Story Evaluation

Elizabeth Clark       4/14/22
Interactive Fiction
The story generation content and examples in today’s talk are mainly from work that is:

- Academic
- From/for the NLP community
- Text-based
- Collaborative
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 Evaluation
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

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

*Evaluation of Text Generation: A Survey* Celikyilmaz et al., 2020
Example: ROUGE

**Candidate:** my favorite food is pineapple

**Reference:** pineapple is my favorite tropical fruit

\[
\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_\text{match}(\text{gram}_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}
\]

- \(n=1\): 4 matches out of 6 \hspace{1cm} \text{ROUGE-1: 0.67}
- \(n=2\): 1 match out of 5 \hspace{1cm} \text{ROUGE-2: 0.20}
- \(n=3\): 0 matches out of 4 \hspace{1cm} \text{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 Kusner et al., 2015
Example: BERTScore

Reference $X$
the weather is cold today

Candidate $\hat{X}$
it is freezing today

Contextual Embedding → Pairwise Cosine Similarity → Maximum Similarity

$R_{BERT} = \frac{(0.713 \times 1.27) + (0.515 \times 7.94) + ...}{1.27 + 7.94 + 1.82 + 7.90 + 8.88}$

BERTScore: Evaluating Text Generation with BERT Zhang et al., 2020
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., 2019
Example: Self-BLEU


Image: [The Curious Case of Neural Text Degeneration](https://www.amegroups.com/article/view/173577) 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., 2019
Example: Fusion model vs. GPT-2

**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.

**GPT-2-117** ($k = 10$):
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: 3.892
Mean verb concreteness: 2.173

Mean noun concreteness: 4.720
Mean verb concreteness: 2.488

**Do Massively Pretrained Language Models Make Better Storytellers?** See et al., 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

[Diagram showing the process of pre-training and fine-tuning with BERT, including steps for synthetic sentence pairs and public human ratings, with an optional step for application-specific human ratings.]

[BLEURT: Learning Robust Metrics for Text Generation Sellam et al., 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.

UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation Guan and Huang, 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.

---

*UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation* Guan and Huang, 2020
Human Evaluation
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 the human evaluation...?:

- Experts?
- In-person?
- Crowdsourced?
- Paid?
- Trained?
- Quality-controlled?
Dimensions of text quality

Is the text...

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

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

Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions
Howcroft et al., 2020
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

PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking Rashkin et al., 2020
Case study: PlotMachines

### PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking
Rashkin et al., 2020

It is Big Bird’s birthday, and he goes to the roller skating rink with his friends. Back at Sesame Street, Maria and Susan take out the big birthday cake and leave it on a table. Cookie Monster sees the cake, but instead of eating it and spoiling the party, he eats a chair and other things all over Sesame Street.

Big Bird and the other skaters return to Sesame Street and are shocked at what Cookie Monster ate, though the cake is safe. Gina and Count Von Count presents the cake to Big Bird. It has 548 candles even though Big Bird is 6 years old. At the end, when Gina announces the sponsors, Cookie Monster eats them along with his cake.
## PlotMachines: Automatic evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Wikiplots</th>
<th>Writing Prompts</th>
<th>New York Times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
<td>R-L</td>
</tr>
<tr>
<td>P&amp;W-Static (Yao et al., 2019)</td>
<td>17.0</td>
<td>3.3</td>
<td>13.6</td>
</tr>
<tr>
<td>Fusion (Fan et al., 2018)</td>
<td>22.7</td>
<td>6.0</td>
<td>17.4</td>
</tr>
<tr>
<td>GroVER (Zellers et al., 2019)</td>
<td>19.6</td>
<td>5.9</td>
<td>12.5</td>
</tr>
<tr>
<td>PlotMachines (GPT)</td>
<td>20.2</td>
<td>5.3</td>
<td>16.0</td>
</tr>
<tr>
<td>– base (GPT) (Radford et al., 2018)</td>
<td>13.2</td>
<td>2.0</td>
<td>7.9</td>
</tr>
<tr>
<td>PlotMachines (GPT-2)</td>
<td>22.8</td>
<td>6.5</td>
<td>17.5</td>
</tr>
<tr>
<td>– PM-NoMEM (GPT-2)</td>
<td>20.5</td>
<td>4.9</td>
<td>15.5</td>
</tr>
<tr>
<td>– PM-NoMEM-NoDisc (GPT-2)</td>
<td>19.3</td>
<td>1.7</td>
<td>13.9</td>
</tr>
<tr>
<td>– base (GPT-2) (Radford et al., 2019)</td>
<td>18.5</td>
<td>3.9</td>
<td>13.3</td>
</tr>
</tbody>
</table>

## PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking

Rashkin et al., 2020
PlotMachines: Human evaluation

% select PLOTMACHINES vs. other model

<table>
<thead>
<tr>
<th>Uses outline more</th>
<th>Not repetitive</th>
<th>Has Transitions</th>
<th>Story relevant</th>
<th>Has beginning</th>
<th>Has ending</th>
<th>Logical order</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
<td>PLOTMACHINES</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Narrative Flow</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rep(↓)</td>
<td>Tran(↑)</td>
</tr>
<tr>
<td>Fusion</td>
<td>2.61</td>
<td>2.98</td>
</tr>
<tr>
<td>GPT</td>
<td><strong>1.39</strong></td>
<td>1.89</td>
</tr>
<tr>
<td>GROVER</td>
<td>1.78</td>
<td>3.00</td>
</tr>
<tr>
<td>PM</td>
<td>1.64</td>
<td><strong>3.02</strong></td>
</tr>
</tbody>
</table>

PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking
Rashkin et al., 2020
Collaborative Story Evaluation
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:

Characters: 0

Click here to submit the finished story and answer evaluation questions: Submit Story


Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories Clark et al., 2018
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

Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories Clark et al., 2018
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


Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories Clark et al., 2018
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


Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories Clark et al., 2018
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

Is the final story:
Creative?
Coherent?
Entertaining?
Grammatical?

Clark et al., 2018
Types of evaluation for collaborative writing

1. Automatic metrics

2. Human evaluations

3. Interaction metrics
   - Edit distance
   - % suggestions accepted
   - Time to complete the story

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

*Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling*
Swanson and Gordon, 2012
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

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
<tr>
<td>Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

Language Models are Unsupervised Multitask Learners Radford et al., 2019
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.”
Experiment setup

Model
- GPT-2
- GPT-3

Domain
- Book
- News
- Recipe

Evaluators
- 130 evaluators

780 evaluators, 3900 judgments

All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text
Clark et al., 2021
Accuracy results

All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text Clark et al., 2021
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.

*Rambles* in a way that make sense.

*There were* personal description[s] a machine wouldn't understand, [like] wanting to be home with his wife and son.

Too *natural* to be AI.

Seems to have run on *thoughts*.

No pirate has a home with his wife and kids unless they're on the ship with him. That is utterly unbelievable.

*Repeating itself* lots.

A human wrote this.

A machine wrote this.

*All That's ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text* Clark et al., 2021
What did evaluators say they based their answers on?

<table>
<thead>
<tr>
<th>Form</th>
<th>Content</th>
<th>Machine abilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar, genre, level of detail</td>
<td>Common sense, factuality, etc.</td>
<td>Writer’s intent or capabilities</td>
</tr>
<tr>
<td>47%</td>
<td>25%</td>
<td>28%</td>
</tr>
</tbody>
</table>

All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text Clark et al., 2021
Can we train 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., 2021
Collaborative story writing

I liked the suggestions I received.

Not at all                                                   Extremely

1) Writer writes a line of the story
2) Model generates a suggestion
3) Writer edits the suggestion
“Choose Your Own Adventure” evaluation

Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models
Clark and Smith, 2021
“Choose Your Own Adventure” evaluation

<table>
<thead>
<tr>
<th>Human-authored text</th>
<th>Machine-generated text</th>
<th>Writer preferences</th>
<th>Writer revisions</th>
</tr>
</thead>
</table>
| It was early morning and the sun was rising up in the sky. | MODEL 1: I was sitting on the couch watching the news. | 1) Writer writes a line of the story | I was sitting on the couch watching the news. 
   having a cup of coffee when I heard a loud noise! |
| MODEL 2: If you were to come to the end, you would have. | 2) 2 models generate a suggestion | 3) Writer chooses 1 suggestion | |

Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models
Clark and Smith, 2021
“Choose Your Own Adventure” evaluation

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

2. How useful are the models’ suggestions?

3. 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 and Smith, 2021
Adira was in the middle of her first mission. She had been given a 'simple' task to escort an airship across river, and then she would take over as pilot while she made some calls for other pilots; disrupting the Germans. As soon they reached their destination - in this case; it being Moscow — Adira took off towards the enemy base on the opposite side. The only thing that mattered right now is getting the women through there without any serious damage; and without being spotted.
The 92nd little pig built a house out of depleted uranium.

Of course, the little pig was smart. He knew he couldn’t live in the house, since it would poison him.

So he lived in the basement.

So he built a big wall around it, and he lived in a little shack on the other side of the wall.

He moved in a couple of miles away.

So he built a little shack next to it, and he painted it and put a little flower garden outside and everything.

The 92nd little pig built a house out of depleted uranium. There are so many pigs to eat, and they make it so easy! But the wolf didn’t know about uranium.

Of course, the little pig was smart. He knew he couldn’t live in the house, since it would poison him. So he built a big wall around it, and he lived in a little, hidden shack on the other side of the wall.

CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities
Lee et al., 2022
Collaborative writing for better model evaluation

Collaborative story writing as:

✍ 1. An engaging and useful tool for writers

👩‍💻 2. An evaluation platform for NLP researchers

---

*Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories* Clark et al., 2018
Recommendations
## Recommendations for designing evaluations

<table>
<thead>
<tr>
<th>Best Practice &amp; Implementation</th>
<th>Yes</th>
<th>No</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Make informed evaluation choices and document them</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluate on multiple datasets</td>
<td>47</td>
<td>9</td>
<td>83.9</td>
</tr>
<tr>
<td>Motivate dataset choice(s)</td>
<td>21</td>
<td>34</td>
<td>38.2</td>
</tr>
<tr>
<td>Motivate metric choice(s)</td>
<td>20</td>
<td>46</td>
<td>30.3</td>
</tr>
<tr>
<td>Evaluate on non-English language</td>
<td>19</td>
<td>47</td>
<td>28.8</td>
</tr>
<tr>
<td><strong>Measure specific generation effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use a combination of metrics from at least two different categories</td>
<td>36</td>
<td>27</td>
<td>57.1</td>
</tr>
<tr>
<td>Avoid claims about overall “quality”</td>
<td>34</td>
<td>31</td>
<td>52.3</td>
</tr>
<tr>
<td>Discuss limitations of using the proposed method</td>
<td>19</td>
<td>46</td>
<td>29.2</td>
</tr>
<tr>
<td><strong>Analyze and address issues in the used dataset(s)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discuss or identify issues with the data</td>
<td>19</td>
<td>47</td>
<td>28.8</td>
</tr>
<tr>
<td>Contribute to the data documentation or create it if it does not yet exist</td>
<td>1</td>
<td>58</td>
<td>1.7</td>
</tr>
<tr>
<td>Address these issues and release an updated version</td>
<td>3</td>
<td>10</td>
<td>23.1</td>
</tr>
<tr>
<td>Create targeted evaluation suite(s)</td>
<td>14</td>
<td>52</td>
<td>21.2</td>
</tr>
<tr>
<td>Release evaluation suite or analysis script</td>
<td>3</td>
<td>63</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>Evaluate in a comparable setting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re-train or -implement most appropriate baselines</td>
<td>40</td>
<td>19</td>
<td>67.8</td>
</tr>
<tr>
<td>Re-compute evaluation metrics in a consistent framework</td>
<td>38</td>
<td>22</td>
<td>63.3</td>
</tr>
<tr>
<td><strong>Run a well-documented human evaluation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run a human evaluation to measure important quality aspects</td>
<td>48</td>
<td>18</td>
<td>72.7</td>
</tr>
<tr>
<td>Document the study setup (questions, measurement instruments, etc.)</td>
<td>40</td>
<td>9</td>
<td>81.6</td>
</tr>
<tr>
<td>Document who is participating in the study</td>
<td>28</td>
<td>20</td>
<td>58.3</td>
</tr>
<tr>
<td><strong>Produce robust human evaluation results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate the effect size and conduct a power analysis</td>
<td>0</td>
<td>48</td>
<td>0.0</td>
</tr>
<tr>
<td>Run significance test(s) on the results</td>
<td>12</td>
<td>36</td>
<td>25.0</td>
</tr>
<tr>
<td>Conduct an analysis of result validity (agreement, comparison to gold ratings)</td>
<td>19</td>
<td>29</td>
<td>39.6</td>
</tr>
<tr>
<td>Discuss the required rater qualification and background</td>
<td>10</td>
<td>38</td>
<td>20.8</td>
</tr>
<tr>
<td><strong>Document results in model cards</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Report disaggregated results for subpopulations</td>
<td>13</td>
<td>53</td>
<td>19.7</td>
</tr>
<tr>
<td>Evaluate on non-i.i.d. test set(s)</td>
<td>14</td>
<td>52</td>
<td>21.2</td>
</tr>
<tr>
<td>Analyze the causal effect of modeling choices on outputs with specific properties</td>
<td>16</td>
<td>50</td>
<td>24.2</td>
</tr>
<tr>
<td>Conduct an error analysis and/or demonstrate failures of a model</td>
<td>15</td>
<td>51</td>
<td>22.7</td>
</tr>
<tr>
<td><strong>Release model outputs and annotations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Release outputs on the validation set</td>
<td>1</td>
<td>65</td>
<td>1.5</td>
</tr>
<tr>
<td>Release outputs on the test set</td>
<td>2</td>
<td>63</td>
<td>3.1</td>
</tr>
<tr>
<td>Release outputs for non-English dataset(s)</td>
<td>1</td>
<td>25</td>
<td>3.8</td>
</tr>
<tr>
<td>Release human evaluation annotations</td>
<td>1</td>
<td>47</td>
<td>2.1</td>
</tr>
</tbody>
</table>
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