Compare without Despair: Reliable Preference Evaluation with Generation SEPARABILITY

Sayan Ghosh
University of Southern California
ghoshsay@usc.edu

Tejas Srinivasan
University of Southern California
tejas.srinivasan@usc.edu

Swabha Swayamdipta
University of Southern California
swabhas@usc.edu

Abstract

Human evaluation of generated language through pairwise preference judgments is pervasive. However, under common scenarios, such as when generations from a model pair are very similar, or when stochastic decoding results in large variations in generations, it results in *inconsistent* preference ratings. We address these challenges by introducing a meta-evaluation measure, SEPARABILITY, which estimates how suitable a test instance is for pairwise preference evaluation. For a candidate test instance, SEPARABILITY samples multiple generations from a pair of models, and measures how *distinguishable* the two sets of generations are. Our experiments show that instances with high SEPARABILITY values yield more consistent preference ratings from both human- and auto-raters. Further, the distribution of SEPARABILITY allows insights into which test benchmarks are more valuable for comparing models. Finally, we incorporate SEPARABILITY into ELO ratings, accounting for how suitable each test instance might be for reliably ranking LLMs. Overall, SEPARABILITY has implications for consistent, efficient and robust preference evaluation of LLMs with both human- and auto-raters.

1 Introduction

As large language models' (LLM's) capabilities have rapidly improved in recent years, evaluation of these capabilities has become increasingly reliant on human preference judgments that compare pairs of model generations. While these judgments offer freedom from gold-standard references (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019), they are far from perfect (Gehrmann et al., 2022).

In particular, human evaluation faces issues including low rater agreements (Goyal et al., 2022), spurious correlations with factors like length (Wu and Aji, 2023; Sun et al., 2019), lack of measurement validity (Ethayarajh and Jurafsky, 2022), and inconsistent usage and interpretation of inter-rater agreement (Amidei et al., 2019; Prabhakaran et al., 2021). Furthermore, for annotation efficiency, human judgments are sometimes replaced with LLM judgments, which have shown high correlation with crowdworker ratings (Dubois et al., 2024; Zheng et al., 2024; Lin et al., 2024; Liu et al., 2023; Zeng et al., 2023); however, it remains unclear whether such auto-evaluations are a step in the right direction or exacerbate existing biases (Zheng et al., 2024; Wang et al., 2023; Wu and Aji, 2023; Chang et al., 2024).

In this work, we focus on the problem of *unreliable* preference judgments from human raters, illustrated in Figure 1. We show that output pairs from any two modern LLMs can often be hard to distinguish from each other; such high **cross-alignment** between models can cause preference judgments to be highly arbitrary. We identify another understudied factor affecting judgments: the *variability within one LLM's generations for the same input*, owing to the stochasticity of popular decoding techniques such as temperature sampling (Giulianelli et al., 2023; Tsvilodub et al., 2024). Such low **self-alignment**, in addition to high cross-alignment between models, may lead to *inconsistent* ratings—highly dependent on the exact sampled pair chosen for preference judgments. As a concrete example, on 100 news articles from CNN/DailyMail (Nallapati et al., 2016), our human evaluations show that when comparing five different summary pairs for each input, raters picked the same model only 46% of the time (§3). These findings raise the question: when can we rely on pairwise judgments to compare generations from two LLMs?

We argue that some test instances might be better suited for human evaluation than others, mirroring insights from prior work (Rodriguez et al., 2021; Vania et al., 2021). We introduce **SEPARABILITY**, a meta-evaluation measure that determines, for a single instance, how distinguishable two sets

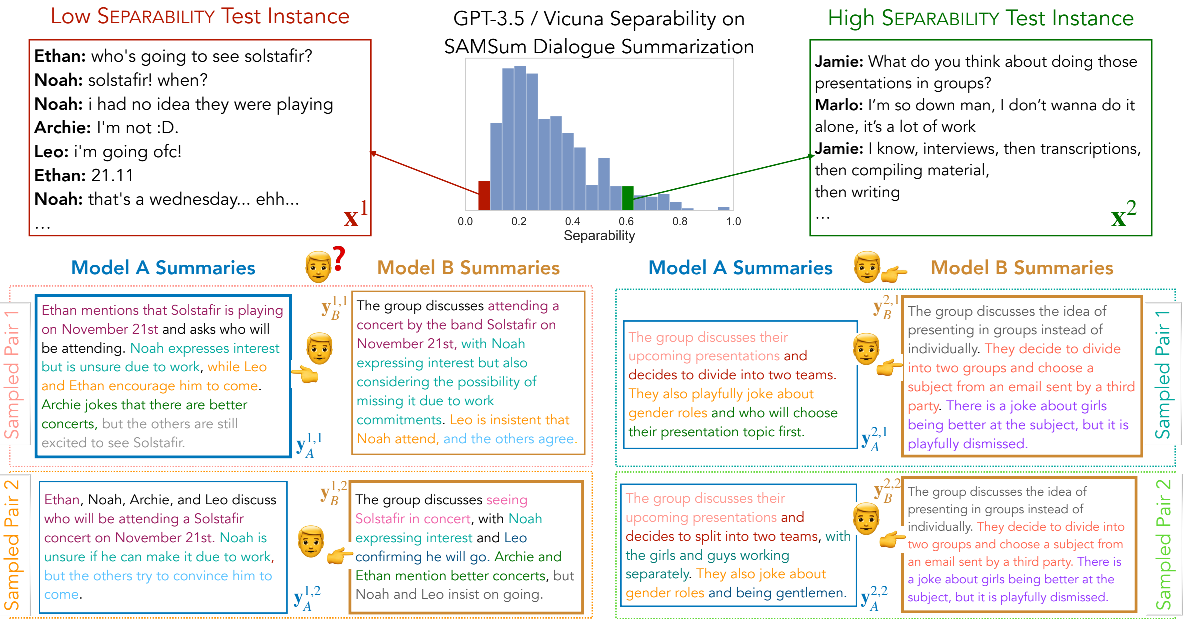

Figure 1: Illustration of SEPARABILITY on SAMSum dialog summarization from our human experiments (§3). Test instances have varying degrees of SEPARABILITY, which lead to different levels of consistency in preference ratings. For lower SEPARABILITY instances, the choice of which pair of sampled generations to show raters affects human rating (raters preferred Model A under Pair 1 and B under Pair 2); hence the overall judgment between model pairs is inconsistent. Human preferences are consistent under higher SEPARABILITY (raters always preferred Model B).

of generations from two models are (§2). SEPARABILITY builds on the intuition that the harder it is to distinguish generations from two models, the less consistent the preference ratings will be. Our formulation of SEPARABILITY combines cross-alignment between generations between pairs of models as well as self-alignment between multiple generations from each given model. We operationalize self- and cross-alignment in SEPARABILITY with a flexible choice of similarity metric depending on the salience of the variability (e.g. lexical, semantic) in the preference judgment.

Our experiments with SEPARABILITY on different model pairs, benchmarks and tasks show that SEPARABILITY can not only identify test instances which are likely to yield consistent preference ratings, but also benchmarks likely to yield consistent comparisons between models. For instance, we show that evaluation sets such as CNN/DailyMail (Nallapati et al., 2016) are not as useful in comparing modern LLMs as they were in comparing earlier summarization-specific models, supporting prior findings (Goyal et al., 2022; Zhang et al., 2023). Through extensive human evaluation, we show that instances with high SEPARABILITY scores tend to result in more consistent preferences (§3). Moreover, we find that LLM-based auto-evaluation systems (Dubois et al., 2024) also have similar patterns of consistency and inconsistency as human raters.

Finally, as a direct application of SEPARABILITY, we extend it to ELO, a rating system based on pairwise comparisons, now used widely for LLM generations (Chiang et al., 2024). By modifying the ELO rating update function to account for the SEPARABILITY of each new instance, our SEPARABILITY-ELO ratings provide more nuanced comparisons (§4). Overall, SEPARABILITY offers a reliable signal in the noisy landscape of generative evaluation via pairwise ratings, provides insights into benchmarks and test instances, and complements existing ranking measures for robust preference evaluation. Our code and data will be publicly released.[1]

2 SEPARABILITY as a Meta-Evaluation

We address the problem of consistency in modern generative evaluation: namely, how suitable a test input $\mathbf{x}_i \in \mathcal{X}$ is for collecting reliable preference ratings between output generations from a pair of models, m_A and m_B. Our approach is based on the key intuition that it is harder to collect consistent preference ratings between m_A and m_B

[1] https://anonymous

if their output generations are, on average, harder to distinguish for a (human) rater. For instance, the distinction is hard when the generations focus on similar content (see summaries in Figure 1, left), or have similar styles; we call this **high cross-alignment** between generations from m_A and m_B. Another factor that may make distinguishing two models' generations harder is large variability within each model's sampled generations, due to stochastic LLM decoding approaches such as temperature and nucleus sampling. Such variability, which we refer to as **low self-alignment**, makes it hard to characterize each model's specific tendencies, which in turn makes it hard to have a consistent preference for a single model. Under the above two conditions, the choice of which generations to use for pairwise comparison influences the preference rating outcome (Figure 1).

Both kinds of alignments, while orthogonal, play a key role in determining how consistent human ratings for an instance might be (§2.1). We introduce **SEPARABILITY**, a meta-evaluation measure that estimates how suitable a test instance is for preference rating by consolidating cross- and self-alignment (§2.2). While SEPARABILITY does not determine which generation is better or more preferable, it helps us understand how much we can trust each preference rating for a given input instance.

At a very high level, there are four common scenarios which may occur in comparing generations from two models, which we highlight in Figure 2. Scenarios 1, 2 and 3 all depict output sets where any sample from model A is expected to be very distant from any sample from model B (i.e. low *cross-alignment*). It is easy to distinguish the two models under scenarios 1 and 3—if two generations are very different, they must be from different models. Under scenario 2, generations *from the same model* are also far from each other (i.e. the *self-alignment* is also low), which makes the overall output sets hard to distinguish. In contrast, scenario 4 depicts a situation where both self- and cross-alignments are high; all generations, regardless of the model they came from, are similar, making it hard for the rater to distinguish the models' output sets.

Formally, given a set of test inputs $\{\mathbf{x}^i\}_{i=1}^{D}$ (e.g. news articles and instructions to summarize them), LLMs m_A and m_B each induce a conditional distribution $p_{m_A}(\mathbf{y}^i \mid \mathbf{x}^i), p_{m_B}(\mathbf{y}^i \mid \mathbf{x}^i)$ over output generations $\mathbf{y}^i \in \mathcal{Y}$ (e.g. summaries of $\mathbf{x}^i$). From this distribution, we can sample K generations using a common stochastic decoding

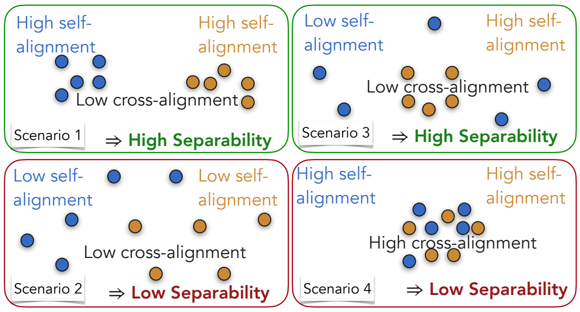

Figure 2: Four scenarios illustrating the intuition behind SEPARABILITY. Blue and gold circles represent generations from models m_A and m_B respectively, and Euclidean distances represent (dis)similarities between them. For a given input, at least one of the two models needs to have higher similarity among its own generations (high self-alignment) to have high SEPARABILITY for that input. High similarity across generations from different models (high cross-alignment) leads to lower SEPARABILITY. High self-alignment corresponds to low spread of a set of same-colored circles and vice-versa. High cross-alignment corresponds to low spread of the entire set of circles and vice-versa.

algorithm such as temperature sampling, yielding sets $\{\hat{\mathbf{y}}_A^{i,j}\}_{j=1}^{K}$ and $\{\hat{\mathbf{y}}_B^{i,l}\}_{l=1}^{K}$.

2.1 Calculating Generation Alignments

We define an ***alignment function***, $\mathcal{A}^i_{A,B}$ that estimates the similarity of two output distributions $p_{m_A}(\mathbf{y}^i \mid \mathbf{x}^i)$, $p_{m_B}(\mathbf{y}^i \mid \mathbf{x}^i)$ produced by LLMs m_A, m_B on an input $\mathbf{x}^i$,

$$\mathcal{A}^i_{A,B} := \mathbb{E}_{\mathbf{y}^i_A \sim p_{m_A}, \mathbf{y}^i_B \sim p_{m_B}}[s(\mathbf{y}^i_A, \mathbf{y}^i_B)], \quad (1)$$

where $s : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a text similarity metric such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), or BLEU (Papineni et al., 2002). Intuitively, the alignment score $\mathcal{A}^i_{A,B}$ measures the expected similarity between an output from m_A and an output from m_B, parameterized by s. A high value of $\mathcal{A}^i_{A,B}$ indicates high similarity (low variability) between the generations of the two models. Different similarity metrics can be used for different tasks. In cases where a user cares about fine-grained lexical differences, a metric such as BLEU may be suitable. On the other hand, if fine-grained lexical differences are less important than coarser semantic differences, metrics such as BERTScore or word embedding cosine similarity would be more suitable. We use a variation of BERTScore with a length-adjustment (defined in Appendix A), unless otherwise noted.

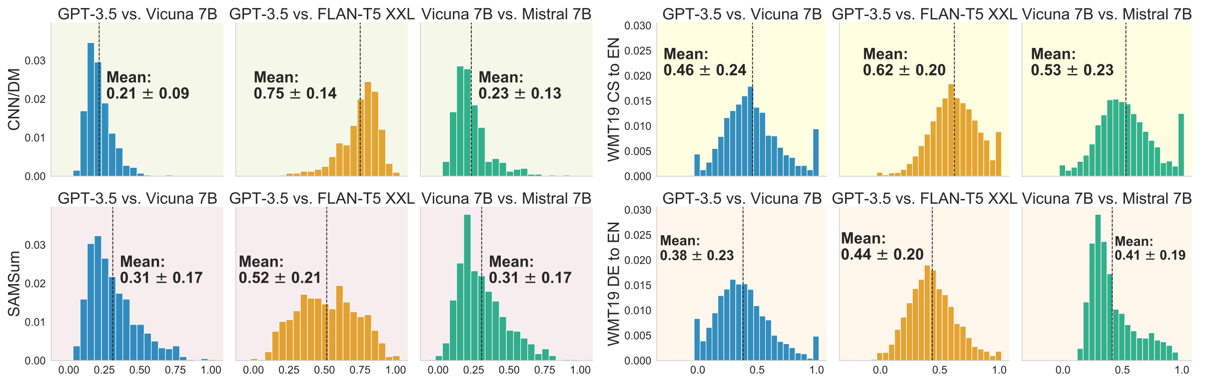

Figure 3: Histograms of SEPARABILITY distributions for **summarization** (Left) and **translation** (Right). For similar model pairs, CNN/DailyMail for news summarization and translation from a high-resource language (German) have lower average SEPARABILITY compared to SAMSum for dialogue summarization and translation from a lower-resource language (Czech). We use length-adjusted BERTScore (Zhang et al., 2019) as the similarity metric (defined in Appendix A) for summarization and BLEU (Papineni et al., 2002) for translation.

Since the space of model generations is intractable to calculating Equation 1 exactly, in practice we use Monte-Carlo samples to approximate the alignment score. We sample K generations from each model, resulting in output sets $\{\hat{\mathbf{y}}_A^{i,j}\}_{j=1}^K$ and $\{\hat{\mathbf{y}}_B^{i,l}\}_{l=1}^K$. We approximate $\mathcal{A}_{A,B}^i$ as:

$$\hat{A}_{A,B}^i = \frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K} s\left(\hat{\mathbf{y}}_A^{i,j}, \hat{\mathbf{y}}_B^{i,l}\right) \qquad (2)$$

When measuring **self-alignment**[2]—the level of variability in an individual model's output—we set $A = B$ in Equation 2. When evaluating the variability between two distinct model, i.e. $A \neq B$, we label this function ***cross-alignment***.[3]

2.2 Calculating Generation SEPARABILITY

Intuitively, in order to determine how distinguishable two models are, we need to measure the difference between the variability within each model's generation sets and the variability of the combined set of generations (i.e. the difference between each self-alignment and the cross-alignment). If the combined set has more variability than the variability within either model's generations, we consider the two generation sets to be *separable*.

We define the generation SEPARABILITY between models A and B for instance i, $\delta_{A,B}^i$ as:

$$\delta_{A,B}^i = \max\left(\mathcal{A}_{A,A}^i, \mathcal{A}_{B,B}^i\right) - \mathcal{A}_{A,B}^i. \qquad (3)$$

[2]For self-alignment, we skip $j = l$ terms in the summation.

[3]We generate $K = 5$ samples using temperature sampling with $\tau = 0.5$ as the default in our experiments; this corresponds to $K^2 = 25$ cross-alignment comparisons.

In Figure 2, under scenarios 1 and 3, generations of at least one model have low variability (and therefore high self-alignment); this combined with the low cross-alignment leads to higher SEPARABILITY than in scenarios 2 and 4.

SEPARABILITY can take values in $[-1, 1]$.[4] In practice, however, cross-alignment is usually lower than self-alignment, making $\delta_{A,B}^i \in [0, 1]$.

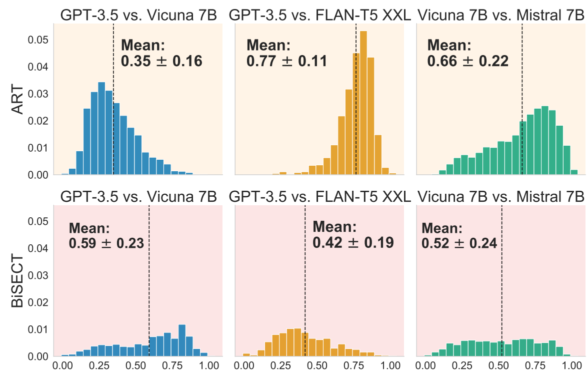

Figure 4: SEPARABILITY distributions for ART and BiSECT. We use length-adjusted BERTScore here (defined in Appendix A) as the similarity metric. SEPARABILITY has higher variance, especially for BiSECT, largely caused by differences in instruction prompt interpretation; see Appendix C.

2.3 Computing SEPARABILITY on Generation Benchmarks

We compute SEPARABILITY for various generation tasks under 6 different benchmarks using 3 model pairs, and demonstrate how SEPARABILITY can allow model developers and users to visualize

[4]We apply min-max normalization to constrain the range of alignments to $[0, 1]$.

and understand how much a model pair's generations differ on a particular dataset. Figure 3 shows the empirical SEPARABILITY distributions on two summarization benchmarks (left): CNN/DailyMail (Nallapati et al., 2016) and SAMSum (Gliwa et al., 2019), and two machine translation benchmarks (right): Czech to English and German to English from the WMT-19 dataset (Barrault et al., 2019). Figure 4 shows the empirical SEPARABILITY distributions for abductive reasoning and sentence simplification; we use the ART (Bhagavatula et al., 2019) and BiSECT (Kim et al., 2021) benchmarks respectively. In each case, we compare three model pairs: `GPT-3.5` vs. `Vicuna-7B` (LMSys, 2023), `Vicuna-7B` vs. `Mistral-7B` (Jiang et al., 2023), and `GPT-3.5` vs. `FLAN-T5-XXL` (Longpre et al., 2023). Appendix A contains additional details about the generation settings. In Appendix C, we present examples of low and high SEPARABILITY generations corresponding to each of these tasks.

We highlight several key takeaways. Models with very different training methods, such as `GPT-3.5` and `FLAN-T5-XXL`, output generations that are, on average, much easier to distinguish than models that are trained similarly, such as `Vicuna-7B` and `Mistral-7B`. Benchmarks such as CNN/DailyMail (Figure 3, top left) have instances with very low SEPARABILITY on average (except `GPT-3.5` versus `FLAN-T5-XXL`). These findings corroborate prior work that suggests CNN/DailyMail may not be useful for comparing modern LLMs (Goyal et al., 2022; Zhang et al., 2023).

Likewise, for machine translation, we see that it is easier to distinguish LLMs on lower-resource language test sets such as Czech→English, compared to high-resource language test sets such as German→English.

Notably, SEPARABILITY distributions for BiSECT are far less peaked (Figure 4), indicating highly variable SEPARABILITY. For both ART and BiSECT, differences in how the models interpreted the instructions (which didn't include explicit length constraints for these benchmarks) led to large differences in generation lengths, contributing to the high SEPARABILITY of certain instances. See Table 5 in Appendix C for examples.

Our formulation of SEPARABILITY is robust to the choice of hyperparameters: K for number of samples, τ for temperature in sampling and the number of samples used in computing cross-alignment, C; Figure 5 shows these ablations.

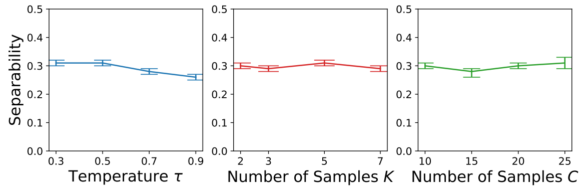

Figure 5: SEPARABILITY is robust to changes in the temperature τ used for generation (Left), the number of samples used to estimate alignments K (Middle), and the number of cross-alignment comparisons C (Right), for `GPT-3.5` vs. `Vicuna-7B` on SAMSum.

3 SEPARABILITY as Rating Consistency

We conduct a human study to verify our formulation for SEPARABILITY as a meta-evaluation measure of preference rating consistency for a given instance $\mathbf{x}^i$ and a pair of generative models, m_A and m_B. Given that generation sets corresponding to low SEPARABILITY instances are harder to distinguish, we hypothesize that preference judgments from raters on those sets will be inconsistent. In other words, raters will not consistently prefer the same model's generation for any pair of generations sampled from low SEPARABILITY instances.

3.1 Rating Consistency

We define consistency of preference judgments as the average ratings from raters over N sampled pairs. For an input instance $\mathbf{x}^i$, we sample N generations $\mathbf{y}^i \in \mathcal{Y}$ each from models A and B to obtain a set of paired generations $\mathcal{P}_{AB}(\mathbf{x}^i) = \left\{ \left(\mathbf{y}_A^{i,j}, \mathbf{y}_B^{i,j} \right) \right\}_{j=1}^{N}$.

We represent a rater (annotator) a by a rating function $r_a : \mathcal{Y} \times \mathcal{Y} \to \{-1, 0, 1\}$ that, for a pair of generations from m_A and m_B, indicates which model's generation was preferred: -1 if m_A's generation was preferred, 1 in case m_B was preferred, and 0 if the rater had no preference. By having rater a make preference judgments for each generation pair in $\mathcal{P}(\mathbf{x}^i)$, we obtain a rating set $\mathcal{R}_a(\mathbf{x}^i) := \left\{ r_a \left(\mathbf{y}_A^{i,j}, \mathbf{y}_B^{i,j} \right) \right\}_{j=1}^{N}$. We define the ***consistency***, $c\left(\mathcal{R}_a(\mathbf{x}^i)\right)$ of that rating set as:

$$\begin{cases} 0, & \text{if } \{-1, 1\} \subset \mathcal{R}_a(\mathbf{x}^i), \\ \text{mean}(\left| r_a(\mathbf{x}^i) \right|_{r_a \in \mathcal{R}_a}), & \text{otherwise} \end{cases} \tag{4}$$

Intuitively, if the rater prefers generations from both models during the course of the N trials, we deem their rating set inconsistent (i.e. $c\left(\mathcal{R}_a(\mathbf{x}_i)\right) = 0$). If the rater only ever picks one

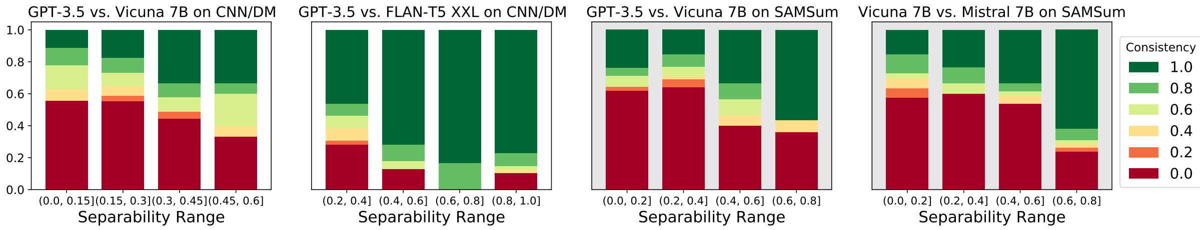

Figure 6: Proportion of rating sets with each value in the range of consistency corresponding to different SEPARABILITY ranges. Ratings are not aggregated across annotators for a test instance here. For each model pair and dataset configuration, the support is divided into equal sized ranges. The proportion of perfectly consistent ratings increases, and the proportion of inconsistent ratings decreases in higher SEPARABILITY ranges.

model's generations or 0 ratings (ties), we deem their rating sets that include fewer 0 ratings as more consistent.

In some cases, we may want to differentiate cases where there are differing degrees of inconsistency. We address these cases through an additional metric called *system preference strength*, with definitions and results in Appendix D.

3.2 Study Protocol and Settings

We conducted a human study with raters hired from Amazon Mechanical Turk. Each human intelligence task (HIT) consisted of reading a source text (in our case, a news article or a dialogue) and $N = 5$ pairs of generated summaries.[5] For each summary pair, raters were asked to select which summary they preferred, with the option of picking no preference. The HIT interface and further details can be found in Appendix B.

Each HIT batch comprised source texts and summaries corresponding to a different model pair and dataset configuration. We chose these configurations such that we had one set of instances with low average SEPARABILITY (~ 0.2), one with high average SEPARABILITY (~ 0.7), and two in-between:

1. Low: `GPT-3.5` vs. `Vicuna-7B` on CNN/DM
2. High: `GPT-3.5` vs. `FLAN-T5-XXL` on CNN/DM
3. Medium: `GPT-3.5` vs. `Vicuna-7B` on SAMSum
4. Medium: `Vicuna-7B` vs. `Mistral-7B` on SAMSum

We collected ratings for 50 HITs for each of the four configurations. Since each HIT was rated by 3 raters, we have 600 total rating sets.

[5]While we performed our experiments with summarization, we expect our results to hold for other tasks as well. In addition, the annotators are performing 5 cross-model comparisons as opposed to 25 when calculating SEPARABILITY, but we find that 5 comparisons suffice.

3.3 Higher SEPARABILITY Instances Receive Consistent Human Ratings

To analyze the relationship between SEPARABILITY ranges and consistency ratings, we bin the support of a SEPARABILITY distribution for each our selected configurations into four equal-width bins. We plot the proportion of rating sets with each possible consistency value in each bin in Figure 6. For each model pair on the two benchmarks, we observe that, as SEPARABILITY increases, the proportion of inconsistent rating sets decreases and the proportion of perfectly consistent ratings increases. For SEPARABILITY $\delta_{A,B} \leq 0.2$, the majority of ratings are inconsistent. When $\delta_{A,B} \approx 0.4$, inconsistent ratings make up less than half of all ratings for all configurations. Ratings for SAMSum tend to be more inconsistent across all ranges. `GPT-3.5` and `FLAN-T5-XXL`, two models with different architectures and capabilities always produce more consistent ratings, even at lower SEPARABILITY ranges. Nonetheless, there are a non-trivial number of inconsistent ratings at the lowest SEPARABILITY range, and perfectly consistent ratings make up less than half of all ratings at this range. These findings indicate that **at higher values of SEPARABILITY, raters are likely to give more reliable preference ratings** that are not dependent on the choice of generation pair that they are shown.

3.4 Higher SEPARABILITY Instances Receive Consistent Auto-Rater Ratings

LLM-based automatic raters have been rising in popularity (Chang et al., 2024) and are being used to replace human raters in many preference evaluation setups. We ask: do auto-raters produce similar patterns of consistency as humans when making preference judgments? We repeat our experiments using the same 600 instances in §3.3 with auto-raters provided by AlpacaEval (Dubois et al., 2024)

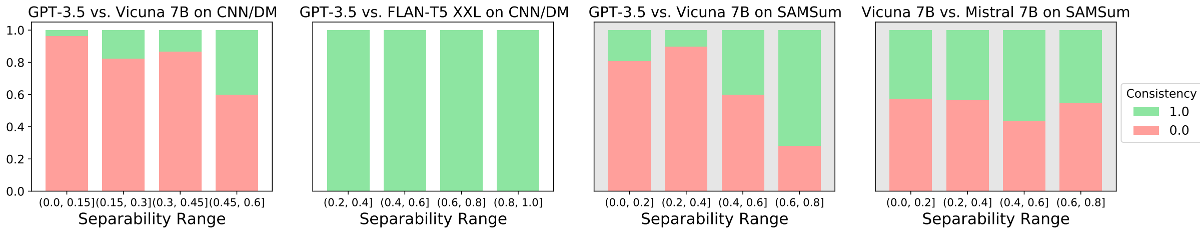

Figure 7: Auto-raters from AlpacaEval produce more consistent ratings at higher SEPARABILITY instances, much like human raters in Figure 6.

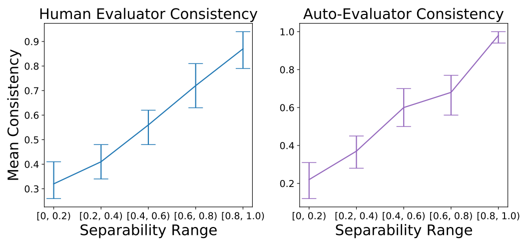

Figure 8: Mean consistency of human and auto-rater preference judgments increases with SEPARABILITY. Mean consistency is computed over all 600 HITs collected. Consistency for a particular test instance is aggregated over annotators by taking the mean of each individual annotator's rating consistency.

Each test instance is judged by three auto-raters, which have the highest correlations with humans.[6] Since these raters cannot give tie judgments, the only possible consistencies are 0 or 1; other details are in Appendix A.

Results in Figure 7 show that, much like humans, auto-raters produce inconsistent ratings for low SEPARABILITY instances under most configurations. For the `GPT-3.5` vs. `FLAN-T5-XXL` comparison, the auto-raters always choose `GPT-3.5`, whereas humans sometimes choose `FLAN-T5-XXL` in lower separability ranges. This phenomenon may be due to auto-raters being biased towards generations from their own model family (Panickssery et al., 2024).[7] In contrast to human raters, auto-raters provide inconsistent ratings between `Vicuna-7B` and `Mistral-7B` even under higher SEPARABILITY. This suggests that the factors influencing human judgments can be subtle and different from those influencing auto-raters.

In Figure 8, we plot the mean consistency for each of five equal-sized SEPARABILITY ranges, aggregating over all four model pair and dataset configurations. Consistency increases with SEPARABILITY for both human- and auto-raters, highlighting that raw SEPARABILITY values can be directly compared across model pairs and datasets. Moreover, human- and auto-rater consistency patterns bear close resemblance with each other, with auto-rater consistency being slightly lower on average. This resemblance suggests that **SEPARABILITY is a valid meta-evaluation measure of the reliability of preference ratings, regardless of the type of rater**.

4 Applying SEPARABILITY to ELO

As another concrete application of SEPARABILITY, we investigate extending a popular novel method for ranking LLMs: ELO ratings (Chiang et al., 2024; Boubdir et al., 2023b). In particular, we weight how much a new preference comparison affects a model's ELO rating using the SEPARABILITY of the test instance for that comparison.

ELO ratings have emerged as a popular method of scoring and comparing LLMs (Chiang et al., 2024; Boubdir et al., 2023b). Originally developed to score and rank Chess players, ELO ratings model the expected win probability of a model in a pairwise comparison. After observing the outcome of a comparison between two models, both models' ratings are updated. The ELO updated rating for a model m_A is given by

$$\mathrm{ELO}'_A = \mathrm{ELO}_A + K^i(S^i_A - E^i_A), \qquad (5)$$

where ELO_A is the original rating, S^i_A is the outcome of the comparison with instance i, E^i_A is the expected win probability (based on the current ELO score), and K^i is a weighting factor which determines how much more recent comparisons should influence the rating. The value of S^i_A is equal to 1 for a win, 0 for a loss, and $\frac{1}{2}$ for a tie. Typically, K^i is set to small values such as $K^i = 4$ for all i in LLM comparisons (Chiang et al., 2024); larger K^i values are used in sports.

We propose incorporating SEPARABILITY into the ELO update in Equation 5 by modifying the

[6] https://github.com/tatsu-lab/alpaca_eval.

[7] In our case, two of the auto-raters are in the `GPT` family.

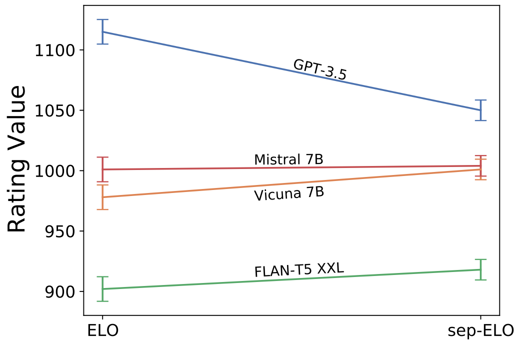

Figure 9: After incorporating SEPARABILITY into ELO, we get narrower gaps in model rankings, reflecting similar capabilities of both `Mistral-7B` and `Vicuna-7B`.

weight K^i for each new comparison based on its SEPARABILITY value. For an update ELO'_A after a comparison on an instance $\mathbf{x}^i$ with SEPARABILITY $\delta^i_{A,B}$, we use the weighting factor:

$$K^i_{\text{SEP}} = K^i \cdot \frac{\alpha}{1 + \exp\left(-\beta(\delta^i_{A,B} - T)\right)}, \quad (6)$$

where T is a chosen threshold, α and β are hyperparameters controlling the how much the weight is updated and how fast. We set $T = 0.4$ and $\alpha = 2$ and $\beta = 6$ in our experiments.[8] Intuitively, this update rule upweights K^i (Equation 5) when the input i has high SEPARABILITY, and vice-versa. When the input's SEPARABILITY value is at the chosen threshold T, K is not updated.

We compute ELO and SEPARABILITY-weighted ELO (SEP-ELO; Equation 6) using data from our 600 human evaluation HITs as preference judgments (§3). To calculate these ratings, we sample one rating for each input from our pool of ratings. We compute confidence intervals with the bootstrap method with 100 trials. Figure 9 shows that SEP-ELO has narrower gaps in model ranking, suggesting that models are more similar under adjustments to consistency of judgments (or, SEPARABILITY). We acknowledge that our results are a pilot due to the limited number of ratings we could use in our computation. However, we expect SEP-ELO can reveal reliable trends even when applied to larger sets of preference data such as LMSYS[9].

Alternative Applications of SEPARABILITY Beyond SEP-ELO, SEPARABILITY values could be used for adversarially filtering test sets (Bras et al., 2020). Not only would this lead to fine-grained comparisons between models, but could also lead to obtaining cost- and time-efficient human ratings. However, some caution is to be urged since such filtering may lead to biases (Schwartz and Stanovsky, 2022), since low SEPARABILITY instances can still contain valuable information. Instead, we recommend importance *weighting* instances by SEPARABILITY when sampling instances for human judgments, in a similar manner as it is used in ELO ratings. Alternatively, a stratified sampling approach from different separability ranges could ensure a more robust preference evaluation scheme.

5 Related Work

Model Output Variability Giulianelli et al. (2023) comprehensively characterize LLM vs human output variability, with a focus on comparing it to human output variability. Suzgun et al. (2022); Bertsch et al. (2023) take advantage of production variability to select more optimal generations using Minimum Bayes Risk (MBR) decoding. In a similar vein, our work incorporates variability in generations into our meta-evaluation measure.

Prioritizing Test Instances Rodriguez et al. (2021); Vania et al. (2021) evaluate test instances on a variety of dimensions such as difficulty and discriminability (similar to our notion of SEPARABILITY) using Item Response Theory (IRT), albeit in a text classification setting. Boubdir et al. (2023a); Ashury-Tahan et al. (2024) study prioritizing test instances for human evaluation. However, their approach relies on access to model logits which are not necessarily available to LLM users. Moreover, we take a more task-centered approach.

6 Conclusion

We present SEPARABILITY, a meta-evaluation measure that estimates how suitable a test instance is for pairwise preference elicitation. We show that instances with high SEPARABILITY yield more consistent human judgments. We show that the test distribution of SEPARABILITY can be used to analyze how useful a benchmark may be for the comparison of two LLMs. We show that SEPARABILITY can be incorporated into ELO scores. Our work shows that SEPARABILITY can help LLM developers and users determine and prioritize evaluation instances and benchmarks. Future work will look at applying SEPARABILITY in building quality filters for preference tuning data for learning from human feedback.

[8] Since we do not have ground-truth regarding true model rankings, these parameters are dependent on user preference

[9] `https://lmsys.org/`

Limitations

We only used SEPARABILITY in tasks that produce English output generations. Due to resource and time constraints, our human evaluation for verifying SEPARABILITY was done on two summarization tasks with five summary pair comparisons for each instance by three annotators. We chose instances with separability values for our human comparisons to highlight different levels of consistency in ratings. We expect that an even larger comparison would reveal more fine-grained variations. Our analysis on applying to SEPARABILITY to ELO also used limited human comparisons and model pairs. Larger scale preference data collection would be needed for more fine-grained analysis. While we expect our conclusions to hold for different tasks, different similarity functions may be optimal for different tasks, since the importance of *what* type of differences are most influential for human judgments can vary by task. Furthermore, we only used 5 human comparisons per pair and 5 samples to compute SEPARABILITY for our main experiments.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

Shir Ashury-Tahan, Benjamin Sznajder, Leshem Choshen, Liat Ein-Dor, Eyal Shnarch, and Ariel Gera. 2024. Label-efficient model selection for text generation. *arXiv preprint arXiv:2402.07891*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. 2023. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In *Proceedings of the Big Picture Workshop*, pages 108–122, Singapore. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2023a. Which prompts make the difference? data prioritization for efficient human llm evaluation. *arXiv preprint arXiv:2310.14424*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023b. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proc. of ICML*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong

Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6209, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

LMSys. 2023. Vicuna: A cloud-native computing service for machine learning workflows.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online. Association for Computational Linguistics.

Roy Schwartz and Gabriel Stanovsky. 2022. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding. *arXiv preprint arXiv:2211.07634*.

Polina Tsvilodub, Hening Wang, Sharon Grosch, and Michael Franke. 2024. Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. *arXiv preprint arXiv:2403.00998*.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. 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 1141–1158, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Experimental Settings and Task Instructions

We compare three model pairs: `GPT-3.5` vs. `Vicuna-7B` (LMSys, 2023), `Vicuna-7B` vs. `Mistral-7B` (Jiang et al., 2023), and `GPT-3.5` vs. `FLAN-T5-XXL` (Longpre et al., 2023).[10] We experiment with a set of instruction-tuned LLMs that can be zero-shot prompted for the chosen tasks. We use identical instructions for each model, with different model-specific system prompts that were used to fine-tune each model during instruction tuning. We prompt each model in a zero-shot manner. These instruction prompts are listed in Appendix A. For our experiments, we use temperature sampling with temperature $\tau = 0.5$. To calculate alignment scores, we use $K = 5$ samples and $C = 25$ cross-alignment comparisons, unless mentioned otherwise. In Table 1, we present the instruction prompts we used for each dataset described in §2.3.

For auto-evaluation experiments, we use: `alpaca_eval_gpt4`, `alpaca_eval_cot_gpt4_turbo_fn`, and `alpaca_eval_llama3_70b_fn` from AlpacaEval (Dubois et al., 2024).

Similarity Functions As our default similarity function s, we use a length-adjusted version of BERTScore (Zhang et al., 2019), using the same length penalty used in BLEU (Papineni et al., 2002).[11] In the case of translation, where more fine-grained lexical differences are important, we use BLEU itself. Appendix E reports results with additional similarity functions: ROUGE-1 F1 (Lin, 2004), BLEU (Papineni et al., 2002), Entity Similarity, and Cosine Similarity of DistilRoBERTa sentence embeddings.[12] Due to the large variance in the range of each of these functions, we apply min-max normalization over the alignment values to constrain them to the range $[0, 1]$.

Some of these metrics (e.g. BLEU) were designed to compare a "candidate" generation to a "reference," we do not make this distinction since we do not use any reference generations. Instead, we arbitrarily choose the longer generation as the reference. We use F1-score variants of these metrics rather than recall or precision-oriented variants. While prior work (Gehrmann et al., 2022) shows that some of these similarity metrics are not optimal for reference-based *evaluation*, we use these as textual *similarity* functions.

B Human Study Details

We hired a pool of 30 raters (workers) from Amazon Mechanical Turk, all of whom were native English speakers. Each rater was hired based on participation in a qualification study. The raters were paid at a rate of $1.20 per HIT, which was equal to roughly $18 per hour using internally calculated time estimates for a single HIT. The order in which models' summaries were shown in each pair was randomized in order to prevent positional bias. We present the HIT interface shown to AWS MTurk workers in Figure 10.

C Qualitative Examples

We present examples of generations corresponding to higher and lower SEPARABILITY instances for the benchmarks used in our experiments in Tables 2 to 7.

D Preference Strength

In addition to consistency (Equation 4), we define define another way to determine how much a rating set shows preference towards one model or another. We call this metric **preference strength**.

[10] Available via HuggingFace ModelHub: `lmsys/vicuna-7b-v1.5`, `mistralai/Mistral-7B-Instruct-v0.2`, and `google/flan-t5-xxl`.

[11] $\mathsf{LP} = \min\left(1, \exp\left(1 - \frac{|\mathbf{y}_A^i|}{|\mathbf{y}_B^i|}\right)\right)$ if $\mathbf{y}_A^i$ is longer than $\mathbf{y}_B^i$

[12] Using `sentence-transformers/all-distilroberta-v1` on HuggingFace ModelHub.

Dataset	Instruction
CNN/DailyMail (Nallapati et al., 2016)	Summarize the following article in 3-4 sentences.
SAMSum (Gliwa et al., 2019)	Summarize the following dialogue in 1-2 sentences.
WMT-19 (Barrault et al., 2019)	Translate the following {Czech, German} sentence into English.
ART (Bhagavatula et al., 2019)	Write a hypothesis that explains the following observations.
BiSECT (Kim et al., 2021)	Write a simplification of the following sentence.

Table 1: Prompt instructions for each benchmark used in §2.3

We recycle the notation from Equation 4 and define **preference strength** of a rating set $\mathcal{R}_a(\mathbf{x}^i)$ as:

$$\textsf{pref-strength}(\mathcal{R}_a(\mathbf{x}^i)) = \frac{\sum_{j=1}^{N} r_a\big((\mathbf{y}_A^{i,j}, \mathbf{y}_B^{i,j})\big)}{N} \tag{7}$$

Intuitively, preference strength is simply the mean of the rating set. Preference strength of $-1, 1$ reflects a rating set where all the ratings were towards model A, B respectively.

We present the proportion of instances with each possible preference strength per SEPARABILITY range in Figures 11 and Figures 12.

E Using other similarity metrics

In Appendix A, we describe how different similarity functions can be used for different tasks, as well as to measure different dimensions of SEPARABILITY.

We present SEPARABILITY distributions in Figures 13 to 16 for ROUGE-1 (Lin, 2004), the original BERTScore (Zhang et al., 2019), entity similarity[13], and cosine similarity[14], and text embedding cosine similarity on the summarization benchmarks used in our experiments.

[13]We calculate entity similarity between two generations by using `spacy` to extract named entities and taking the Jaccard Similarity of the set of entities from each generation

[14]Using `sentence-transformers/all-distilroberta-v1` on HuggingFace ModelHub as our text embedding model.

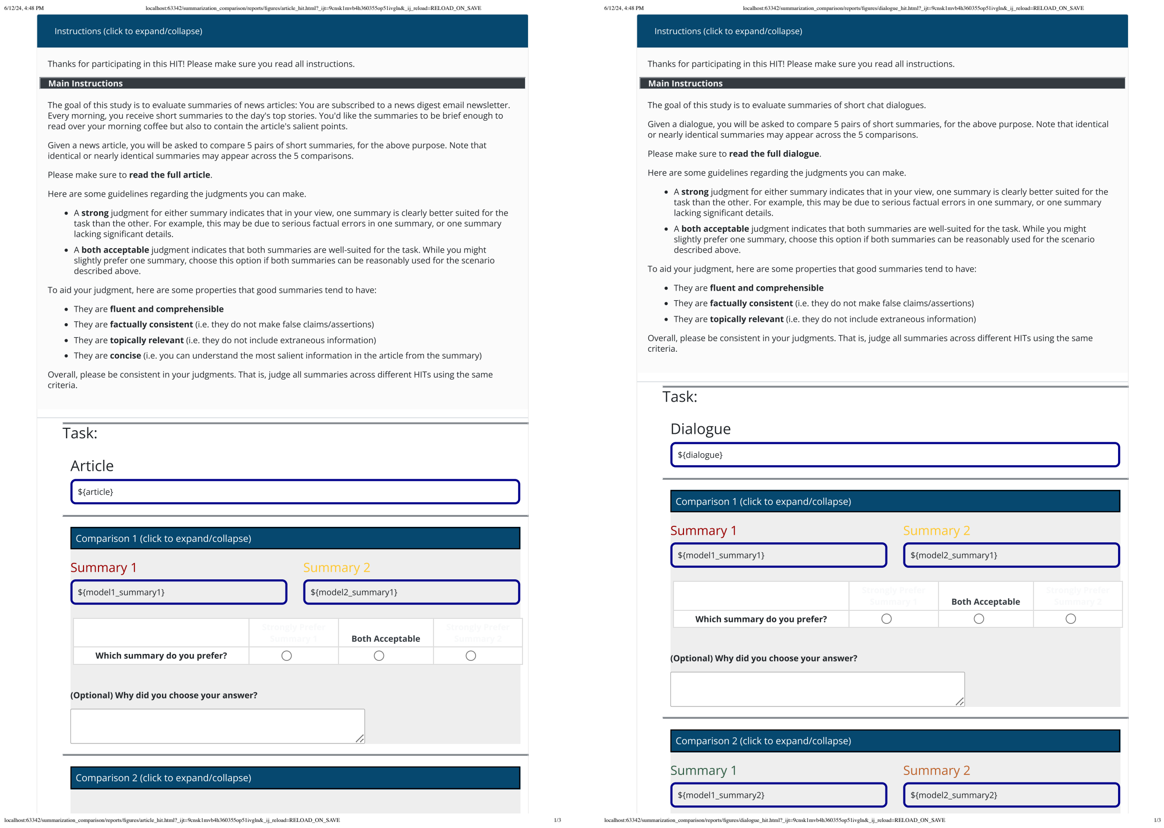

Figure 10: MTurk HIT Interface for news article (Left) and dialogue (Right) summary evaluation used in §3

.

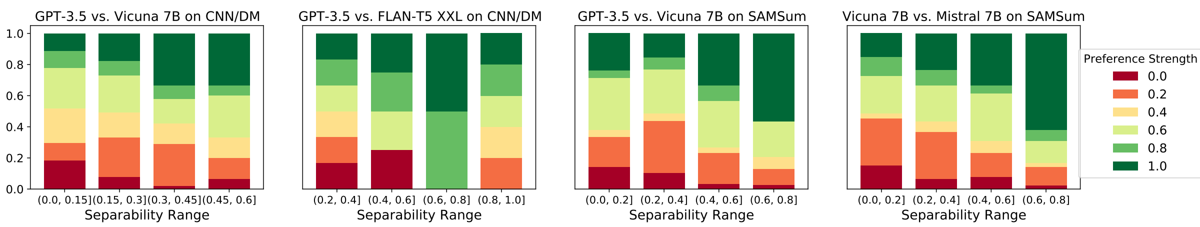

Figure 11: Proportion of instances with each possible preference strength value in a SEPARABILITY range, with human raters

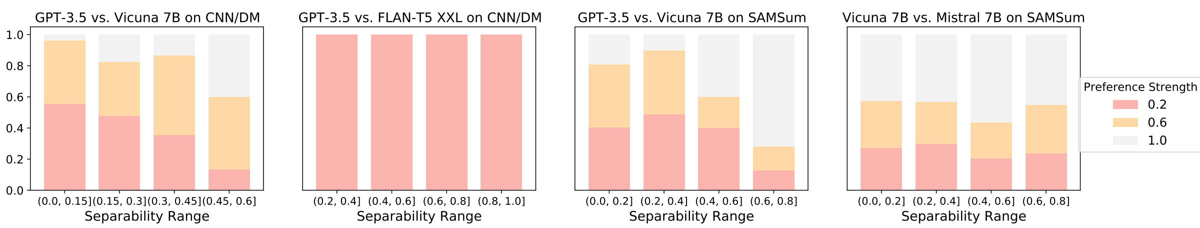

Figure 12: Proportion of instances with each possible preference strength value in a SEPARABILITY range, with auto-raters.

	Lower Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
CNN/DailyMail	1. A Palestinian teenager's name was recently added to an Israeli memorial commemorating fallen soldiers and victims of terrorism, causing controversy and complaints from his family and an organization representing victims of terror. The teen's name, Mohammed Abu Khdeir, was ultimately removed from the memorial after his father's objections and a demand from the organization. The three suspects in Abu Khdeir's killing are currently on trial. 2. The name of a Palestinian teenager, Mohammed Abu Khdeir, was included on an Israeli memorial for fallen soldiers and victims of terrorism. His family and an organization representing victims of terror objected to his inclusion, and his name will now be removed from the memorial. The teenager was beaten and burned alive by three Israelis in July, and the suspects are currently on trial. 3. The name of a Palestinian teenager, Mohammed Abu Khdeir, was included on an Israeli memorial wall for fallen soldiers and victims of terrorism. His family and an organization representing victims of terror objected to his inclusion and demanded his name be removed. The National Insurance Institute of Israel has agreed to remove his name from the memorial. The teenager was beaten and burned alive by three Israelis in July and the suspects are currently going through a trial.	A Palestinian teenager's name was added to an Israeli memorial for fallen soldiers and terrorism victims, but his family and a victim advocacy group objected, leading to the removal of his name. The teenager, Mohammed Abu Khdeir, was killed by three Israelis in July, and his family did not give permission for his name to be included. The National Insurance Institute of Israel will remove his name from the memorial following the complaints. A Palestinian teenager's name was added to an Israeli memorial for fallen soldiers and terrorism victims, but his family and a victim advocacy group objected. The father of the teenager, who was killed in a brutal way, refused permission for his son's name to be listed among soldiers of the occupation. The National Insurance Institute of Israel will remove the teenager's name from the memorial following the complaints. The teenager's name has already been removed from the organization's website. A Palestinian teenager's name was added to an Israeli memorial for fallen soldiers and terrorism victims, but his family and a victim advocacy group objected. The father of the teenager, who was killed in a brutal way, said his son's name should not be listed among soldiers of the occupation. The National Insurance Institute of Israel, which maintains the memorial site, will remove the teenager's name following the complaints. The judicial hearing for the suspects in the teenager's killing began in January.
SAMSum	1. Ursula shares her success of getting a 93 on her French exam, while her friends congratulate her and discuss their own experiences with the exam. 2. Ursula shares her achievement of getting a 93 on her French exam, and her friends congratulate her while discussing their own experiences with the exam. 3. Ursula shares her success of getting a 93 on her French exam, while Vaughn admits to getting a 65 without studying. Jana and Bob congratulate Ursula while Vaughn jokes about his own score.	1. Ursula got a 93 on her French exam, while Vaughn got a 65. Jana is impressed with Ursula's score and asks how she managed it, to which Ursula replies that she studied hard. Vaughn admits that he didn't study and only got a passing grade. 2. Ursula got a 93 on her French exam, Bob congratulated her, Jana was impressed, and Vaughn got a 65 but didn't study for it. 3. Ursula got a 93 on her French exam, Bob congratulated her, Jana was impressed, and Vaughn got a 65 but didn't study for it.

Table 2: Examples of lower SEPARABILITY instances for summarization benchmarks used in our experiments

	Higher Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
CNN/DailyMail	1. Ciudad Juarez, Mexico, which was once known as the murder capital of the world with an average of 8.5 killings per day in 2010, is now making efforts to attract foreign tourists and investors back to the city. The city launched the tourism campaign "Juarez is Waiting for You" this month. Violence in the city has decreased significantly, with only 89 killings so far this year compared to over 3,000 in 2014. 2. Ciudad Juarez, Mexico, which was once known as the murder capital of the world, is now making efforts to lure tourists and investors back with a new tourism campaign called "Juarez is Waiting for You." In 2010, the city averaged 8.5 killings per day, but local officials claim that it is much safer now. The violent past, which included a turf war between cartels and high levels of corruption within the local police force, has left deep scars. 3. Ciudad Juarez, Mexico, which was once known as the murder capital of the world with an average of 8.5 killings per day in 2010, is now making efforts to lure foreign tourists and investors back. The city launched the tourism campaign "Juarez is Waiting for You" this month, and local officials claim that the city is much safer than it was five years ago. In 2014, there were 424 homicides, but this year, there have been 89 killings, a significant decrease in violence.	1. Ciudad Juarez, Mexico, was once the murder capital of the world, with an average of 8.5 killings per day in 2010. However, local officials say the city is much safer now, and plans are underway to lure foreign tourists and investors back to Juarez. 2. Ciudad Juarez, Mexico, was once the murder capital of the world, with an average of 8.5 killings per day in 2010. However, local officials say the city is much safer now, and plans are underway to lure foreign tourists and investors back to Juarez. 3. Ciudad Juarez, Mexico, was once the murder capital of the world, with an average of 8.5 killings per day in 2010. However, local officials say the city is much safer now, and plans are underway to lure foreign tourists and investors back to Juarez.
SAMSum	1. Tom agrees to pick up Susan's daughter Becky from school because she has a fever and Susan cannot leave the office. Tom also needs to return to the office soon, so he will take Becky to his mother's house. 2. Tom agrees to pick up Susan's daughter Becky from school because she has a fever and Susan cannot leave the office. Tom also needs to return to the office to finish a project. 3. Tom will pick up Susan's daughter Becky from school because she has a fever and Susan cannot leave the office. Tom also needs to return to the office to finish a project.	1. Susan's day was disappointing due to a failed presentation, and she asked Tom to pick up their sick daughter Becky from school since she couldn't leave work early. Tom agreed and planned to take Becky to his mother's house. Susan expressed her frustration over a disappointing day, including a failed presentation, and requested Tom's help in picking up their sick daughter Becky from school due to her inability to leave work early. Tom agreed and shared his own work pressure, intending to take Becky to his mother's house. 2. Susan shares that her day was disappointing due to a failed presentation, and reveals that her daughter Becky is sick at school, requiring Tom to pick her up and take her to his mother's house, as Susan is unable to leave work early and doesn't have her car. Tom mentions the pressure he's under to complete a project on time. 3. Susan's day was disappointing due to a failed presentation, and she asked Tom to pick up their sick daughter Becky from school as she couldn't leave work early and didn't have her car. Tom agreed, mentioning his own work pressure. Susan had a frustrating day with her presentation going wrong, and she asked Tom to pick up their sick daughter Becky from school as she couldn't leave work early and didn't have her car; Tom agreed, mentioning his own work pressure and offering to take Becky to his mother's house.

Table 3: Examples of higher SEPARABILITY instances for summarization benchmarks used in our experiments

	Lower Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
ART	1. Regular practice and exposure to rock climbing with friends can improve an individual's climbing ability and confidence, leading to better performance on a real rock wall. 2. My hypothesis is that the practice and experience gained from indoor rock climbing with friends improved my skills and confidence, leading to better performance on a real rock wall the following day. 3. Participating in indoor rock climbing with friends can improve an individual's climbing abilities and confidence, leading to better performance on a real outdoor rock wall.	1. The experience of rock climbing with friends and the subsequent visit to the real rock wall helped me develop a better understanding of the techniques and skills required for rock climbing, which improved my performance on the real rock wall. 2. The experience of rock climbing with friends, even if it was just pretend, helped improve my skills and confidence in real rock climbing. 3. The experience of rock climbing with friends, even if it was just pretend, helped me develop the skills and confidence needed to successfully climb a real rock wall.
BiSECT	1. Stiripentol is absorbed quickly, reaching its peak concentration in about 1.5 hours. 2. Stiripentol is absorbed quickly, reaching its peak concentration in about 1.5 hours. 3. Stiripentol is rapidly absorbed, reaching its peak plasma concentration in 1.5 hours.	1. Stiripentol is rapidly absorbed, reaching its peak concentration in approximately 1.5 hours. 2. Stiripentol is rapidly absorbed, reaching its peak concentration in approximately 1.5 hours. 3. Stiripentol is rapidly absorbed, reaching its peak concentration in approximately 1.5 hours.

Table 4: Examples of lower SEPARABILITY instances for ART and BiSECT

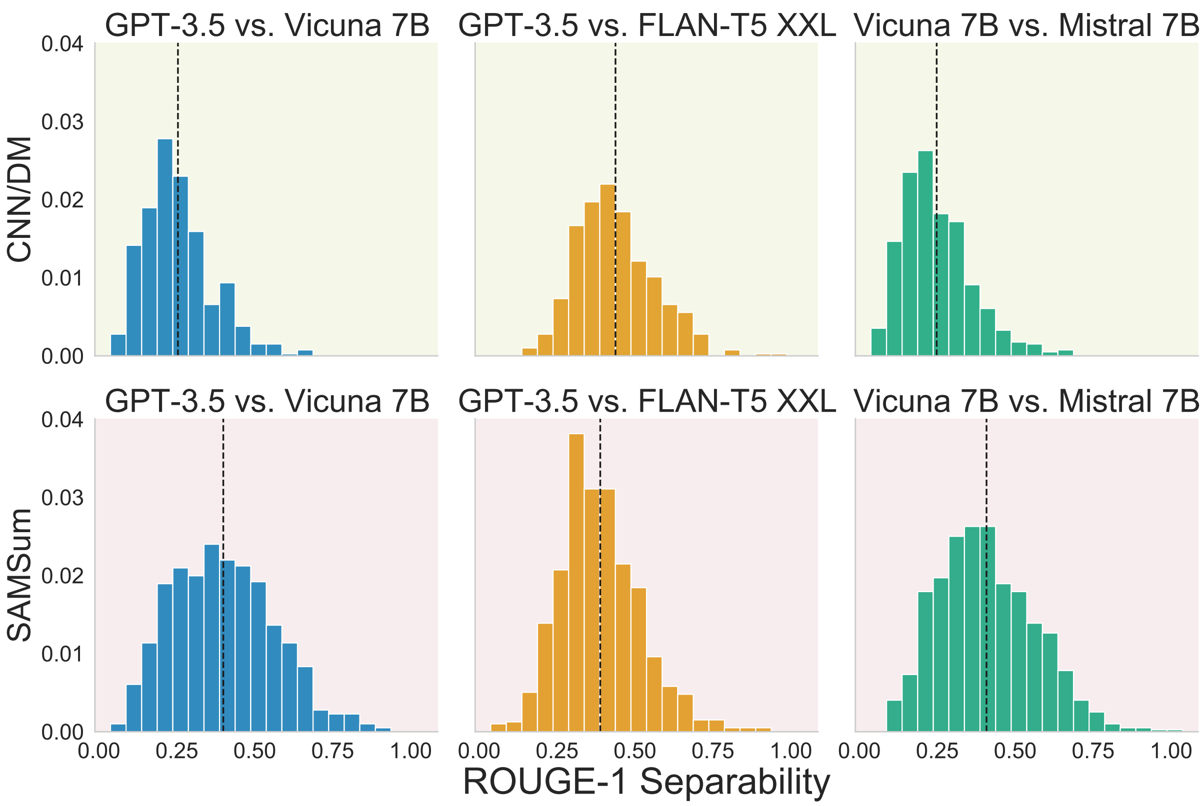

Figure 13: SEPARABILITY distributions using ROUGE-1 F1 as a similarity metric.

	Higher Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
ART	1. Joey's desire to color in coloring books and his wish for his parents to appreciate him may be related to his need for attention and validation. 2. Joey's desire to color in coloring books and his wish for his parents to appreciate him may be related to his need for attention and validation. 3. Joey's desire to color in coloring books and his wish for his parents to appreciate him may be related to his need for attention and validation.	1. Hypothesis: Joey's engagement in coloring activities in his coloring books could be an expression of his innate need for creativity and self-expression, which may serve as a coping mechanism to deal with his unmet emotional need for appreciation from his parents. 2. Explanation: Joey's enjoyment of coloring might provide him with a sense of accomplishment, relaxation, and control, which can help him feel good about himself. 2. Hypothesis: Joey's engagement in coloring as a second-grader could be an expression of his desire for approval and appreciation from his parents. The act of coloring may provide him with a sense of accomplishment, creativity, and relaxation, which in turn, could be a way for him to express his emotions and seek validation from his caregivers. However, the fact that he expressed a wish for his parents to appreciate him more suggests that he may not feel fully valued or recognized for his efforts and talents outside of this activity. 3. Hypothesis: Joey's engagement in coloring activities in his coloring books could be an expression of his desire for recognition and appreciation from his parents. This hypothesis is based on the observation that Joey, as a second-grader, enjoys coloring, but also expressed a wish for his parents to value and acknowledge him. It's possible that the act of coloring provides Joey with a sense of accomplishment and pride, and he may hope that his parents will notice and praise him for his efforts.
BiSECT	1. The rise in cheap imports harms Community producers by putting downward pressure on prices and decreasing their sales. 2. The rise in cheap imports harms Community producers by putting pressure on prices and decreasing their sales volume in the market. 3. The increase in low-priced imports is harming Community producers by putting pressure on prices and decreasing their sales volume in the Community market.	1. There is a correlation between the rise in imports at low prices and the harm suffered by European producers. The increase in imports has had negative effects on prices and sales volume for European producers in the European market. 2. There is a connection between the rise in imports at low prices and the harm suffered by European producers. This increase in imports has negatively impacted prices and reduced the volume sold by European producers in the European market. 3. There is a connection between the rise in low-priced imports and the harm suffered by European producers. This increase in imports has caused negative effects, such as pushing down prices and reducing the amount of goods sold by European producers in their home market.

Table 5: Examples of higher SEPARABILITY instances for ART and BiSECT

	Lower Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
WMT-19 DE to EN	1. If the offer arrives, we will expand it. 2. If the offer arrives, we will expand it. 3. If the offer arrives, we will expand it.	1. If the offer comes, we will expand it. 2. If the offer comes, we'll expand it. 3. If the offer comes, we'll expand it.
WMT19 CS to EN	1. Papuans rarely have the necessary capital and education to start a business. 2. Papuans rarely have the necessary capital and education to start a business. 3. Papuans rarely have the necessary capital and education to start a business.	1. Papuans rarely have the necessary capital and education to start a business. 2. Papuans rarely have the necessary capital and education to start a business. 3. Papuans rarely have the necessary capital and education to start a business.

Table 6: Examples of lower SEPARABILITY instances for translation benchmarks used in our experiments

	Higher Separability Instance Outputs	
Dataset	**Model A Generations**	**Model B Generations**
WMT-19 DE to EN	1. In 2013, a total of 25 plots were fenced in by eleven participants through the pasture fence promotion. 2. In 2013, a total of 25 plots were fenced in by eleven participants through the pasture fence promotion. 3. In 2013, a total of 25 plots were fenced in by eleven participants through the pasture fence promotion.	1. 11 participants enclosed 25 parcels with fencing in 2013. 2. 11 participants enclosed 25 parcels with fencing in 2013. 3. 11 participants enclosed 25 parcels through the pasture fencing promotion in the year 2013.
WMT19 CS to EN	1. They shoot less, and even when they have a half-empty goal in front of them, they prefer to pass. 2. They shoot less, and even when they have a half-empty goal in front of them, they prefer to pass. 3. They shoot less, and even when they have a half-empty goal in front of them, they prefer to pass.	1. I prefer to shoot when there is an empty space in front of me, even if it means taking a risk. 2. I prefer to shoot when there is an empty space in front of me, even if it means taking a risk. 3. I prefer to shoot when there is an empty space in front of me, rather than when I have an empty space behind me.

Table 7: Examples of higher SEPARABILITY instances for translation benchmarks used in our experiments

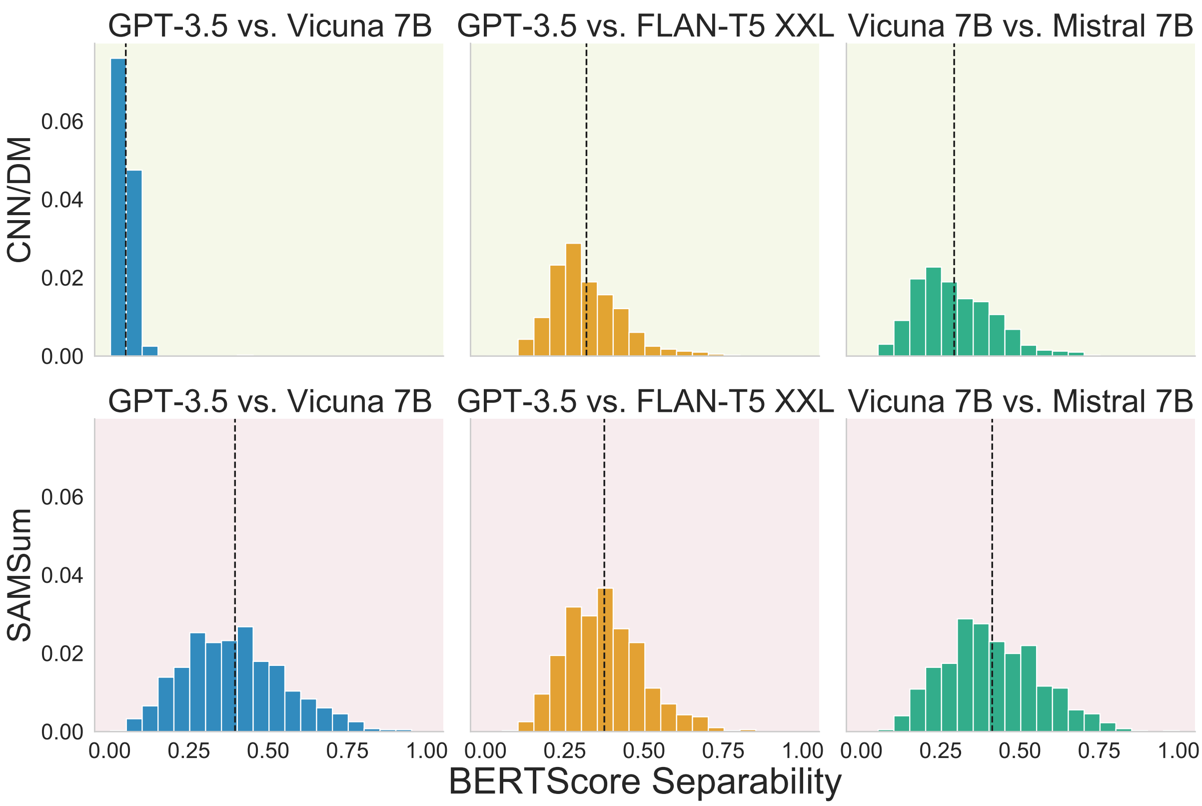

Figure 14: SEPARABILITY distributions using vanilla BERTScore as a similarity metric.

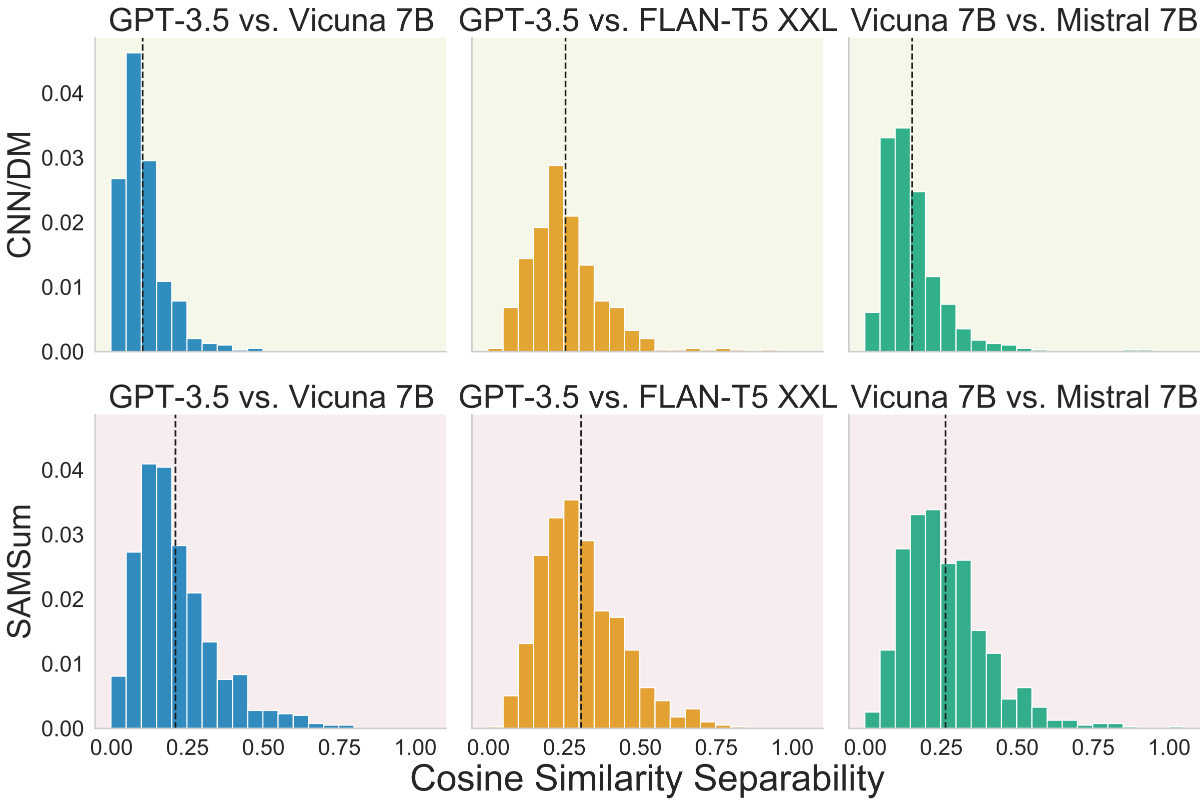

Figure 15: SEPARABILITY distributions using cosine similarity as a similarity metric.

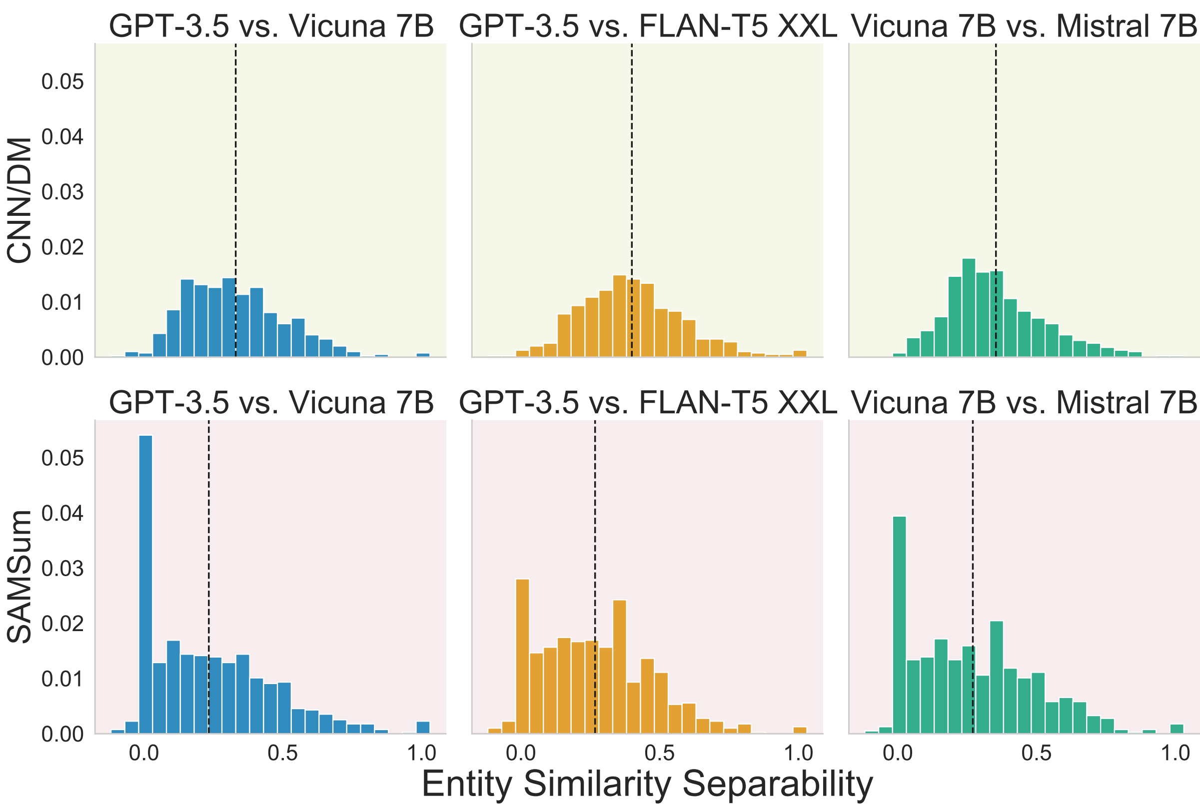

Figure 16: SEPARABILITY distributions using entity similarity as a similarity metric.