How Human is Human Evaluation?
Improving the Gold Standard for NLG with Utility Theory

Kawin Ethayarajh
Stanford University
kawin@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Human ratings are treated as the gold standard in NLG evaluation. The standard protocol is to collect ratings of generated text, average across annotators, and then rank NLG systems by their average scores. However, little consideration has been given as to whether this approach faithfully captures human preferences. In this work, we analyze this standard protocol through the lens of utility theory in economics. We first identify the implicit assumptions it makes about annotators and find that these assumptions are often violated in practice, in which case annotator ratings become an unfaithful reflection of 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 evaluation protocol called *system-level probabilistic assessment* (SPA). In our experiments, we find that according to SPA, annotators prefer larger GPT-3 variants to smaller ones – as expected – with all comparisons being statistically significant. In contrast, the standard protocol only yields significant results half the time.

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 rarely used 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). 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-writing ability. Given that larger variants generate more coherent, grammatical, and creative text, annotators (in aggregate[1]) should prefer each GPT-3 variant to the next smallest one, which gives 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. SPA is able to recover all 5 out of 5 expected preferences, with statistically significant results. However, the standard protocol only recovers 2/5 and also yields the surprising – and likely incorrect – result that human-written text is slightly (but significantly) *less* preferred to `davinci`. This suggests that the theoretical limitations of the standard protocol have practical consequences, and the flexibility of SPA makes it the better option in most intrinsic evaluation settings.

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.

For example, an NLG system that places higher probabilities on coherent text across all prompts will usually be preferred to one centered on incoherent text, although different agents can have different preferences over the same two systems. This means that determining whether an annotator prefers one NLG system to another is an instance of a common economics problem: 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 space of output text is so large, we cannot ask an annotator to categorically determine which of two lotteries they prefer. What is

---

[1] At the population-level, since individual annotators may have aberrant preferences.

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 from which those outcomes are drawn (§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). A ordinary 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 u, v 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 cardinal utility of a lottery. Although straightforward to implement, OAA requires strong – and at times untenable – assumptions about agent preferences (§2.4).

Estimating the cardinal utility of a lottery as the expected utility of its outcomes is possible in theory 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. To derive an ordinal utility function over the lotteries, we would have to ask the annotator to categorically decide whether they prefer lottery X or Y, which is not viable when the outcome space is as large as it is in NLG.

2.4 Agent Rationality

Though rarely used nowadays compared to OAA, outcome-level relative assessment (ORA) explicitly encodes its assumptions about annotator preferences in a comparison model (e.g., Bradley-Terry (Bradley and Terry, 1952), Thurstone (Thurstone, 1927), etc.). 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). 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 $X \succ_i Y \iff \mathbb{E}[u_i(X')] > \mathbb{E}[u_i(Y')]$. 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 $X \succ_i Y$ and $Y \succ_i Z$, then $X \succ_i Z$. The same holds for the indifference relation $\sim_i$.

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 $X \succ_i Y \iff pX + (1 - p)Z \succ_i Y + (1 - p)Z$.

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). We argue later in the paper that the continuity axiom in particular is unlikely to hold when humans evaluate NLG systems (§3, Remark 3).

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). 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 (§6).

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

1. Likert scales, which discretize and normalize the utility into an integer from 1-to-k (usually 1-to-5 or 1-to-7). To be more specific, a Likert scale is a collection of Likert items, each of which is a discrete rating from 1-to-k.
2. Continuous scales (a.k.a., continuous direct assessment), which normalize the utility from 0-to-k (k usually being 100).

Our first two critiques of scoring 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 utility can violate tenets of cardinal 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 very concern and used continuous scales instead (Gatt and Belz, 2009). Past empirical work, both in NLG and psychology, also found annotators strongly preferred continuous scales to discrete scales, as they permitted more nuanced judgments (Svensson, 2000; Belz and Kow, 2011).

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.).[2] 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.

Thus the distribution of the *intra*-tier relations – and whether they agree with the *inter*-tier relations – will determine the direction and magnitude of the bias. For example, say X-generated texts and Y-generated texts were identically distributed across tiers, but the former were always preferred within a tier. This advantage is erased by applying the ceiling function, so both NLG systems end up with an identical distribution of Likert scores. 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 under-estimate 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

---

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

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**)**.** The independence of lotteries is an axiom of VNM-rationality but often fails to hold in practice for NLG systems.

Collecting annotator ratings of text outputs and averaging them to get a score for the NLG system is only viable if the annotators are VNM-rational (§2.4). 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 for instance that X, Y place zero mass on offensive text (e.g., swear words, racial slurs, etc.). This would be fairly 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 being unusable. Even if an annotator has a non-zero tolerance for offensive content, under lottery independence, the preference must hold regardless of the proportion in which Z is mixed with X and Y. Mixing Z with X and Y in equal proportion does not change the direction of the expected utility inequality, but it can change the relation between the lotteries from preference $X \succ_i Y$) to indifference ($X \sim_i Y$) if both systems become unusable. 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.

Aside from our theoretical argument, there is also some empirical evidence that the continuity axiom does not hold in the real world. Consumer preferences are partially revealed by the NLG systems deployed in industry, since these companies optimize for user satisfaction. Those deployed are often template-based even though they produce less varied and less interesting text than the latest open-ended NLG systems, suggesting that a model that behaves reliably is preferred over one that produces higher-quality output on average (Khadpe et al., 2020).

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.rt. 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

$$\text{EEU}[X] = \int_U \mathbb{E}[u_i(X')] p(u_i) \tag{1}$$

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 the range 90-to-100. Averaging across these two agents would bias the agent 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 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). There are post hoc approaches to standardizing scores as well – such as item response theory (Embretson and Reise, 2013) – that can make empirically useful adjustments. However, such approaches make assumptions about the distribution of the underlying choices that are unverifiable; averaging over the adjusted scores does not give us an unbiased estimate of the EEU.

Boutilier (2003) formally proved that *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. 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. This need was pointed out even in early shared tasks for NLG (Gatt and Belz, 2009).

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, extractive summarization, etc.). Doing so can manipulate agents' utility functions, which may be desired when evaluating w.r.t. a particular attribute such as fluency, but may not be desired when trying to capture true human preferences. If extremum equivalence cannot be enforced, only one annotator should be used for all ratings, though this risks yielding conclusions that are unrepresentative of the population at large.

3. To satisfy lottery independence, there should be no outcome that can make an NLG system unusable (e.g., because any such outcomes have been filtered out or because the system is only used by a limited set of users whose utility functions are bounded).

The WMT competition for machine translation – which has experimented with many evaluation schemes over the years – 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 positive (i.e., maximum-utility) references, and the agent population comprises only translators hired for the competition, meaning lottery independence is safe to assume. These changes were driven by empirical evidence (e.g., higher inter-annotator agreement) instead of the theoretical arguments that motivated our suggestions, suggesting that the theoretical problems we identified indeed have practical consequences. WMT does not provide negative (i.e., minimal-utility) references, however, so there remains room for improvement.

Still, this theoretically sounder protocol cannot be applied for all NLG tasks and settings. Maximal- and minimal-utility references are not available in open-ended tasks where there is no singular notion of correctness or where maximal-utility outcomes can be very diverse (e.g., story generation, chitchat dialogue, etc.). Moreover, the assumption of lottery independence is likely be violated in the real-world; for example, consumers of a chatbot may find it completely unusable if it produces any offensive content. Such tasks and settings demand an entirely new evaluation protocol, which we propose in (§5).

5 System-level Probabilistic Assessment

We observe that the limitations of both ORA and OAA stem from trying to aggregate preferences over outcomes into a preferences 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 comparison of outputs over 50% of the time, but another may choose by comparing the worst-case output generated by each. Therefore we propose a middle ground: directly estimate the probability $P[X \succ Y]$ that a preference holds across two systems, a protocol we call *system-level probabilistic assessment* (SPA).

5.1 Theory

Where $p(\succ_i)$ is the frequency (among the agent population) of preferences $\succ_i$, we can expand

$P[X \succ Y]$ similarly to EEU in (1):

$$P[X \succ Y] = \int_{\succ_i} P[X \succ_i Y] p(\succ_i) \qquad (2)$$

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. 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.

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? We take the view that every time an annotator's preference probability is updated, they effectively become a new agent (i.e., an agent no longer represents 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., $P[X \succ_{i,t=n} Y] = 1$ for some large n), but at this point the annotator is no longer the same agent that was split between the two options. This is a useful property – the agent population will be constantly evolving in the real world and with it the aggregate utility of the system will too. By composing different agent populations during the evaluation stage – for example, by exposing annotators to different amounts of output – the aggregate preference of these different populations can be estimated *in vitro*.

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 one assumption:

Definition 7 (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 $\hat{P}[X \succ_i Y]$.

5.2 Implementing SPA

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

1. Choose the prior for your desired task and provide a sample of m prompts/inputs drawn from this prior. Alternatively, allow agents to use their own under the assumption that they share a prior over the prompt/input space.

2. Find n_A unique annotators who are representative of the agent population whose preferences you want to model. Allow agents to directly interact with each system or show them a fixed number of randomly sampled outputs from each system based on the chosen prior. We take the latter approach in our experiments (§5.3), as we include human-written references as well.

3. Then ask each annotator a variation of the following 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?" If needed, swap X and Y and repeat the question. Note that $\hat{P}[X \succ_i Y]$ does not necessarily need to equal $1 - \hat{P}[Y \succ_i X]$ since the agent's estimates are noisy and they can be indifferent to the two.

4. Estimate the aggregate probability $P[X \succ Y]$ by averaging over $\{\hat{P}[X \succ_i Y]\}$. Use a Student's t-test to determine whether $P[X \succ Y]$ is significantly different from chance (0.5).

In Appendix A, we provide details of the SPA implementation we use in our experiments in §5.3.

5.3 Experiments

To test our proposed protocol, we ask 100 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). 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

System X	System Y	Expected Preference	$P[X \succ Y]$	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 human preferences for story generation, using both system-level probabilistic assessment (SPA) and the standard protocol with 5-point Likert ratings. We use a Student's t-test with Holm-Bonferroni-corrected statistical significance at $\alpha = 0.10(^*), 0.05(^{**}), 0.01(^{***})$. SPA yields a significant result in the expected direction for all comparisons, while Likert ratings only do so for 2 (green). Notably, the standard protocol yields the unintuitive result 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). Insignificant results are in yellow.

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 each annotator will take longer to complete each task. We balance these concerns by choosing $m = 5$.

Given that larger GPT-3 variants generate more coherent, grammatical, and creative text, annotators should overall prefer larger variants to smaller ones: 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 between the latter 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 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 from one another. As we are making 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 surprisingly the probability of a human being a better writer than `davinci` is significantly less than chance, despite past work suggesting that annotators cannot tell the difference between the two (Clark et al., 2021).

5.4 Limitations

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

1. SPA does not measure the magnitude of a preference, only the probability that it exists. This is a necessary trade-off for SPA to be applicable to open-ended NLG tasks, for which extremum equivalence (§3) cannot be satisfied when measuring magnitude.

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. 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). We discuss possible strategies in Appendix A and show in Figure 1 how adding filtering techniques in future work may make SPA even more effective.

6 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 there is much inconsistency is how these terms are defined and explained to annotators (Howcroft et al., 2020; van der Lee et al., 2021). This is most often done in the NLG literature by having annotators

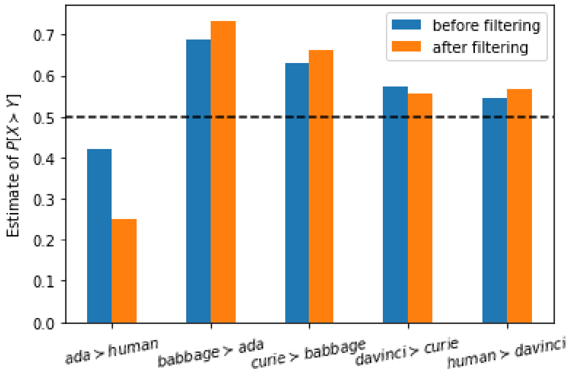


Figure 1: Only considering annotators whose stated probabilities concur with the Likert ratings they assign to the system outputs – despite being a noisy filter for annotator quality – gives us preference probabilities that more strongly go in the expected direction. The main results in Table 1 are given without this filter.

assign a Likert rating from 1-to-k, where k is usually 3 or 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; Zhou et al., 2022).

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 is far less common nowadays, in large part because making pairwise comparisons of outcomes is more expensive than outcome-level ratings at scale (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, but little of it extends to open-ended NLG, as the latter lacks strict references. 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 rankings of MT systems 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 competitions in 2017 and used thereafter (Bojar et al., 2017; Specia et al., 2021).

To our knowledge, utility theory has only been applied in NLP to design leaderboards (Ethayarajh and Jurafsky, 2020; Ma et al., 2021). However, past work has hinted at utilitarian issues without framing them as such, such as whether crowdworkers are sufficiently motivated (Belz and Reiter, 2006; Dugan et al., 2020). Some work has recommended best practices to align annotator's incentives with that of potential users of the NLG system (Oppenheimer et al., 2009; Daniel et al., 2018), which can include training crowdworkers to become better evaluators (Mitra et al., 2015; Clark et al., 2021).

7 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 about annotator preferences – 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 derived using the standard protocol and seeing which conclusions in the NLG literature stand up to scrutiny.

Acknowledgements

Many thanks to Kaitlyn Zhou for helpful discussion and feedback. We also thank Clara Meister and Simran Arora for answering queries about annotation on MTurk.

Acknowledgements

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.

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*, pages 313–320.

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.

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. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

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.

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.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

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*.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

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.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34.

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

Tanushree Mitra, Clayton 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.

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.

Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872.

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.

Elisabeth Svensson. 2000. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(4):417–434.

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 Experiment Details

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, and a result, the human-written stories were 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 2). 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 $15/hour. Each annotator was presented with the instructions in Figure 3 and then asked to provide 5 preference probabilities $P[X \succ Y]$, 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 be onerous and unnecessary; for an apples-to-apples comparison of SPA and the standard evaluation protocol, we should have an equal sample size for each. This way, for every pair of systems that were being compared, we would have 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.

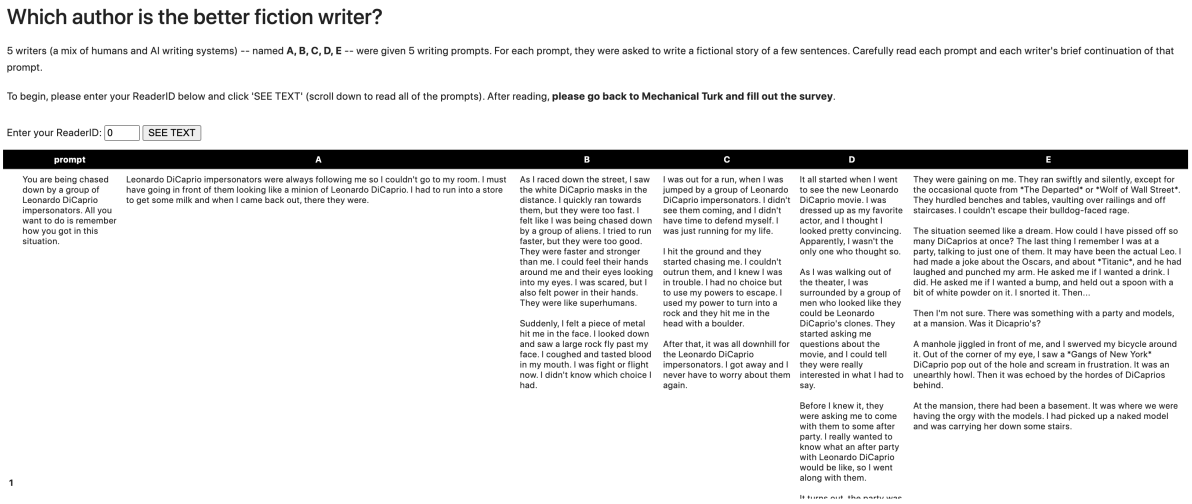
After the annotators provided their annotations, we excluded those who: (1) said they were not native English speakers; (2) did not follow our instructions and submitted multiple HITs. 10% of annotators were excluded, leaving 90 whose annotations we used. Submitting multiple HITS was an issue because 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 thee same annotators from making multiple submissions.

A.3 Future Work

Our design choices in our experiments raise two interesting directions for future work:

1. How do the results change as m, the number of samples seen, changes? We chose $m = 5$ to trade-off the greater uncertainty from viewing too few samples from each writing system against the greater cost and time needed to read many samples. However, what does this trade-off look like precisely? Is there are Pareto-optimal choice for m?

2. How can we better filter out insincere and inattentive annotators? After all, just because an annotator states they're a native English speaker and only submitted one HIT does not necessarily mean they read all the writing samples thoroughly, as instructed. One idea is to hire annotators personally, but this naturally does not scale as well as MTurk.



Which author is the better fiction writer?

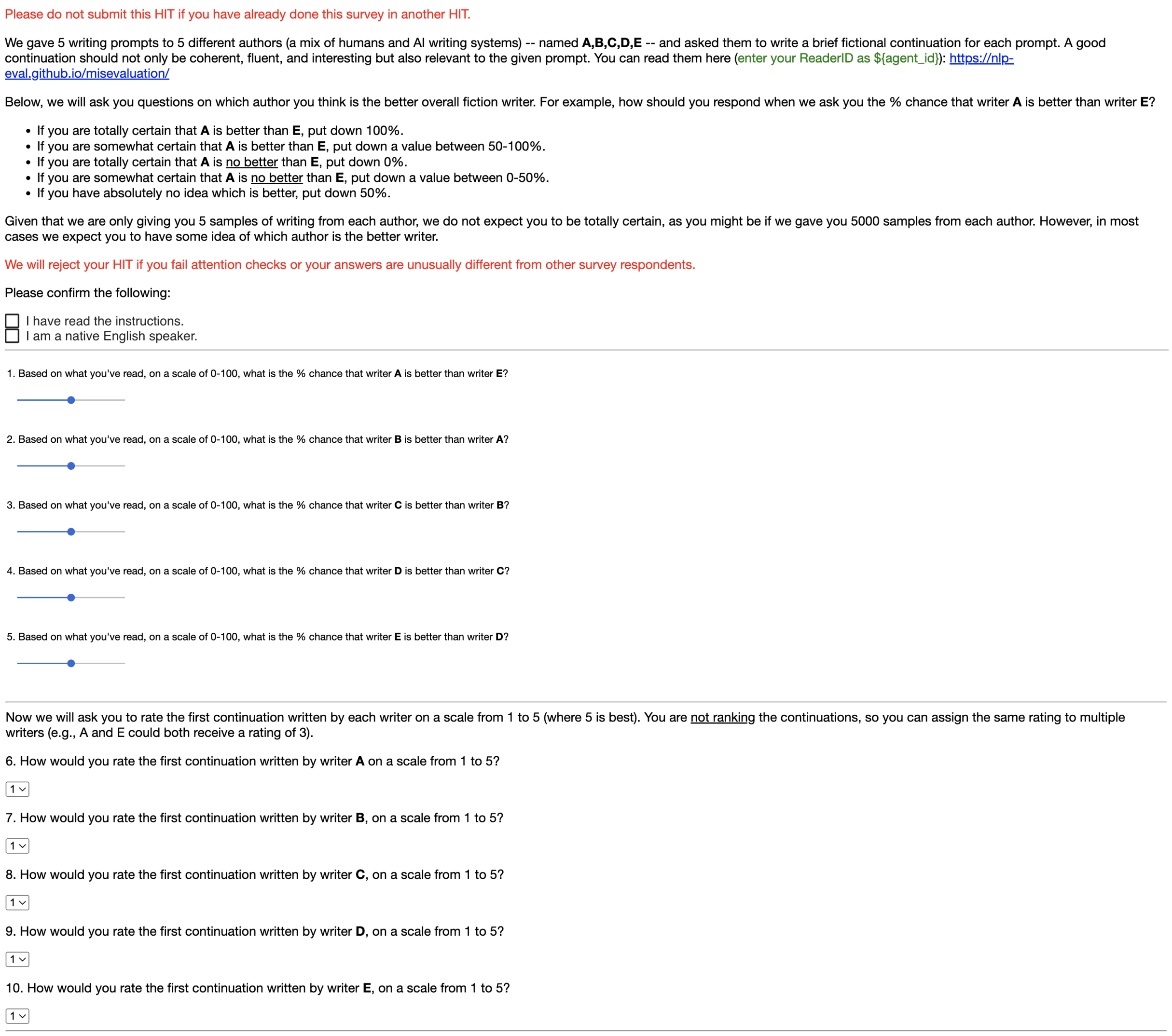
5 writers (a mix of humans and AI writing systems) -- named **A, B, C, D, E** -- were given 5 writing prompts. For each prompt, they were asked to write a fictional story of a few sentences. Carefully read each prompt and each writer's brief continuation of that prompt.

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 2: 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 3: The instructions given to annotators on Amazon Mechanical Turk.