All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

Elizabeth Clark[1] Tal August[1] Sofia Serrano[1] Nikita Haduong[1] Suchin Gururangan[1] Noah A. Smith[1,2]

[1] Paul G. Allen School of Computer Science & Engineering, University of Washington

[2] Allen Institute for Artificial Intelligence

{eaclark7,taugust,sofias6,qu,sg01,nasmith}@cs.washington.edu

Abstract

Human evaluations are typically considered the gold standard in natural language generation, but as models' fluency improves, how well can evaluators detect and judge machine-generated text? We run a study assessing non-experts' ability to distinguish between human- and machine-authored text (GPT2 and GPT3) in three domains (stories, news articles, and recipes). We find that, without training, evaluators distinguished between GPT3- and human-authored text at random chance level. We explore three approaches for quickly training evaluators to better identify GPT3-authored text (detailed instructions, annotated examples, and paired examples) and find that while evaluators' accuracy improved up to 55%, it did not significantly improve across the three domains. Given the inconsistent results across text domains and the often contradictory reasons evaluators gave for their judgments, we examine the role untrained human evaluations play in NLG evaluation and pro- vide recommendations to NLG researchers for improving human evaluations of text generated from state-of-the-art models.

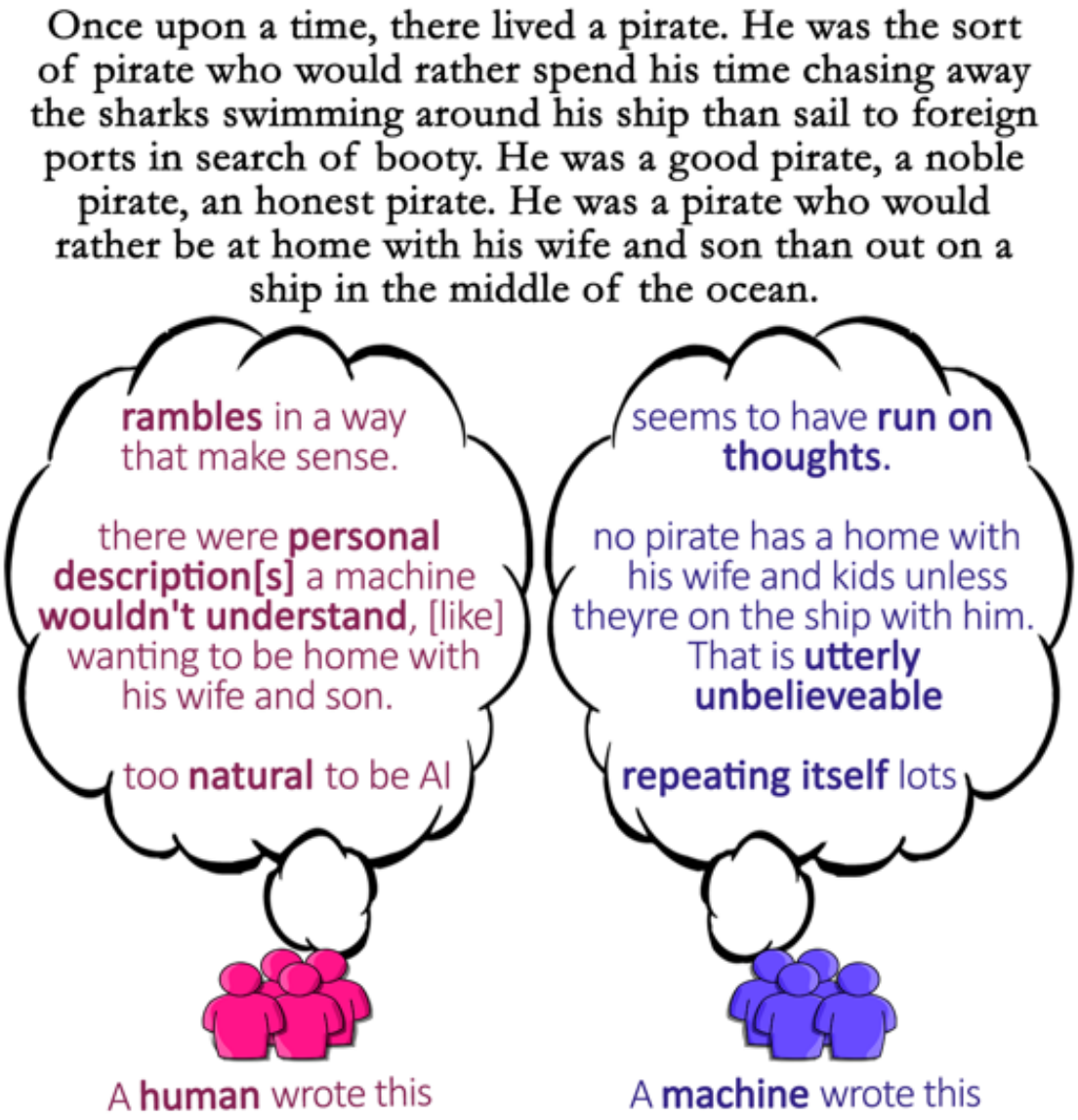


Figure 1: Excerpts from human evaluators' explanations for why they believe a GPT3-generated story (also excerpted) was written by a human (left) or a machine (right). The evaluators point to a wide range of text attributes to make their decisions, sometimes using the same aspect of the text to come to opposite conclusions.

1 Introduction

Human-quality text has long been a holy grail for the output of natural language generation (NLG) systems, serving as an upper bound on their performance. Since

we lack a good way of encoding many aspects of what constitutes human-quality output in an automated method, we often must rely on human evaluation for our models. Though evaluations with end-users in an applied setting are encouraged (Belz and Reiter, 2006), in practice, most human evaluations instead ask people to rate generated text's *intrinsic* quality (van der Lee et al., 2019; Howcroft et al., 2020). Sometimes the generated text is explicitly compared to human-authored text (e.g., Liu et al., 2016; Zellers et al., 2021; Zhang et al., 2020), but even when no human-authored text is evaluated, evaluators implicitly compare the generated text to their knowledge of language and norms within specific domains.

Evaluators are often asked to assess a text holistically, e.g., based on its overall quality, naturalness, or humanlikeness (van der Lee et al., 2021; Howcroft et al., 2020), where the exact evaluation criteria is left to the discretion of the evaluator. Though other evaluations are broken down along specific dimensions of text quality (e.g., grammaticality, coherence, etc.), Novikova et al. (2017, 2018) and Callison-Burch et al. (2007) found that these dimensions are often correlated and may be conflated in some evaluation settings. This is concerning because, as NLG models improve, evaluators are asked to read longer passages of text conditioned on large amounts of context. In these cases, fluency-related aspects of quality (i.e., the ones that don't require careful reading of the context and meaning of the passage) are the easiest to assess, particularly in small-batch evaluations with non-expert evaluators where speed is incentivized. This poses a challenge when collecting human evaluations for state-of-the-art language models, as errors are often content- based (e.g., factual inaccuracies or inconsistencies with the context) rather than fluency-based (Brown et al., 2020), so a superficial read may not be sufficient to catch model errors. For accurate assessments of generated text, we need human evaluations that are designed to encourage a sufficiently careful reading of the text to examine these subtler aspects of text quality.

We asked non-expert evaluators to assess the humanlikeness (operationalized as how believably human an evaluator finds a text) of text generated by current NLG models (GPT2 and GPT3) to test what current human evaluation practices can reveal about the models' quality (§2). We found that evaluators were unable to distinguish between GPT3- and human-authored text across story, news, and recipe domains. However, when we categorized the aspects of text the evaluators used to make their judgments, we found they primarily focused on the grammar, spelling, and style of the text. The evaluators' responses also indicated that they

underestimated the quality of text current models are capable of generating (as seen in Figure 1). To our knowledge, this paper is the first to evaluate human evaluations of GPT3-generated text across multiple domains.

We then looked at three different evaluator training methods—providing detailed instructions, annotated examples, and human-machine paired examples—to test whether we could improve evaluators' accuracy (§3). While we found including examples in the task increased the set of texts evaluators thought could be machine-generated and increased their focus on textual content, no training method significantly increased evaluators' performance consistently across domains.

Based on our results (discussed in §4), we recommend moving away from small-batch evaluations with little training when collecting human evaluations of NLG models (§5). We also encourage practitioners to consider alternative evaluation frameworks that capture the usefulness of generated text in downstream settings rather than its humanlikeness.

2 How well can untrained evaluators identify machine- generated text?

In our first study, we ask how well untrained evaluators can distinguish between human- and machine-generated text. This task format, inspired by the Turing (1950) Test, is used to compare the quality of machine-generated text to human-authored text and, as models' fluency improves, to analyze NLG models' ability to "fool" readers (Garbacea et al., 2019; Ippolito et al., 2020; Brown et al., 2020).

By asking evaluators to assess the humanlikeness of the text with only minimal instructions (see Figure 2), we observe how well untrained evaluators can detect state-of-the-art machine-generated text and which attributes evaluators focus on and think are important for detecting machine-generated text.

2.1 The Task

We gave evaluators 5 text passages, some of which were written by people and some generated by a model. We asked them to rate the text on a 4-point scale (Ippolito et al., 2020):

1. Definitely human-written

2. Possibly human-written
3. Possibly machine-generated
4. Definitely machine-generated

If they selected option 1, we asked them: “Why did you select this rating?” Otherwise, they were asked, “What would you change to make it seem more human-like?” The interface is shown in Figure 2.

2.2 Data

We considered human- and machine-generated text in three different domains: stories, news articles, and recipes. In all three cases, we collected 50 human-authored texts in English and generated 50 texts from both the 175B parameter GPT3 model (also known as Davinci; Brown et al., 2020)[1] and GPT2-XL (Radford et al., 2019).[2] Evaluators were assigned to one domain and one model; the texts read by any given evaluator included some human-authored texts and some texts generated by their assigned model. We only considered texts 100 words or longer, and after reaching 100 words, all texts were truncated at the end of the next sentence.[3]

[1] beta.openai.com/

[2] huggingface.co/gpt2-xl

[3] Using NLTK; www.nltk.org/

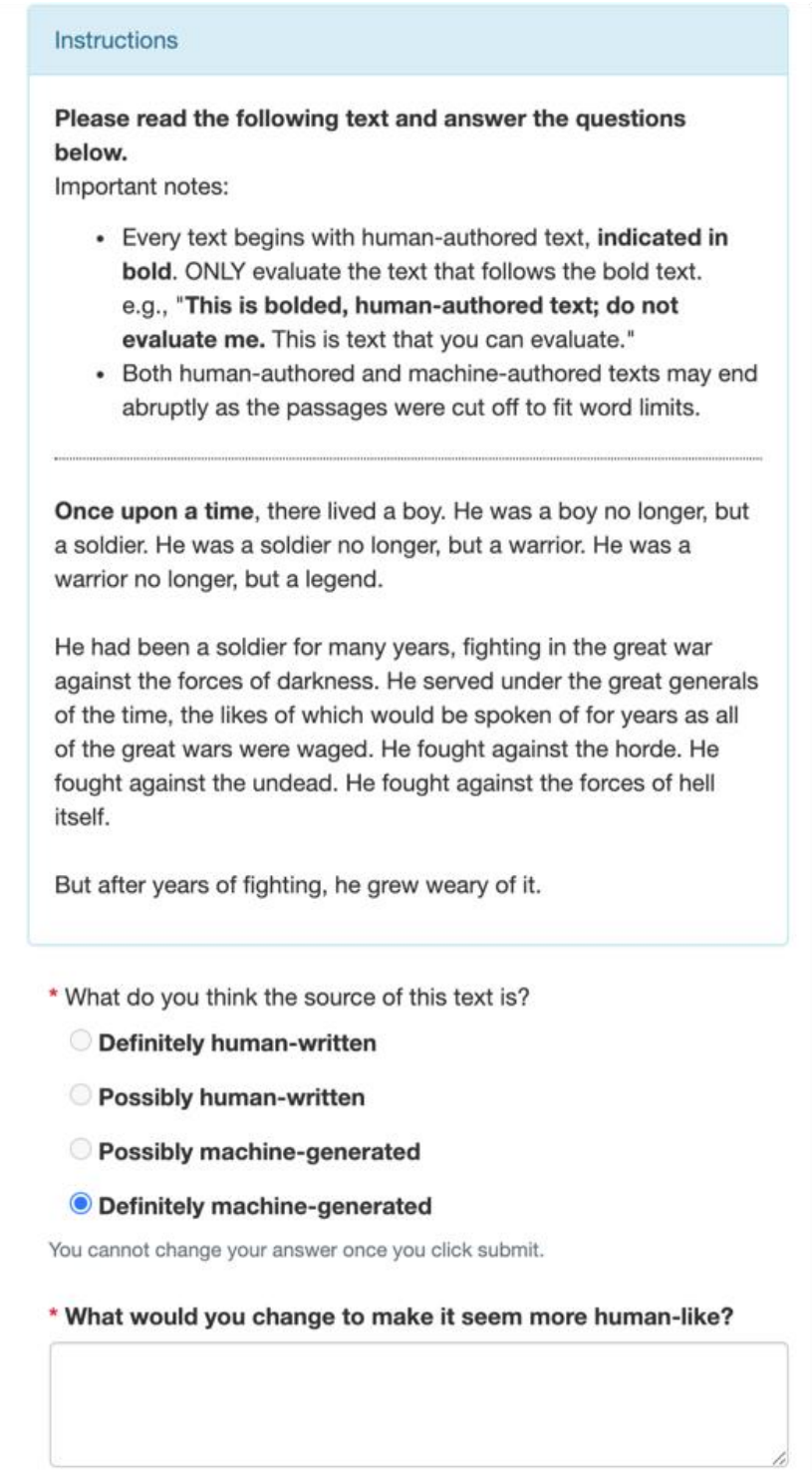


Figure 2: The task interface (story domain)

To generate text, we used the "three-shot" setting described in Brown et al. (2020), conditioning the text on three additional samples of in-domain, human-authored text, which we refer to as the *priming texts* (all priming texts are in the supplementary materials and at ark.cs.washington.edu/human_evals_ACL21). While this setting is not typically how GPT2 is used in practice, we held this approach constant to directly compare how model quality changes evaluators' ability to distinguish between texts. For each domain, each generated text was conditioned on the same set of priming texts. The texts were delimited with an ⟨EOS⟩ token and generated using the default GPT3 generation settings (i.e., sampling with temperature = 0.7).

2.2.1 Stories

The human-authored texts came from the Reddit WritingPrompts dataset (Fan et al., 2018).[4] We collected all the stories that began with *Once upon a time* (255 stories total) and randomly chose 50 human-authored stories from this set. For the machine-generated text, we conditioned the models on the three priming texts and on the phrase *Once upon a time*. We removed generated stories that directly copied a priming text (with > 80% overlap) and regenerated those texts (9 instances with GPT2, 2 with GPT3).

This is the most open-ended of the three domains, as the story's content is virtually unrestricted, and the only creative domain. It is also the noisiest of the human-authored datasets, as the stories were originally collected from social media comments with no quality-based filtering.

2.2.2 News Articles

We collected 2,111 recent local news articles from 15 different newspapers using Newspaper3k[5] (details in Appendix A.1). After filtering out articles under 100 words, we manually filtered out articles that weren't local news or that referenced the coronavirus pandemic. We randomly chose 50 articles to use as our human- authored news articles and another 50 to use as prompts for our generation models. We conditioned each generated text on the headline and first sentence from the prompt articles, along with the three priming texts.

---

[4] github.com/pytorch/fairseq/tree/master/examples/stories

[5] github.com/codelucas/newspaper

Because the title and the first sentence of a news article often summarize its contents, the generated content must adhere to the topics they introduce. By using local, recent news, we also limit the models' ability to copy from their training data. The models seemed to have the most trouble with this dataset structurally, e.g., generating new headlines without ending the current article or outputting invalid end-of-file tags.

2.2.3 Recipes

We collected 50 human-authored recipes from the RecipeNLG dataset (Bień et al., 2020), which contains 2,231,142 recipes scraped from the web. We randomly chose an additional 50 recipes and used their titles and ingredient lists as prompts, appending them to the end of the priming texts.

This is the most closed of the three domains, as the recipe must incorporate the listed ingredients and result in the dish described by the title. Recipes are typically written in clear commands, leaving little room for surprising or unexpected text.

2.3 Participants

We used Amazon Mechanical Turk (AMT) to collect the text evaluations with non-expert evaluators, commonly used in NLG evaluations (van der Lee et al., 2019). To have adequate power in our analyses (based on a power analysis with $\beta = 0.8$; Card et al., 2020), we had 130 different evaluators for each of the 6 task settings (3 domains × 2 models). Each participant evaluated 5 texts each, giving us a total of 780 participants and 3,900 text evaluations.

We paid evaluators US$1.25 for completing the task. Following common best practice on AMT (Berinsky et al., 2012), evaluators had to have over a 95% acceptance rate, be in the United States, and have completed over 1,000 HITs (AMT tasks). We excluded evaluators' work if their explanations were directly copied text from the task, did not match their responses, did not follow the instructions, or were short, vague, or otherwise uninterpretable. Across experiments, 445 participants (18.6%) were rejected and not included in the §2 results (780 approved participants) and §3 results (1,170 approved participants).

2.4 Results

Model	Overall acc.	Domain	Acc.	F1	Prec.	Recall	Kripp. α	% human	% confident
GPT2	*0.58	Stories	*0.62	0.60	0.64	0.56	0.10	55.23	52.00
		News	*0.57	0.52	0.60	0.47	0.09	60.46	51.38
		Recipes	0.55	0.48	0.59	0.40	0.03	65.08	50.31
GPT3	0.50	Stories	0.48	0.40	0.47	0.36	0.03	62.15	47.69
		News	0.51	0.44	0.54	0.37	0.05	65.54	52.46
		Recipes	0.50	0.41	0.50	0.34	0.00	66.15	50.62

Table 1: §2 results, broken down by domain and model, along with the F_1, precision, and recall at identifying machine-generated text, Krippendorff's α, % human-written guesses, and % confident guesses (i.e., *Definitely* machine- or human-authored). * indicates the accuracies significantly better than random (two-sided t-test, for Bonferroni-corrected $p < 0.00333$).

Overall, evaluators choosing between human and GPT2-generated text correctly identified the author of the text 57.9% of the time,[6] but the evaluators choosing between human- and GPT3-generated text only guessed correctly 49.9% of the time (Table 1), compared to 50% random chance. While the accuracy of classifying GPT2- vs. human-authored text is significantly[7] different from chance, evaluators' accuracy distinguishing GPT3- and human-authored text is not.[8] This remains the case regardless of text domain; we failed to find any evidence that evaluators' accuracy on any one domain for GPT3 differs from the overall GPT3 accuracy of ≈ 50%.[9] The story texts saw the biggest drop in evaluator accuracy from GPT2 to GPT3 (62% to 48%, Cohen's d = 0.57). The distribution of evaluators' scores are shown in Appendix A.2.

[6] Unless otherwise noted, all analyses binned the responses into 2 categories (*human* and *machine*).

[7] $t_{388} = 6.58$, $p < 0.0001$

[8] $t_{388} = -0.09$, $p = 0.93$

[9] ANOVA with $F_{2,390} = 0.78$, $p = 0.46$

In Table 1, we see other statistics worsen as well between GPT2 and GPT3: how well evaluators identified the machine-generated text (F_1, precision, and recall), evaluators' agreement (Krippendorff's α, a measure of annotator agreement that corrects for the probability of random agreement), and the percent of guesses that the text was human-written (% human). Given that the texts are equally likely to be human- and machine-written, there are disproportionately many *human* guesses, making up two thirds of the responses in the GPT3 experiments. Despite the significantly lower scores, evaluators' confidence (the percent of *Definitely* responses) remains fairly constant across conditions.

2.5 Analysis

Taken on its own, the evaluators' difficulty identifying GPT3-generated text compared to GPT2 points to the improvement of new NLG models. However, it also points to concerns about extending current human evaluation methodologies to state-of-the-art text generation. In particular, the evaluators' explanations reveal underlying confusion and misconceptions about state-of-the-art NLG.

To better understand what untrained evaluators focused on in the text to make their decisions, the authors annotated 150 random responses from the evaluators who distinguished between human- and GPT3-generated text (see Appendix A.3 for annotation details). We divided the text annotation labels into three categories: *form*, *content*, and *machine capabilities*. *Form* qualities focus on the format, style, and tone of the text, while *content* focuses on the text's meaning. We also coded for comments that explicitly referenced people's perceptions of what types of language machines are capable (or incapable) of generating (*machine capabilities*).

We found nearly twice as many comments about the form of the text than the content (*form*: 47% of labels, *content*: 25%). Evaluators in our sample focused most on the spelling, grammar, or punctuation of the texts (45 out of 150 comments) and the style or tone of the text (24 out of 150 comments). However, these dimensions of text are unlikely to be helpful in identifying text generated by current models, considering that GPT3 has already been shown to generate fluent text and to adapt easily to new generation domains (Brown et al., 2020).

We also found that the reasons evaluators gave for their answers often contradicted each other. The formality of the text, spelling and grammar errors,

and clarity were all cited to justify both *human* and *machine* judgments. This was also reflected in the low agreement scores between evaluators, with Krippendorff's $\alpha \approx 0$ across domains.

Evaluators' expectations about what NLG models are capable of ranged from thinking their text is already indistinguishable from human-authored text ("I have no idea if a human wrote anything these days. No idea at all.") to doubting machines' ability to use basic language ("Usually AI has terrible grammer [sic] and messes up."). But overall we found most evaluators' beliefs about generated language underestimated or misunderstood current NLG models, as seen in Appendix A.4.

3 Can we train evaluators to better identify machine-generated text?

Given evaluators' inability to distinguish GPT3- and human-authored text and their inconsistent reasoning for their decisions, we investigated whether there were simple ways of improving evaluators' ability to spot attributes of GPT3-generated text. Inspired by crowdsourcing research on guiding workers on writing or other subjective tasks (Kim et al., 2017; Mitra et al., 2015), we tested three *lightweight* evaluator-training methods to see if we could improve people's ability to identify machine-generated text while maintaining the short, low-cost nature of the evaluations.

3.1 Evaluator Training Methods

We considered 3 evaluator trainings that can be added to the beginning of a human evaluation task, at most requiring only 3 extra samples of human- and machine- generated text. To test the effectiveness of each type of training, we re-ran the experiments from §2, but this time, we prepended one of three evaluator-training methods to the evaluation task: an *instruction-based* training, an *example-based* training, and a *comparison-based* training. Screenshots of the training interfaces are in Appendix A.6; the full set of training materials are in the supplementary materials and at ark.cs.washington.edu/human_evals_ACL21.

Other than the training, the task setup was identical to the GPT3-based tasks in §2. We again ran the task on Amazon Mechanical Turk across three domains (stories, news, and recipes), using the same texts. As each individual participant

was only permitted to complete one set of evaluations, the set of evaluators who received these trainings was completely disjoint from the set of evaluators from our first study. The participants were subject to the same restrictions described in §2.3 and excluded according the same criteria; we did not use the trainings to filter out evaluators. For each domain and training method pair, we had 130 unique evaluators complete the task, giving us 5,850 text annotations from 1,170 evaluators.

3.1.1 Training with Instructions

To give evaluators a better sense of which parts of the text to pay attention to, we extended the original task instructions to include dimensions of the text that could be helpful for identifying machine-generated text (repetition and factuality) and ones that could be misleading (grammar, spelling, and style). We chose these dimensions based on previous work (Ippolito et al., 2020) and evaluators' comments in a pilot study (see Appendix A.5).

The Instructions training was the simplest of our 3 evaluator training methods. It was general enough to be applied across the 3 domains but provided little information about the quality and domain of text the evaluator would be rating. It did not increase the cost of collecting evaluations (US$1.25 per HIT) because it does not require any extra work on the part of the evaluator, though this also made it the easiest training to ignore. The instruction-based training is the most prescriptive of the training methods, as the researcher has to choose the dimensions they want the evaluators to focus on.

3.1.2 Training with Examples

Our Examples training consisted of 3 practice rounds of the actual task: given a text, guess if it is machine- or human-authored. We collected 3 additional texts in the same manner described in §2.2 and wrote a short explanation of which aspects of the text hinted at its source. After an evaluator makes their guess, the correct answer and explanation are shown. Each domain had its own set of examples and explanations.

By showing examples, this training helps set the evaluators' expectations about the quality of the human- and machine-generated text. We paid evaluators more for completing this task (US$1.75 per HIT) to compensate for the extra texts they needed to read. As with the instruction-based training, while pointing out specific

text dimensions can help evaluators focus on important features, it may also restrict their search space.

3.1.3 Training with Comparison

In the Comparison training, we took the example passages from the Examples training and paired them with a text from the opposite source (machine or human) that began with the same prompt. We asked evaluators to guess which of the two texts was the machine-generated one. We then provided the correct answer to the evaluator, along with the same explanations used in the Examples training.

This training allows evaluators to directly compare human and machine texts written from the same prompt. It is also the most expensive training, as it required evaluators to read three more passages than the Examples training; we paid evaluators US$2.25 per HIT.

3.2 Results

Training	Overall acc.	Domain	Acc.	F1	Prec.	Recall	Kripp. α	% human	% confident
None	0.50	Stories	0.48	0.40	0.47	0.36	0.03	62.15	47.69
		News	0.51	0.44	0.54	0.37	0.05	65.54	52.46
		Recipes	0.50	0.41	0.50	0.34	0.00	66.15	50.62
Instructions	0.52	Stories	0.50	0.45	0.49	0.42	0.11	57.69	45.54
		News	0.56	0.48	0.55	0.43	0.05	62.77	52.15
		Recipes	0.50	0.41	0.52	0.33	0.07	67.69	49.85
Examples	*0.55	Stories	0.57	0.55	0.58	0.53	0.06	53.69	64.31
		News	0.53	0.48	0.52	0.45	0.05	58.00	65.69
		Recipes	0.56	0.56	0.61	0.51	0.06	55.23	64.00
Comparison	0.53	Stories	0.56	0.56	0.55	0.57	0.07	48.46	56.62
		News	0.52	0.51	0.53	0.48	0.08	53.85	50.31
		Recipes	0.51	0.49	0.52	0.46	0.06	54.31	53.54

Table 2: §3 results, broken down by domain and training method, along with the F_1, precision, and recall at identifying machine-generated text, Krippendorff's α, % human-written guesses, and % confident guesses (i.e., *Definitely* machine- or human- authored). "None" training refers to the GPT3 results from §2. * indicates accuracies significantly better than None (no training; two-sided t-test, for Bonferroni-corrected $p < 0.00333$).

We found that while all 3 training methods improved evaluators' accuracy at identifying machine- vs. human-authored text over the no-training accuracy, the Examples training was the only one that showed significant improvement (see Table 2).[10]

Breaking down the results by domain, however, we find the Examples accuracy did not significantly increase over the no-training accuracy when considering any of the three domains individually. Even so, the significant difference in overall performance is mainly contributed by the story domain; when comparing evaluators' performance with no training to its Examples training counterpart, we see a change of 0.019 and 0.062 mean accuracy in the news and recipe domains, respectively, versus 0.086 on the story domain. This is perhaps due to the examples helping override the preconception that machines cannot generate "creative" text.

Across all 3 domains, the Examples and Comparison trainings produced the highest recall and F_1 scores for evaluators' judgments and decreased the percentage of texts they guessed were human-written, which indicate that evaluators were willing to consider a broader set of texts to be machine-generated than the evaluators in §2. However, despite the trainings and the increased proportion of confident responses, evaluator agreement remained low across domain and training settings ($\alpha \leq 0.11$), and higher agreement did not correspond to higher accuracy.

Training	Form	Content	Machine Capabilities
None	47.1	24.6	28.3
Examples	32.5	50.0	17.5

Table 3: % of annotation labels that reference the text's form and content and the evaluator's perception of machines' capabilities

[10] Tukey's HSD adjusted $p < 0.003$ for distinguishing between the Examples training and no training, $d = 0.25$

3.3 Analysis

We again annotated 150 comments along the dimensions listed in Appendix A.3, divided into *form*, *content*, and *machine capabilities* categories, this time from evaluators who received the best-performing Examples training. As shown in Table 3, we found that the proportion of *form* comments dropped in the sample of evaluators who went through the Examples training, while the proportion of *content* comments doubled. We also saw a drop in the number of comments mentioning evaluators' expectations of machine-generated text. While this change in focus doesn't necessarily correspond to correct judgments, *content* reasons are more in-line with current NLG model capabilities (Brown et al., 2020).

4 Discussion

Overall, none of our three training methods significantly improved evaluators' ability to detect machine-generated text reliably across text domains while still maintaining the small-batch nature of Amazon Mechanical Turk. This speaks to the improving quality of NLG models, but we also found that untrained evaluators mainly focused on the format of the text, deciding if it was human or machine-generated based on whether the text was grammatically or stylistically correct. This, combined with the high percentage of *human* guesses, the low recall scores for the *machine* guesses, and the evaluators' comments on their expectations of NLG models, indicates a systematic underestimation by the evaluators of the quality of machine-generated text. Evaluators who were trained with examples had higher expectations of machine-generated text and focused more on the text's content; however, the training was not sufficient to significantly raise evaluators' scores across all three domains.

Many of the explanations given by evaluators included references to the text that reflected human attributes or intent that they suspected machines could not generate (e.g., "personal description a machine wouldn't understand, [like a pirate] wanting to be home with his wife and son" from Figure 1 and the examples in Appendix A.4). However, current NLG models are capable of generating text with at least superficial reference to human attributes or intent, as seen in the generated story in Figure 1. This assumption that machines can't generate text with these aspects of humanlikeness led many evaluators astray, and we suspect it is one cause of the low accuracy we found.

Crowdsourcing studies dealing only with human-authored texts often include extensive training, quality checks, or coordination (Kittur and Kraut, 2008; Kim et al., 2017; Bernstein et al., 2010). NLG evaluations usually forego such structures, based, we suspect, on the assumption that evaluating machine-generated text requires only fluency in the language the text is generated in. Our results suggest otherwise. Evaluators often mistook machine-generated text as human, citing superficial textual features that machine generation has surpassed (Brown et al., 2020). One potential remedy for this is to focus evaluator training on debunking this misconception. We did see evidence that the increase in accuracy we saw with our Examples training was associated with fewer explanations mistakenly referencing machine capabilities, even though the training did not specifically focus on this.

5 Recommendations

Based on our findings, if NLG researchers must run human evaluations as small-batch evaluations on Amazon Mechanical Turk or similar platforms, we recommend they train evaluators with examples. This will help calibrate the evaluators' expectations of generated text and indicate the careful reading they may need to do to properly assess the text's quality. Our experiments also indicate the importance of confirming with evaluators why they have made the decisions they have, as the criteria they might implicitly be evaluating may be mismatched with researchers' intended criteria. However, other evaluation setups may be more successful on Amazon Mechanical Turk, such as long-term evaluations with qualified evaluators who have gone through an extended training (like those in Kittur and Kraut, 2008; Zellers et al., 2019a) or third-party evaluator quality tools (e.g., Positly, used by Brown et al., 2020).

However, given the increasing length of text NLG models can handle and the careful reading needed to detect many errors in generated text, we encourage NLG researchers to move away from standalone, intrinsic human evaluation tasks. We found that, by default, our evaluators in this evaluation setting were most likely to focus on surface-level, fluency-related aspects of quality. We join past work (Belz and Reiter, 2006; van der Lee et al., 2021) in recommending a move towards evaluation settings where evaluators are better motivated to carefully consider the content and usefulness of generated text. For example, TuringAdvice (Zellers et al., 2021) asks evaluators to rate NLG models by their ability to generate helpful advice, and RoFT (Dugan et al., 2020) engages evaluators

through a guessing game to determine the boundary between human- and machine-generated text. Other evaluation methods ask the evaluators to directly interact with the generated text; for example, Choose Your Own Adventure (Clark and Smith, 2021) and Storium (Akoury et al., 2020) evaluate story generation models by having people write stories with the help of generated text.[11] We see that GPT3 can successfully mimic human-authored text across several domains, renewing the importance of evaluations that push beyond surface-level notions of quality and consider whether a text is helpful in a downstream setting or has attributes that people would want from machine-generated text.

Finally, given the mixed effect we found different trainings can have on evaluators' performance and the lack of human evaluation details typically presented in NLG papers (van der Lee et al., 2019; Howcroft et al., 2020), we encourage NLG researchers to include details of any instructions and training they gave evaluators in their publications. This, along with efforts to standardize human evaluation design (Belz et al., 2020; Howcroft et al., 2020) and deployment (Khashabi et al., 2021; Gehrmann et al., 2021), will support future development of evaluator training procedures and the comparison of human evaluation results in future NLG evaluation work.

6 Related Work

A subfield of NLG analyzes the role of human evaluations, including discussions of the tradeoffs of human and automatic evaluations (Belz and Reiter, 2006; Hashimoto et al., 2019). There are critiques and recommendations for different aspects of human evaluations, like the evaluation design (Novikova et al., 2018; Santhanam and Shaikh, 2019), question framing (Schoch et al., 2020), and evaluation measures like agreement (Amidei et al., 2018), as well as analyses of past NLG papers' human evaluations (van der Lee et al., 2021; Howcroft et al., 2020). Additionally, crowdsourcing literature has work on effectively using

[11] Note that we initially tried a fourth training condition along these lines, where we asked evaluators to directly interact with the generated text by rewriting it to be more humanlike. We found we were unable to successfully recruit evaluators to complete this task. The rate of retention was less than 30%, and the rejection rate was over 50%. We found AMT was not a good platform for this type of task, at least not for the format and the price point we explored in this work.

platforms like Amazon Mechanical Turk (e.g., Daniel et al., 2018; Oppenheimer et al., 2009; Weld et al., 2014; Mitra et al., 2015). In this work, we focus on the role evaluator training can play for producing better accuracy at distinguishing human- and machine-generated text, though other quality control methods are worth exploring.

Previous work has asked evaluators to distinguish between human- and machine-authored text. For example, Ippolito et al. (2020) found that trained evaluators were able to detect open-ended GPT2-L-generated text 71.4% of the time, Garbacea et al. (2019) reported that individual evaluators guessed correctly 66.6% of the time when evaluating product reviews, and Brown et al. (2020) found evaluators could guess GPT3-davinci-generated news articles' source with 52% accuracy, though these results are not directly comparable to ours due to differences in the evaluation setup, data, and participants.

Finally, our findings that untrained evaluators are not well equipped to detect machine-generated text point to the importance of researching the safe deployment of NLG systems. Gehrmann et al. (2019) proposed visualization techniques to help readers detect generated text, and work like Zellers et al. (2019b), Ippolito et al. (2020), and Uchendu et al. (2020) investigated large language models' ability to detect generated text.

7 Conclusion

We found that untrained evaluators were unable to distinguish between human- and GPT3-generated text from three domains. However, we also found that the evaluators focused on surface-level text qualities to make these decisions and underestimated current NLG models' capabilities. We experimented with three methods for training evaluators, and while example-based trainings led to increases in recall and the amount of content-based evaluations, they did not lead to significant improvements in accuracy across all domains. Given that evaluators struggled to distinguish between human- and machine-generated text in this setting, we should shift how we think about collecting human evaluations for current NLG models.

Acknowledgments

This research was supported in part by the Office of Naval Research under the MURI grant N00014-18-1-2670. The authors would like to thank OpenAI, specifically Bianca Martin and Miles Brundage, for providing access to GPT3 through the OpenAI API Academic Access Program. The authors would also like to thank Katharina Reinecke, the members of the CSE 599 crowdsourcing class, and the ARK group for their feedback, the reviewers for their helpful comments, and the participants who took part in our study.

Ethical considerations All experiments in this paper were approved by our institution's internal review board. Evaluators' responses were collected and stored anonymously. Evaluators were paid based on an estimated US$10 per hour rate; we raised the price of the task in proportion to the added difficulty of our 3 training methods. For each dataset we considered, its source and language are included, along with any other details we believed would be relevant to evaluators' ability to read and understand the text. Evaluators were warned about possible risks before starting the task, namely that NLG models can generate text with harmful language or themes, and were able to leave comments about their experience at the end of the study.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. In *Political Analysis*, volume 20, pages 351–368. Cambridge University Press.

Michael Bernstein, Greg Little, Robert Miller, Björn Hartmann, Mark Ackerman, David Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A word processor with a crowd inside. In *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, volume 58, pages 313–322.

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th Inter- national Conference on Natural Language Generation*, pages 22–28, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-)evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Elizabeth Clark and Noah A. Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575, Online. Association for Computational Linguistics.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. In *ACM Computing Surveys*, volume 51. Association for Computing Machinery.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3968–3981, Hong Kong, China. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Mad- dela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *ArXiv*, abs/2102.01672.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. *ArXiv*, abs/2101.06561.

Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 233–245. Association for Computing Machinery.

Aniket Kittur and Robert E. Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, page 37–46, New York, NY, USA. Association for Computing Machinery.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical

study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing person- and process-centric strategies for obtaining quality data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1345–1354. Association for Computing Machinery.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. In *Journal of Experimental Social Psychology*, volume 45, pages 867–872. Elsevier.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. “This is a problem, don’t you agree?” Framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Alan Turing. 1950. Computing Machinery and Intelligence. In *Mind*, volume LIX, pages 433–460. Oxford University Press (OUP).

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Daniel S. Weld, Mausam, Christopher H. Lin, and Jonathan Bragg. 2014. Artificial intelligence and collective intelligence. In *Handbook of Collective Intelligence*, chapter 3. MIT Press.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2021. TuringAdvice: A generative and dynamic evaluation of language use. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4856–4880, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019b. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339. PMLR.

A Appendices

A.1 Newspapers

Each newspaper came from a randomly chosen U.S. state and was selected from Wikipedia's lists of newspapers by state (en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States#By_state_and_territory). The human- authored news articles and prompts came from the following states and websites:

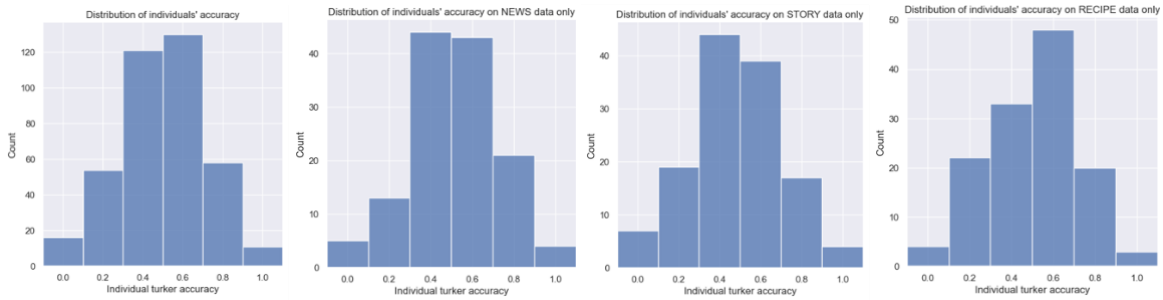
- HI: www.westhawaiitoday.com
- CT: www.greenwichtime.com/
- WA: www.vashonbeachcomber.com/ • SD: www.argusleader.com/
- CA: www.redding.com/
- MA: www.lowellsun.com/
- NE: starherald.com/
- VA: dailyprogress.com/
- WV: www.theintermountain.com/ • NM: www.lcsun-news.com/
- LA: www.nola.com/
- IA: qctimes.com/
- NY: www.pressconnects.com/
- IN: www.pal-item.com/
- NJ: www.northjersey.com/

A.2 Score Frequencies

The frequency of the scores (out of 5) received by evaluators is shown in Figures 3 (for GPT2 experiments) and 4 (for GPT3 experiments).

(a) GPT2 overall (b) GPT2 story (c) GPT2 news (d) GPT2 recipe

Figure 3: Histogram of scores classifying human and GPT2 texts.



(a) GPT3 overall (b) GPT3 story (c) GPT3 news (d) GPT3 recipe

Figure 4: Histogram of scores classifying human and GPT3 texts.

A.3 Annotation Details

The authors annotated 300 comments (150 from the No Training experiment and 150 from the Examples experiment). For each experiment, we randomly chose 50 authors from each setting and randomly added 1 of their responses to the annotation set. Each comment was annotated by 2 of the authors. The annotation labels are shown in Table 4. To create the set of annotation labels, the authors created a candidate list of labels, annotated a subset of the data collected in the pilot study (Appendix A.5) together, then another subset separately, and finally refined the labels based on feedback from that process. Because evaluators' responses often contained more than one reason for their choice, comments could receive more than one label.

A.4 Evaluators' Expectations of Generated Text

Because we asked evaluators whether they thought the text was human- or machine- authored, they often justified their choices by explaining what types of human language they believed machines could (or could not) generate. We took note of these comments and annotated for them in our data annotation process (Appendix A.3) because they demonstrate the expectations evaluators have for the quality of machine-generated text. Some example comments shown in Table 5.

Category	Label	Description	Example
Form	Grammar	The spelling and grammar of the text, punctuation/formatting issues	I would make the text more grammatical by adding more punctuation where necessary.
	Level of detail	Is the text simple or does it go more in-depth?	i would include more examples and explanations of the statements. The author needs to elaborate more on the topic.
	Genre	If the text is the genre/domain/style/ formality that the reader expects, adheres to style norms	written exactly the way a human will tell a story
Content	Repetition	Words/phrases/content repeated itself	Repeating “or some would say” seemed very unnatural.
	Factuality	The accuracy of the text, whether it describes things that are “true.”	The article lists many facts that make the information seem like it was machine-generated.
	Consistency	How the text relates to the context and other pieces of the text	The subject of the article follows the headline well without repeating it exactly.
	Common sense	Whether the text “makes sense” within the world that it is written	Change the “bake in the preheated oven for 20 minutes on top of the stove.” You can’t bake on top of the stove but to bake in the oven.
	Coherence	The structure and coherence of the text. Order issues go here.	More cohesion between sentences. Feel loosely related, but wording is strange.
Machine capabili-ties	Writer intent and expression	Speculating about writer’s intent or capabilities (e.g., ability to express emotions)	The text is thorough and tries to cover all basis of the situation. It is very inclusive and humans worry about being inclusive not machines.
Null	Miscellaneous	Everything else	too many dialogue-like things, and make it less gender-dicey.
	Null/Vague	No reasons given, or too vague to be considered a real reason	i selected this rating because it is definitely written by human

Table 4: The annotation labels, along with an example of each label. Note that some example sentences would also be labeled with additional labels. We did not use the Null category in the paper’s analyses.

Punctuation is perfect as well as the flow of the text. There is also more complex punctuation, such as quotes, that I think a computer would get wrong.
“fried anyone to a crisp.” That is a human if I’ve ever seen one. a bot or AI is more proper, they wouldn’t write so casual.
Because it talked about love which robots know nothing about.
Lack of oxford comma. A computer would know better.
The article flows properly, has appropriate English and multiple quotes. This would seem to be more than a bot could create. How would a bot create random quotes?
This was more of a ramble which humans do, not computers.
There are details and key phrases used in this article that computer generated text would not have in it, such as “came up short”, “put together a solid drive”, “put up any points”. These are human specific terms and are not generally able to be programmed into a text program.
This piece quotes the host and I don’t believe AI can interview people yet so this has to be human written.
It has a lot of detail in an emotional description that a machine isn’t capable of giving to its readers.
The way some words are phrased here again shows the human uncertainty, “let the apples marinate for about 30 minutes”. If this was machine-generated, it would most likely just say marinate for 30 minutes.
It seems to know when to use semicolns very well. This could be a human or a really smart computer.
I don’t think AIs are capable of writing recipes on their own just yet.
I don’t believe a machine could come up with this level of whimsy or creativity and have it make sense.
I don’t think AI would use the term ‘literally’.
There is a lot of every day language written in this recipe that I couldn’t see a machine possibly replicating.
It adds that she is both nervous and excited whereas a machine wouldn’t care what emotions are involved.
The writer used proper grammar and punctuation. No bot could write this,
I’m not sure if a computer would get the concept or use the word “your” where the recipe begins with “Start by doing your prep.”

Table 5: Example reasons evaluators gave for their decisions that spoke to their beliefs about current NLG capabilities.

A.5 Pilot Study

Before running the experiments described in the paper, we ran a smaller-scale version with both Amazon Mechanical Turk (n = 22) and "expert" evaluators (NLP graduate students; n = 11). We asked the evaluators to distinguish between stories authored by humans, GPT2, and GPT3 and to explain their reasoning. When we coded and analyzed their responses, we found that the most accurate evaluators focused on textual aspects like repetition and were less likely to mention aspects like style. The AMT evaluators mentioned grammar and spelling far more frequently than the expert evaluators, who were more likely to mention the repetition, factuality, and commonsense of the passage.

A.6 Training and Instructions

Figure 5 shows the basic instructions that were shown to all evaluators, in both §2 and §3, regardless of training or domain. All training information occurred after receiving the basic instructions.

A.6.1 Instruction Training

The training shown to evaluators in the Instruction training condition is shown in Figure 6.

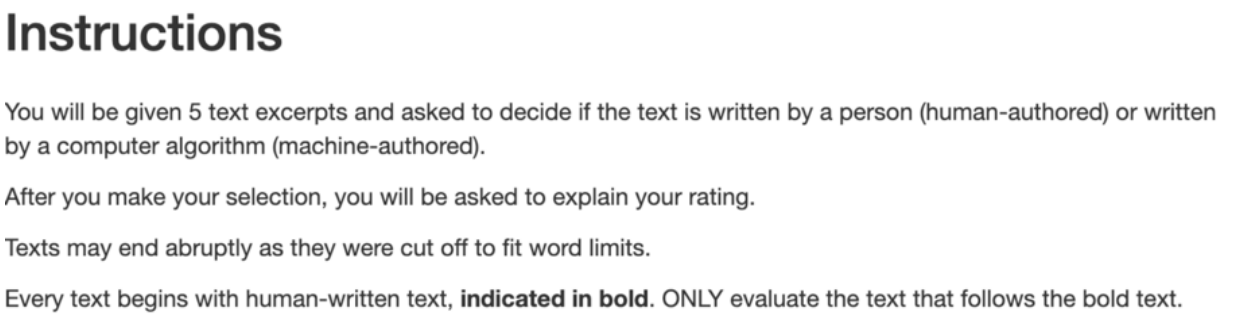


Figure 5: Basic instructions shown to all evaluators.

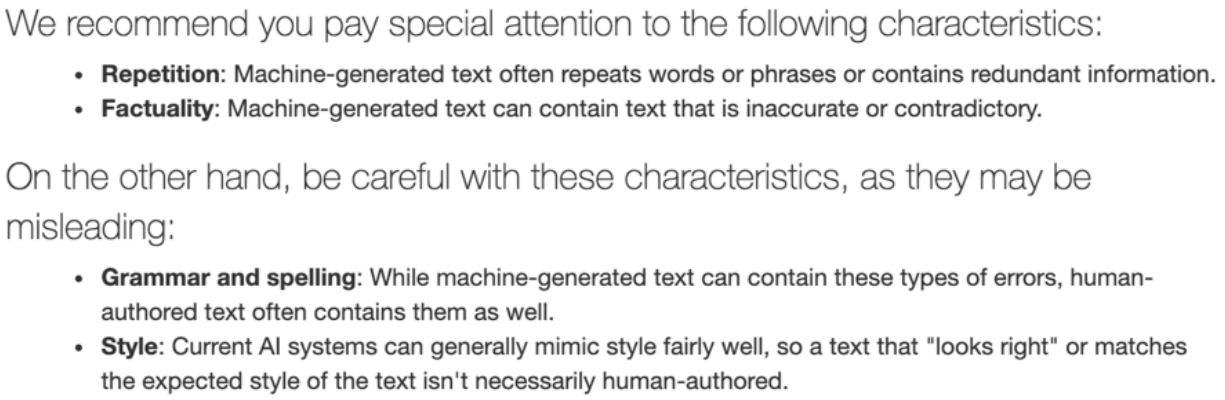


Figure 6: The Instruction training.

A.6.2 Example Training

A screenshot of the Examples and Comparison training is in Figure 7. The full set of examples and annotations used in the Examples and Comparison trainings can be found in the supplementary materials and at ark.cs.washington.edu/human_evals_ACL21.

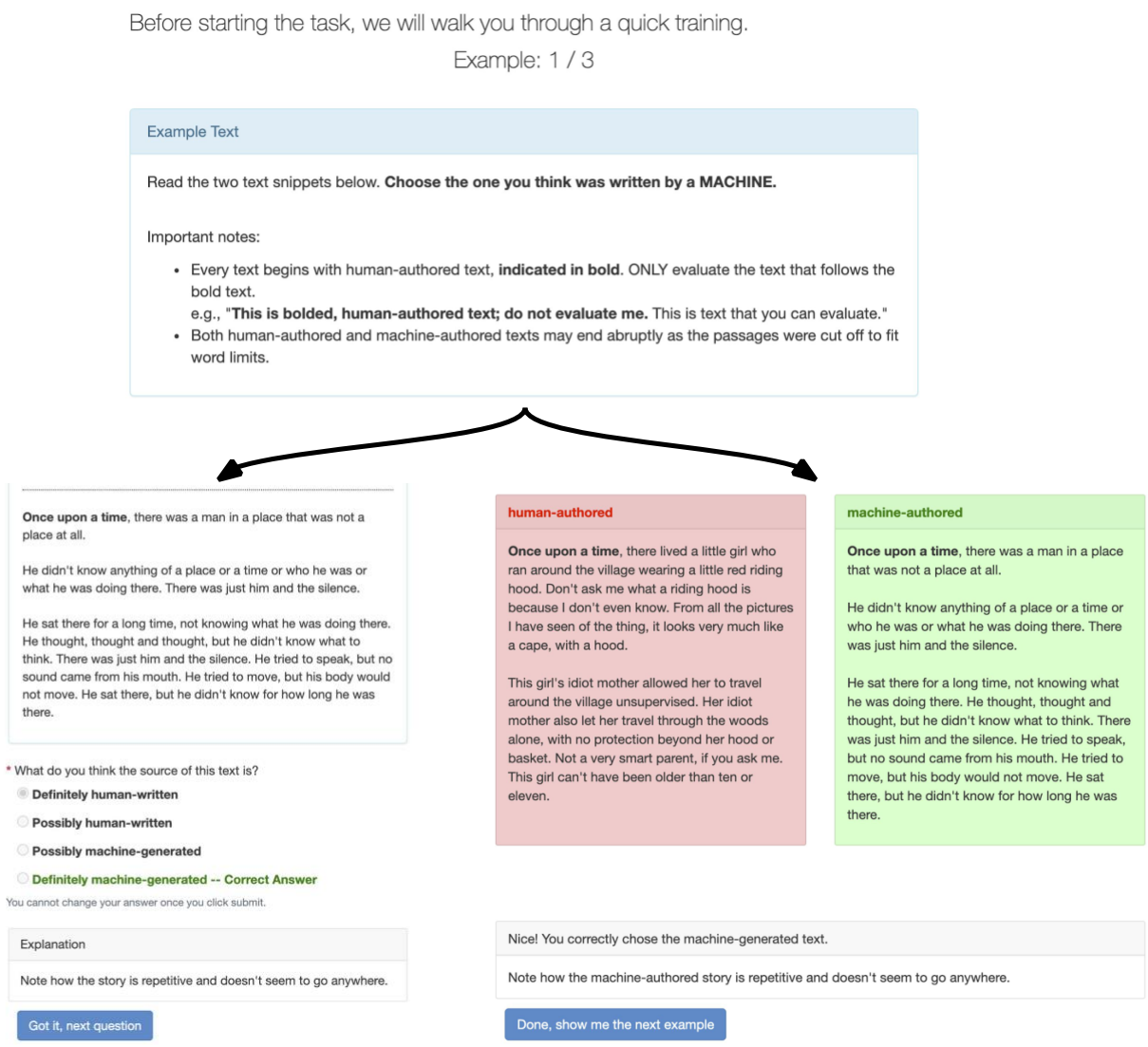


Figure 7: The Example training (left) and Comparison training (right) in the story domain. The instructions are the same for both, except “Choose the one you think was written by a machine.” was in Comparison only.