1 Scientific Background

The advent of Large Language Models (LLMs) has dramatically transformed the landscape of natural language processing (NLP), providing unprecedented abilities in understanding, generating, and evaluating human-like text [40, 4, 25]. These models are now increasingly being employed not just as tools for text generation but also as evaluators of models (a.k.a., judges) or annotators of data in a variety of NLP tasks [10, 9]. This new utilization aims to leverage the special advantages of AI models to improve data annotation processes (the process of giving labels to data instances) and model evaluation, which are crucial for training and refining machine learning models.

Given the increasing use of LLMs in annotation and evaluation tasks, it is essential to rigorously and thoroughly assess their performance in these roles. This research proposes a novel methodology for evaluating LLMs as annotators, complemented by statistical tests that measure how closely LLM outputs align with human annotations. Our goal is to determine to what extent LLMs can reliably replace human annotators in various tasks, whether human annotators are not expected to be experts or in tasks where nuanced understanding and expert judgment are crucial.

In this research, we will introduce a novel methodology for calculating the quality score of an LLM as an annotator. Our approach is based on the hypothesis that an LLM should not significantly alter the distribution of annotations compared to the extent to which a human annotator changes a given distribution of annotations. Rather than measuring the agreement between LLM and an average human label that is being calculated from a collection of human annotations, we evaluate the impact of replacing each human annotator with an LLM. This provides a more reliable measure of the LLM's performance, as it mimics how the LLM would function as a substitute for a human annotator.

For this research, we assume that the evaluation process of a judge is analogous to that of an annotator, even though the action of annotation typically involves producing a correct output for a data instance, and a judgment action involves assigning scores or comparing given outputs, which is a less laborious task.

Figure 1 illustrates an example comparison between a distribution of labels from a partial set of human annotators (orange bars), an LLM (blue bars), and an extracted human annotator (green bars) across three categories. Each subplot represents a different human annotator being extracted, allowing us to observe how closely the LLM's annotations align with the collective distribution of human annotation compared to the extracted human annotator's annotations. The extracted human annotators (green bars) sometimes have distributions that differ more drastically from the remaining human annotators compared to the LLM. In such cases, the LLM might be performing as well as or better than the individual human annotator in representing the consensus. For example, when Annotator 2 is extracted, the LLM's distribution is closer to the rest of the annotators than Annotator 2's distribution.

When an LLM can effectively act as an annotator, we expect that substituting a human annotator with an LLM will lead to only minor differences in the labeling distributions. For example, we anticipate that the level of disagreement between Annotator 3 and the partial set of annotators $\{1, 2, 4\}$ will closely resemble the level of disagreement between the LLM and the partial set of annotators $\{1, 2, 4\}$. In the scenario depicted in Figure 1, we will perform the following comparisons (where Δ denotes some measure of disagreement):

1. $\Delta(LLM, \{1, 2, 3\})$ versus $\Delta(4, \{1, 2, 3\})$
2. $\Delta(LLM, \{1, 2, 4\})$ versus $\Delta(3, \{1, 2, 4\})$

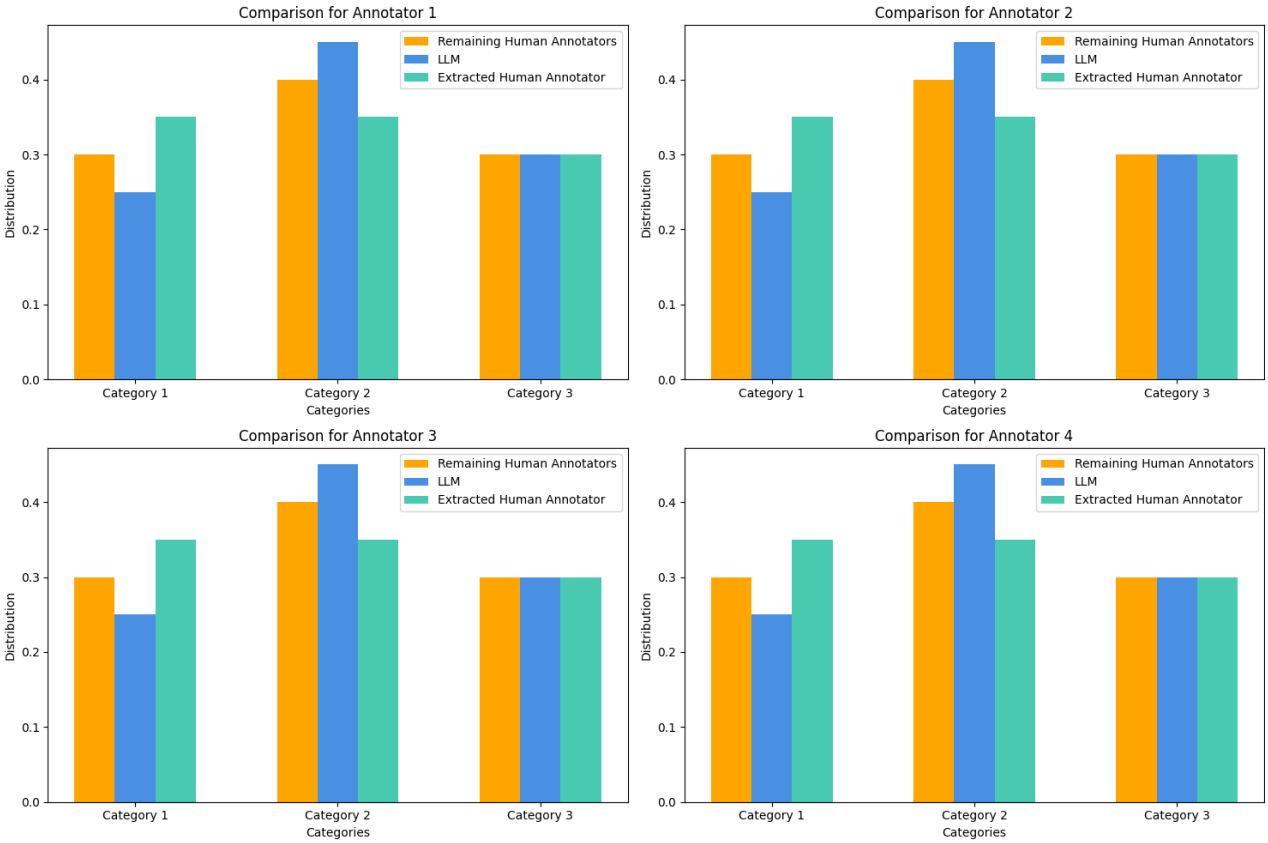

Figure 1: An example comparison of label distributions across three categories for four different annotators. Orange bars show the distribution of the remaining three human annotators, blue bars represent the LLM's annotation distribution, and green bars show the distribution from an extracted human annotator. Each subplot highlights a different human annotator, illustrating how closely the LLM aligns with the human consensus.

3. $\Delta(LLM, \{1, 3, 4\})$ versus $\Delta(2, \{1, 3, 4\})$
4. $\Delta(LLM, \{2, 3, 4\})$ versus $\Delta(1, \{2, 3, 4\})$

These comparisons allow us to quantify the discrepancy in disagreements and derive a score reflecting the quality of the LLM as an annotator. If a significant difference is observed, then using the LLM as an annotator or a judge may not be advisable, as its annotations do not align with "good" annotations (under the assumption that human annotations are generally desirable and of high quality).

1.1 Annotation Paradigms in NLP

Different annotation paradigms have been developed in NLP to address the challenges of data labeling, particularly in tasks that require subjective judgments or involve complex linguistic phenomena. These paradigms vary in their approaches to quality control, scalability, and the degree of human involvement. An important aspect is the method used to aggregate annotations from multiple annotators, which can significantly impact the final quality of the annotations.

Traditional Expert Annotation involves a small group of highly trained annotators who meticulously label data according to predefined guidelines [2, 20]. This approach is typically used in tasks that require deep linguistic knowledge, such as syntactic parsing or semantic role labeling. The primary advantage of expert annotation is the high quality and consistency of the labels, as experts

are more likely to understand the nuances of the task and apply the guidelines accurately. However, this paradigm is costly and time-consuming, limiting its scalability. Moreover, even experts are not immune to biases and subjective interpretations, which can introduce variability into the annotations.

Crowdsourcing has emerged as a popular alternative to traditional expert annotation, particularly for tasks that require large volumes of labeled data [34, 6, 29]. In this paradigm, annotations are collected from a large pool of non-expert annotators, often through platforms like Amazon Mechanical Turk (`https://www.mturk.com/`) or Prolific (`https://www.prolific.com/`). Crowdsourcing offers advantages in terms of speed and cost, enabling the rapid collection of diverse annotations. However, the quality of the annotations can vary significantly, as non-experts may misinterpret the task or lack the necessary background knowledge. To address these challenges, various methods are used to aggregate annotations from multiple annotators:

- **Majority Vote:** The most straightforward method, majority voting, involves selecting the label that the majority of annotators agree upon. This method is simple and effective for tasks with clear-cut answers, but it may not capture the nuances of more subjective tasks.
- **Average:** For tasks where annotations are numerical or ordinal (such as rating scales), the average of the annotators' scores can be used to obtain a *vox populi* label (wisdom of the crowd). This helps smoothing out extreme values and providing a balanced view of the collective judgment.
- **Weighted Average:** In some cases, annotators are assigned different weights based on their expertise, reliability, or consistency, giving more influence to higher-quality annotators.
- **Complete Agreement:** This method requires that all annotators agree on a label for it to be accepted. While this ensures high confidence in the label, it can also lead to a significant reduction in the amount of usable data, especially in tasks where disagreement is common.

Collaborative Annotation is a paradigm in which multiple human annotators work together, either simultaneously or sequentially, to label a dataset. Collaborative annotation can improve the reliability of annotations by incorporating diverse perspectives and reducing individual biases. Aggregation methods in collaborative annotation often include:

- **Consensus Through Discussion:** Annotators discuss and debate difficult cases to reach a final consensus label, ensuring that all perspectives are considered.
- **Sequential Refinement:** Annotations are refined over multiple rounds, with each annotator reviewing and adjusting previous labels. The final label is determined by the last annotator in the sequence or through a weighted voting system that considers the changes made by each annotator.

Comparative Annotation is a method in which annotators are not asked to label an instance directly but are instead asked to compare a set of different possible labels and decide which one is better or more appropriate for a given instance. This paradigm is particularly useful in tasks where choosing a single label might be challenging due to the subjective nature of the content or the task complexity (tasks such as summarization). In comparative annotation, the final label is determined based on the total rank of the given labels by the annotators. This approach can reduce the cognitive

load on annotators, as comparing options is often easier than generating a label. However, this paradigm requires the generation of possible labels in advance.

Machine-Assisted Annotation represents a hybrid approach, where initial labels are generated by a machine learning model and then reviewed or refined by human annotators [33]. This paradigm leverages the strengths of both machines and humans: the efficiency and consistency of automated labeling and the nuanced judgment of human annotators. Machine-assisted annotation can significantly reduce the time and effort required for data labeling, particularly in tasks where the model can achieve high accuracy. However, it also presents challenges, such as the potential for humans to over-rely on the machine-generated labels, leading to confirmation bias. Ensuring that humans critically review the labels rather than passively accept them is crucial to maintaining annotation quality. Aggregation methods in machine-assisted annotation often involve:

- **Model-Annotator Agreement:** Labels are accepted if both the model and the annotator agree. Discrepancies may be flagged for further review.
- **Human-in-the-Loop Adjustment:** In cases where the model's suggestions are off-target, human annotators can adjust the labels, and these adjustments can be used to further train and refine the label-generating model, creating a feedback loop that improves performance over time.

Interactive Annotation involves a dynamic interaction between the annotator and the automatic annotation system, where the system actively learns from the annotator's inputs and adapts its suggestions accordingly [31, 35]. This paradigm is particularly useful in tasks that are complex or ambiguous, as it allows the system to iteratively refine its understanding based on real-time feedback. For example, in active learning setups [38, 23], the system might query the annotator for the most uncertain instances, thereby improving its performance on difficult cases. A key distinction between this approach and the earlier machine-assisted methodology is that in the former, the human annotator sets the final labels, whereas in this scenario, the model establishes the labels with iterative improvements from human feedback.

The choice of annotation paradigm and aggregation method depends on the specific requirements of the NLP task, including the need for accuracy, the desired level of agreement between annotators, the size of the data, and the available resources. Each method offers unique advantages and challenges, and often, a combination of approaches is used to balance quality with efficiency. This research aims to explore and evaluate a new annotation paradigm that emerged with the advent of LLMs, which is **LLM as a judge**.

1.2 LLMs for Data Labeling

LLMs, such as GPT-4o by OpenAI [25], Claude 3.5 Sonnet by Anthropic (`https://www.anthropic.com`), Llama 3 by Meta [39], and others, are built upon deep learning models, primarily leveraging Transformer architectures [40] that have demonstrated exceptional proficiency in capturing the nuances of human language. Trained on vast corpora encompassing diverse textual data, these models can perform a myriad of tasks, including but not limited to text generation, translation, summarization, and question-answering [4].

In the context of data annotation, LLMs can be prompted to generate labels for given data instances or to select the most appropriate label from a set of options. For example, as demonstrated by

Ravid and Dror [27], an LLM was used to annotate data on public opinions on the justification of punishments expressed on Twitter regarding four homicide cases. The LLM was provided with definitions of punishment theories (retribution, utilitarianism, expressiveness, and restoration) and was asked to annotate tweets with the most appropriate approach expressed in that tweet. This annotation approach differs from the traditional human annotation procedure, where individuals read and manually label each instance following detailed annotation guidelines, or from machine-augmented annotation approaches where models are trained for a specific task of annotation. In Ravid and Dror [27], the LLM's annotations were validated for accuracy by expert graduate law students.

LLMs offer numerous advantages when they are utilized as annotators:

1. **Scalability and Speed:** Traditional annotation processes, especially for large-scale datasets, require substantial human resources and time. Depending on the complexity of the task, annotation can be a time-consuming and labor-intensive endeavor. LLMs, on the other hand, can process and annotate vast amounts of data in a fraction of the time.

2. **Cost Efficiency:** LLMs offer a significant cost reduction compared to human annotators. While human annotation requires continuous labor and coordination, LLMs only require computational resources, which are increasingly cost-effective as cloud computing and AI infrastructure become more accessible. This makes LLMs particularly useful for large-scale projects where the budget for human annotators may be prohibitive.

3. **Consistency** Unlike humans, whose judgments may vary due to fatigue, bias, or differing interpretations of guidelines, LLMs provide consistent annotations, adhering strictly to their training and the prompts provided.

Despite their advantages, the use of LLMs as annotators is not without challenges:

1. **Bias and Fairness:** LLMs inherit biases present in their training data. These biases can manifest in every text the LLM generates, including in annotations. Studies have highlighted instances where LLMs exhibit gender, racial, or cultural biases [43, 32, 28].

2. **Interpretability:** Understanding the rationale behind an LLM's annotation can be challenging. Unlike human annotators, who can explain their reasoning, LLMs operate as black boxes, making it difficult to trace the source of a particular annotation decision, and even if the LLM is prompted to explain its decisions, it is not clear whether this explanation is the true reason for generating a certain label [30, 5].

3. **Accuracy:** While LLMs excel in many tasks, their accuracy in complex or nuanced annotation tasks may not match that of human experts, especially in domains requiring specialized knowledge [27].

The utilization of LLMs as annotators in NLP signifies a major stride towards automating and scaling the data labeling process. While offering notable advantages in speed, consistency, and scalability, challenges related to bias, interpretability, and accuracy persist. A balanced approach, integrating LLM capabilities with human expertise and ethical considerations, is essential to harness the full potential of this emerging trend, yet it should be accompanied by a methodology for evaluating the abilities of LLMs in these roles.

1.3 Evaluation of LLM as A Judge

Evaluating the potential of LLMs as reliable judges or annotators, replacing or completing human effort, has embarked in recent studies that followed the proposal and demonstration of LLMs being used as judges [10, 44]. Chiang and Lee [9] investigated the feasibility of LLMs as alternatives to human evaluations. Their work demonstrated that LLMs could reliably substitute humans in certain evaluative tasks, albeit with some limitations, particularly in complex, subjective contexts.

In a related vein, Dong et al. [15] explored the concept of personalized LLM judges. Their research suggested that LLMs could be fine-tuned to reflect individual preferences or judgments, offering a personalized evaluation framework that could enhance user-specific tasks. Building on this, Verga et al. [41] proposed an approach of using a panel of diverse LLM models to evaluate outputs, shifting the paradigm from a single authoritative judge to a jury of models.

The question of bias in LLM evaluations has been a prominent concern in recent literature. Jung et al. [22] focused on ensuring provable guarantees for human agreement when using LLMs as judges. Their work introduced mechanisms to align LLM outputs with human consensus, aiming to mitigate judgment discrepancies due to model biases. Similarly, Chen et al. [8] conducted an analysis on judgment biases in LLMs, comparing human and LLM decision-making processes. They emphasized the need for bias-mitigation strategies to improve fairness and reliability in automated judgments.

The intersection of LLMs and human-centered design was explored by Pan et al. [26]. They provided recommendations for designing LLMs as judges in a way that enhances user trust and interpretability, focusing on the need for transparency in decision-making processes.

Chen et al. [7] extended the evaluation of LLMs into multimodal tasks, assessing how LLMs perform as judges when handling both text and visual data. Finally, Thakur et al. [37] investigated the vulnerabilities in LLM judgments, particularly focusing on alignment issues between LLM-generated outputs and expected human decisions. Their study highlights the risks of over-reliance on LLMs as judges, especially in sensitive or high-stakes tasks.

In all of these studies, these researchers aim to establish a consensus-based benchmark, they assume that the aggregated human judgment is the correct label and compare the LLM generated label to that one. In this research, we claim that this assumption is flawed because it ignores the nuances and potential biases in individual human judgments, which can lead to an inaccurate assessment of the true quality or expertise of the LLM. In this study, we devise an alternative approach for comparing human annotations with LLM annotations that evaluates the LLM's capability to replace each human annotator separately, and then combine these assessments to draw a conclusion about the LLM's overall ability to function as a judge or annotator.

2 Research Objectives and Expected Significance

The overarching goal of this research is to develop a valid and robust evaluation framework that quantifies the capability of LLMs to serve as annotators or judges. The aim is to determine whether LLMs can reliably replace human annotators by preserving the distribution of annotations in complex datasets and tasks. We will focus on three main research objectives:

RO1 **Develop a Protocol for Evaluating the Distribution of Annotations When Replacing Human Annotators with LLMs:** For this objective, we will design and implement a systematic protocol for comparing the distribution of annotations produced by LLMs with those

generated by human annotators. The protocol will focus on understanding how the inclusion of an LLM alters the distribution of labels in various tasks. Rather than relying solely on traditional measures of agreement (e.g., accuracy or inter-annotator agreement), this protocol will examine how well LLMs maintain the overall structure and balance of the label distribution when substituting human annotators. We aim to ensure that LLMs not only match human judgments but also reflect the diversity and complexity inherent in human annotation processes.

RO2 **Implement and Develop Statistical Tests to Compare LLM Annotations to Human Annotators Across Various NLP Tasks:** This objective involves the development and validation of robust statistical methods to measure the alignment between LLM annotations and human annotations. We will design tests that go beyond simple accuracy or majority voting approaches, leveraging advanced metrics to capture more nuanced comparisons between outputs. These tests will account for both individual instance-level agreement and aggregate distribution-level agreement, allowing us to quantify how closely LLMs emulate human decision-making.

RO3 **Rank LLMs and Human Annotators:** We will create a ranking system to assess both LLMs and human annotators based on how well their annotations align with the collective judgments of other human annotators. This ranking system will provide a more granular understanding of LLM performance, enabling us to determine whether LLMs can surpass human-level performance in certain tasks and will constitute a benchmark for choosing an LLM for these applications.

2.1 Expected Significance

This research has the potential to fundamentally transform the evaluation of NLP models. Introducing innovative evaluation frameworks of LLMs in the roles of judges and annotators will provide more accurate assessments of labeled datasets, annotator quality, and model performance, leading to NLP models that better align with human reasoning and judgment. The research also addresses a critical gap in current industry practices, where many companies release LLMs without standardized evaluation procedures. In some cases, companies bypass detailed evaluation by opting for informal releases or preprints on platforms like arXiv or blogs. This lack of standardization poses significant risks, as it can lead to the deployment of AI products that have not been thoroughly vetted for safety, fairness, or generalizability.

A key innovation of this work is the development of a standardized protocol for dataset collection and model evaluation, ensuring that AI products are reliable, safe, and ethically sound. This research goes beyond determining which model is state-of-the-art; it seeks to verify that models operate safely and without bias, exceeding expectations in positive ways rather than introducing harmful outcomes. By setting higher standards for evaluation, this research will contribute to the creation of better and safer AI products that align with human values with greater fidelity and fairness. The insights gained will also clarify the role of LLMs as reliable substitute or complementary to human annotators, advancing the field of NLP in both technical and ethical dimensions.

3 Detailed Description of the Proposed Research

We aim to rigorously evaluate the effectiveness of LLMs as annotators, i.e., labeling data instances, and judges, i.e., evaluating outputs of models, by comparing their performance to that of human

annotators and judges. We will focus on quantifying how closely LLMs align with human annotators and whether they can replace or complement humans in NLP tasks.

3.1 Methodology

Our approach compares the label distribution generated by human annotators with the distribution produced when an LLM replaces a human annotator. Instead of directly substituting a human annotator with an LLM, we evaluate the LLM's output against the collective distribution of the remaining human annotators. We evaluate the LLM's alignment with this partial set of human annotators and compare it to the alignment of the excluded human annotator with the same group. Both LLM and human annotators are ranked based on these alignment scores, allowing us to determine whether the LLM matches or exceeds human performance. We argue that a good LLM judge can effectively emulate the actions and decisions of a human annotator, thereby making it a suitable replacement in the annotation process.

Notations and Definitions For a dataset of n instances $\{x_1, \ldots, x_n\}$, we denote the annotations of m human annotators as $h_j(x_i)$, where $j = 1, \ldots, m$ is the index of the human annotator and $i = 1, \ldots, n$ is the index of the data instance. The annotations of the LLM will be denoted as $f(x_i)$. In addition, we denote by $[-j]$ the set of indices from 1 to m without the jth index, i.e., $[-j] = \{1, \ldots, j-1, j+1, \ldots, m\}$. The set of indices of the instances annotated by the human annotator j is denoted by $\mathbb{I}_j$. Similarly, $\mathbb{H}_i$ is the set of indices of human annotators that annotated instance x_i. For example, if we have 3 instances in our data and 4 human annotators, and each human annotator annotated two instances, then $\mathbb{I}_2 = \{2, 3\}$ means that the second annotator annotated instances 2 and 3, and $\mathbb{H}_1 = \{1, 3, 4\}$ means that the first instance was annotated by the first, third, and fourth annotators.

To calculate the quality score of LLM as a judge, we start by examining the removal of each human annotator, j, in turn, and compute a score that measures the alignment between the annotations of the $[-j]$ human annotators and the annotation of the LLM for instance x_i. We use $S(f, x_i, j)$ to denote the alignment scoring metric between the LLM annotation, $f(x_i)$, and the annotations of the human annotators of $h(x_i)$, excluding annotator j. For example, S could be `RMSE` (root mean squared error) in regression tasks (continuous numerical labels) or `ACC` (accuracy) in classification tasks (categorical or rank labels). In generation tasks (e.g., machine translation), S can be computed using a relevant evaluation metric that typically measures the similarity between the LLM-generated output and the human-generated output, such as BERTScore [42]. In this case, we denote the evaluation metric by `sim` and S by `SIM`. For convenience, higher values of S indicate a better alignment between an LLM and the human annotator; thus, we use negative `RMSE`. Below, we formally define the mentioned variants of S:

$$-\texttt{RMSE}(f, x_i, j) = -\sqrt{\frac{1}{|\mathbb{H}_i| - 1} \sum_{k \in \mathbb{H}_i[-j]} (f(x_i) - h_k(x_i))^2}$$

$$\texttt{ACC}(f, x_i, j) = \frac{1}{|\mathbb{H}_i| - 1} \sum_{k \in \mathbb{H}_i[-j]} \mathbf{1}_{f(x_i) = h_k(x_i)}$$

$$\texttt{SIM}(f, x_i, j) = \frac{1}{|\mathbb{H}_i| - 1} \sum_{k \in \mathbb{H}_i[-j]} \texttt{sim}(f(x_i), h_k(x_i))$$

Note that $-\texttt{RMSE}(h_j, x_i, j)$, $\texttt{ACC}(h_j, x_i, j)$ and $\texttt{SIM}(h_j, x_i, j)$ represent score differences between annotator j and all the other annotators. Consequently, we are interested in comparing $S(f, x_i, j)$ to $S(h_j, x_i, j)$.

After calculating the similarity score per instance, we want to aggregate the scores over the entire dataset. In this research, we plan to focus on two aggregation methods defined as follows:

Average Score We can calculate the average score over all examples, i.e.,

$$\mu_{f_j} = S(f, j) := \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} S(f, x_i, j) \qquad \mu_{h_j} = S(h_j, j) := \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} S(h_j, x_i, j)$$

The average is the most common method for estimating the mean value of a parameter. It is an unbiased estimator for the mean and, as such, does not give any special meaning or weight to any of the test instances. It is suitable for cases where we do not expect an imbalance of the labels given to the test instances by the annotators.

Statistical Test for Average For each human annotator, j, we can perform a one-sided t-test to compare the average performance of the human annotator j and the LLM over the entire dataset [17]. I.e., we would like to test the following hypothesis testing problem[1]:

$$H_{0j} : \mu_{f_j} \leq \mu_{h_j} \quad versus \quad H_{1j} : \mu_{f_j} > \mu_{h_j}$$

Notice that in our case, we want to prove that the LLM judge does **not** diverge much from the decisions of all human annotators; therefore, we aim to observe a low p-value in every test to prove that the similarity between the LLM and the set of $[-j]$ human annotators is greater than the similarity between the jth human annotator and the rest of the human annotators. The output of these comparisons will be m p-values (one for each human annotator) which we will use in the next step to calculate the final quality score and determine whether the LLM is comparable or even superior to human judges.

Score Distribution A more general way to aggregate the scores is to take the entire empirical score distribution defined as

$$\widehat{P}_{f_j}(t) = \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} \mathbf{1}_{S(f, x_i, j) \leq t} \qquad \widehat{P}_{h_j}(t) = \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} \mathbf{1}_{S(h_j, x_i, j) \leq t}$$

Considering the entire score distribution rather than just the average enables us to account for variance, skewness, and other measures of shape and dispersion, highlighting the robustness of LLMs as annotators and not just the average tendency.

Statistical Test for Distribution To compare score distributions, we propose using the statistical test proposed by Dror et al. [18] to test the hypothesis that there is a relation of almost stochastic dominance (ASD) between the two score distributions. As in the case of the average score, we wish to prove that the LLM score distribution almost stochastically dominates the score distribution generated

[1] A t-test is appropriate only if the number of test instances is large enough to match the normality assumption, i.e., $n > 30$.

by the jth human annotator to prove that the LLM agrees more with the rest of the human annotators. The hypothesis testing problem that we examine for each human annotator j is[2]:

$$H_{0j} : \widehat{P}_{f_j} \preceq \widehat{P}_{h_j} \quad versus \quad H_{1j} : \widehat{P}_{f_j} \succ \widehat{P}_{h_j}$$

Final Quality Score The final step in calculating the quality score of the LLM as a judge is to measure how many human judges it had "won" in the hypothesis testing problems, i.e., how many null hypotheses we were able to reject. It is problematic to simply count the number of p-values for which we were able to reject the null hypothesis ($p-val < 0.05$), since there is an accumulative type-I error when performing multiple hypothesis testing, especially when the hypotheses are dependent.

In this study, we recommend using a method that controls the false discovery rate (FDR), which is the expected proportion of false positives (incorrect rejections of null hypotheses) among all the hypotheses that were rejected in the multiple hypothesis testing scenario. Specifically, FDR is used to control the proportion of false discoveries among the significant results to balance the trade-off between identifying true effects and limiting incorrect conclusions.

In our case, we recommend the Benjamini-Yekutieli (BY) procedure [3] to control the FDR because our null hypotheses are dependent (we measure agreement on the same dataset using partially overlapping sets of annotators). The BY procedure is designed to handle complex dependency structures between hypotheses and adjusts the significance thresholds accordingly. By incorporating these adjustments, the BY procedure provides a more accurate control over the FDR, ensuring that our evaluation of the LLM's performance as a judge is not unduly influenced by the interdependence of hypotheses. This approach allows for a more reliable assessment of how many human judges can be replaced by the LLM, effectively mitigating the risk of overestimating its performance due to the accumulated type-I error.

The final quality score that we propose is the **percentage of null hypotheses that were rejected**, based on the findings produced by the BY procedure mentioned above. This score will reflect the proportion of cases for which the LLM was as aligned with human annotators or even better aligned with human annotators than actual human annotators. Essentially, it indicates how effectively the LLM aligns with human judgment standards, showcasing its capability to replicate or exceed human-level decision-making accuracy in the context evaluated.

3.2 Preliminary Results

As an example, we annotated the dataset of CEBaB [1] using six large language models (LLMs), namely Gemini-Flesh, Gemini-Pro [36], Llama-31 [39], GPT-4o, GPT-4o-Mini [25], and Mistral-v03 [21], and compared their performance to examine if they can replace the ten human annotators that originally annotated the dataset. The CEBaB dataset contains 711 restaurant reviews and star ranking (1-5 stars). For our calculations, we used the negative RMSE measure to calculate the difference in star ratings between the human annotations and the annotation given by the LLM. To give the star rank, the LLM was prompted with the following prompt: "You will be provided with a restaurant review. Your task is to analyze the review and determine the sentiment for the following four aspects: food, service, ambiance, and noise, as well as the number of stars (1-5). The sentiment for each aspect can only be: 'Positive', 'Negative', or 'unknown'. The number of stars must be 1, 2, 3, 4, or 5."

[2]For clarity, the statistical test is presented in terms of stochastic dominance rather than ASD. However, the actual statistical test coded for implementation is ASD. Further information is available in Dror et al. [18].

LLM	p-values	Win Rate	Wins according to BY
Gemini-Flesh	[0.99, 0.61, 0.77, 0.00, 0.64, 1.09e-37, 0.26, 0.99, 0.96, 0.34]	0.1	['w40']
Gemini-Pro	[0.67, 0.06, 0.42, 0.00, 0.988, 2.54e-37, 0.01, 0.89, 0.55, 0.01]	0.2	['w162', 'w40']
GPT-4o	[0.53, 0.00, 0.04, 1.67e-05, 0.92, 1.27e-38, 1.77e-05, 0.09, 0.14, 0.00]	0.5	['w168', 'w197', 'w198', 'w162', 'w40']
Llama-31	[0.74, 0.21, 0.25, 0.00, 0.95, 1.58e-35, 0.25, 0.75, 0.84, 0.29]	0.2	['w162', 'w40']
GPT-4o-Mini	[0.23, 0.00, 0.02, 8.54e-05, 0.93, 4.86e-40, 7.40e-05, 0.40, 0.25, 0.02]	0.4	['w197', 'w162', 'w198', 'w40']
Mistral-v03	[0.96, 0.61, 0.80, 0.01, 0.80, 3.56e-38, 0.04, 0.96, 0.99, 0.53]	0.1	['w40']

Table 1: Results of LLM as a Judge on CEBaB Dataset.

Table 1 shows the results of our experiment with CEBaB. It can be observed that the winning rates of the LLMs are relatively low, leading to the conclusion that most of them are not suitable to function as annotators instead of humans. GPT-4o, on the other hand, was able to win half of the human annotators, from which we can conclude that it may be a good substitution for this task.

Figure 2 presents the distribution of RMSE scores and the different LLMs. The figure presents a mixed result where, in some cases, the LLMs demonstrate a smaller divergence from the human annotators, and in some cases, the human annotators get a higher score, i.e., agree less with the annotations of the rest. This plot is consistent with the outcome we calculated using the p-values and the win rates shown in Table 1.

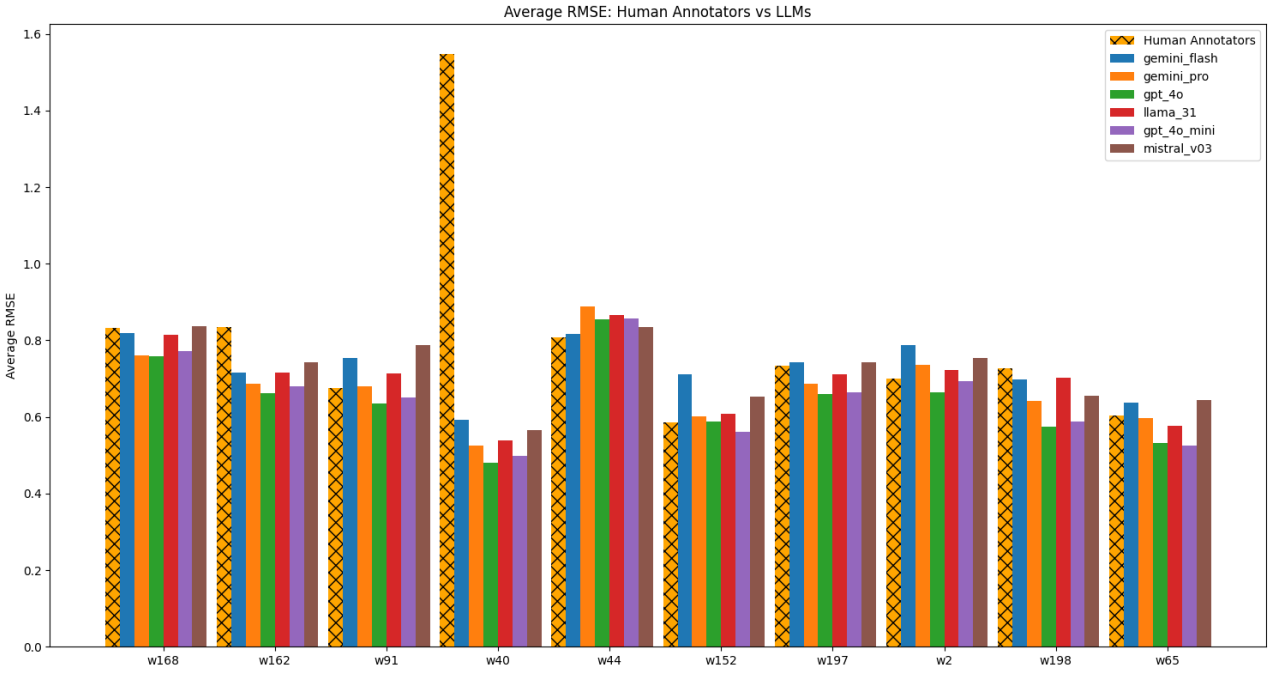

Figure 2: A comparison of the RMSE scores of the LLMs and the RMSE scores of human annotators.

3.2.1 Expert Annotations

In developing the aforementioned quality score, a key consideration was to ensure that the opinion of every human annotator is valued, even if they disagree with each other. This means that in order to get

a high quality score, the LLM should practically sample annotations from the human label distribution. Our approach assesses the likelihood that the annotations produced by the LLM originate from the label distribution generated by human annotators.

However, when expert annotations are provided, the LLM should get a high quality score if it specifically emulates that expert's input because we have higher confidence that these labels are indeed gold labels. An example scenario of when we should make an adjustment to the LLM quality score is when an expert has annotated a subset of examples, and we need to annotate additional examples, we face a decision: Should we use an LLM for these new annotations, or should we recruit a human annotator (not an expert)? In such a scenario, the instance RMSE will be calculated only with regard to the expert annotations and not with other labels. The corrections to the formulas and calculations are as follows:

Single Instance Score

$$-\texttt{RMSE}(f, x_i, exp) = -|f(x_i) - h_{exp}(x_i))|$$
$$\texttt{ACC}(f, x_i, exp) = \mathbf{1}_{f(x_i)=h_{exp}(x_i)}$$

Aggregation The methods for aggregating the scores across the entire datasets remain unchanged. We have the option to either compute the average score across all examples or take the whole score distribution into account when determining our quality score.

Quality Score of LLM as an Expert Given that we are dealing with just a single hypothesis testing scenario, the final score will be defined as the p-value associated with that particular test, or alternatively, the minimal-ϵ value calculated in the ASD statistical test. Using the p-value or the minimal-ϵ value can effectively convey the statistical significance of our decision to use LLMs as annotators instead of humans. Although in this case, a lower score is better. This approach ensures that the quality score accurately reflects the expert's standards.

While it may be challenging to definitively prove that an LLM is as proficient as a human expert, the suggested approach allows for demonstrating that the LLM outperforms a non-expert human annotator. To enhance the validity of conclusions, we propose including multiple human annotators in the evaluation. By comparing the performance of human annotators versus experts and LLMs versus experts, we can increase the number of hypothesis tests, thereby providing a more comprehensive assessment. Additionally, using multiple expert annotators in this methodology would offer the optimal validation, ensuring that the LLM is not just comparable to a single expert but can generalize its performance across different expert standards.

A Note on a Single Annotator While it is true that the calculations for the expert scenario also apply when there is only one annotator, not an expert, we advise against using these formulas in such cases. This is because checking if the LLM can overfit a single annotator who might not be a good annotator is not ideal. Our recommendation in cases where we only have a single human annotator is to recruit more human annotators to label a small portion of the dataset. This approach allows the LLM to be evaluated based on different perspectives on what the gold label should be, and the quality score will accurately reflect the LLMs' capability to assign labels in a manner similar to human annotators.

3.3 Expected Results and Pitfalls

The outcomes from this research are expected to significantly enhance our understanding of the capabilities and limitations of LLMs when they are used as annotators or judges in NLP tasks. By conducting the outlined experiments, we aim to achieve the following:

1. **A Comparison of Evaluation Approaches:** We will compare the average-based evaluation method with the distribution-based approach to determine which is more statistically powerful.
2. **A Demonstration of LLM Abilities:** Through demonstration experiments with various annotated datasets, we will assess how often the LLM performs better than the human annotator. We anticipate that the LLM will outperform human annotators in a significant number of cases, providing evidence of the potential of LLMs to serve as reliable annotators. Additionally, we will quantify the percentage of human annotators that are "defeated" by the LLM and assess how closely this aligns with IAA (inter-annotator agreement) values and FDR values, offering insights into the robustness of LLMs across different datasets.
3. **LLM Performance Ranking:** By comparing the FDR values of different LLMs, we will establish a ranking system that accurately reflects the relative performance of these models in annotator roles. This will be compared against traditional accuracy-based rankings, such as those derived from majority vote comparisons. We expect to find that our rankings offer a more robust and reliable measure of LLM performance. This will be validated by comparing our conclusions to existing evaluations, such as those from Chatbot Arena [11] or similar benchmarks.

Despite the potential advantages of using LLMs as annotators and judges, several challenges and risks could impact the success of this research:

1. **Generalization Across Domains:** LLMs are highly flexible and can perform well on a variety of tasks, but their performance may vary depending on the specific domain of the annotation task. There is a risk that the model may excel in certain types of annotation tasks (e.g., sentiment analysis or summarization) but fail to generalize to more specialized or nuanced domains. This potential limitation will be addressed by evaluating LLM performance across a diverse range of tasks and domains, as well as through fine-tuning or retraining models for domain-specific use cases when necessary.
2. **Inconsistent Performance Across Models:** Different LLMs may perform inconsistently when compared to human annotators. Variability in performance across models such as GPT-4, Gemini, or Llama could complicate the evaluation and ranking processes. To mitigate this, the research will include multiple LLM models in the evaluation framework and account for differences in model architecture and training data. Performance metrics will be carefully designed to accommodate model-specific characteristics.
3. **Computational and Resource Constraints:** Evaluating large-scale LLMs across multiple datasets and tasks requires significant computational resources. There is a risk of running into resource limitations, which could slow down or restrict the scope of experiments. This will be mitigated by optimizing the computational pipeline and using efficient model evaluation techniques, while ensuring access to adequate computing infrastructure.

Addressing these pitfalls requires a careful balance of methodological rigor, domain-specific considerations, and computational efficiency to ensure the reliability and validity of the research outcomes.

4 Resources

4.1 Data

We define a few guiding principles to follow when collecting, generating, and annotating data. In our research, the guiding principles for data collection and annotation revolve around capturing the diversity and subjectivity of human judgment in complex tasks. The datasets we select will showcase a broad array of annotation formats, including binary labels, multi-class categories, Likert scales, structured labels, and free-form generated text, to cover a wide spectrum of subjective and objective data types. By exploring datasets with varying levels of difficulty and ambiguity, we aim to understand how different tasks affect the distribution of human annotations.

A key aspect of our approach is ensuring that the datasets reflect the natural ambiguity in some tasks. For example, translation tasks often involve multiple correct translations due to the lack of direct word-for-word equivalence between languages or the introduction of grammatical genders when translating from a genderless language. Similarly, in summarization tasks, the quality of an annotation is often debated, as criteria for what makes a "good" summary can vary significantly across individuals. In these ambiguous and subjective tasks, we actively seek datasets where the diversity in annotations reflects the complexity of the task, making it possible to explore how personal interpretation and task ambiguity contribute to varied outcomes. For comparison and control, we will also include datasets for clearly objective tasks, such as named entity recognition or part-of-speech tagging, where there is typically only one correct label per instance.

To ensure the robustness and validity of the research, all datasets will be annotated by multiple human annotators. When working with existing datasets, we will prioritize those that include multiple annotations per data instance and clearly indicate the annotator ID, allowing us to trace how individual annotators contribute to the diversity of labels and how consistent annotators are with their own perspective. In cases where we must create our annotations, we will rely on crowdsourcing platforms such as Amazon Mechanical Turk or Prolific to recruit annotators. Each data instance will be annotated by several annotators, providing a rich pool of annotations that reflects a wide range of viewpoints.

4.2 LLMs

In this research, we will utilize several state-of-the-art LLMs to evaluate their effectiveness as annotators and judges. The choice of LLMs will be based on specific criteria designed to capture a broad range of capabilities and performance across different domains and tasks. The selection of LLMs for this research will be based on the following criteria:

1. Task Versatility: The chosen LLMs must demonstrate the ability to handle a wide variety of tasks, including but not limited to text classification, summarization, translation, and sentiment analysis. This ensures that the models are robust across different types of NLP tasks and domains.

2. Performance in Multimodal Tasks: Since some of the datasets involve both text and other data types (such as images), the LLMs must be capable of handling multimodal inputs. Models like Gemini Pro, which support multimodal learning, will be especially important for tasks requiring more than just text processing.

3. Open-Source and Flexibility for Fine-Tuning: Open-source models like Llama 3 allow for flexibility in terms of customization and fine-tuning, making them valuable for domain-specific tasks where additional training may be required. Open-source models will also allow for greater transparency and adaptability in the research process.

4. Computational Efficiency: While model performance is critical, it must be balanced with the practical consideration of computational resources. We will prioritize models that are efficient to run without sacrificing performance, ensuring that the research can be conducted within the available computational limits.

4.3 Expertise and Relevant Work of PI

Dr. Rotem Dror has established herself as a leading researcher in the field of NLP with a particular focus on evaluation methodologies, statistical analysis, and the replicability of research. Her extensive research portfolio showcases a deep understanding of the complexities involved in evaluating NLP models and systems.

Her contributions include significant advances in statistical methodologies for evaluating NLP models, such as the development of robust testing frameworks that ensure reliable model evaluations across datasets [16]. Her work has set new benchmarks in the field by addressing limitations in widely-used models and introducing innovative solutions, notably in the comparison of deep neural models and the evaluation of text generation models like summarization and translation [18, 14].

Dr. Dror's research on evaluating NLP models has been groundbreaking, particularly her statistical analysis of summarization and translation evaluation metrics, which has led to more accurate and fair model comparisons [12, 13]. Additionally, she has contributed to understanding the robustness of LLMs to prompt paraphrasing, which directly influences the current research direction aimed at further uncovering the capabilities of these models [24].

Throughout her career, Dr. Dror has consistently pushed the boundaries of NLP evaluation, ensuring that the methods she develops are statistically rigorous and practically applicable [17, 19]. Her work has had a lasting impact on the field, and her expertise in statistical evaluation, bias assessment, and model robustness will be crucial to achieving the objectives of the proposed research. Dr. Dror is dedicated to advancing NLP through meticulous research and by making all findings, code, data, and results publicly available, fostering transparency and collaboration within the community.

References

[1] Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596, 2022.

[2] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

[3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Nitay Calderon and Roi Reichart. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. *arXiv preprint arXiv:2407.19200*, 2024.

[6] Chris Callison-Burch and Mark Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 1–12, 2010.

[7] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

[8] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*, 2024.

[9] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *ArXiv*, abs/2403.04132, 2024. URL https://api.semanticscholar.org/CorpusID:268264163.

[12] Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146, 2021.

[13] Daniel Deutsch, Rotem Dror, and Dan Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.442. URL https://aclanthology.org/2022.naacl-main.442.

[14] Daniel Deutsch, Rotem Dror, and Dan Roth. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.753. URL https://aclanthology.org/2022.emnlp-main.753.

[15] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*, 2024.

[16] Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017.

[17] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392, 2018.

[18] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, 2019.

[19] Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. *Statistical significance testing for natural language processing*. Springer, 2020.

[20] Eduard Hovy and Julia Lavid. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36, 2010.

[21] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego De las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven LeScao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*, 2023.

[22] Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*, 2024.

[23] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 08 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00681. URL `https://doi.org/10.1162/tacl_a_00681`.

[25] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL `https://arxiv.org/abs/2303.08774`.

[26] Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. Human-centered design recommendations for llm-as-a-judge. *arXiv preprint arXiv:2407.03479*, 2024.

[27] Itay Ravid and Rotem Dror. 140 characters of justice? the promise and perils of using social media to reveal lay punishment perspectives. *U. Ill. L. Rev.*, page 1473, 2023.

[28] Yuval Reif and Roy Schwartz. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.378. URL `https://aclanthology.org/2024.naacl-long.378`.

[29] Marta Sabou, Kalina Bontcheva, and Arno Scharl. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8, 2012.

[30] Tulika Saha, Debasis Ganguly, Sriparna Saha, and Prasenjit Mitra. Workshop on large language models' interpretability and trustworthiness (llmit). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5290–5293, 2023.

[31] Burr Settles. Active learning literature survey. 2009.

[32] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language 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 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1339. URL `https://aclanthology.org/D19-1339`.

[33] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, 2008.

[34] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

[35] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[36] Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[37] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

[38] Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 45–48, 2009.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL `https://aclanthology.org/D17-1323`.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. 2024.