

Evaluation

Lara J. Martin (she/they)

<https://laramartin.net/interactive-fiction-class>

Slides adapted from Elizabeth Clark

Learning Objectives

Determine what evaluation to use for your projects

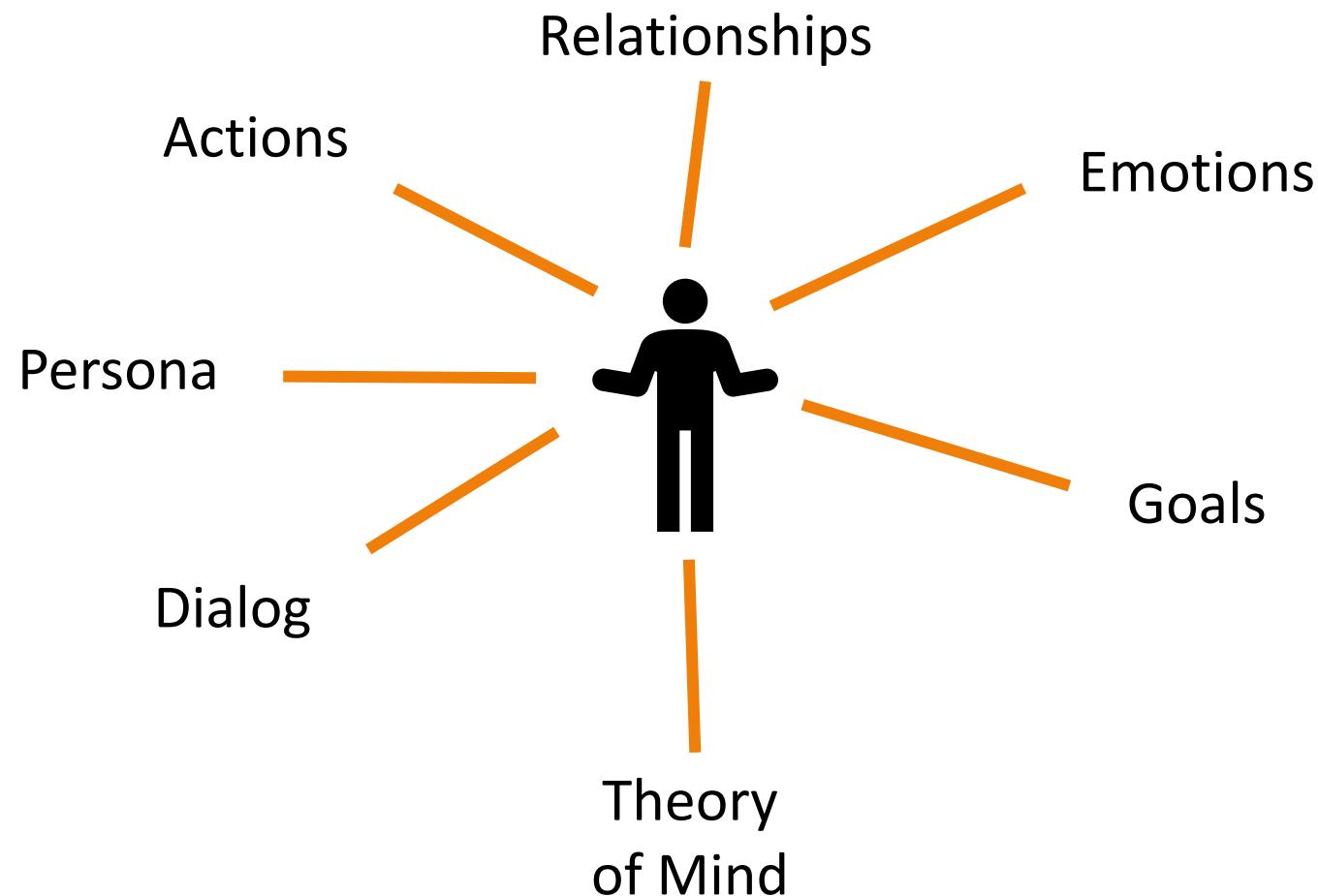
Enumerate automated story evaluation techniques and when they'd be useful

Define what creativity is

Describe the procedures for running human evaluations of stories

Find uses of human-AI collaboration in evaluation

Review: What makes up a character?



What else might you want to model about a character?

Why we need strong evaluations for story generation

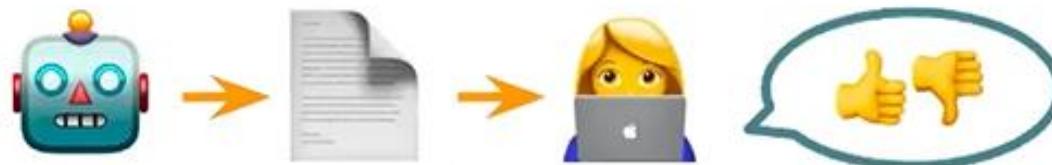
- Validate research hypotheses
- Compare results with other systems
- Understand a model's strengths and weaknesses
- Supports future research and model development
- Well-defined and well-scoped research questions and evaluations allow measurable progress

Outline

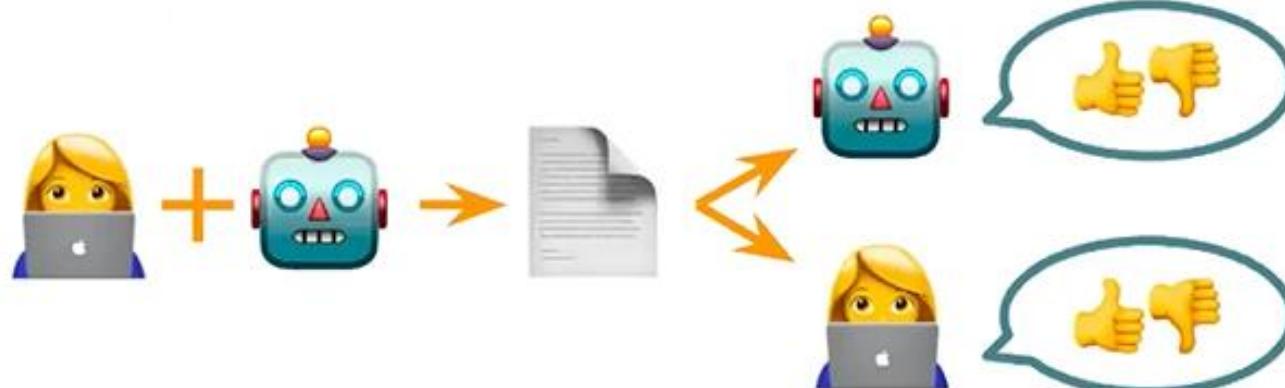
1. Automatic evaluation
of generated stories



2. Human evaluation
of generated stories



3. Evaluation of
human-machine
collaborative stories

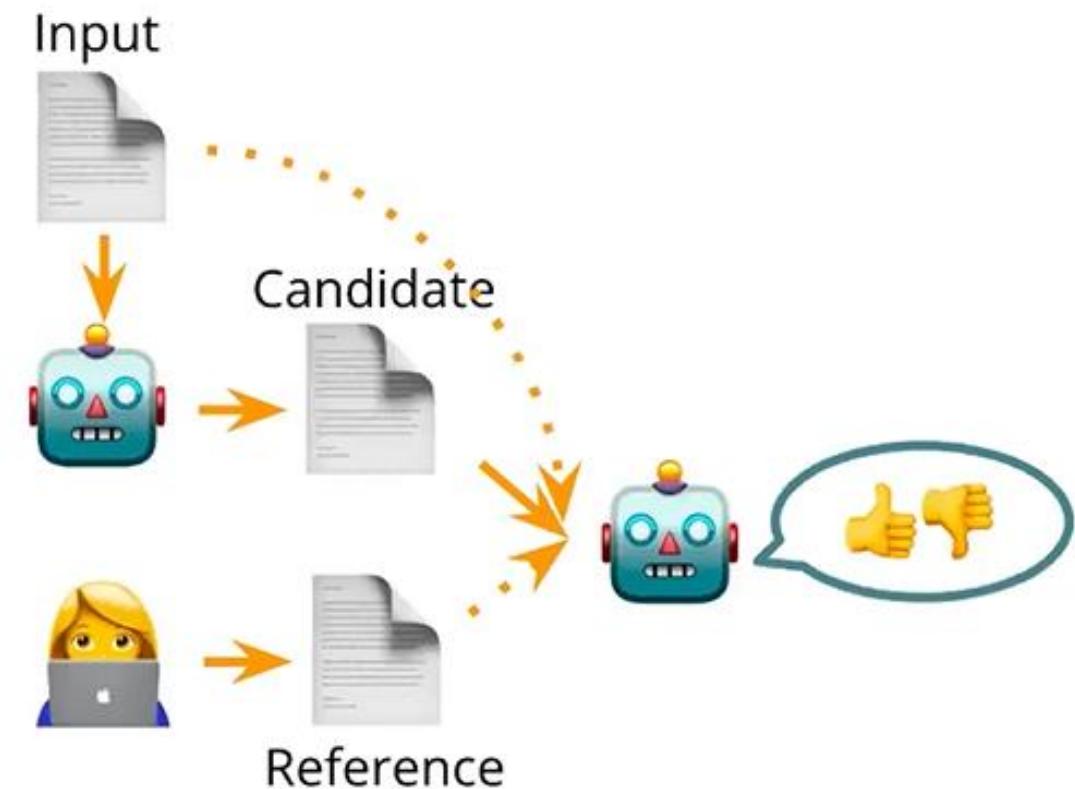


Automatic Story Evaluation

Given a generated story (and optionally additional context), automatically assess its quality

Pros: does not require the time/\$\$ of human evaluations, can compare and benchmark results

Cons: a metric's definition of "quality" may not align with a person's definition



Lexical Overlap Metrics

Measure the n -grams shared between two texts

Compares a candidate text to a reference text

n -gram overlap	Metric	Property
	F-SCORE	precision and recall
	BLEU	n -gram precision
	METEOR	n -gram w/ synonym match
	CIDER	$tf-idf$ weighted n -gram sim.
	NIST	n -gram precision
	GTM	n -gram metrics
	HLEPOR	unigrams harmonic mean
	RIBES	unigrams harmonic mean
	MASI	attribute overlap
	WER	% of insert, delete, replace
	TER	translation edit rate
	ROUGE	n -gram recall
	DICE	attribute overlap

Celikyilmaz et al. "[Evaluation of Text Generation: A Survey](#)" 2020

Example: ROUGE

ROUGE-N

$$= \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Candidate: my favorite food is pineapple

Reference: pineapple is my favorite tropical fruit

n=1: 4 matches out of 6 ROUGE-1: 0.67

n=2: 1 match out of 5 ROUGE-2: 0.20

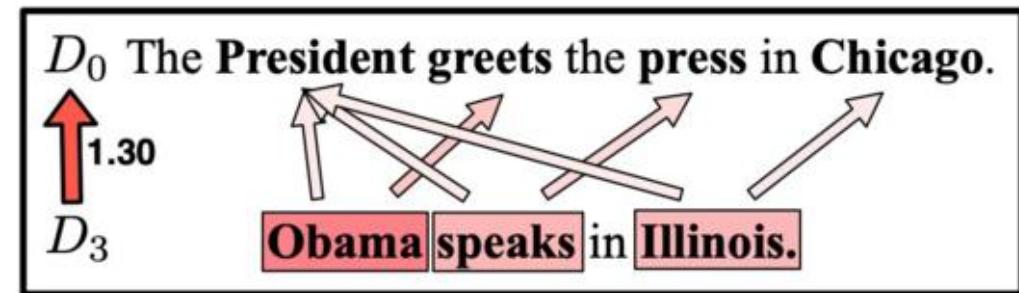
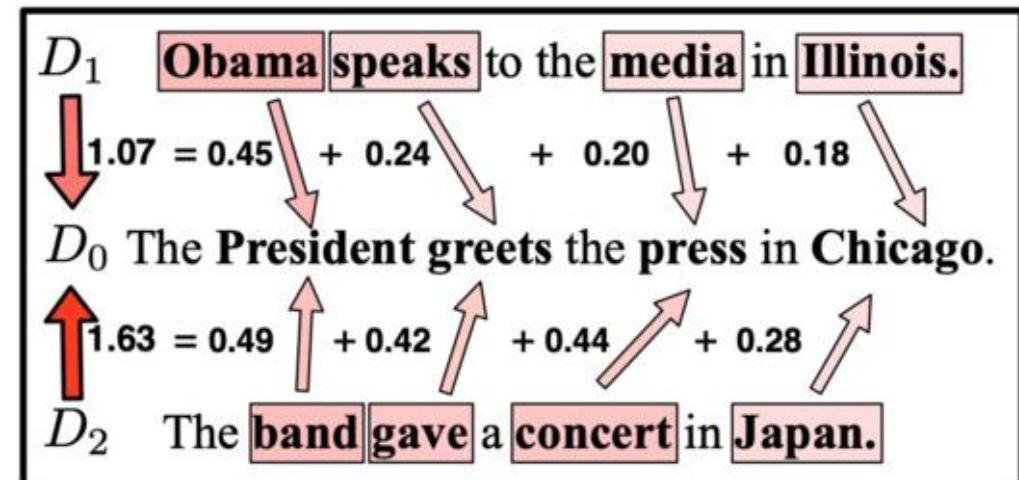
n=3: 0 matches out of 4 ROUGE-3: 0.00

[ROUGE: A Package for Automatic Evaluation of Summaries](#) (Lin, 2004)

Embedding-based metrics

Measure a candidate's similarity to a reference text based on their embeddings

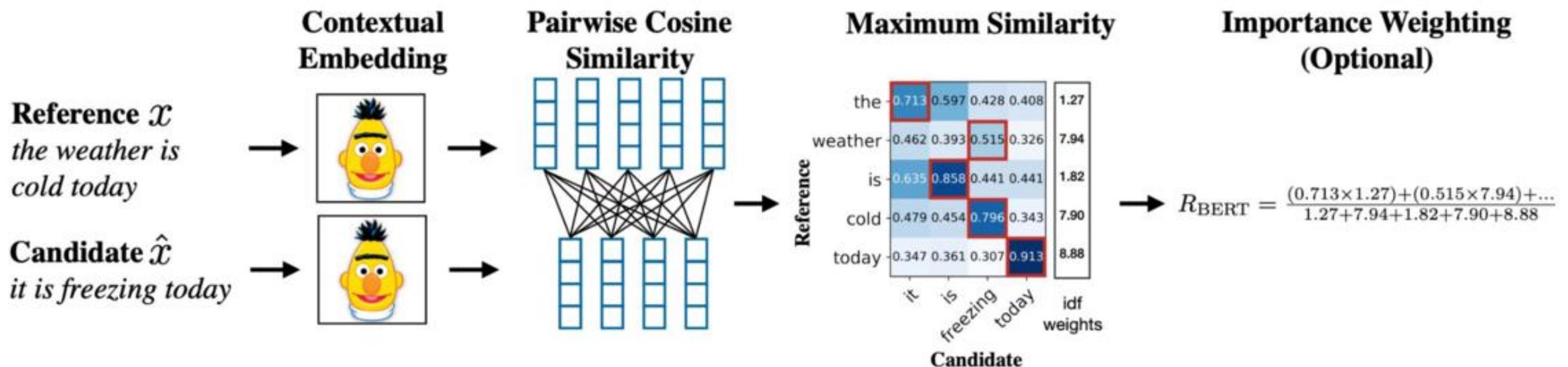
Take advantage of ever-improving pretrained NLP models



[From Word Embeddings To Document Distances](#)

Matt Kusner, et al. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:957-966, 2015.

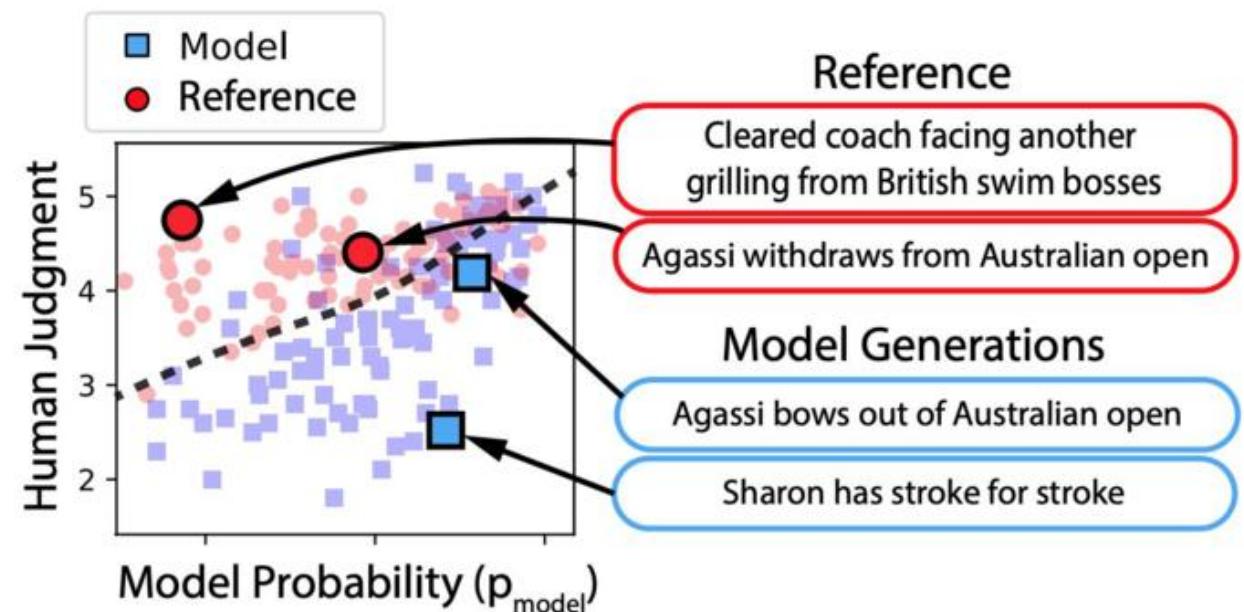
Example: BERTScore



Diversity Metrics

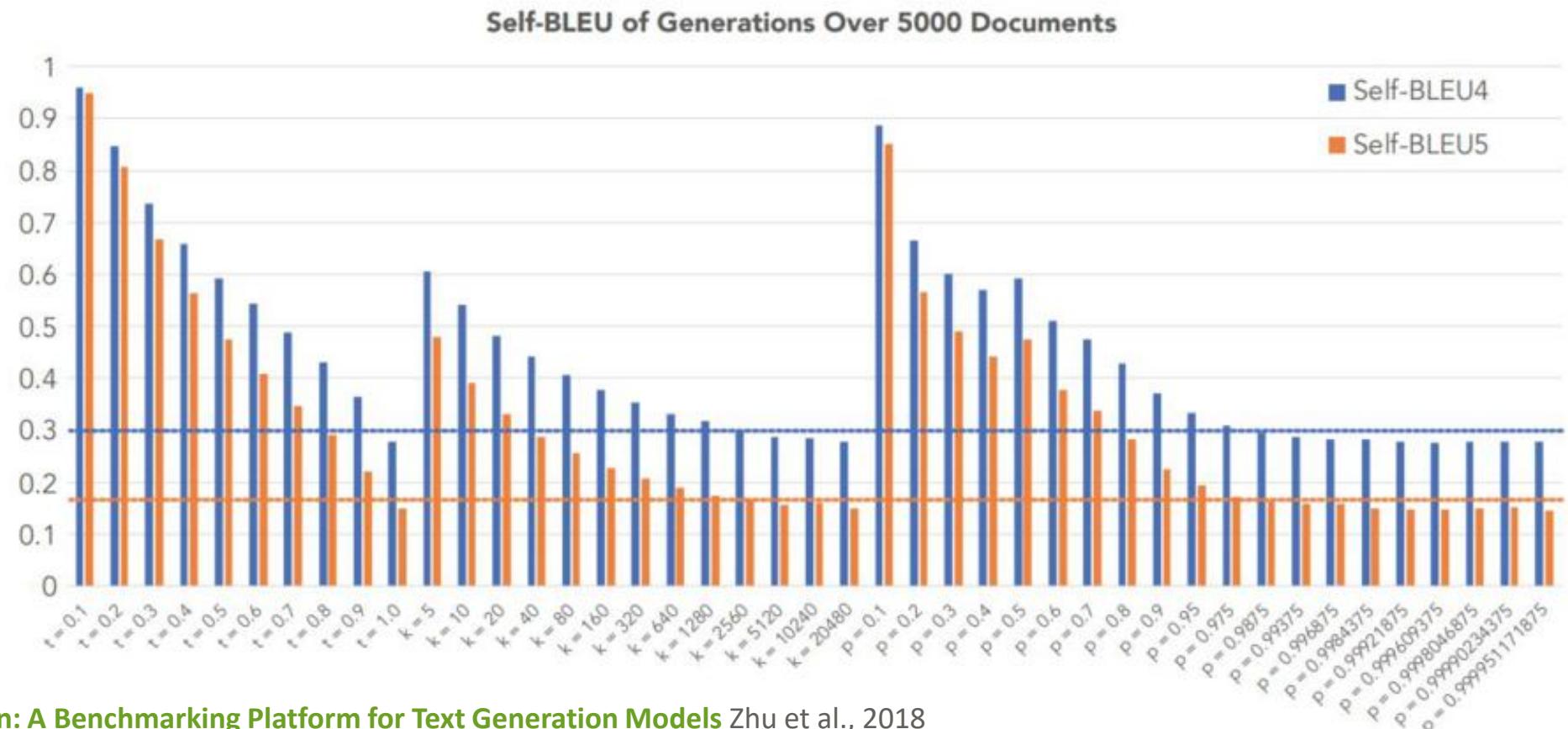
How unique is the generated text?

Trade-off between text that is high-quality and text that is diverse



[Unifying Human and Statistical Evaluation for Natural Language Generation](#) (Hashimoto et al., NAACL 2019)

Example: Self-BLEU



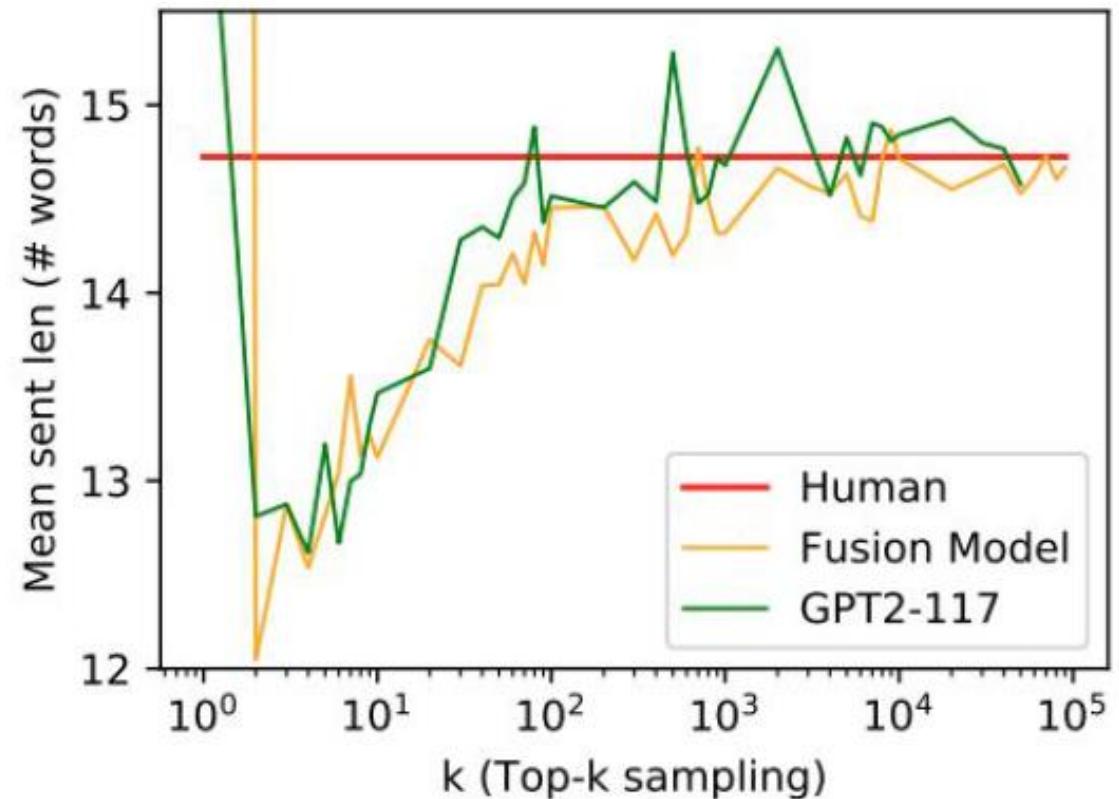
Self-BLEU: [Txygen: A Benchmarking Platform for Text Generation Models](#) Zhu et al., 2018

Image: [The Curious Case of Neural Text Degeneration](#) Holtzman et al., 2020

Descriptive Metrics

Analysis of the type of language model produces

The results may not mean much alone, but can be compared to results on reference texts or text generated by other models



[Do Massively Pretrained Language Models Make Better Storytellers?](#) (See et al., CoNLL 2019)

Example: Output comparison (Fusion model vs GPT-2)

Generated story vs.
the prompt

Coherence

Repetition and
rareness

Syntactic style and
complexity

Surprisingness

Concreteness

Prompt: A man finally discovers his superpower... well into his 80's.

Fusion Model ($k = 10$):

I be a child . bear into a world of darkness , and in the world , there be only a few people . My mother and I be always alone . She have never be like this . But , she have no idea what we be together to be . We be always alone . We be always together . And , of course , she have never exist . But , I never think about it , because she be always alone . In the world , it was n't like she have a power . I be always alone , and there be a time when the light would turn on . There be a time when I could see the light , and I could see it in her eyes , and I could see the light , and I could see it

Mean noun concreteness: 3.892

Mean verb concreteness: 2.173

GPT2-117 ($k = 10$):

A man in an old white coat sit in a corner. He wear a black jacket with a blue tie. His hair be grey. His eyes be black, but his voice be deep and clear. He look to where the man be sit, and he see a smile on his lips. It be a smile he know he would see from his own eyes. But he be too late. He be on the sidewalk by the river when the man come. He be wear a black coat with a purple tie. He have a black tie and a white shirt. But he be still wear a white suit. And it seem he would look back at him. A smile on his face. A look his friend do n't recognize. He have no

Mean noun concreteness: 4.720

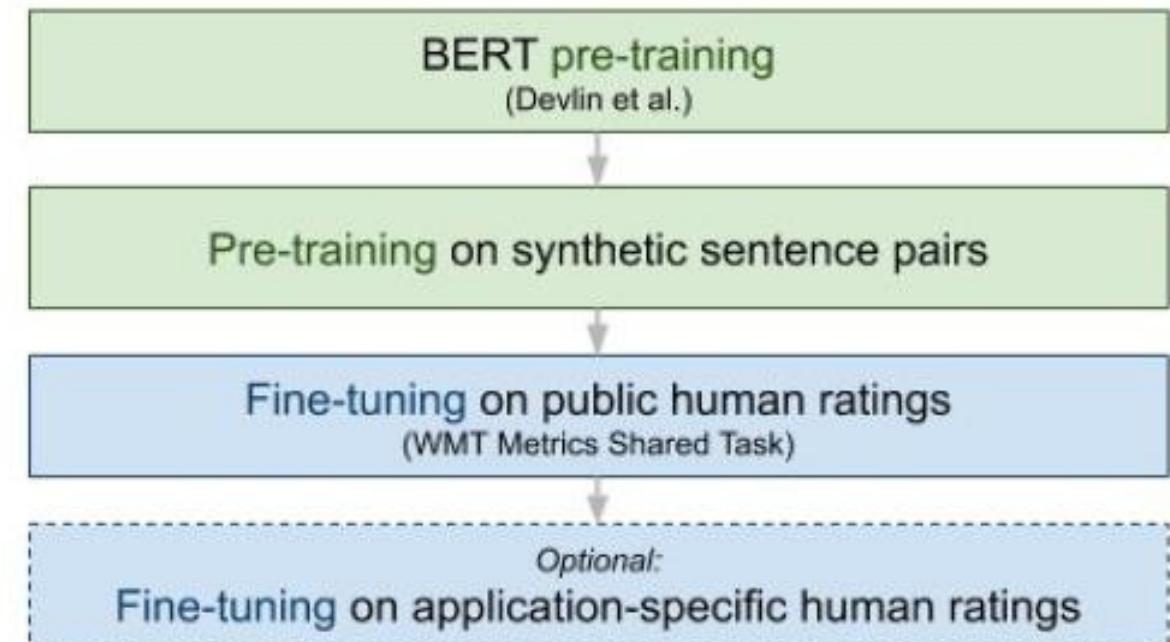
Mean verb concreteness: 2.488

[Do Massively Pretrained Language Models Make Better Storytellers?](#) (See et al., CoNLL 2019)

Learned Metrics

Train a model on to predict a score of the text's quality

A metric is usually evaluated by its correlation with human judgments



[BLEURT: Learning Robust Metrics for Text Generation](#) (Sellam et al., ACL 2020)

Example: UNION

Leading Context

Jack was at the bar.

Reference By Human

He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.

Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00)

On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box.

Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)

He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars.

Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)

He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.

B: BLEU

M: MoverScore

U: Union

Example: UNION

Leading Context

Jack was at the bar.

Reference By Human

He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.

Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00)

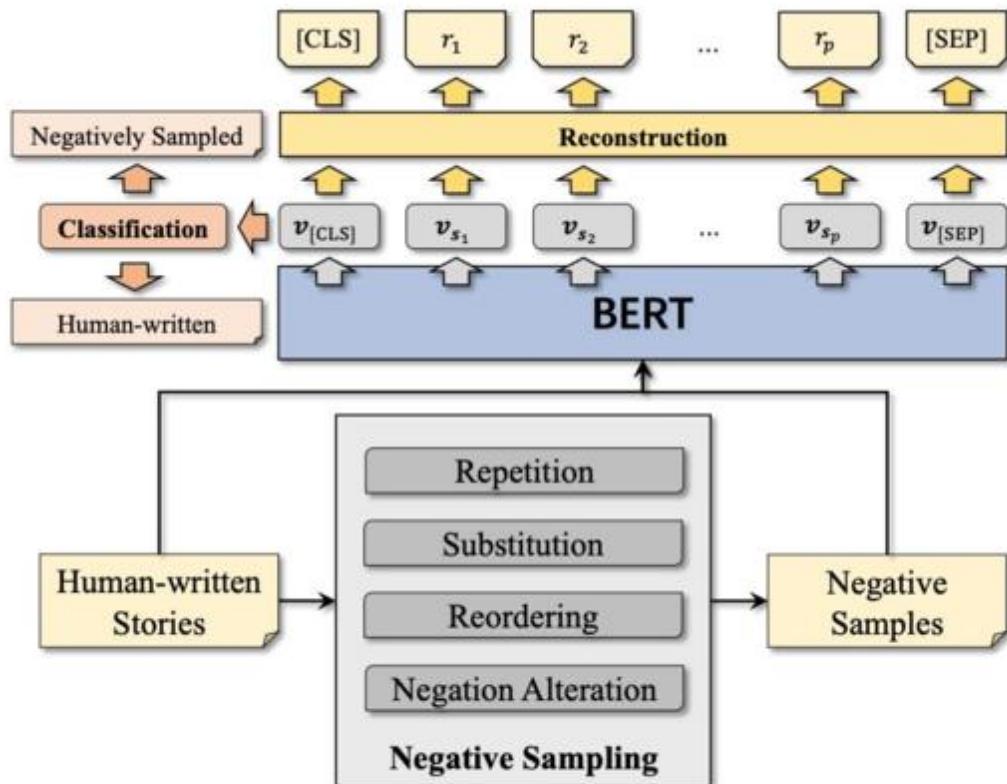
On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box.

Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)

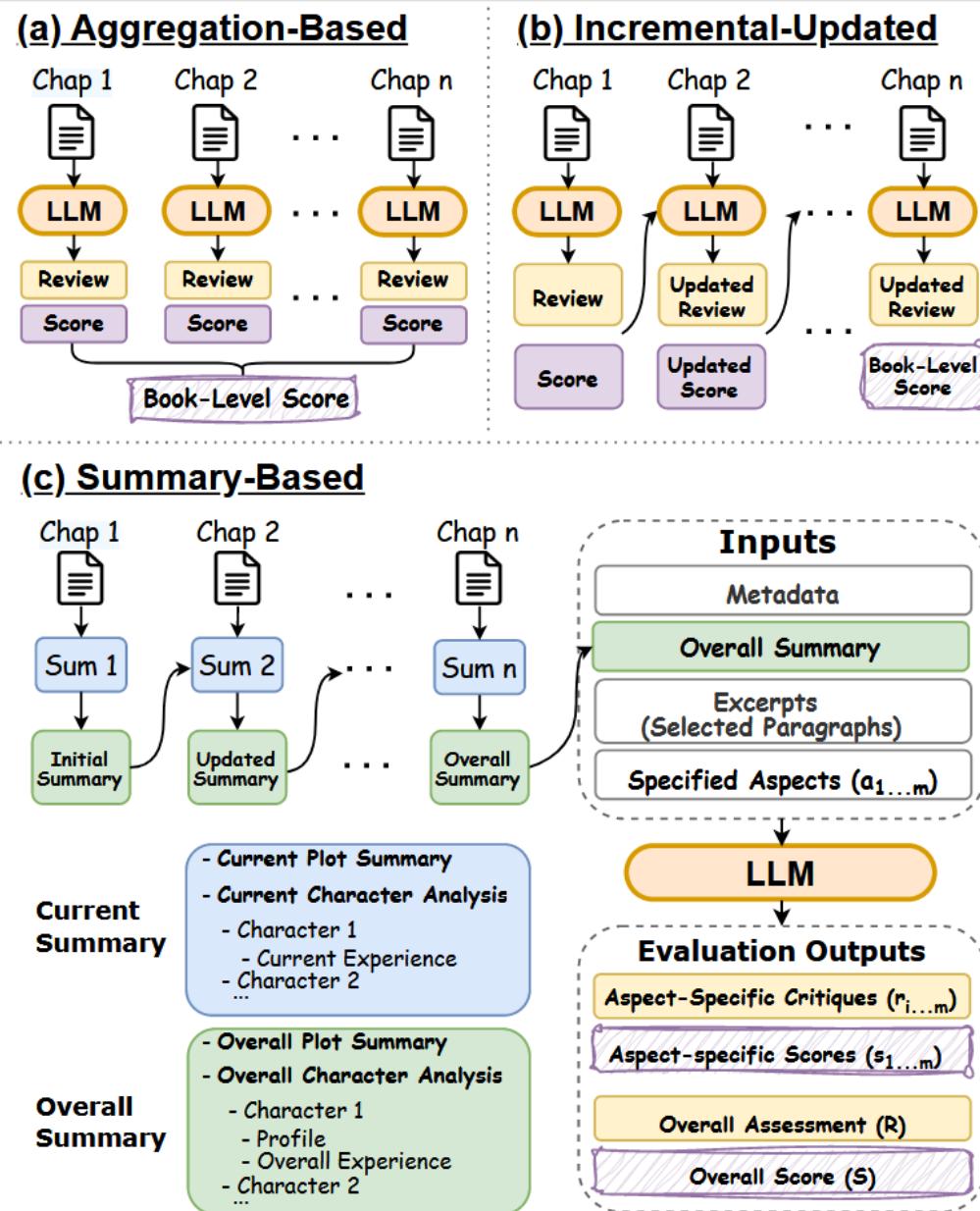
He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars.

Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)

He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.



LLM-as-Judge



		PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP	Overall
One-Pass (Subset)	GPT-4o	3.3	4.1	7.9	0.8	3.3	-1.2	-3.2	8.4	5.5
	DeepSeek-v2.5	4.4	3.5	4.8	-0.9	3.3	-1.1	-1.3	9.4	4.8
Aggregation -Based	GPT-4o	14.3	16.7	10.2	7.9	10.4	9.7	9.1	14.1	15.2
	DeepSeek-v2.5	17.2	15.8	7.0	7.1	11.0	14.2	11.1	16.7	15.1
	GPT-4o-mini	14.2	17.2	7.2	4.4	9.5	8.9	8.1	15.1	12.3
	Llama 3.1-70B	19.6	13.8	2.3	13.8	13.4	7.7	11.5	18.9	13.8
	Llama 3.1-8B	15.5	8.5	-1.4	2.8	12.3	7.5	7.0	13.7	11.6
	Mixtral 8×7B	9.5	4.0	2.5	-0.2	8.9	9.5	10.2	6.8	9.0
Incremental -Updated	GPT-4o	8.0	9.1	9.1	11.7	10.5	12.3	12.1	11.5	10.9
	DeepSeek-v2.5	8.9	12.2	9.0	8.6	12.5	12.3	6.6	12.2	11.6
	GPT-4o-mini	7.9	10.8	6.7	7.4	8.5	11.6	8.5	10.7	9.3
	Llama 3.1-70B	9.3	13.3	4.1	1.7	8.7	4.9	4.6	6.1	9.9
	Llama 3.1-8B	7.0	7.1	4.4	2.5	1.9	8.0	7.8	5.1	6.7
	Mixtral 8×7B	4.2	10.8	4.4	6.6	5.8	2.3	5.8	2.6	4.2
Summary -Based	GPT-4o	15.3	17.8	4.5	5.0	7.2	12.6	11.8	14.0	13.4
	DeepSeek-v2.5	13.4	12.2	1.8	-3.8	7.1	8.9	13.2	15.1	14.4
	GPT-4o-mini	8.7	7.5	5.4	4.8	11.1	11.6	8.3	7.9	9.7
	Llama 3.1-70B	11.2	10.8	-1.6	5.3	12.4	9.2	11.4	14.5	13.0
	Llama 3.1-8B	10.4	14.1	4.9	9.1	9.6	15.3	14.5	12.3	12.4
	Mixtral 8×7B	7.8	7.4	7.1	-0.5	-4.0	5.6	9.4	6.7	8.3
	NovelCritique-8B	21.4	20.8	15.1	11.2	18.5	21.1	22.8	20.5	20.1

Table 2: The system-level Kendall correlations between the human-assigned scores and model-generated evaluations. We report the correlation between aspect-specific scores and the overall score.

Are any of these relevant to your project?

Lexical Overlap

Embedding-based Metrics

Diversity Metrics

Descriptive Metrics

Learned Metrics

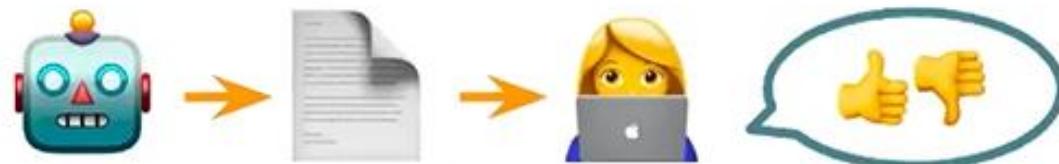
LLM-as-Judge

Outline

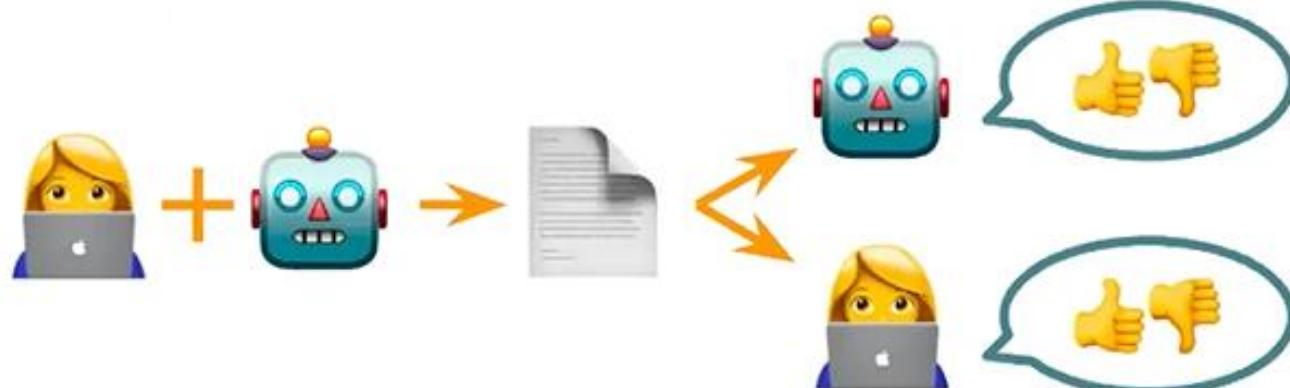
1. Automatic evaluation
of generated stories



2. Human evaluation
of generated stories



3. Evaluation of
human-machine
collaborative stories



(One) Definition of Creativity

“Creativity is the ability to come up with ideas or artefacts that are **new, surprising, and valuable**”

1. New to who?
 - a. P-creativity vs H-creativity
2. Surprising
 - a. Unfamiliar
 - b. Unexpected realization that idea X could be considered concept Y
 - c. Impossible idea
3. Valuable
 - a. Interesting, meaningful, etc.

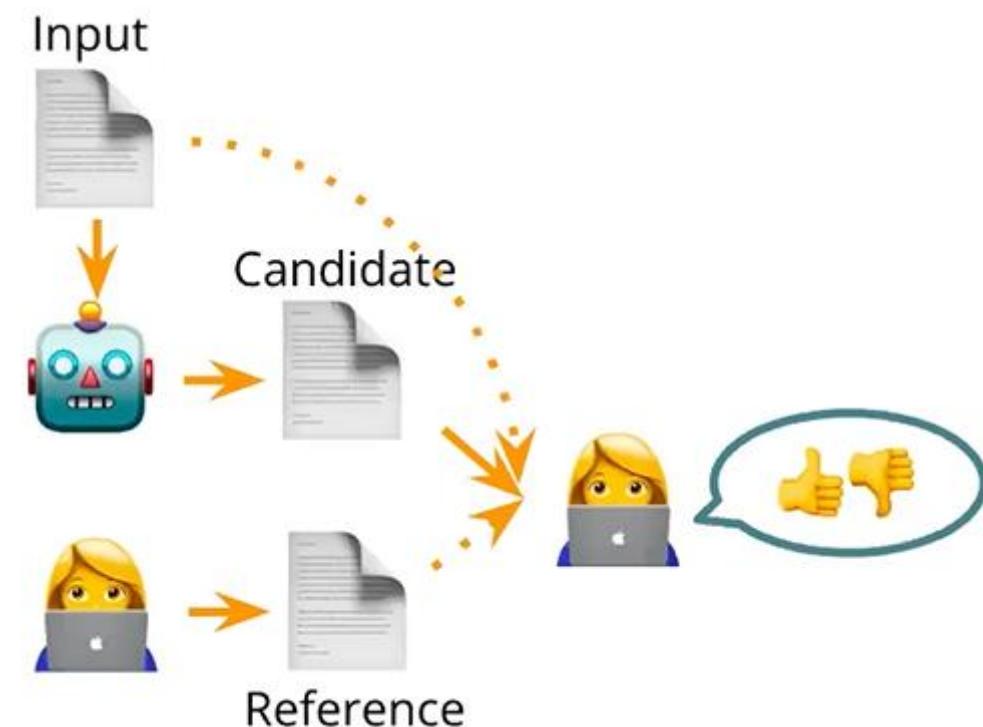
Human Story Evaluation

People read generated story text and judge their quality

Judgments can be about overall quality or broken down into specific criteria

Pros: aligned with modeling goals, can be more specific/nuanced

Cons: collecting reliable evaluations can be difficult, especially when text is long or complex



Participants

Are the participants in human evaluation...

- Experts?
- In-Person?
- Crowdsourced?
- Paid?
- Trained?
- Quality-controlled?



Dimensions of Text Quality

Is the text...

- Grammatical?
- Fluent?
- Coherent?
- Creative?
- Surprising?
- Entertaining?

Criterion Parphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
goodness of outputs in their own right (both form and content)	13
referent resolvability	11
usefulness (nonspecific)	11
appropriateness (content)	10
naturalness	10
user satisfaction	10
wellorderedness	10
correctness of outputs in their own right (form)	9
correctness of outputs relative to external frame of reference (content)	8
ease of communication	7
humanlikeness	7
appropriateness	6
understandability	6
nonredundancy (content)	6
goodness of outputs relative to system use	5
appropriateness (both form and content)	5

Types of Human Feedback

Is this generated story...

- Good or bad?
- Good on a scale from 1 to 5?
- Better than another story?

Q1: Which do you think is better at utilizing the keywords?

- Story 1
- Story 2

Q2: Which do you think is more repetitive?

- Story 1
- Story 2

Q3: Which do you think has better transitions?

- Story 1
- Story 2

Q4: Which do you think is better at following a single storyline?

- Story 1
- Story 2

Q5: Which do you think has a better introduction?

- Story 1
- Story 2

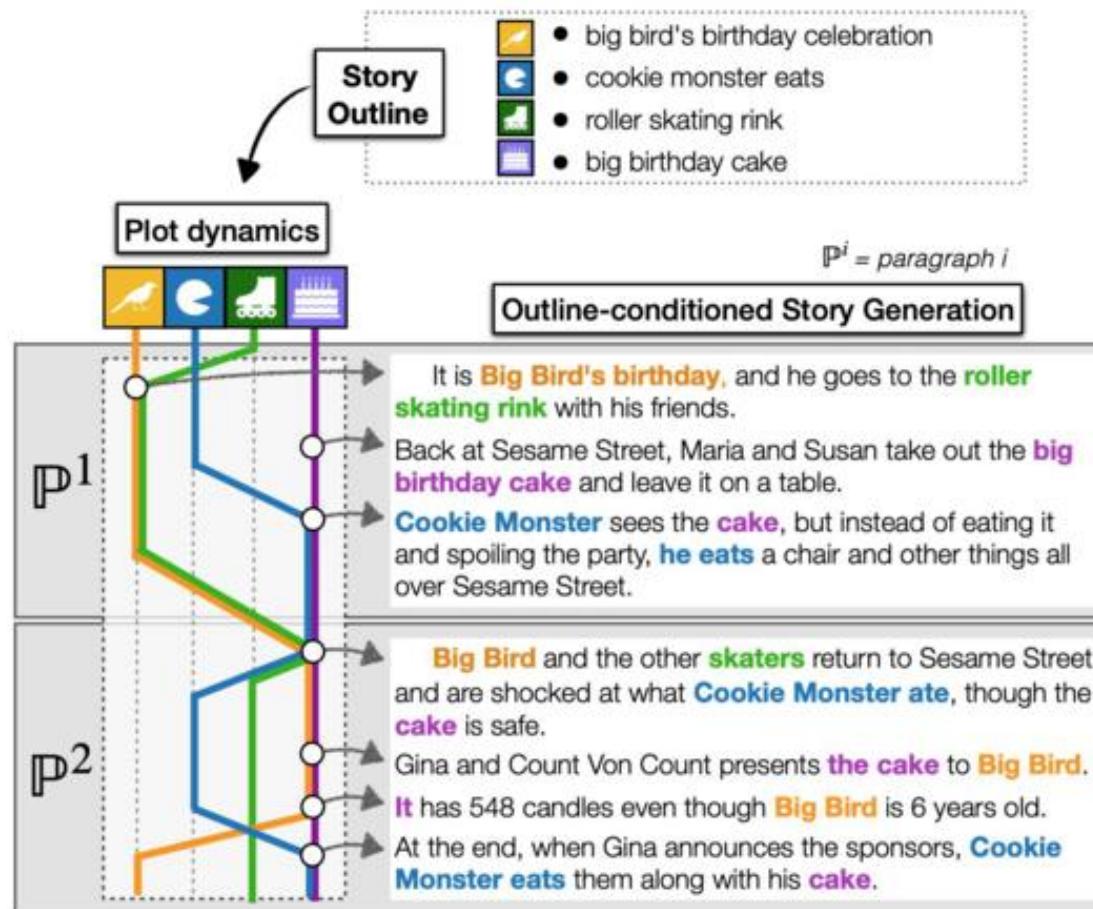
Q6: Which do you think has a better conclusion?

- Story 1
- Story 2

Q7: Which do you think has a clear order of events?

- Story 1
- Story 2

Case Study: PlotMachines

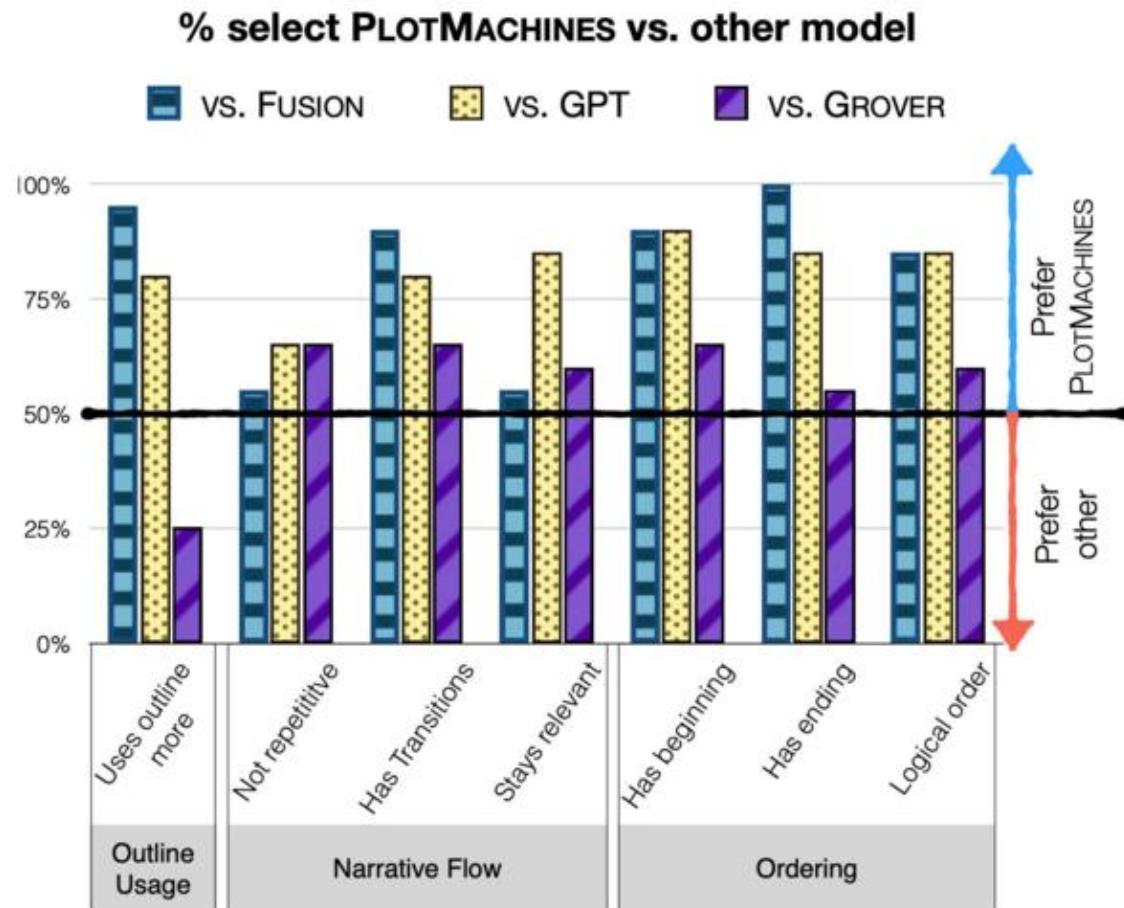


PlotMachines: Automatic Evaluation

Model	Wikiplots			WritingPrompts			New York Times		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
P&W-Static (Yao et al., 2019)	17.0	3.3	13.6	19.2	3.6	14.4	19.3	4.6	15.6
Fusion (Fan et al., 2018)	22.7	6.0	17.4	14.3	1.7	9.6	23.2	7.2	18.1
GROVER (Zellers et al., 2019)	19.6	5.9	12.5	23.7	5.3	17.2	20.0	5.8	14.2
PLOTMACHINES (GPT)	20.2	5.3	16.0	30.5	5.3	25.4	21.2	5.0	15.5
– base (GPT) (Radford et al., 2018)	13.2	2.0	7.9	22.1	2.7	14.3	13.9	1.6	8.3
PLOTMACHINES (GPT-2)	22.8	6.5	17.5	31.1	6.7	26.1	22.1	6.4	16.5
– PM-NoMem (GPT-2)	20.5	4.9	15.5	26.6	3.7	23.5	20.0	5.4	14.4
– PM-NoMem-NoDisc (GPT-2)	19.3	1.7	13.9	26.8	4.5	23.2	18.4	3.4	14.2
– base (GPT-2) (Radford et al., 2019)	18.5	3.9	13.3	26.5	4.6	20.5	19.2	4.7	13.6

Model	Wikiplots					Writing Prompts					NY Times				
	AvgL	B-2	B-3	B-4	B-5	AvgL	B-2	B-3	B-4	B-5	AvgL	B-2	B-3	B-4	B-5
Gold Test	330	.74	.50	.29	.15	661	.82	.61	.40	.25	315	.73	.50	.32	.21
P&W-Static	352	.93	.85	.75	.64	675	.97	.94	.89	.85	352	.93	.85	.74	.63
Fusion	191	.84	.71	.58	.48	197	.93	.85	.75	.65	171	.89	.80	.70	.60
GROVER	835	.72	.49	.48	.37	997	.88	.72	.52	.34	719	.79	.57	.38	.25
GPT	909	.77	.47	.25	.11	799	.73	.40	.19	.08	739	.68	.36	.27	.08
GPT-2	910	.60	.26	.10	.03	799	.74	.41	.19	.08	756	.69	.36	.17	.08
PLOTMACHINES (GPT)	682	.77	.58	.40	.27	850	.89	.81	.72	.63	537	.85	.69	.53	.40
PLOTMACHINES (GPT-2)	553	.56	.19	.07	.02	799	.83	.56	.30	.14	455	.79	.57	.37	.23

PlotMachines: Human Evaluation



Model	Narrative Flow			Order Acc(\uparrow)
	Rep(\downarrow)	Tran(\uparrow)	Rel(\uparrow)	
Fusion	2.61	2.98	3.36	73
GPT	1.39	1.89	2.06	42
GROVER	1.78	3.00	3.29	62
PM	1.64	3.02	3.39	59

Human Evaluation for the Project

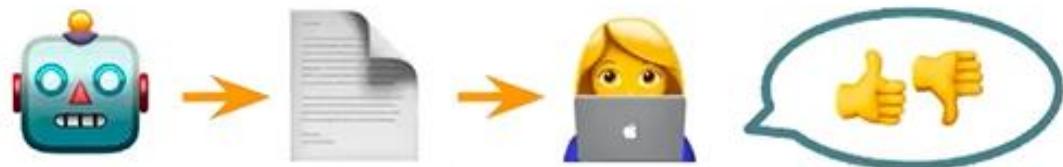
- How would you have people evaluate your project?
 - What are you measuring?
 - Who would participate?
 - How will you keep things fairly consistent across participants?

Outline

1. Automatic evaluation
of generated stories



2. Human evaluation
of generated stories



3. Evaluation of
human-machine
collaborative stories

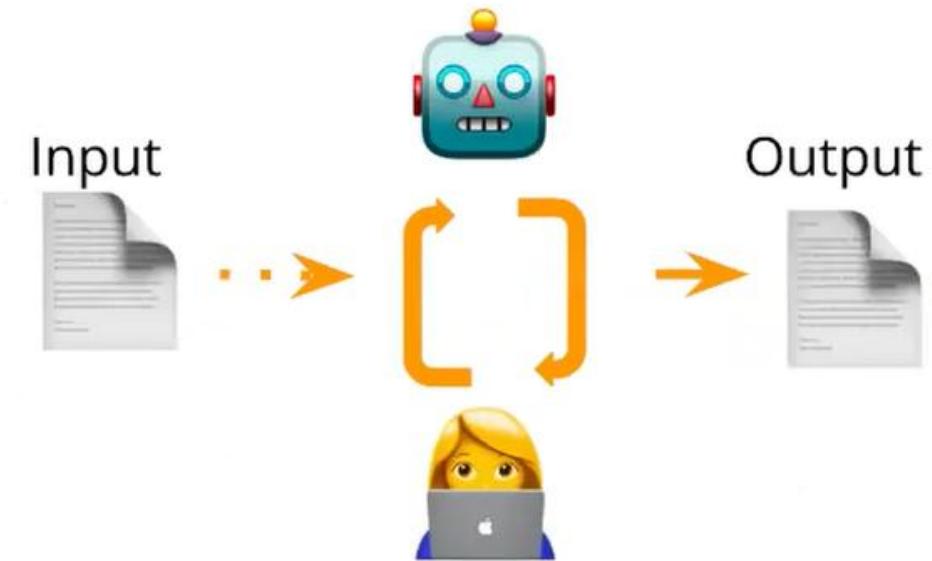


Collaborative Story Generation

A person works with model output to write a story together

This collaboration can take many forms, e.g.,:

- Auto-complete
- Incorporating keywords or concepts
- Turn-taking
- Offering suggestions or improvements



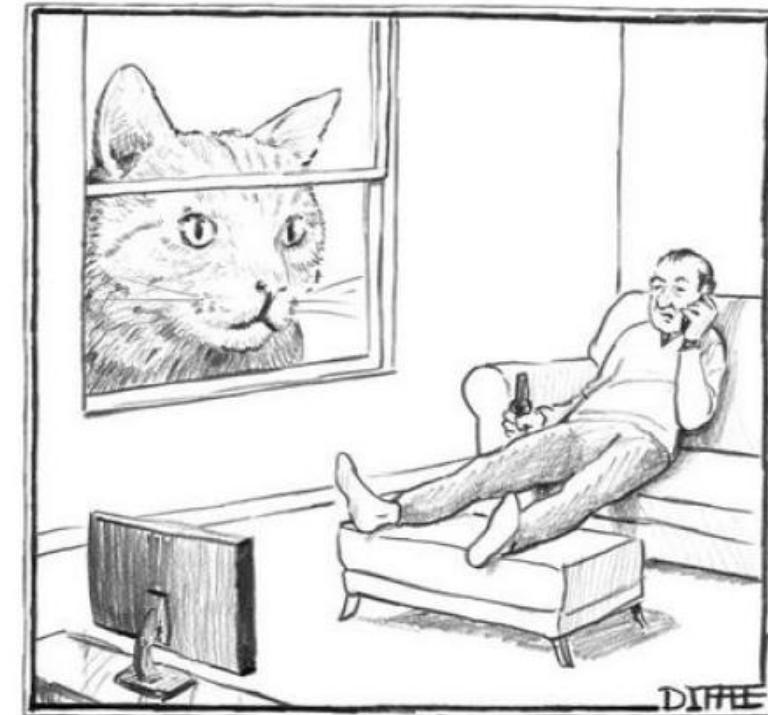
Example: Turn-taking collaborative writing

Add a sentence to the story:

Add Line to Story

Characters: 0

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)



Diffee, Matthew. *The New Yorker*. 11 Aug 2014.

Elizabeth Clark, et al. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 329–340. <https://doi.org/10.1145/3172944.3172983>

Example: Turn-taking collaborative writing

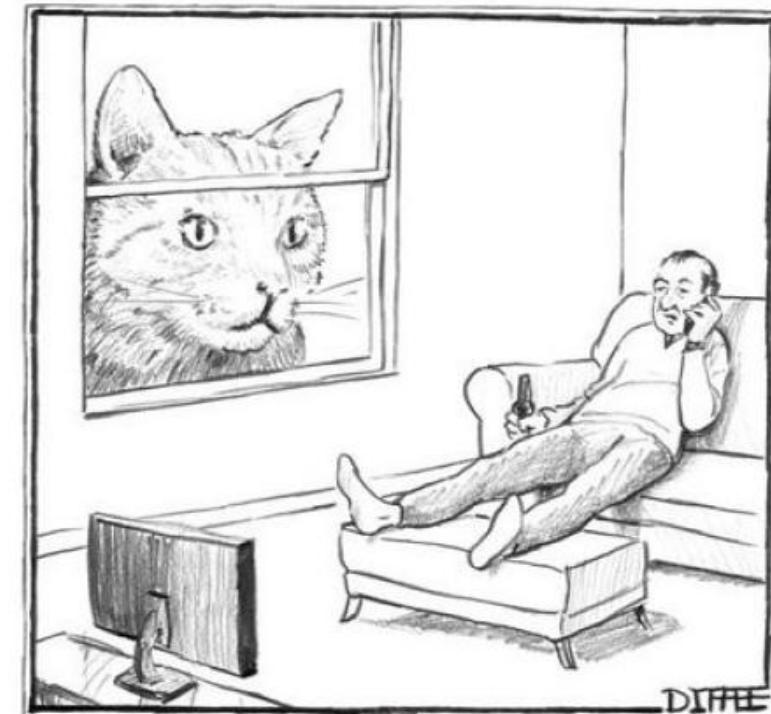
Add a sentence to the story:

Phil woke up on the couch with a huge hangover.

Add Line to Story

Characters: 47

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)



Diffee, Matthew. *The New Yorker*. 11 Aug 2014.

Elizabeth Clark, et al. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 329–340. <https://doi.org/10.1145/3172944.3172983>

Example: Turn-taking collaborative writing

The prompt will appear below.

You can edit it as much as you like before adding it to the story.

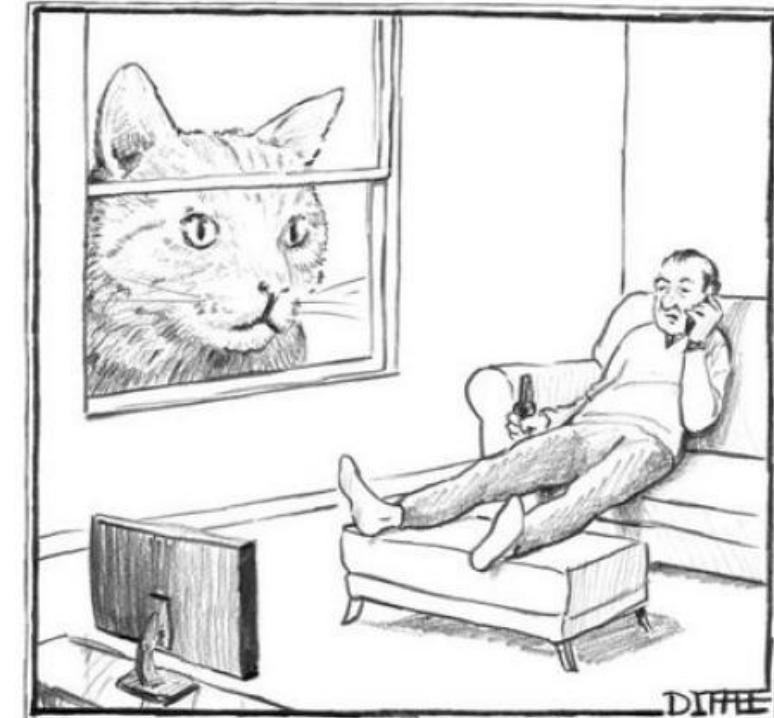
Phil woke up on the couch with a huge hangover.

Now he looked at Anne.

Add Line to Story

Characters: 22

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)



Diffee, Matthew. *The New Yorker*. 11 Aug 2014.

Elizabeth Clark, et al. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 329–340. <https://doi.org/10.1145/3172944.3172983>

Example: Turn-taking collaborative writing

The prompt will appear below.
You can edit it as much as you like before adding it to the story.

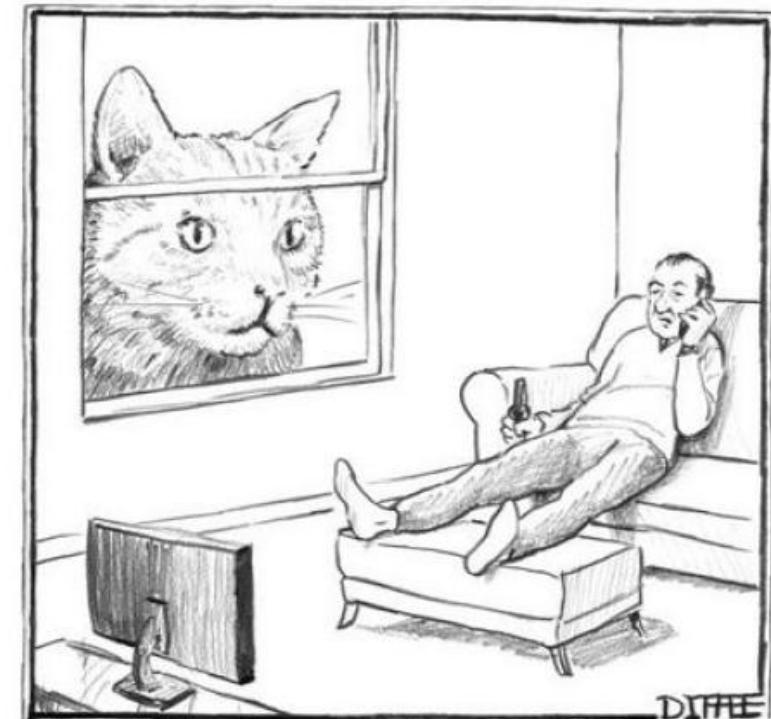
Phil woke up on the couch with a huge hangover.

He looked out the window at Anne, the neighbor's cat.

Add Line to Story

Characters: 53

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)



Diffee, Matthew. *The New Yorker*. 11 Aug 2014.

Elizabeth Clark, et al. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 329–340. <https://doi.org/10.1145/3172944.3172983>

How does evaluation change?

Reference texts are much rarer

Text can be a mix of human- and machine-generated text

“Experience” becomes important, not just the generated text

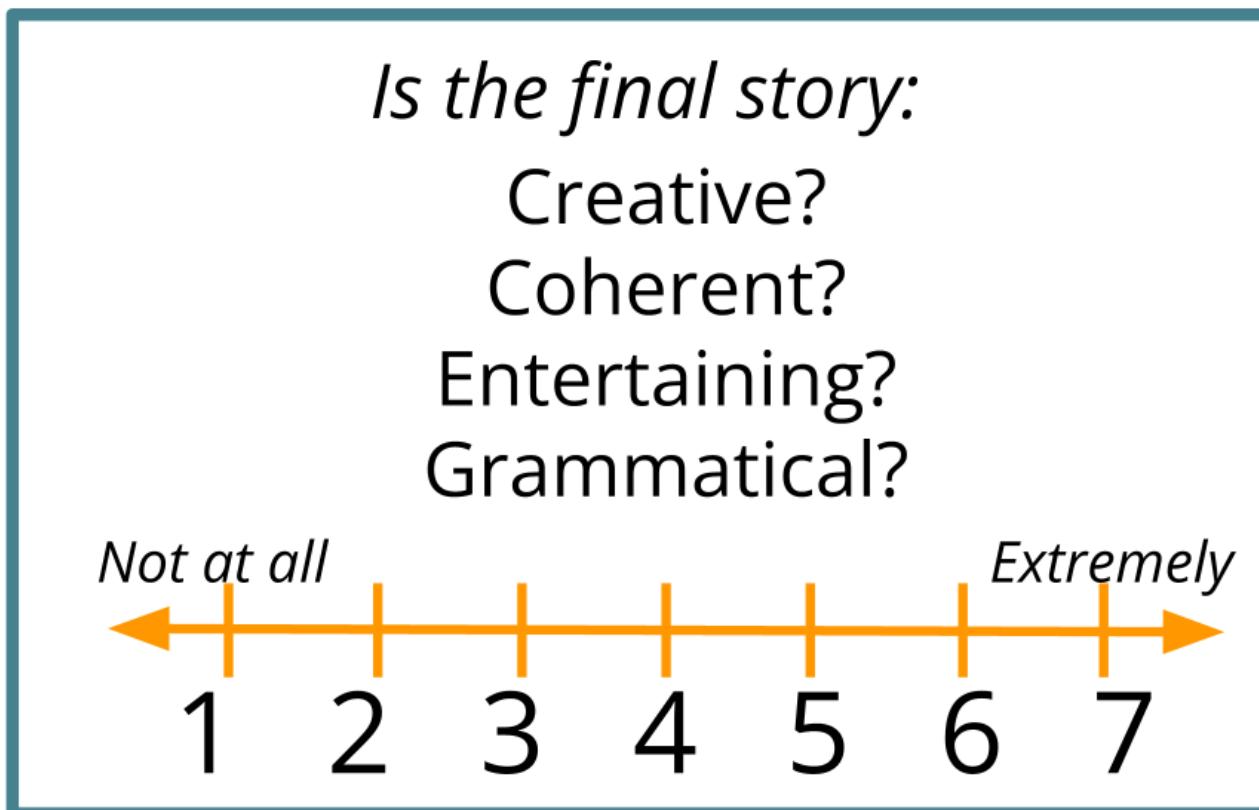
Evaluations can be from the writer’s perspective or the reader’s perspective

“Did you find the generated text helpful?”

vs.

“Did the generated text help produce a high-quality output?”

Example: Two human evaluation perspectives



Elizabeth Clark, et al. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18), 329–340. <https://doi.org/10.1145/3172944.3172983>

Types of evaluation for collaborative writing

1. Automatic metrics

2. Human evaluations

3. Interaction metrics

- Edit distance
- % suggestions accepted
- Time to complete the story

Model	Max Len	Avg Len	% Top	MRR	Time(s)	Time(s)/Sen
Unigram	27	9.41 ± 2.31	0.08 ± 0.09	0.36 ± 0.30	460.5 ± 411.8	44.9 ± 32.0
Bigram	25	9.50 ± 2.51	0.09 ± 0.10	0.34 ± 0.29	492.4 ± 463.7	47.9 ± 35.6
Reranking	27	9.54 ± 2.68	0.07 ± 0.08	0.28 ± 0.07	399.2 ± 294.3	40.1 ± 22.8
Adaptation	36	9.63 ± 3.07	0.04 ± 0.04	0.23 ± 0.04	406.1 ± 286.5	39.3 ± 20.6

Reid Swanson and Andrew S. Gordon. 2012. Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. ACM Trans. Interact. Intell. Syst. 2, 3, Article 16 (September 2012), 35 pages. <https://doi.org/10.1145/2362394.2362398>

Challenges in human evaluation with today's models

Text generation models have improved, and generated text is more fluent and higher quality than ever before

Crowdsourced evaluations are increasingly common

The easiest evaluation is not always the best evaluation

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

“[Language Models are Unsupervised Multitask Learners](#)” Radford et al., 2019

GPT-3

(Brown et al., 2020)

Once upon a time, in a land not so far away, there was a lovely young maiden named Charlotte. She had many admirers, but none as devoted as the prince. They were to be married, and she was the happiest girl in the world. One day, while she was walking in the forest, she came upon a fairy who offered her three wishes. She thought for a long time and then said, “I wish for a million dollars.”

“Your wish is granted,” said the fairy. “But you must pay a terrible price for it.”

“I don’t care,” said Charlotte. “I’ll do anything to be rich.”

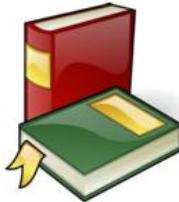
Definitely human-generated

This looks like something I’d read in a book

Possibly machine-generated

It seems kind of weird for a fantasy character to wish for something as concrete as a million dollars.

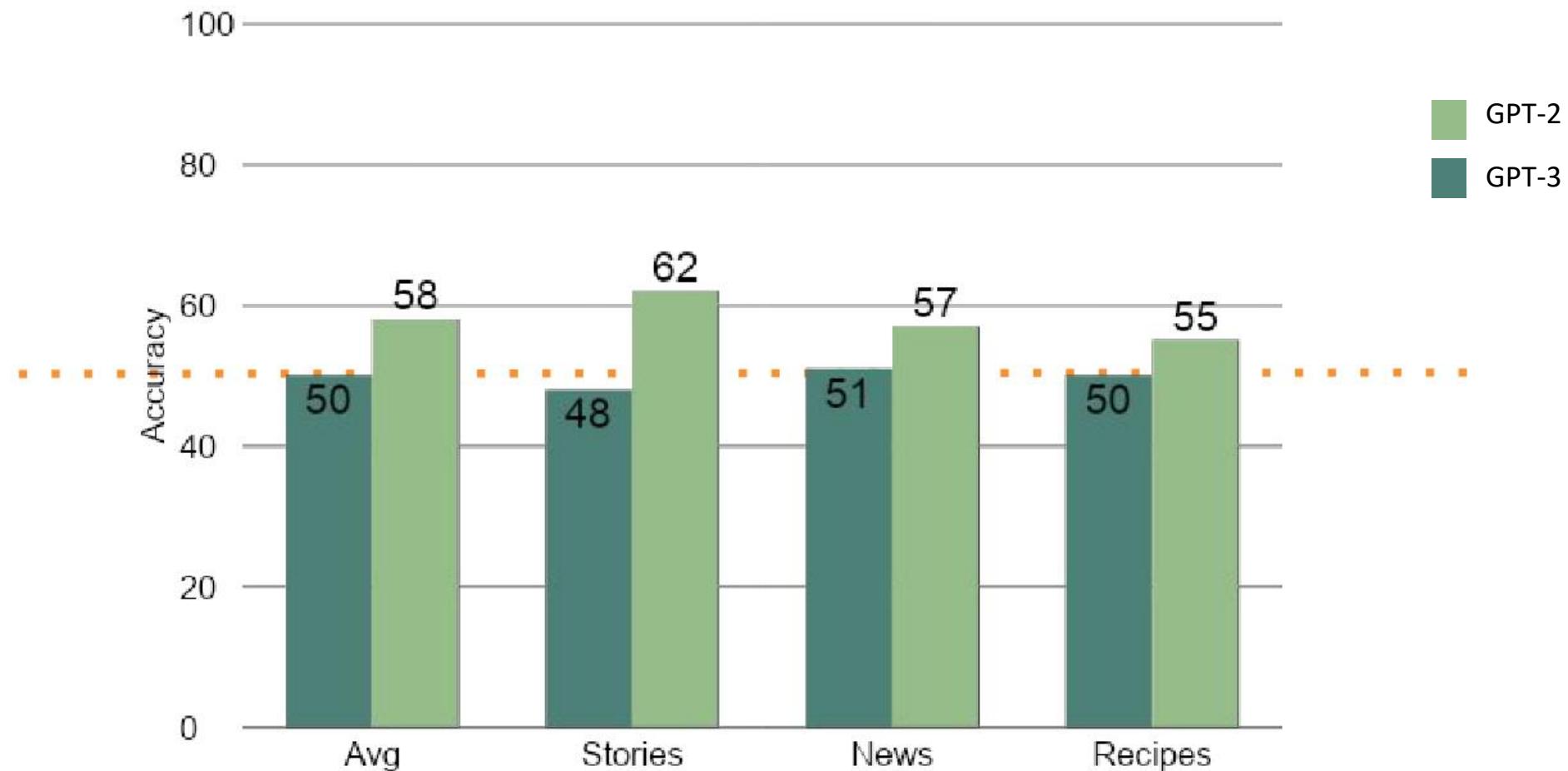
Experiment Setup

Model	GPT-2	GPT-3
Domain		
Evaluators	130 evaluators 	

780 evaluators, 3900 judgments

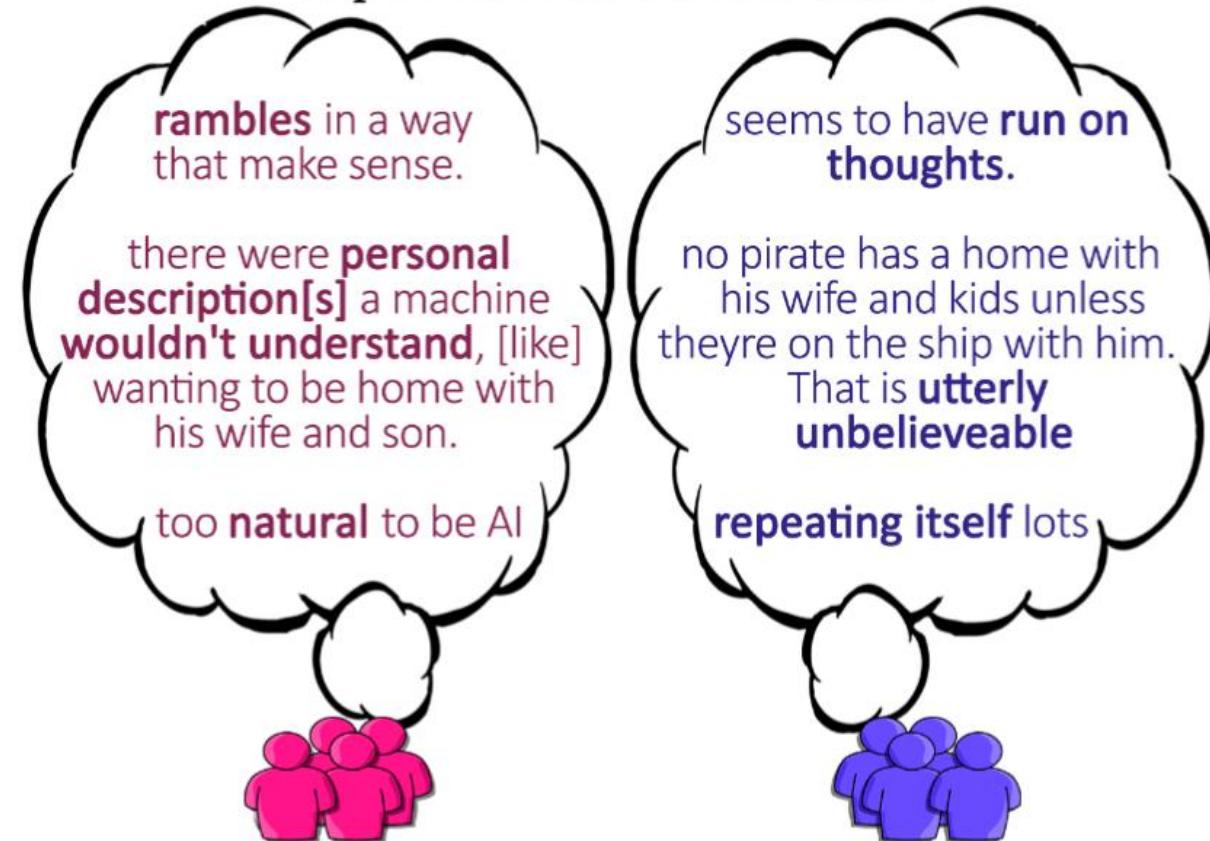
[All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text](#) (Clark et al., ACL-IJCNLP 2021)

Accuracy of Identifying Machine-Generated Text

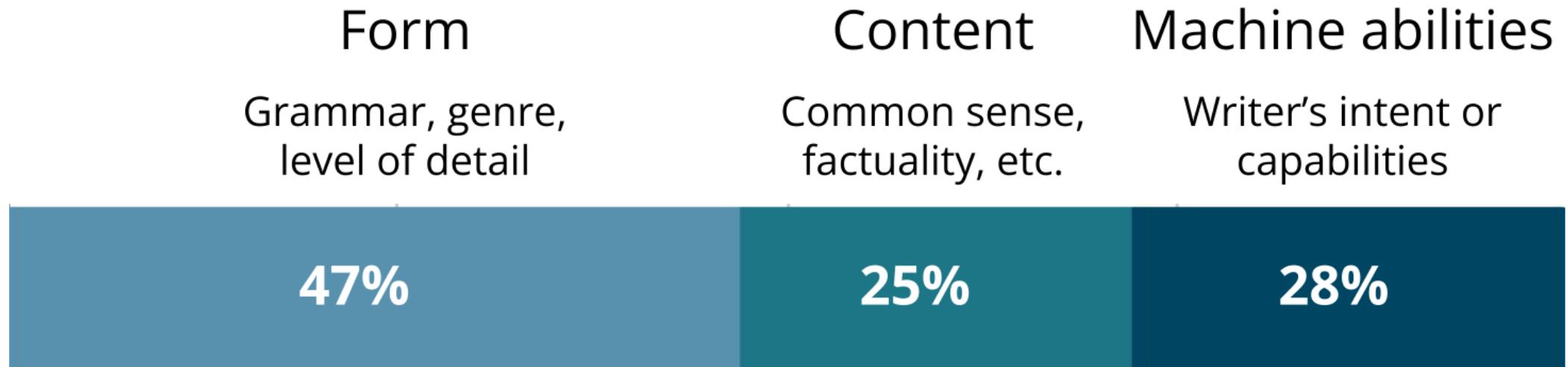


Contradicting Opinions

Once upon a time, there lived a pirate. He was the sort of pirate who would rather spend his time chasing away the sharks swimming around his ship than sail to foreign ports in search of booty. He was a good pirate, a noble pirate, an honest pirate. He was a pirate who would rather be at home with his wife and son than out on a ship in the middle of the ocean.



What did evaluators say they based their answers on?



[All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text](#) (Clark et al., ACL-IJCNLP 2021)

Can we train human evaluators to do better?

Once upon a time, there was a man in a place that was not a place at all.

He didn't know anything of a place or a time or who he was or what he was doing there. There was just him and the silence.

He sat there for a long time, not knowing what he was doing there. He thought, thought and thought, but he didn't know what to think. There was just him and the silence. He tried to speak, but no sound came from his mouth. He tried to move, but his body would not move. He sat there, but he didn't know for how long he was there.

* What do you think the source of this text is?

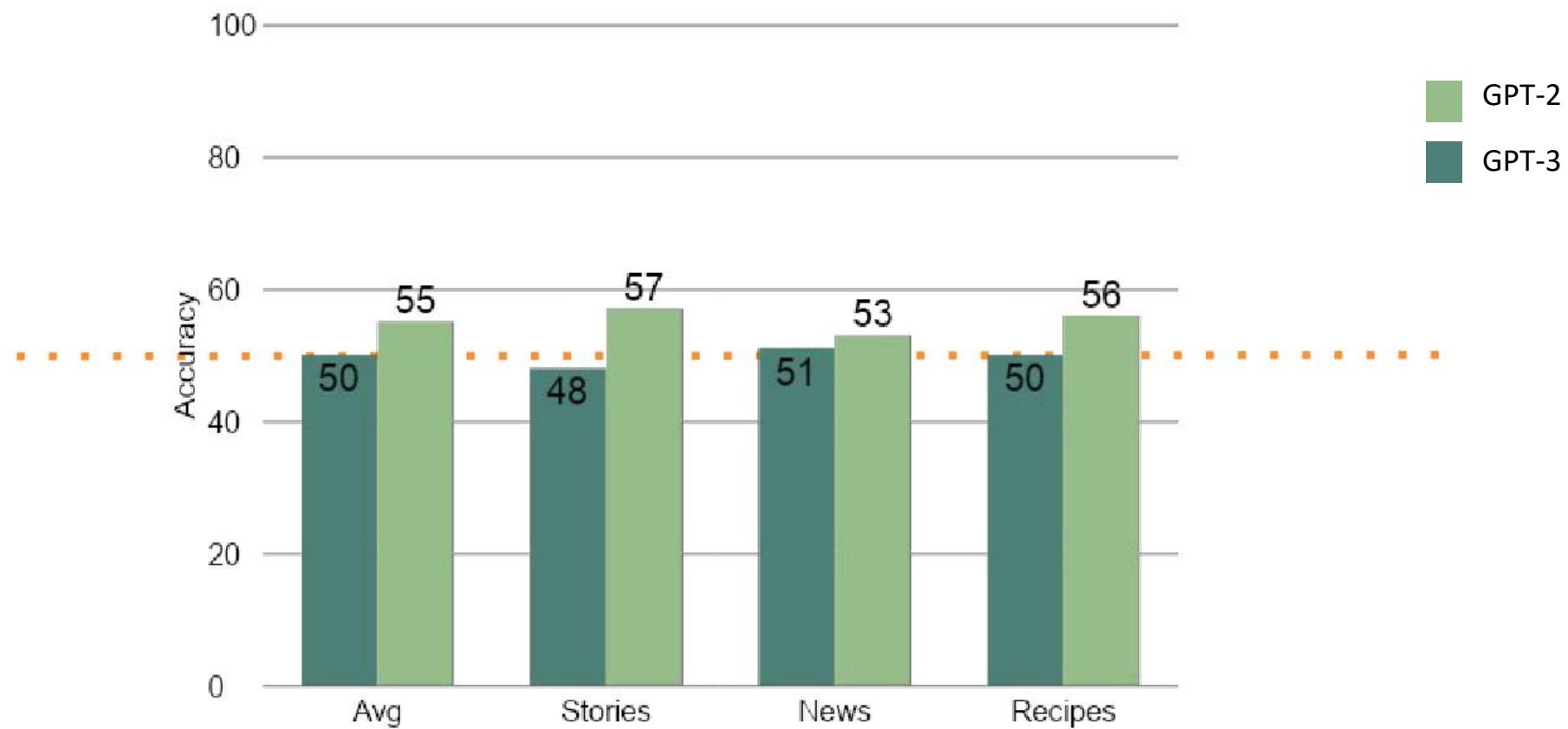
- Definitely human-written**
- Possibly human-written**
- Possibly machine-generated**
- Definitely machine-generated -- Correct Answer**

You cannot change your answer once you click submit.

Explanation

Note how the story is repetitive and doesn't seem to go anywhere.

Accuracy After Training



All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text (Clark et al., ACL-IJCNLP 2021)

“People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text”

METRIC	NONEXPERTS	EXPERTS
Avg. TPR	56.7	92.7
Avg. FPR	51.7	4.0
Avg. Confidence	4.027	4.394

Table 2: On average, nonexperts perform similar to random chance at detecting AI-generated text, while experts are highly accurate.

In Alaska, a pilot drops turkeys to rural homes for Thanksgiving

A half-dozen villagers in Napakiak, on the Kuskokwim River’s west bank, gathered near a gravel airstrip last Thursday to watch a small plane circle overhead. ... This crowd was waiting for a seasoned pilot who had a tradition: dropping Thanksgiving turkeys to homes scattered across miles of tundra and frozen waterways.

The pilot, 47-year-old Alaskan flyer Erik Fosnes, has been doing this for nearly a decade, working with volunteers from a regional nonprofit called Delta North Outreach. “We tried shipping turkeys one year by cargo, but half never made it in time,” said Fosnes, running a hand through the frost on his jacket sleeve after landing. “So I said, ‘What if I just fly them in myself?’” He shrugged as if that were the most ordinary idea, then laughed. “Folks around here have gotten used to it.”

Looks human-written

Looks AI-generated



Annotator #4
content writer,
frequently uses ChatGPT

Annotator's Decision



AI-generated

Confidence

1

(Least Confident)

4

5

(Most Confident)

Explanation

... Lots of the quotes felt realistic, but many of the quotes did not need a narration alongside it such as with “He shrugged as if that were the most ordinary idea, then laughed.” ... could have been shortened to get more facts in about what people in Alaska face and why they face such limited transportation from the rest of the world. Also, it got sentimental and corny at times too.

Collaborative Story Writing



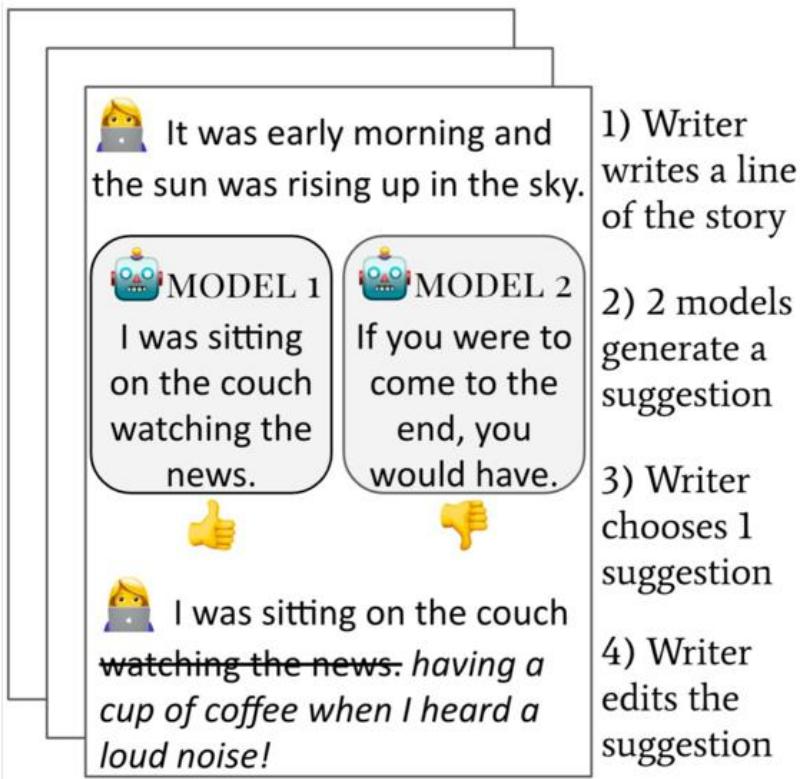
- 1) Writer writes a line of the story
- 2) Model generates a suggestion
- 3) Writer edits the suggestion

I liked the suggestions I received.



All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text (Clark et al., ACL-IJCNLP 2021)

“Choose Your Own Adventure” evaluation



[Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models](#) (Clark & Smith, NAACL 2021)

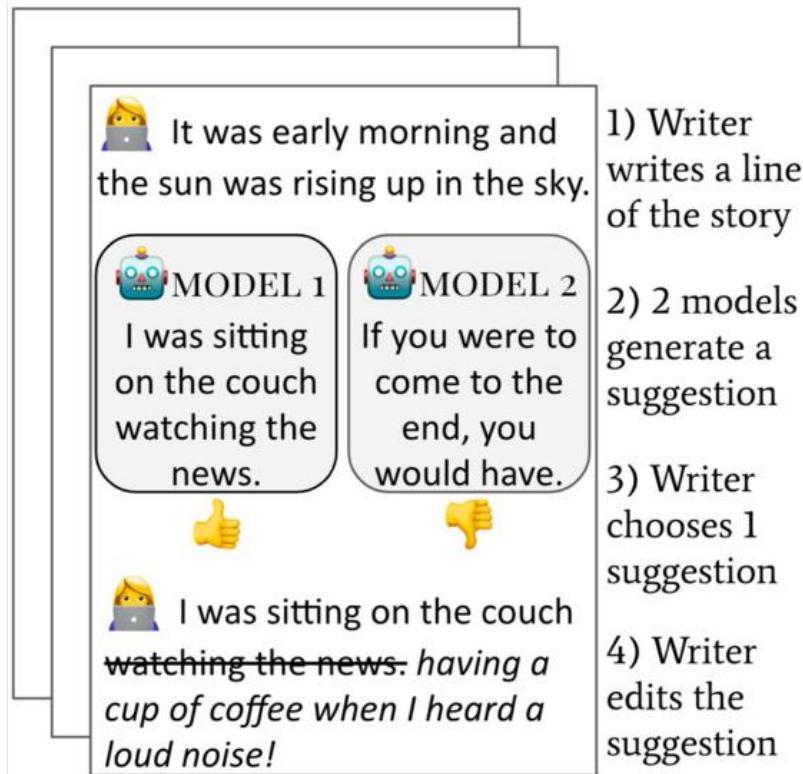
“Choose Your Own Adventure” evaluation

Human-authored
text

Machine-generated
text

Writer preference

Writer revisions



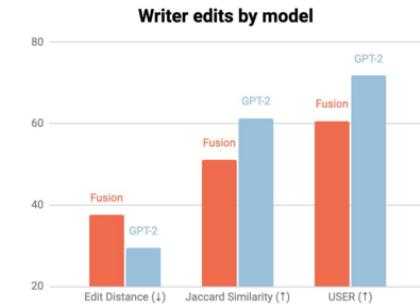
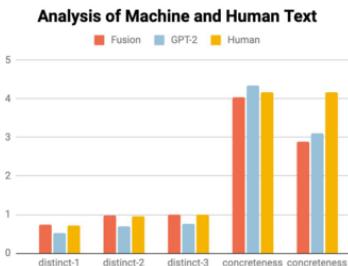
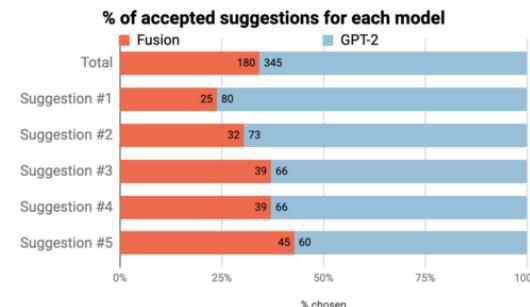
[Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models](#) (Clark & Smith, NAACL 2021)

“Choose Your Own Adventure” evaluation

Is my model better at generating
story suggestions than a
baseline model?

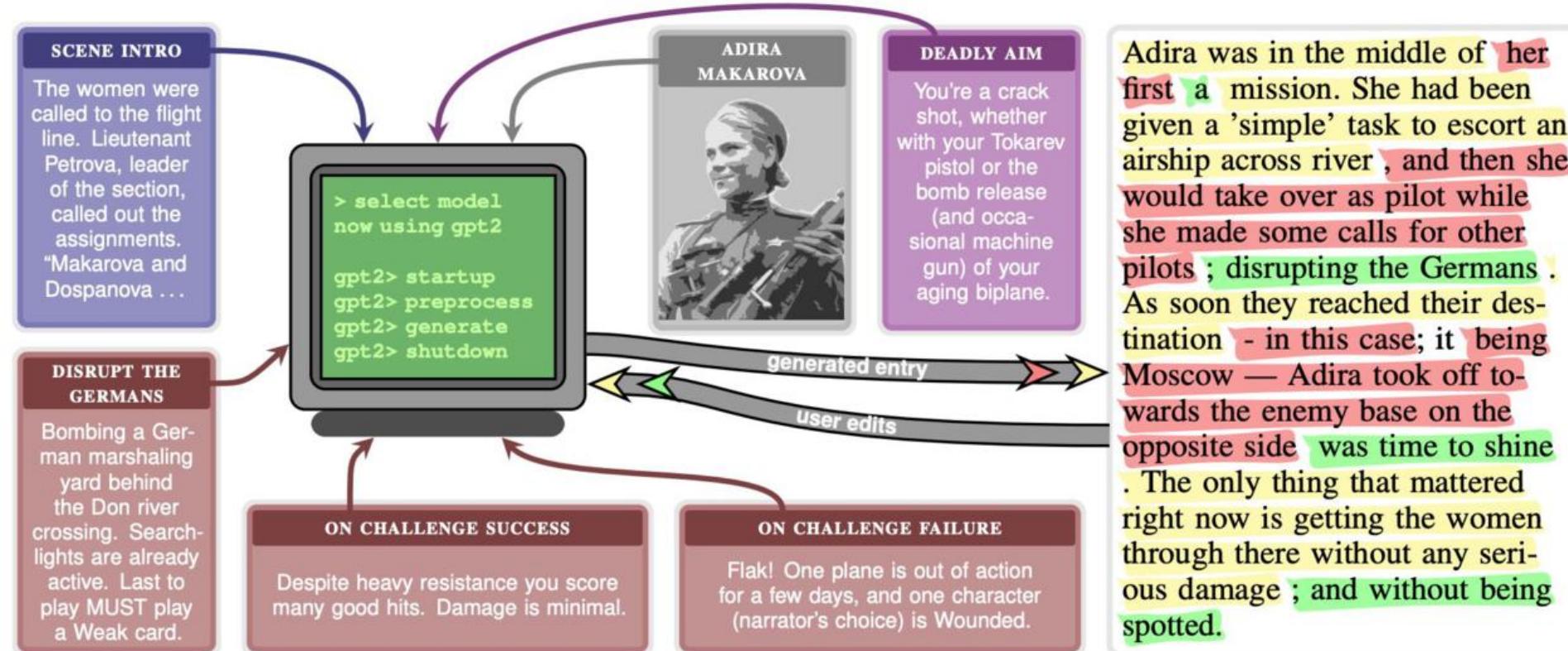
How useful are the models'
suggestions?

How does the model-generated
text compare to human-
authored text?



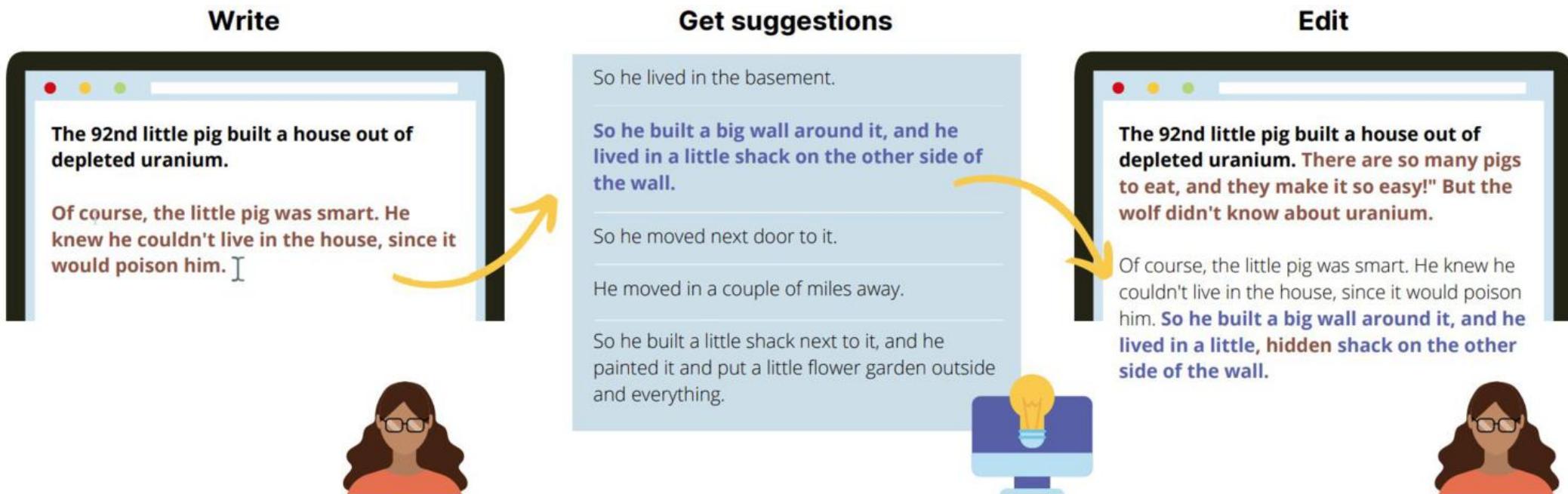
[Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models](#) (Clark & Smith, NAACL 2021)

STORIUM



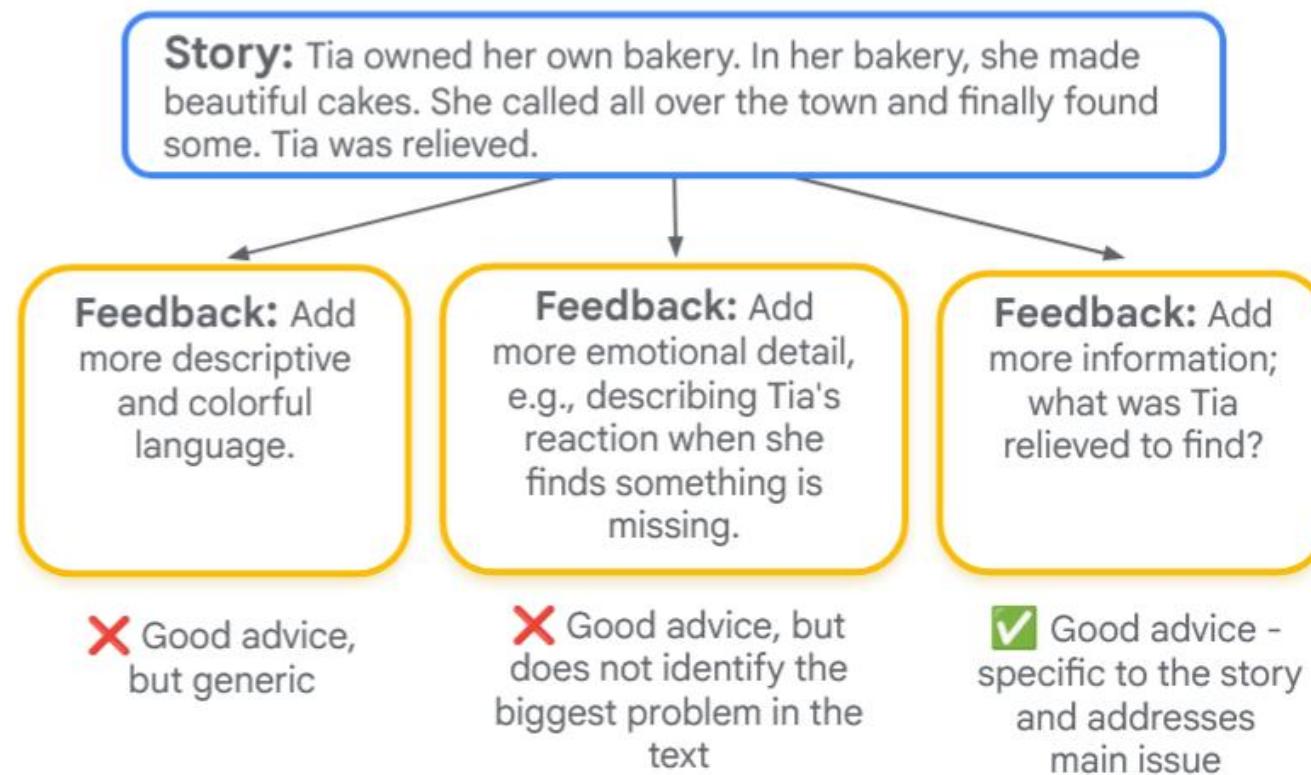
STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation (Akoury et al., EMNLP 2020)

CoAuthor



Mina Lee, et al. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Article 388, 1–19.
<https://doi.org/10.1145/3491102.3502030>

Generating Feedback



[Help Me Write a Story: Evaluating LLMs' Ability to Generate Writing Feedback](#) (Rashkin et al., ACL 2025)

Best Practice & Implementation	Yes	No	%
Make informed evaluation choices and document them			
Evaluate on multiple datasets	47	9	83.9
Motivate dataset choice(s)	21	34	38.2
Motivate metric choice(s)	20	46	30.3
Evaluate on non-English language	19	47	28.8
Measure specific generation effects			
Use a combination of metrics from at least two different categories	36	27	57.1
Avoid claims about overall “quality”	34	31	52.3
Discuss limitations of using the proposed method	19	46	29.2
Analyze and address issues in the used dataset(s)			
Discuss or identify issues with the data	19	47	28.8
Contribute to the data documentation or create it if it does not yet exist	1	58	1.7
Address these issues and release an updated version	3	10	23.1
Create targeted evaluation suite(s)	14	52	21.2
Release evaluation suite or analysis script	3	63	4.5
Evaluate in a comparable setting			
Re-train or -implement most appropriate baselines	40	19	67.8
Re-compute evaluation metrics in a consistent framework	38	22	63.3
Run a well-documented human evaluation			
Run a human evaluation to measure important quality aspects	48	18	72.7
Document the study setup (questions, measurement instruments, etc.)	40	9	81.6
Document who is participating in the study	28	20	58.3
Produce robust human evaluation results			
Estimate the effect size and conduct a power analysis	0	48	0.0
Run significance test(s) on the results	12	36	25.0
Conduct an analysis of result validity (agreement, comparison to gold ratings)	19	29	39.6
Discuss the required rater qualification and background	10	38	20.8
Document results in model cards			
Report disaggregated results for subpopulations	13	53	19.7
Evaluate on non-i.i.d. test set(s)	14	52	21.2
Analyze the causal effect of modeling choices on outputs with specific properties	16	50	24.2
Conduct an error analysis and/or demonstrate failures of a model	15	51	22.7

Considerations for collaborative story evaluation design

What aspects of the generated text do you care about evaluating most?

What collaborative role is the model playing?

Who is the audience for the model?

Tradeoffs between quality of the evaluation and the quality of the writing experience

Combinations of evaluation types and methods

Comparisons to previous methods Investigate errors and potential weaknesses

When reporting evaluation results, explain:

- What you did
- Why you did it
- Possible shortcomings