The Authenticity Gap in Human Evaluation

Kawin Ethayarajh
Stanford University
kawin@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Human ratings are the gold standard in NLG evaluation. The standard protocol is to collect ratings of generated text, average across annotators, and rank NLG systems by their average scores. However, little consideration has been given as to whether this approach faithfully captures human preferences. Analyzing this standard protocol through the lens of utility theory in economics, we identify the implicit assumptions it makes about annotators. These assumptions are often violated in practice, in which case annotator ratings cease to reflect their preferences. The most egregious violations come from using Likert scales, which provably reverse the direction of the true preference in certain cases. We suggest improvements to the standard protocol to make it more theoretically sound, but even in its improved form, it cannot be used to evaluate open-ended tasks like story generation. For the latter, we propose a new human evaluation protocol called *system-level probabilistic assessment* (SPA). When human evaluation of stories is done with SPA, we can recover the ordering of GPT-3 models by size, with statistically significant results. However, when human evaluation is done with the standard protocol, less than half of the expected preferences can be recovered (e.g., there is no significant difference between `curie` and `davinci`, despite using a highly powered test).

1 Introduction

Human ratings are treated as the gold standard in NLG evaluation (Zhou et al., 2022). For example, say one wants to claim that that their NLG model X is better than the current state-of-the-art Y and Z for story generation. The standard protocol is **outcome-level absolute assessment** (OAA): hire crowdworkers as annotators, collect individual ratings of a sample of stories generated by each model, and then claim that X is the best because its average rating is the highest (Celikyilmaz et al., 2020). There is inconsistency in how this is implemented in the literature: terms such as 'text quality' are often left undefined when instructing annotators (Howcroft et al., 2020) and different papers use different rating scales (Amidei et al., 2019). However, such criticism has been restricted to the implementations—little to no consideration has been given as to whether OAA can faithfully capture human preferences to begin with.

We start by analyzing the standard evaluation protocol through the lens of utility theory from economics (§2). We find that OAA can only capture an annotator's preferences under certain assumptions, which are unstated in the NLG literature and often violated in practice. In such cases, annotator ratings become an unfaithful reflection of their preferences (§3). For example, by framing ratings as utility estimates, we extend a result from Boutilier (2003) to prove that using the same scale is insufficient for aggregating ratings *across* annotators—they must agree on the maximum- and minimum-utility outcomes as well. This precludes annotator ratings from being averaged unless they are given both maximally "correct" and "incorrect" references, which are available for some NLG tasks (e.g., machine translation) but not for open-ended ones (e.g, story generation), since the space of high-quality outputs is too diverse. We provide concrete suggestions on how to improve the standard protocol to a point where it can faithfully capture human preferences in *some* NLG tasks and settings (§4); however, for open-ended generation, an entirely new evaluation protocol is needed.

Though uncommon nowadays, a historic alternative to OAA was **outcome-level relative assessment** (ORA): create random pairs containing an output from X and Y, ask annotators to pick the one they prefer in each, infer a score for X and Y that explains the results—based on a comparison model such as Bradley-Terry (Bradley and Terry, 1952)—and argue that X is better because its estimated score is higher (Sakaguchi et al., 2014).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6056 - 6070
December 7-11, 2022

However, this also makes untenable assumptions about annotators; for example, even if X's outputs are preferred to Y's over 50% of the time, X may be less preferred to Y if it has a tendency to fail catastrophically. We observe that the main limitation of both OAA and ORA is their reliance on outcome-level judgments.

To this end, we propose **system-level probabilistic assessment** (SPA), which can be used for both open- and close-ended NLG tasks (§5). SPA's key insight is that while an annotator cannot categorically determine whether they prefer system X or Y—because the output space is too large for them to observe—they can estimate the *probability* with which they prefer X or Y based on some fixed level of exposure to both. SPA obviates assumptions about annotator preferences by delegating the responsibility of aggregating preferences over individual outputs into a preference over the underlying systems to the annotator themself, acknowledging that there is no canonical way to do so. Because we are working with probabilities, aggregating across annotators is also straightforward.

We then ask annotators to use both the standard protocol (with a 5-point Likert scale) and SPA to express their preferences about the different GPT-3 variants w.r.t. their story-generation ability. Given that larger GPT-3 variants generate more coherent, grammatical, and creative text, annotators[1] should prefer each GPT-3 variant to the next smallest one, giving us 3 ground-truth preferences (Brown et al., 2020). Past work also suggests that annotators can distinguish human-written text from `ada` (the smallest variant) but not from `davinci` (the largest) (Clark et al., 2021), which gives us two additional ground-truth preferences, for a total of 5.

When human evaluation is mediated by SPA, we can recover all 5 out of 5 expected preferences, with statistically significant results. However, when human evaluation is done with the standard protocol, we can only recover 2 of the 5 preferences, despite all tests having statistical power ≈ 1 for $\alpha = 0.001$. The standard protocol also yields the surprising—and likely incorrect—result that human-written text is significantly *less* preferred to `davinci`. The failures of the standard protocol suggest that its theoretical limitations have practical consequences, and the flexibility of SPA makes it the better option in most intrinsic evaluation settings.

[1] In aggregate, since individual annotators may have aberrant preferences.

2 Reframing Human Evaluation

To understand what causes human preferences to be misrepresented, we will analyze NLG evaluation through the lens of economic theory on preference modeling. In doing so, we find that comparing NLG systems is an instance of a common problem in utility theory. To begin, let X, Y denote the NLG systems to be compared and annotator a_i be the *agent* making the comparisons. As we are drawing from the economics literature, we will primarily use economic terms such as *lottery* and *utility* in our framing, which we will define as we go along.

2.1 NLG Systems as Lotteries

Definition 1 (Lottery). A *lottery* is a probability distribution over a space of finite outcomes (Boutilier, 2003). Given a (possibly empty) prompt or input, an NLG system induces a lottery over all possible output text.

Given that there is also a discrete distribution over the prompts/inputs used, the lottery that the NLG system induces over the output text is itself the outcome of a lottery over the prompts/inputs. This means that X and Y are *compound lotteries*: a lottery of a lottery, which can be reduced to a simple lottery over the output text by marginalization. Thus for a known prior over the prompts/inputs, we can think of any NLG system as a simple lottery over all possible output text.

2.2 Choices as Preference Relations

Definition 2 (Preference Relations). The *relation* $X \succ_i Y$ means that the agent a_i strictly prefers X to Y; the relation $X \prec_i Y$ means that the agent strictly prefers Y to X; the relation $X \sim_i Y$ means that the agent is indifferent to the two. Relations without the subscript i denote the aggregate preference across all agents.

This means that determining whether an annotator prefers one NLG system to another is an instance of a common problem in economics: determining which of two lotteries the agent prefers. For most such problems in the real world, we could ask the agent to directly compare the two lotteries (e.g., we could ask an investor what split of stocks to bonds they would invest in) (Mankiw, 2020). However, because in NLG the output space is so large, we cannot ask an annotator to categorically determine which of two lotteries they prefer. What is feasible is asking an annotator to compare two individual output texts, but there is no assumption-free

means of aggregating preferences over individual outcomes into preferences over the lotteries (§2.4).

2.3 Text Quality as Utility

Definition 3 (Utility). Abstractly, the *utility* of a good denotes the benefit that an agent receives from it. The *utility function* $u_i : S \to \mathbb{R}$ for agent a_i maps outcomes S to real values based on the utility derived (Mankiw, 2020). For NLG, the utility of a text is how good the agent perceives it to be, optionally w.r.t. some attribute such as coherence.

Definition 4 (Ordinal Utility). Under the theory of *ordinal utility*, only the ranking induced by u_i matters; the magnitude of the difference between the values do not (Mankiw, 2020). An ordinal utility function u_i *represents* $\succ_i$ if it preserves the ranking the latter induces: $X \succeq_i Y \iff u_i(X) \geq u_i(Y)$. Two utility functions u, v are *ordinally equivalent* if they induce the same preference ordering.

Definition 5 (Cardinal Utility). Under the theory of *cardinal utility*, the magnitude of the difference between two outcomes' utility does matter. Two utility functions f, g are *cardinally equivalent* up to a positive affine transformation (Dybvig and Polemarchakis, 1981).

Estimating cardinal utility is the approach that has been implicitly taken by the standard evaluation protocol for NLG (a.k.a., outcome-level absolute assessment (OAA)). When an annotator rates an example, they are estimating the cardinal utility $u_i(x)$ they get from an outcome x. When those ratings are averaged to score the system X that produced those examples, one is estimating the expected utility of a lottery. Estimating the cardinal utility of a lottery as the expected utility of its outcomes is possible because of the von Neumann-Morgenstern theorem (Morgenstern and Von Neumann, 1953). No similar result exists for estimating the ordinal utility, however—we cannot average rankings.

2.4 Assumptions of Agent Rationality

Outcome-level relative assessment (ORA) explicitly encodes its assumptions about annotator preferences in a comparison model such as Bradley-Terry (Bradley and Terry, 1952) or Thurstone (Thurstone, 1927). These assumptions are easy to identify and invalidate, so we refer the reader to prior work on its limitations (Sakaguchi et al., 2014; Bojar et al., 2016). ORA has also declined in popularity in recent years, with OAA now making up a supermajority of human evaluation (van der Lee et al., 2021). Because it does not use a comparison model, the now widely-used OAA may seem as though it makes no assumption about annotators. However, ranking systems by their average rating only captures annotator preferences if they are Von Neumann-Morgenstern-rational agents (Morgenstern and Von Neumann, 1953):

Definition 6 (VNM Rationality). Let X', Y' denote random variables representing the outcomes of lottery X, Y respectively. For any *von Neumann-Morgenstern-rational* agent, there exists a utility function u_i such that $X \succeq_i Y \iff \mathbb{E}[u_i(X')] \geq \mathbb{E}[u_i(Y')]$. In other words, VNM-rational agents always choose to maximize their expected utility. In order for an annotator a_i to be a VNM-rational agent, their preferences must satisfy the following four axioms for any NLG systems X, Y, Z:

Axiom 1 (Completeness). For any X, Y, exactly one of the following holds for each agent a_i: $X \succ_i Y, X \prec_i Y$ or $X \sim_i Y$ (i.e., the agent prefers X, prefers Y, or is indifferent respectively).

Axiom 2 (Transitivity). If $X \succeq_i Y$ and $Y \succeq_i Z$, then $X \succeq_i Z$.

Axiom 3 (Continuity). If $X \succeq_i Y \succeq_i Z, \exists\, p \in [0, 1]$ such that $pX + (1-p)Z \sim_i Y$.

Axiom 4 (Independence). For any Z and $p \in (0, 1]$, we have $X \succeq_i Y \iff pX + (1-p)Z \succeq_i Y + (1-p)Z$.

Although it may seem intuitive that any agent would maximize their expected utility, work in behavioral economics has identified many situations where agents choose not to do so (Samuelson, 1977; Kahneman and Tversky, 1979; Allais, 1979).

3 Causes of Misrepresentation

By framing human evaluation in terms of utility theory, we found that the standard protocol in NLG evaluation serves to estimate the cardinal utility of a system via outcome-level absolute assessment (§2.3). We then listed the assumptions that agent preferences need to satisfy in order to make this estimation valid (§2.4). In this section, we discuss how these assumptions are often violated in NLG evaluation, and how this begets misrepresentation of an annotator's true preferences. We limit our criticism to OAA in this section, since ORA has already been criticized in prior work (Sakaguchi et al., 2014) and has, over the past several years, become far less common than OAA (van der Lee et al., 2021).

We begin by noting that rating generated text is done one of two ways (Celikyilmaz et al., 2020):

1. Likert scales[2], which discretize the utility into an integer from 1-to-k, usually 1-to-5.

2. Continuous scales (a.k.a., continuous direct assessment), which normalize the utility from 0-to-k (k usually being 100).

Our first critiques apply only to Likert scales, but our last two apply to the standard protocol at-large.

Remark 1 (**Ordinal-Cardinal Conflation**). Averaging ordinal Likert ratings to estimate cardinal utility can violate tenets of utility theory.

The Likert scale is ordinal: an outcome with a higher score is preferred to one with a lower score, but the distance between the points is not significant. In contrast, the intervals *are* significant in cardinal utility. Averaging Likert ratings to estimate cardinal utility thus assumes that the annotator has perceived the distance between each point to be the same, which is impossible to verify. At best, annotators can be steered into an interval-based interpretation through careful wording of the question, but there is no guarantee that they will interpret the distances as intended. In a survey of the NLG literature, Amidei et al. (2019) found that 31 of 38 papers using Likert scales took an interval-based interpretation of them, but only 1 paper provided justification for this interpretation.

This problem is not solved by normalization methods such as z-scoring, as they do not work when the interval widths are asymmetric (e.g., the annotator might perceive the jump between 1-to-2 to be larger than the jump from 2-to-3 on a 3-point scale). This is not a novel observation either; there is extensive work on the limitations of averaging over Likert ratings (Jamieson, 2004; Sullivan and Artino Jr, 2013; Pimentel and Pimentel, 2019). Even early shared tasks for NLG expressed this concern and used continuous scales instead (Gatt and Belz, 2009).

Remark 2 (**Biased Estimation**). Averaging Likert ratings can be a biased estimator of the expected utility, potentially reversing the direction of the true preference over two NLG systems.

Building upon Remark 1, let us make a best-case assumption that the annotator perceives the intervals between the points on the Likert scale to be equal. As such, they determine the Likert score by normalizing their utility to [0,5] and then applying the ceiling function (e.g., $[0, 1] \to 1; (1, 2] \to 2$, etc.).[3] This effectively replaces a subset of preference relations $\succ_i$ with indifference relations. That is, if two texts both have utilities in the tier $(i, i+1]$, the annotator becomes indifferent to them because of this transformation.

This can be stated more generally:

Proposition 1. *Let* $r_i(s) := \lceil u_i(s) \rceil - u_i(s)$. *Without loss of generality, if* $\mathbb{E}_{s\sim X}[r_i] > \mathbb{E}_{s\sim Y}[r_i]$, *then Likert ratings over-estimate the utility of lottery* X *relative to* Y*; if* $\mathbb{E}_X[r_i] < \mathbb{E}_Y[r_i]$, *they underestimate the utility of* X *relative to* Y.

Proposition 2. *Let* $\mathbb{E}[u_i(X')] > \mathbb{E}[u_i(Y')]$ *without loss of generality. If* $(\mathbb{E}[u_i(X')] - \mathbb{E}[u_i(Y)']) < (\mathbb{E}_Y[r_i] - \mathbb{E}_X[r_i])$, *then averaging Likert ratings reverses the direction of the true preference.*

Since our annotator is implicitly assumed to be VNM-rational, by the von Neumann-Morgenstern theorem, $X \succ_i Y \iff \mathbb{E}[u_i(X')] > \mathbb{E}[u_i(Y')]$. Including the residuals can potentially change the direction of the inequality between the expected utilities. Thus by the VNM theorem, it can also change the direction of the preference relation. Since $r \in [0, 1]$, the difference $|\mathbb{E}_Y[r_i] - \mathbb{E}_X[r_i]| \leq 1$, meaning that a reversal of preference could only occur when the annotator perceived both NLG systems to produce outcomes of similar utility on average. This is a common situation in practice, as proposed systems are often an incremental improvement over the state-of-the-art (Card et al., 2020).

Remark 3 (**Non-Independent Lotteries**). Lottery independence is an axiom of VNM-rationality but often fails to hold in practice for NLG systems.

One of the conditions that needs to be satisfied for VNM-rationality is independence over lotteries, as defined in Axiom 4. Put simply, the preference $X \succ_i Y$ should not change if another lottery Z is mixed with both in equal proportion. However, this assumption is often violated in the real world. Say that X, Y place zero mass on offensive text (e.g., swear words). This is typical for consumer-facing NLG systems, which may explicitly filter out such outputs to avoid public outcry, the loss of users, and a potential lawsuit (Zhou et al., 2022). If lottery Z places any mass on offensive output, adding it to either X or Y may result in the system

[2]To be more specific, a Likert scale is a collection of Likert items, each of which is a discrete rating from 1-to-k.

[3]Using a window of 0.5 around each number and rounding would make the 1-star bucket larger than the rest.

being unusable. If both systems become unusable, the relation between the lotteries would change from preference ($X \succ_i Y$) to indifference ($X \sim_i Y$), despite the direction of the expected utility inequality remaining the same. In such a case, the agent would not be VNM-rational, meaning that their preference could not be inferred by comparing the expected utility of each NLG system.

Remark 4 (**Inter-Agent Incomparability**). Using the same scale across annotators is insufficient for aggregating their cardinal utility (i.e., estimating the *expected expected utility*) due to differences in the magnitude of utility.

When ranking NLG systems, we do not want to rank them according to just one individual, since that individual's preferences may be unrepresentative of the population. In other words, there is a distribution over utility functions, and we want to estimate the expected utility w.r.t. this distribution. This quantity is also known as the expected expected utility (EEU): $\mathbb{E}_i[\mathbb{E}[u_i(X')]]$ (Boutilier, 2003), which can be expanded as

$$\text{EEU}[X] = \int \mathbb{E}[u_i(X')]p(u_i)du_i \qquad (1)$$

Then we could infer the direction of the aggregate preference over the entire agent population, since $X \succ Y \iff \text{EEU}[X] > \text{EEU}[Y]$.

Estimating the EEU is not as straightforward as averaging the expected utility estimates of different agents. Given a continuous scale from 0-to-100, one agent may score in the range 0-to-10 while another may score in 90-to-100. Averaging across the two agents would bias the one with a greater magnitude of scoring. In technical terms, EEU is not invariant to the choice of utility function in a set of cardinally equivalent utility functions. This has been observed empirically in NLP and been framed as annotators being too strict or too forgiving (Zemlyanskiy and Sha, 2018; Kulikov et al., 2019).

Presenting all annotators with the same scale does not necessarily solve this problem, since it does not force annotators to adopt the same magnitudes. Z-scoring does not necessarily solve this problem either, since the annotator scores are not guaranteed to be normally distributed. Relative magnitude estimation (Moskowitz, 1977; Novikova et al., 2018), where the annotator provides the score of an outcome relative to some reference, *partially* addresses this problem, but using a single arbitrary reference point is not provably sufficient.

Boutilier (2003) formally proved that in addition to the continuity axiom (§2.4), *extremum equivalence* is sufficient to estimate EEU, which he defined as: (1) all agents agree on the most and least preferred outcomes; (2) all agents assign their most and least preferred outcomes the utility $c_{\text{max}}, c_{\text{min}}$ respectively, where $c_{\text{max}} > c_{\text{min}} \geq 0$. These conditions might be satisfied in machine translation, for example; one could argue that providing "correct" and "incorrect" references forces all annotators to share utility function endpoints. But when there are no references or the space of high-quality outputs is diverse, as in open-ended NLG tasks (e.g., chitchat dialogue), this condition cannot be satisfied.

4 Improving the Standard Protocol

By making some minor changes, the OAA-based standard evaluation protocol can be improved to a point where it can adequately capture human preferences in *some* NLG tasks and settings:

1. Continuous scales should be used instead of Likert scales to avoid ordinal-cardinal conflation and potentially biased estimation.

2. To satisfy extremum equivalence (§3, Remark 4), both maximal- and minimal-utility references should be provided, effectively forcing all annotators' utility functions to share endpoints. This can only be done when the space of ideal outcomes for a given input is small and well-defined (e.g., machine translation).

3. To satisfy lottery independence, there should be no outcome that can make an NLG system unusable (e.g., because the system is only used by a limited set of users whose utility is bounded).

The WMT competition for machine translation—which has experimented with many evaluation schemes—has had, since 2017, a protocol that follows many of these suggestions (Bojar et al., 2017; Specia et al., 2021). It uses continuous scales, provides maximum-utility references, and hires translators, meaning lottery independence is safe to assume. Still, this improved protocol cannot be applied to open-ended tasks where there is no singular notion of correctness, tasks where maximal-utility outcomes can be diverse (e.g., story generation), or when lottery independence is likely to be violated in the real-world (e.g., offensive chatbots). Such tasks and settings demand an entirely new evaluation protocol (§5).

5 System-level Probabilistic Assessment

The limitations of both ORA and OAA stem from trying to aggregate preferences over outcomes into a preference over systems, despite there being no canonical way to do so. For example, one annotator may prefer X to Y only if the former wins head-to-head comparisons of outputs over 50% of the time, but another annotator may choose by comparing the worst-case output from each system. Therefore we propose directly asking annotators to estimate the probability $P[X \succ_i Y]$ that a preference holds across two systems, a protocol we call *system-level probabilistic assessment* (SPA).

5.1 Theory

Let $P[X \succ Y]$ denote the aggregate preference probability of $X \succ Y$ for a population of agents. Where $p(\succ_i)$ is the frequency of preferences $\succ_i$, we can expand $P[X \succ Y]$ similarly to EEU (1):

$$P[X \succ Y] = \int P[X \succ_i Y] p(\succ_i) d \succ_i \quad (2)$$

Since $P[X \succ_i Y] \in [0, 1]$ for all a_i, the values are inherently comparable across annotators, making inter-annotator aggregation easy. As in comparison models (Bradley and Terry, 1952), this one measure is sufficient to infer the direction of the preference: annotators are indifferent iff $P[X \succ Y] = 0.5$ (i.e., no different than chance); X is more(less) preferred to Y if $P[X \succ Y]$ is greater(less) than 0.5. In practice, however, statistical significance is important to consider (see §5.2 for details).

If we assumed preferences were complete, then $P[X \succ_i Y]$ could only take a value in $\{0, 1\}$, but doing so would be unrealistic, since annotators are almost never exposed to the entirety of an NLG system's output in practice, precluding them from preferring one system with absolute certainty. Therefore we model preferences as stochastic. Modeling preferences as stochastic is not new (Bradley and Terry, 1952; Thurstone, 1927), but the approach has traditionally been to use categorical preference labels to learn stochastic models (Chu and Ghahramani, 2005). The novelty of our approach is that we ask the annotators themselves to estimate the preference probability.

However, an annotator's preferences change as they are exposed to more output while $P[X \succ_i Y]$ is a fixed value. How can we reconcile this? Every time an annotator's preference probability is updated, they effectively become a new agent (i.e., an agent is not an individual annotator, but a specific iteration of an annotator with fixed beliefs). For example, at the start, an annotator has no knowledge of the systems, so $P[X \succ_{i,t=0} Y] = 0.5$. As they are exposed to more outputs, they may develop a preference for one system (e.g., $P[X \succ_{i,t=1} Y] = 0.7$). At some point they will become certain about their choice (e.g., $P[X \succ_{i,t=\infty} Y] = 1$), but at this point the annotator is no longer the same agent that was split between the two options. In other words, agent a_i is uniquely defined by an annotator a and their level of exposure t.

The standard protocol in NLG evaluation requires that annotators be VNM-rational and have preferences that are complete, transitive, independent, continuous, and extremum equivalent (§2.4). SPA obviates those assumptions by delegating the responsibility of aggregating preferences over outcomes into a preference over the underlying lotteries to the agent themself, acknowledging that there is no canonical way to do so. Estimating $P[X \succ Y]$ only requires two assumptions:

Assumption 1 (Unbiased Stated Preferences). An agent a_i has unbiased stated preferences if, when asked to estimate the probability of their preference for lottery X over Y, they provide an unbiased estimate $\hat{P}[X \succ_i Y]$ (i.e., the noise has expectation zero).

Assumption 2 (Indifference). $X \sim_i Y \iff P[X \succ_i Y] = P[X \prec_i Y] = 0.5$ (i.e., an agent is indifferent if and only if the probability of preferring a system is no different from chance).

5.2 Implementing SPA

If you want to use SPA to compare two NLG systems X, Y, you should do as follows:

1. Find n_A unique annotators who are representative of the agent population whose preferences you want to model. Choose the prior for your desired task and draw m prompts/inputs from this prior. Give each annotator m randomly sampled outputs from each system, one per prompt. It is possible to use a different set of prompts for X and Y, but this will make it harder for the agent to do an apples-to-apples comparison, making them less certain about their preference.

2. Ask each annotator a variation of the question: "Based on what you've read, from 0 to 100, what is the % chance that system X is a better writer than system Y?" Swapping X with Y and then

asking the question will not necessarily equal $1 - \hat{P}[X \succ_i Y]$, since the estimates are noisy.

3. (**optional**) To filter out annotators with a poor understanding of probability, ask annotators to estimate both $\hat{P}[X \succ_i Y]$ and $\hat{P}[Y \succ_i X]$ and exclude those for whom $\hat{P}[X \succ_i Y] + \hat{P}[Y \succ_i X] > \tau = 1.1$. We set $\tau \leftarrow 1.1$ instead of 1.0 to account for noisy estimates. If multiple systems are being compared, exclude annotators that fail this condition even once.

4. Estimate the aggregate probability $P[X \succ Y]$ by averaging over $\{\hat{P}[X \succ_i Y]\}$. Use a two-sided Student's t-test to determine whether it is significantly different from chance (0.5). If $P[X \succ Y]$ is significantly higher(lower) than 0.5 at level α, then you can conclude that X is better(worse) than Y with at least probability $1-\alpha$. If $P[X \succ Y]$ is not significantly different from 0.5, then the null hypothesis that $X \sim Y$ cannot be rejected.

In the Appendix, we provide details of the SPA implementation we use in our experiments in §6.

6 Experiments

6.1 GPT-3 Story Generation

To test our proposed protocol, we ask 90 unique crowdworkers to use both the standard protocol (with a 5-point Likert scale) and SPA to express their preferences about the different GPT-3 variants w.r.t. their story-writing ability (see Appendix A for details). The story prompts are drawn from the WritingPrompts dataset (Fan et al., 2018) and each annotator is given: m randomly drawn prompts, stories generated by each GPT-3 variant for those prompts, and a human-written story for each prompt. The annotator is not told which of the 5 systems is a human. With SPA, they are asked to compare the systems themselves, while with the standard protocol, they are just asked to rate the outputs. The smaller m is, the more uncertain annotators will be about their preference, making it hard to elicit a statistically significant result in SPA. The larger m is, the higher the per annotator cost, since the task will take longer to complete. We balance these concerns by choosing $m = 5$.

Given that larger GPT-3 variants generate more coherent and creative text, annotators in aggregate should prefer larger variants: i.e., `davinci` $\succ$ `curie` $\succ$ `babbage` $\succ$ `ada` (Brown et al., 2020). Clark et al. (2021) also found that annotators can distinguish between GPT2- and human-written text, but not at all between human- and `davinci`-written text. Since `ada` is not much larger than GPT2, this implies that the following preferences should also hold: `human` $\succ$ `ada` and `human` $\sim$ `davinci`. For SPA, we use a two-sided Student's t-test to measure whether each probabilistic preference is significantly different from chance ($P[X \succ Y] = 0.5$). For the standard protocol, we use a paired t-test to determine whether the Likert ratings of two systems' outputs are significantly different. Since we make multiple comparisons, we apply the Holm-Bonferroni correction (Holm, 1979).

As seen in Table 1, SPA recovers 5/5 expected preferences: each GPT-3 variant is significantly preferred to the next smallest one; `ada` is significantly less preferred to human-written text; and annotators are indifferent to human- and `davinci`-written text. However, the standard protocol only recovers 2/5 expected preferences: `curie` and `babbage` are not significantly preferred to the next smallest GPT-3 variant, and the human writer is significantly *less* preferred to `davinci`, despite past work suggesting that annotators cannot tell the difference between the two (Clark et al., 2021).

6.2 DALL-E Image Generation

Though the focus of this work is NLG evaluation, SPA can also be used to evaluate other types of generated content. We run a similar experiment with image generation, where the system `DALL-E-scrambled` scrambles the prompt before feeding it to DALL-E[4] and the system `DALL-E-raw` does not (see Appendix B for details). For example, given the prompt *'ball on a chair'*, `DALL-E-raw` feeds the prompt as is to DALL-E while `DALL-E-scrambled` may feed it *'chair on a ball'*. Annotators are then asked which system's images are better, where the goodness of an image is defined as how interesting, coherent, and relevant it is to the original (unscrambled) prompt. Given that scrambling the text affects its compositionality, images generated using the scrambled prompt should be no better than those generated with the original prompt; we expect to recover `DALL-E-raw` $\succ$ `DALL-E-scrambled`.

Our goal with this experiment is to understand how design choices may affect the conclusion drawn with SPA. For one, as shown in Figure 1,

[4]https://openai.com/blog/dall-e/

System X	System Y	Expected Preference	$P[X \succ Y]$ (SPA)	Likert Rating Δ
GPT-3-ada	human	$X \prec Y$	0.420*	−0.822***
GPT-3-babbage	GPT-3-ada	$X \succ Y$	0.688***	0.644***
GPT-3-curie	GPT-3-babbage	$X \succ Y$	0.630***	0.322
GPT-3-davinci	GPT-3-curie	$X \succ Y$	0.575***	0.244
human	GPT-3-davinci	$X \sim Y$	0.544	−0.389*

Table 1: Eliciting preferences for story generation, using both system-level probabilistic assessment (SPA) and the standard protocol with 5-point Likert ratings. We use two-sided Student's t-tests with Holm-Bonferroni-corrected significance at $\alpha = 0.10(^*), 0.05(^{**}), 0.01(^{***})$. SPA consistently yields a significant result in the expected direction; the standard protocol, only twice (green). Notably, the latter suggests that human-written text is significantly *less* preferred to davinci-written text (red), although past work has found that annotators cannot tell the difference (Clark et al., 2021). SPA finds $P[\texttt{human} \succ \texttt{davinci}]$ to not be significantly different from chance.

we find that increasing m, the number of examples shown to the agent, makes them more certain in their preference, pushing the probability to 0 or 1. But there are diminishing returns: the jump from $4 \rightarrow 8$ examples yields less of a change in the preference probability than from $2 \rightarrow 4$. Note the outlier at $m = 1$: surprisingly, agents are quite certain about which system is better when they only see one example from each. We believe that this over-confidence stems from failing to imagine the variance in image quality produced by a single system, variance that the agent is exposed to at $m > 1$. This reaffirms our choice of $m = 5$ in §6.1, and we encourage readers to use $m \geq 4$ in practice.

Secondly, it should not matter whether we ask annotators to estimate $P[X \succ_i Y]$ or $P[Y \succ_i X]$, since by Assumption 2, $P[X \succ_i Y] = 1 - P[Y \succ_i X]$. As seen in Figure 1, this holds in practice; for all values of m, the absolute distance of $P[\texttt{DALL-E-scrambled} \succ \texttt{DALL-E-raw}]$ and $P[\texttt{DALL-E-raw} \succ \texttt{DALL-E-scrambled}]$ from 0.5—a measure of the annotators' conviction in their preference—is not significantly different. This is important, as it implies that practitioners cannot "hack" SPA by changing the ordering of the systems in the question posed to the annotators.

6.3 Discussion

Why does SPA work better than the standard protocol, successfully recovering curie ≻ babbage and babbage ≻ ada while the latter does not? In Figure 2, we show that this is *not* due to statistical power; since we use $n_A = 90$ (after excluding the 10 annotators that did not follow instructions), the power of all our statistical tests—both using SPA and the standard protocol—is ≈ 1 for $\alpha = 0.001$. If the null hypothesis (e.g., curie ∼ babbage) is false and the probability of correctly rejecting the null hypothesis is ≈ 1, then why does the standard protocol fail to do so? We contend that this is because the elicited Likert ratings do not represent the annotators' true preferences to begin with (§3). Just as replacing annotator judgments with random noise would preclude us from rejecting the null hypothesis, the judgments collected via the standard protocol are so distorted that a highly powerful test fails to recover the true preference.

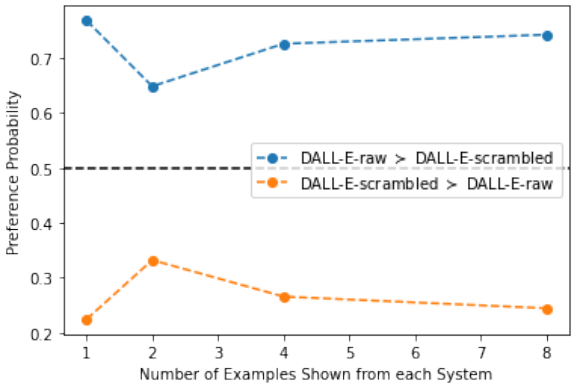


Figure 1: According to SPA, scrambling the prompt before using DALL-E creates significantly worse images than using the original prompt, as expected ($\alpha = 0.01$, Holm-Bonferroni-corrected). The conclusion is the same regardless of whether we ask agents to estimate $P[\texttt{DALL-E-raw} \succ \texttt{DALL-E-scrambled}]$ or vice-versa.

7 Related Work

Soliciting humans to directly evaluate the quality of generated text is known as *intrinsic evaluation*. The text can be judged for its overall quality or along a specific dimension such as coherence, though these terms are not consistently defined (Howcroft et al., 2020; van der Lee et al., 2021). This is

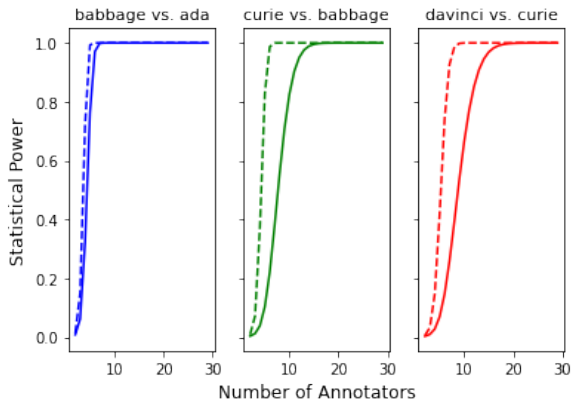


Figure 2: The statistical power of our experiments as a function of the number of annotators n_A, both for SPA (dashed) and the standard protocol (solid), assuming the observed effect size stays constant and $\alpha = 0.001$. Differences in statistical power do not explain why SPA works better than the standard protocol: since we use $n_A = 90$, the power of all our tests is ≈ 1.

most often done in the NLG literature by having annotators assign a Likert rating from 1-to-k, where k is usually 5 (Van Der Lee et al., 2019). Given the ubiquity of this standard protocol, little justification is given when it is used and implicit assumptions, such as equal intervals for Likert scales, are entirely omitted (Amidei et al., 2019).

The earliest shared tasks in NLG, such as the TUNA (Gatt and Belz, 2009) and GREC (Belz and Kow, 2011) challenges for expression generation, used a continuous scale for scoring, explicitly noting that annotators may not perceive the intervals on a Likert scale to be equal. In contrast, early modeling work—such as the STORYBOOK system for narrative prose generation (Callaway and Lester, 2002)—used discrete ratings. This difference in evaluation protocol between shared challenges and individual modeling papers continued over the years. For example, the E2E NLG challenge (Dušek et al., 2018) used continuous scores based on relative magnitude estimation (Novikova et al., 2018; Bard et al., 1996). However, these challenges have not served as a bulwark against the popularity of Likert-based OAA. Even recent attempts to standardize human evaluation in NLG—using evaluation platforms—collect Likert ratings (Khashabi et al., 2021; Gehrmann et al., 2021).

Compared to OAA, outcome-level relative assessment (ORA) is far less common nowadays, in large part because the cost of pairwise output comparisons grows combinatorially as you evaluate more systems (Celikyilmaz et al., 2020). Recall that given binary outcome-level preferences (e.g., $x_i \succ y_i$) as labels, ORA uses a preference model such as Bradley-Terry to estimate the scores of the systems, analogous to how ELO scores are calculated for chess players (Chu and Ghahramani, 2005). In explicitly stating its assumptions about annotator preferences using a preference model, ORA was easier to criticize than OAA, which contributed to the former's decline (Sakaguchi et al., 2014). The one area in which comparison-based evaluation still prevails is when conducting a Turing test—seeing whether annotators do better than chance when guessing whether a text is human- or machine-generated (Garbacea et al., 2019; Ippolito et al., 2020; Brown et al., 2020; Clark et al., 2021). This is acceptable, since what is being measured is not annotator preference but rather discriminability.

Over the years, machine translation (MT) has had spirited debate about evaluation. Callison-Burch et al. (2007) found that compared to ranking outputs, annotators took more time and agreed less when providing Likert scores. Citing this, Sakaguchi et al. (2014) use the TrueSkill algorithm (Herbrich et al., 2006) to estimate scores for NLG systems based on pairwise preferences of their output. This approach, called *relative ranking* (RR) was used in the WMT competition until 2016, when *direct assessment* (DA) on a 0-to-100 continuous scale were trialled and found to produce systems rankings that strongly correlated with RR (Bojar et al., 2016). DA also had the advantage of providing an absolute measure of quality, so it was adopted as the standard for WMT in 2017 and used thereafter (Bojar et al., 2017; Specia et al., 2021).

8 Conclusion

We analyzed the standard evaluation protocol in NLG through the lens of utility theory, finding that it makes untenable assumptions about annotator preferences. When these assumptions are violated, annotator ratings become an unfaithful reflection of their preferences, both in theory and in practice. We proposed a new evaluation protocol called SPA that makes minimal assumptions—not only is it more theoretically sound than the standard protocol, but it performs better in practice as well, consistently recovering the expected preference with statistically significant results. An important direction of future work will be re-evaluating conclusions in the NLG literature with SPA and seeing which ones stand up to scrutiny.

Limitations

Although SPA does not suffer from the existential limitations of the standard evaluation protocol (§3), it does have three notable limitations.

1. SPA does not measure the magnitude of a preference, only the probability that it exists. The magnitude of a preference is useful for understanding the trade-offs involved in deploying one NLG system over another—even if a new system is certainly more preferred to an older one, it might not be worth deploying if the magnitude of the preference is small. This is a necessary trade-off for SPA to be applicable to open-ended NLG tasks, for which extremum equivalence (§3)—a condition necessary for aggregating utility across annotators—cannot be satisfied. However, magnitude estimation (on a continuous scale) is still possible when using *a single annotator*, since extremum equivalence only applies across annotators.

2. Annotators may not understand the notion of probability or may not read the outputs assigned to them, providing noisy and biased annotations. This problem is not unique to SPA, but since human preferences are inherently subjective, identifying insincere annotators is more difficult. The filtering strategy of asking annotators to estimate both $P[X \succ_i Y]$ and $P[Y \succ_i X]$ and excluding those for whom $\hat{P}[X \succ_i Y] + \hat{P}[Y \succ_i X] > \tau = 1.1$ proved to be successful in our DALL-E experiments, though there may be even better strategies. Also, since we want to estimate the aggregate preference of an agent population, we have to use n_A unique agents, instead of letting a few talented annotators do most of the work, as is common in NLP (Geva et al., 2019).

3. There is no simple way to aggregate preference probabilities along multiple axes (e.g., is X more coherent/factual/grammatical than Y?). Assuming that these factors are independent is not realistic, since one may be downstream of another. When doing OAA, the standard practice is to simply take an unweighted average of the factors' Likert scores, but this presumes that equal importance should be given to each factor. Under the principles of preference modeling discussed in this paper, practitioners should delegate the task of creating an overall preference to the annotator themself. That is, in addition to judging whether X is more coherent/factual/grammatical than Y, annotators should also directly judge whether they prefer X to Y.

Ethics Statement

Accurately reflecting the preferences of users is an ethical imperative when building NLG systems. Our work can help practitioners be more cognizant of the assumptions and limitations in their evaluation protocol and the broader risks of deploying improperly tested NLG systems. Our own experiments were conducted with English-speaking US residents, whose preferences are not necessarily representative of the broader population that interfaces with some form of NLG system.

Acknowledgements

We thank Kaitlyn Zhou and Tatsunori Hashimoto for feedback on this work. We also thank Clara Meister and Simran Arora for answering queries about annotation on MTurk. KE was supported by a Facebook Fellowship. This work was also supported by NSF award IIS-2128145.

References

Maurice Allais. 1979. The so-called allais paradox and rational decisions under uncertainty. In *Expected utility hypotheses and the Allais paradox*, pages 437–681. Springer.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402.

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, pages 32–68.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Craig Boutilier. 2003. On the foundations of expected expected utility. In *IJCAI*, volume 3, pages 285–290. Citeseer.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Charles B Callaway and James C Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

Chris Callison-Burch, Cameron Shaw 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.

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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Wei Chu and Zoubin Ghahramani. 2005. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328.

Philip Dybvig and Heraklis Polemarchakis. 1981. Recovering cardinal utility. *The Review of Economic Studies*, 48(1):159–166.

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.

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.

Albert Gatt and Anja Belz. 2009. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

David M Howcroft, Anja 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.

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.

Susan Jamieson. 2004. Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218.

D Kahneman and A Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*.

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 preprint arXiv:2101.06561*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87.

N Gregory Mankiw. 2020. *Principles of economics*. Cengage Learning.

Oskar Morgenstern and John Von Neumann. 1953. *Theory of games and economic behavior*. Princeton university press.

Howard R Moskowitz. 1977. Magnitude estimation: notes on what, how, when, and why to use it. *Journal of Food Quality*, 1(3):195–227.

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.

J Pimentel and JL Pimentel. 2019. Some biases in likert scaling usage and its correction. *International Journal of Science: Basic and Applied Research (IJSBAR)*, 45(1):183–191.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11.

Paul A Samuelson. 1977. St. petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature*, 15(1):24–55.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics.

Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

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.

Yury Zemlyanskiy and Fei Sha. 2018. Aiming to know you better perhaps makes me a more engaging dialogue partner. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 551–561.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. *arXiv preprint arXiv:2205.06828*.

A GPT-3 Experiment

A.1 Story Generation

For each annotator, we first randomly sampled $m = 5$ story prompts from the WritingPrompts datasets after filtering out any prompt that: (1) did not begin with [WP]; (2) contained a question mark; (3) did not end in punctuation. This was done so that the writing prompts were all of a consistent format and style. We observed that prompts ending in questions sometimes elicited opinion essays from GPT-3, as opposed to a fictional continuation. In our trial runs, this confused some annotators, who thought all writers were writing fictional continuations. We thus over-corrected by excluding all prompts with questions.

For each writing prompt, we generated a story by each of the four GPT-3 variants: `davinci-002`, `curie-001`, `babbage-001`, `ada-001`, which we anonymized as writers D,C,B,A respectively. We set the following hyperparameters for all models: a maximum of 1600 tokens, top-p of 1, and a temperature of 0.9. Each story prompt came with a human-written continuation as well, which we anonymized as writer E. In practice, the GPT-3 models usually generated far fewer than the allowable 1600 tokens, resulting in the human-written stories being longer than their machine-written counterparts. To prevent annotators from using story length as a proxy for quality, we trimmed—to the nearest whole sentence—the human-written story for each prompt so that it was no longer than the longest machine-written story for that prompt.

The 25 continuations (5 writers $\times$ 5 prompts) that each annotator had to read were put up on a static website, where the annotator would input their assigned ID to read the batch that was assigned to them (see Figure 3). The annotator was informed that there were a mix of human and AI writing systems, but we did not reveal which writers were which or how many of each there were.

A.2 Filtering Annotators

We recruited $n_A = 100$ annotators on Amazon Mechanical Turk, filtering for those who were in the US, had a HIT approval rate $> 98\%$, and who had completed at least 100 HITs. Each annotator was paid \$5 for approximately 20 minutes of work, working out to roughly USD \$15/hour. Each annotator was presented with the instructions in Figure 4 and then asked to provide 5 preference probabilities $P[X \succ_i Y]$, one for each comparison of interest. They were asked to evaluate each writing system on the basis of how coherent, fluent, interesting and relevant to the prompt the stories were.

They were then asked to provide a Likert rating of the first continuation written by each writer. We did not ask for a rating of all 25 continuations because that would have been onerous and unnecessary; for an apples-to-apples comparison of SPA and the standard evaluation protocol, we had an equal sample size for each, giving us 100 probability estimates of the preference and 100 Likert rating deltas that we could feed into a Student's t-test. Since the order of the prompts was random, asking the annotator to provide a Likert rating for the first continuation (as opposed to say, the second or third) made no systematic difference.

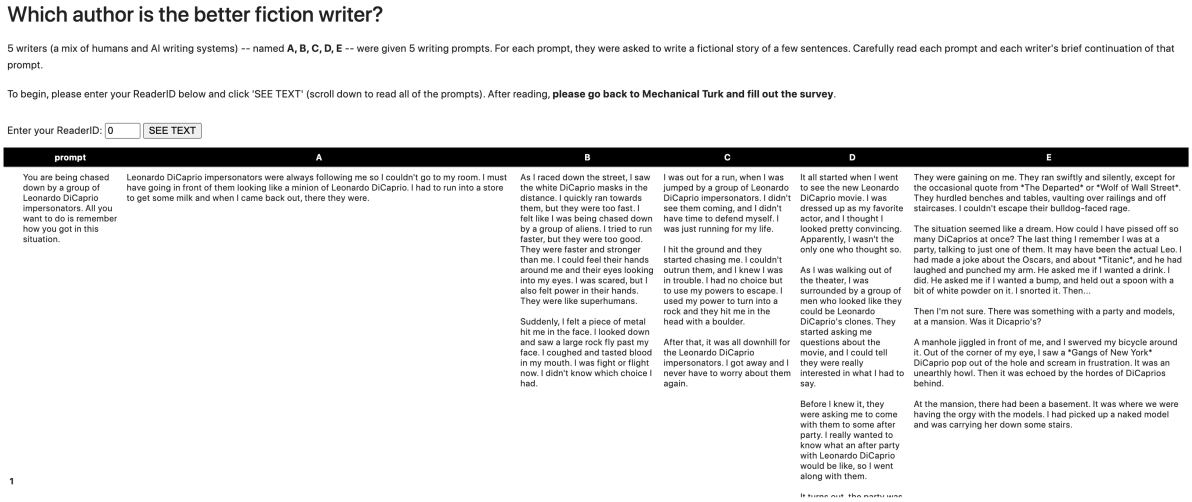
After the annotators provided their annotations, we excluded those who: (1) said they were not native English speakers; (2) did not follow our instructions and submitted multiple HITs. 10% of annotators were excluded, leaving 90 whose annotations we used. Submitting multiple HITS was an issue because we wanted to control the amount of exposure that the annotator had to each writing system, which is why we provided exactly 5 samples from each. Annotations were collected in small batches to prevent the same annotators from making multiple submissions.

Note that we did not implement the additional filtering suggested in §5.2, namely excluding annotators for whom $\hat{P}[X \succ_i Y] + \hat{P}[Y \succ_i X] > \tau = 1.1$ for any pair of systems (X, Y) being compared. This was to see how well SPA would work with minimal annotator filtering. It is only in the DALL-E experiment (Appendix B) that we explored changes to the experiment design, finding that filtering out these self-contradicting annotators is indeed beneficial.

B DALL-E Experiment

B.1 Image Generation

We started with 15 image prompts, collected from the DALL-E website itself. Each prompt was scrambled by randomly re-ordering its tokens. Both the original prompt and the scrambled prompt were fed to DALL-E, generating 4 images for each. We call the pipeline that scrambles the prompt before feeding it to DALL-E `DALL-E-scrambled` and the pipeline that feeds the prompt as is `DALL-E-raw`. If the original (unscrambled) prompt is used as the reference, we would expect the latter system



Which author is the better fiction writer?

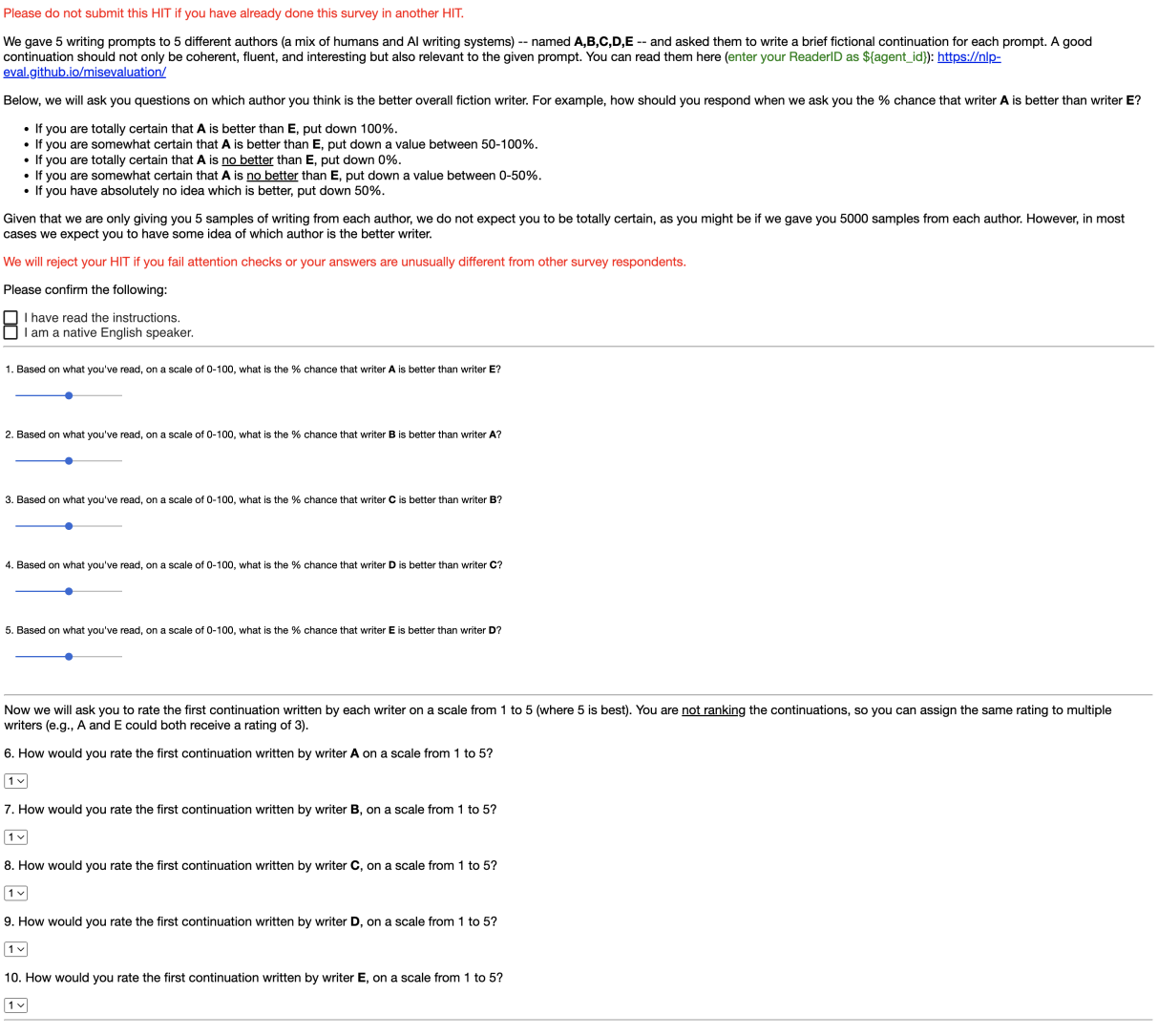
5 writers (a mix of humans and AI writing systems) -- named **A, B, C, D, E** -- were given 5 writing prompts. For each prompt, they were asked to write a fictional story of a few sentences. Carefully read each prompt and each writer's brief continuation of that prompt.

To begin, please enter your ReaderID below and click 'SEE TEXT' (scroll down to read all of the prompts). After reading, **please go back to Mechanical Turk and fill out the survey**.

Enter your ReaderID: 0 SEE TEXT

	prompt	A	B	C	D	E
1	You are being chased down by a group of Leonardo DiCaprio impersonators. All you want to do is remember how you got in this situation.	Leonardo DiCaprio impersonators were always following me so I couldn't go to my room. I must have going in front of them looking like a minion of Leonardo DiCaprio. I had to run into a store to get some milk and when I came back out, there they were.	As I raced down the street, I saw the white DiCaprio masks in the distance. I quickly ran towards them, but they were too fast. I felt like I was being chased down by a group of aliens. I tried to run faster, but they were too good. They were faster and stronger than me. I could feel their hands around me and their eyes looking into my eyes. I was scared, but I also felt power in their hands. They were like superhumans. Suddenly, I felt a piece of metal hit me in the face. I looked down and saw a large rock fly past my face. I coughed and tasted blood in my mouth. I was fight or flight now. I didn't know which choice I had.	I was out for a run, when I was jumped by a group of Leonardo DiCaprio impersonators. I didn't see them coming, and I didn't have time to defend myself. I was just running for my life. I hit the ground and they started chasing me. I couldn't outrun them, and I knew I was in trouble. I had no choice but to use my powers to escape. I used my power to turn into a rock and they hit me in the head with a boulder. After that, it was all downhill for the Leonardo DiCaprio impersonators. I got away and I never have to worry about them again.	It all started when I went to see the new Leonardo DiCaprio movie. I was dressed up as my favorite actor, and I thought I looked pretty convincing. Apparently, I wasn't the only one who thought so. As I was walking out of the theater, I was surrounded by a group of men who looked like they could be Leonardo DiCaprio's clones. They started asking me questions about the movie, and I could tell they were really interested in what I had to say. Before I knew it, they were asking me to come with them to some after party. I really wanted to know what an after party with Leonardo DiCaprio would be like, so I went along with them. It turns out, the party was	They were gaining on me. They ran swiftly and silently, except for the occasional quote from *The Departed* or *Wolf of Wall Street*. They hurdled benches and tables, vaulting over railings and off staircases. I couldn't escape their bulldog-faced rage. The situation seemed like a dream. How could I have pissed off so many DiCaprios at once? The last thing I remember I was at a party, talking to just one of them. It may have been the actual Leo. I had made a joke about the Oscars, and about *Titanic*, and he had laughed and punched my arm. He asked me if I wanted a drink. I did. He asked me if I wanted a bump, and held out a spoon with a bit of white powder on it. I snorted it. Then... Then I'm not sure. There was something with a party and models, at a mansion. Was it Dicaprio's? A manhole jiggled in front of me, and I swerved my bicycle around it. Out of the corner of my eye, I saw a *Gangs of New York* DiCaprio pop out of the hole and scream in frustration. It was an unearthly howl. Then it was echoed by the hordes of DiCaprios behind. At the mansion, there had been a basement. It was where we were having the orgy with the models. I had picked up a naked model and was carrying her down some stairs.

Figure 3: The interface to the generated stories. The continuations generated by the GPT-3 models (A,B,C,D) and the human-written continuation (E) were placed side-by-side.



Please do not submit this HIT if you have already done this survey in another HIT.

We gave 5 writing prompts to 5 different authors (a mix of humans and AI writing systems) -- named **A,B,C,D,E** -- and asked them to write a brief fictional continuation for each prompt. A good continuation should not only be coherent, fluent, and interesting but also relevant to the given prompt. You can read them here (enter your ReaderID as ${agent_id}): https://nlp-eval.github.io/misevaluation/

Below, we will ask you questions on which author you think is the better overall fiction writer. For example, how should you respond when we ask you the % chance that writer **A** is better than writer **E**?

- If you are totally certain that **A** is better than **E**, put down 100%.
- If you are somewhat certain that **A** is better than **E**, put down a value between 50-100%.
- If you are totally certain that **A** is no better than **E**, put down 0%.
- If you are somewhat certain that **A** is no better than **E**, put down a value between 0-50%.
- If you have absolutely no idea which is better, put down 50%.

Given that we are only giving you 5 samples of writing from each author, we do not expect you to be totally certain, as you might be if we gave you 5000 samples from each author. However, in most cases we expect you to have some idea of which author is the better writer.

We will reject your HIT if you fail attention checks or your answers are unusually different from other survey respondents.

Please confirm the following:

☐ I have read the instructions.
☐ I am a native English speaker.

1. Based on what you've read, on a scale of 0-100, what is the % chance that writer **A** is better than writer **E**?

2. Based on what you've read, on a scale of 0-100, what is the % chance that writer **B** is better than writer **A**?

3. Based on what you've read, on a scale of 0-100, what is the % chance that writer **C** is better than writer **B**?

4. Based on what you've read, on a scale of 0-100, what is the % chance that writer **D** is better than writer **C**?

5. Based on what you've read, on a scale of 0-100, what is the % chance that writer **E** is better than writer **D**?

Now we will ask you to rate the first continuation written by each writer on a scale from 1 to 5 (where 5 is best). You are not ranking the continuations, so you can assign the same rating to multiple writers (e.g., A and E could both receive a rating of 3).

6. How would you rate the first continuation written by writer **A** on a scale from 1 to 5?

1

7. How would you rate the first continuation written by writer **B**, on a scale from 1 to 5?

1

8. How would you rate the first continuation written by writer **C**, on a scale from 1 to 5?

1

9. How would you rate the first continuation written by writer **D**, on a scale from 1 to 5?

1

10. How would you rate the first continuation written by writer **E**, on a scale from 1 to 5?

1

Figure 4: The instructions given to annotators on Amazon Mechanical Turk.

to be preferred in aggregate (i.e., `DALL-E-raw` $\succ$ `DALL-E-scrambled`), since scrambling destroys some compositional concepts. For example, if the original prompt is *'ball on a chair'*, then an image containing a ball atop a chair is preferable to one that contains a chair atop a ball.

To understand the role that m, the number of examples shown, plays in preference probability, we asked the annotator to make four comparisons:

1. **A vs. B**, where A is `DALL-E-raw` and B is `DALL-E-scrambled`. We randomly sampled (without replacement) **1** of the 15 original prompts and then uniformly randomly sampled 1 of the 4 original-prompt-based images and 1 of the 4 scrambled-prompt-based images.

2. **C vs. D**, where C is `DALL-E-raw` and D is `DALL-E-scrambled`. We randomly sampled (without replacement) **2** of the 15 original prompts and then uniformly randomly sampled 1 of the 4 original-prompt-based images and scrambled-prompt-based images for each.

3. **E vs. F**, where E is `DALL-E-raw` and F is `DALL-E-scrambled`. We randomly sampled (without replacement) **4** of the 15 original prompts and then uniformly randomly sampled 1 of the 4 original-prompt-based images and scrambled-prompt-based images for each.

4. **G vs. H**, where G is `DALL-E-raw` and H is `DALL-E-scrambled`. We randomly sampled (without replacement) **8** of the 15 original prompts and then uniformly randomly sampled 1 of the 4 original-prompt-based images and scrambled-prompt-based images for each.

Thus the annotator was given the impression that they were seeing images from 8 unique image-generation systems. This concealment is necessary; if the annotator knew that systems C and A were the same, then they may have used the images in the A vs. B comparison when judging C vs. D, which would have precluded us from studying the effect of m.

B.2 Filtering Annotators

We recruited $n_A = 60$ annotators on Amazon Mechanical Turk, filtering for those who were in the US, had a HIT approval rate > 98%, and who had completed at least 100 HITs. Each annotator was paid 3 for approximately 10 minutes of work, working out to roughly USD $18/hour. For each comparison of the form X vs. Y, we asked the annotator to estimate both $P[X \succ_i Y]$ and $P[Y \succ_i X]$. Since there are four comparisons and two estimates per comparison, a total of eight questions were answered by each annotator.

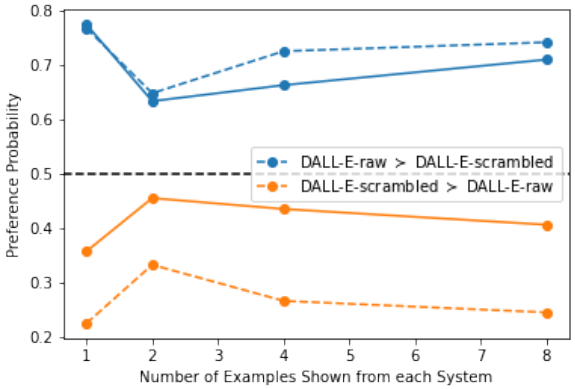


Figure 5: Preferences probabilities for the DALL-E image-generation experiment before filtering annotators (solid) and after filtering (dashed). Though all preference probabilities are in the right direction, filtering out annotators with a poor understanding of probability increased effect sizes and restores the symmetry that should exist under Assumption 2 (i.e., $\hat{P}[X \succ_i Y] \approx 1 - \hat{P}[Y \succ_i X]$).

As in the GPT-3 experiment, we excluded those who said they were not native English speakers or who did not follow our instructions and submitted multiple hits. We also applied the additional filtering step suggested in §5.1 for filtering out annotators with a poor understanding of probability: excluding those for whom $\hat{P}[X \succ_i Y] + \hat{P}[Y \succ_i X] > \tau = 1.1$ for any $(X, Y) \in \{(A, B), (C, D), (E, F), (G, H)\}$. The filtering left 29 eligible annotators.

As seen in Figure 5, the benefits of this additional filtering step were two-fold:

1. The remaining annotators are more certain about their preference, leading to a larger effect size.

2. The excluded annotators are less willing to commit to a preference than a dispreference (e.g., less willing say that $\hat{P}[Y \succ_i X] < 0.5$ than say that $\hat{P}[X \succ_i Y] > 0.5$, though both are semantically equivalent).

For these reasons, we recommend practitioners follow the optional filtering step suggested in §5.2.