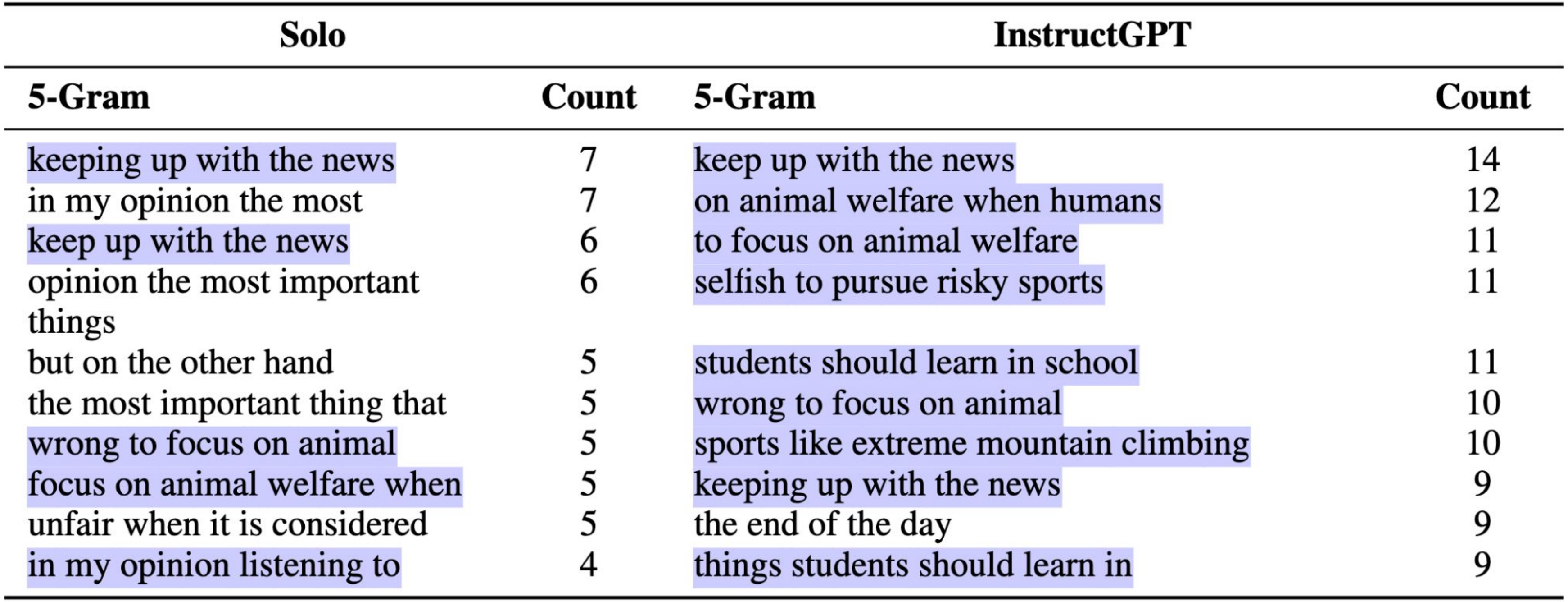
The reduced lexical diversity with InstructGPT is also manifested in frequent repetition of higher-order N-grams.



(a) Unigram Distribution

(b) 5-gram Distribution

Writing with InstructGPT leads to repeated 5-Grams containing topic-specific phrases.



Solo		InstructGPT	
5-Gram	**Count**	**5-Gram**	**Count**
keeping up with the news	7	keep up with the news	14
in my opinion the most	7	on animal welfare when humans	12
keep up with the news	6	to focus on animal welfare	11
opinion the most important things	6	selfish to pursue risky sports	11
but on the other hand	5	students should learn in school	11
the most important thing that	5	wrong to focus on animal	10
wrong to focus on animal	5	sports like extreme mountain climbing	10
focus on animal welfare when	5	keeping up with the news	9
unfair when it is considered	5	the end of the day	9
in my opinion listening to	4	things students should learn in	9

Limitations

- Single interactions with users so unclear how long term behavior changes
- Ablations into the kind of interactions with users
- The writing setting is still not natural i.e. we hire folks to perform tasks for us

How do real users feel about this assistive technology?

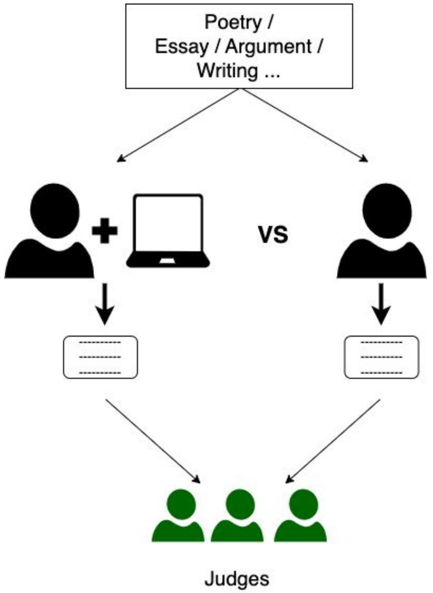
Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers

Overview - Collaborative Writing

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

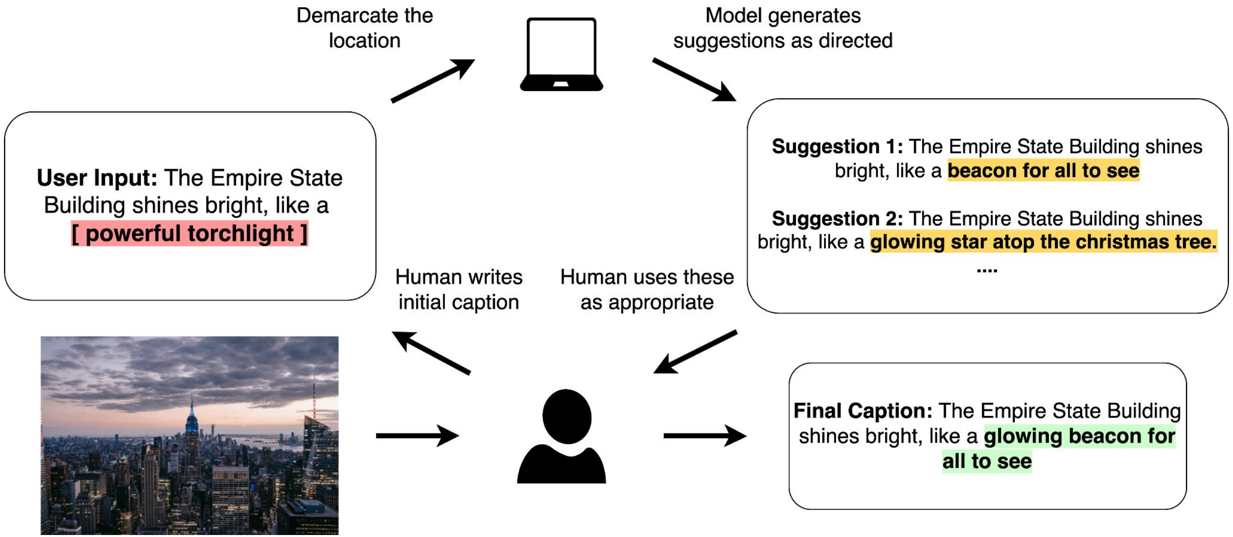
Overview (Upto 2022)

Phase 1: Train models to solve specific user needs and demonstrate how writers achieve their goals in these interactive setups



Overview (Upto 2022)

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



Demarcate the location

Model generates suggestions as directed

User Input: The Empire State Building shines bright, like a **[powerful torchlight]**

Suggestion 1: The Empire State Building shines bright, like a **beacon for all to see**

Suggestion 2: The Empire State Building shines bright, like a **glowing star atop the christmas tree.**

....

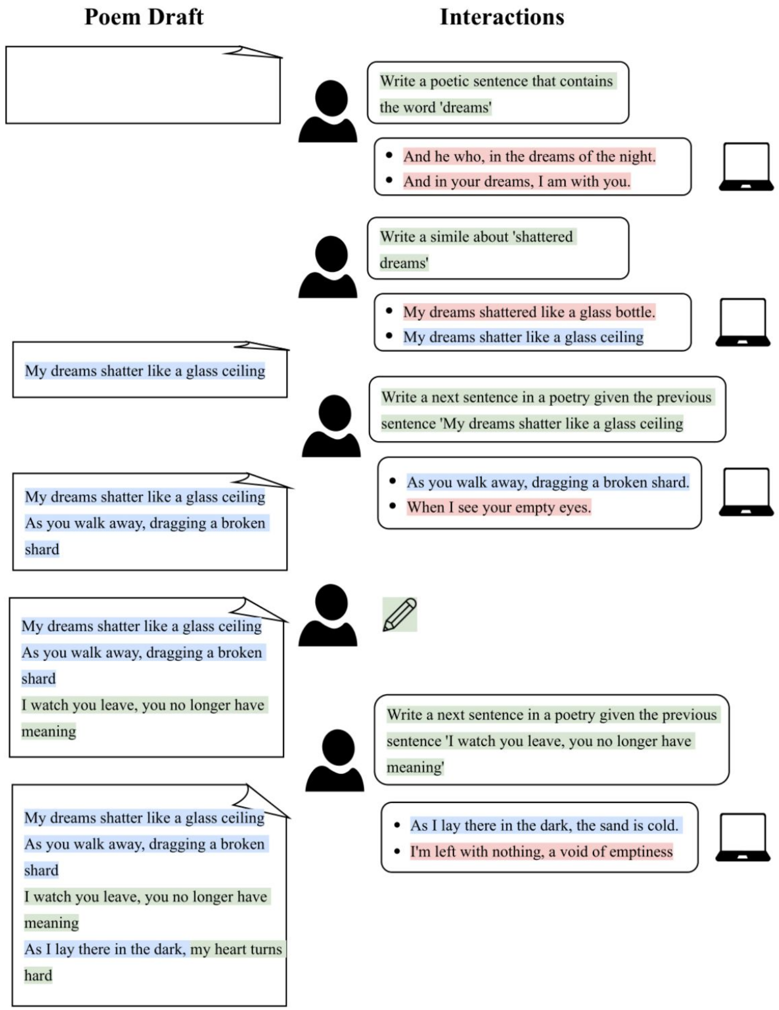
Human writes initial caption

Human uses these as appropriate

Final Caption: The Empire State Building shines bright, like a **glowing beacon for all to see**

Overview (Upto 2022)

- **Machine-in-the-Loop Rewriting for Creative Image Captioning**
  - NAACL 2022
- **Collaborative Poetry Writing with Instruction Tuning**
  - EMNLP 2022



Poem Draft

Interactions

Write a poetic sentence that contains the word 'dreams'

- And he who, in the dreams of the night.
- And in your dreams, I am with you.

Write a simile about 'shattered dreams'

- My dreams shattered like a glass bottle.
- My dreams shatter like a glass ceiling

My dreams shatter like a glass ceiling

Write a next sentence in a poetry given the previous sentence 'My dreams shatter like a glass ceiling

- As you walk away, dragging a broken shard.
- When I see your empty eyes.

My dreams shatter like a glass ceiling
As you walk away, dragging a broken shard

My dreams shatter like a glass ceiling
As you walk away, dragging a broken shard
I watch you leave, you no longer have meaning

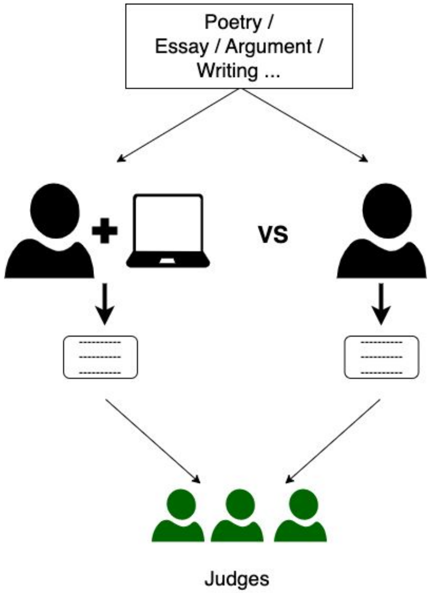
Write a next sentence in a poetry given the previous sentence 'I watch you leave, you no longer have meaning'

- As I lay there in the dark, the sand is cold.
- I'm left with nothing, a void of emptiness

My dreams shatter like a glass ceiling
As you walk away, dragging a broken shard
I watch you leave, you no longer have meaning
As I lay there in the dark, my heart turns hard

Overview (Upto 2022)

Phase 1: Train models to solve specific user needs and demonstrate how writers achieve their goals in these interactive setups



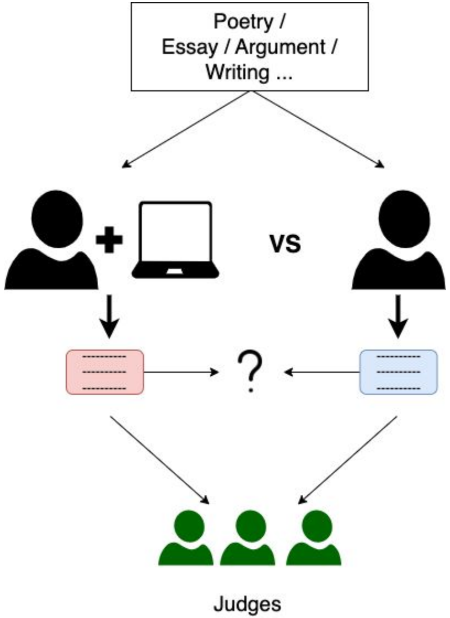
Overview

Phase 1: Train models to solve specific user needs and demonstrate how writers achieve their goals in these interactive setups

Phase 2: If collaborative writing is mainstream, what is the impact of writing with model help individually and collectively?

Overview (2022 Onwards)

Phase 2: If collaborative writing is mainstream, what is the impact of writing with model help individually and collectively?



Overview (2022 Onwards)

Phase 2: If collaborative writing is mainstream, what is the impact of writing with model help individually and collectively?

- Does Writing With Language Model Reduce Content Diversity?
- Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers

Overview (2022 Onwards)

Phase 2: If collaborative writing is mainstream, what is the impact of writing with model help individually and collectively?

- **Does Writing With Language Models Reduce Content Diversity?**
- Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers