1 Scientific Background

Consider changing the font.

The field of Natural Language Processing (NLP) has undergone remarkable transformations with the rise of large language models (LLMs). These models are now widely used as text generators across a range of applications, sometimes even being anthropomorphized and attributed with human-like qualities such as intelligence, empathy, and humor. While many applications of LLMs are constructive and beneficial, others raise concerns of misuse or even harm to humanity as a whole and to the scientific community in particular [7, 56].

For researchers, it can be disconcerting to observe that a single methodology or tool is able to rise to completely dominate a research field, and LLMs have clearly become the de facto standard in state-of-the-art NLP research. For example, the terms "LLM," "large language models," and other related expressions appeared in 47% of the titles of long papers accepted to ACL 2025, the leading conference of the NLP community. However, this pattern is not unprecedented. The NLP community has experienced similar takeovers with the introduction of word embeddings such as Word2Vec in 2013 [39], the adoption of deep learning and LSTMs in 2015 [60], the emergence of attention mechanisms and transformers in 2018 [53], and now with the advent of LLMs [33, 1, 35, 57]. Even prior to the neural network era, comparable transitions occurred with the rise of probabilistic graphical models [34] and the broader shift from linguistics-driven to statistical approaches in NLP.

Indeed, NLP has a long history of such "methodological conquests." However, through each and every one of them, one principle remained constant: **human judgment and opinion was always regarded as the ultimate gold standard** and what researchers defined as their objective when optimizing and designing models for language processing. **In this research, I wish to explore whether this foundational principle is still true and, in particular, to investigate if humans are still required for annotating data and judging model outputs**.

Labeled data is fundamental in NLP because it provides the ground truth that models learn from and are evaluated against. Labeled data enables supervised learning by supplying input-output pairs, supporting reliable evaluation and benchmarking, and capturing complex aspects and attributes of text such as sentiment, intent and semantic roles, or linguistic phenomena, that models cannot infer from raw text alone. However, **with the advent of LLMs, which can generate, summarize, and reason over text with no task-specific supervision or labels, the question of the necessity and role of labeled data annotated by humans has resurfaced**.

The position taken in this proposal is that labeled data remains essential. While LLMs achieve impressive performance on widely known benchmarks, **they still are generalists**—they are not pre-trained to optimize performance for every task or domain—and real-world applications often require domain-specific knowledge and expertise, such as analyzing medical reports, legal texts, or understanding dialects and texts written in low-resource languages. Data annotations provide the labeled data needed to fine-tune or adapt LLMs for such specialized use cases. Furthermore, without evaluation against high-quality labels, it is impossible to know how well a model performs or to diagnose its failures; labeled data serves as the gold standard for measuring accuracy, comparing models, and understanding errors. Finally, human oversight is critical to assure fairness and robustness of models, help detecting bias, preserve social values, and ensure inclusion [25]. By decorating the text with complex linguistic, pragmatic, and cultural nuances, we guide models toward outputs that are reliable, interpretable, and aligned with human judgment and ethics. In short, even in the era of LLMs, human-labeled data provide both the foundation and the compass for responsible, high-quality NLP research and applications.

At the same time, traditional manual annotation and model evaluation are slow, costly, and often inconsistent processes. Almost every researcher who has curated, generated data, or supervised model output evaluations has had to navigate it: drafting clear guidelines, accounting for labor costs, ethics approvals, human fatigue, and other practical considerations [40, 38]. Once the annotations are collected, they must be thoroughly reviewed for consistency and noisy annotators may need to be filtered out. Today, however, we have the luxury of leveraging LLMs for annotation and evaluation. This requires crafting an effective prompt, which can itself be assisted by LLMs, and implementing a straightforward interface with the model's API [43]. The result is a rapid, large-scale annotation at minimal cost and effort. While not perfect and occasionally prone to hallucinations, this approach drastically reduces the time, effort, and expense associated with traditional manual annotation.

So, while I unequivocally believe that we need humans in the loop of data annotation and model evaluation, the role of human annotators has evolved. We still rely on human intelligence and preferences when designing and guiding models, but the sheer volume of manual annotations required is no longer as extensive as it once was. Instead, human supervision can focus on monitoring the evaluation of model outputs, ensuring that automated labeling is accurate, and providing guidance where the models struggle. This shift allows us to concentrate on quality rather than quantity. Instead of annotating entire datasets, it is more effective to engage a smaller number of highly skilled annotators to produce a high-quality sample that can guide automatic annotation and evaluation.

In this research proposal, my goal is to investigate how and when LLMs can effectively and reliably replace human annotators and judges, as well as to develop methods for assessing the quality of annotators in both subjective and objective NLP tasks. To achieve this, we will conduct a thorough examination of the LLM-as-a-judge framework [36], explore potential improvements to this paradigm, and propose new methodologies for implementing automatic annotators and judges. In addition, we will design evaluation methodologies applicable to both human and machine annotators and judges. This research will be complemented by large-scale empirical studies across diverse NLP tasks and datasets, with the goal of identifying when and how LLM-based annotation and evaluation can serve as a reliable substitute for human input, and when human expertise remains indispensable.

1.1 LLM-as-A-Judge and LLM-as-An-Annotator

LLM-as-a-judge is a recently new paradigm in NLP that is increasingly being employed in both research and industry [36, 49, 12, 13]. According to the initial definition of the paradigm, LLMs are used as evaluators of model outputs [36]. The term LLM-as-an-annotator was coined in our recent work [9], to denote the general paradigm that uses LLMs for annotation, evaluation, or labeling tasks that are traditionally performed by humans. Thus, making LLM-as-a-judge a special case of LLM-as-an-annotator.

LLMs are extensively used in NLP research, taking on a pivotal role once filled by humans. They are employed to annotate new datasets [26, 50], or refine existing ones [42, 45], and commonly serve as evaluators for benchmarking models and methods [2, 28, 36]. LLMs' influence extends far beyond the NLP field. They annotate papers for literature reviews [8, 30] and in social science, researchers leverage LLM annotations to uncover social insights [54, 61]. Accordingly, LLMs directly shape the results, findings, and insights of studies and guide the direction of scientific inquiry.

Despite the advantages of the LLM-as-a-judge paradigm, research shows that LLMs amplify biases, leading to unfair or inconsistent judgments [4, 10, 58] and that they may struggle with tasks that re-

quire deep contextual understanding or domain-specific expertise [46, 48]. These weaknesses highlight the need for rigorous evaluation and transparency when relying on LLM annotations in research.

1.2 Evaluation of LLMs as Judges and Annotators

Evaluating the potential of LLMs as reliable judges or annotators, replacing or completing human effort, has embarked in recent studies that proposed and demonstrated the usage of LLMs as judges [13, 59]. Chiang and Lee [12] investigated the feasibility of LLMs as alternatives to human evaluations. Their work demonstrated that LLMs can replace humans in certain evaluative tasks, albeit with some limitations, particularly in complex, subjective contexts.

In a related vein, Dong et al. [19] 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 can enhance user-specific tasks . Building on this, Verga et al. [55] 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, a concept that was suggested before by Gordon et al. [27] in the pre-LLM era.

The question of bias in LLM evaluations has been a prominent concern in the literature. Jung et al. [31] 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 arising from model biases. Similarly, Chen et al. [10] 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.

Still, many studies employing LLM annotations do not explicitly measure the alignment between LLMs and humans, and those that do typically use traditional measures such as accuracy (% agreement), F1 score, inter-annotator agreement (IAA) kappas [14, 24], and correlations [37], which have limitations. To start, IAA measures assess agreement among a group of annotators, while our goal is to compare the LLM to the group of human annotators. Other measures frequently rely on majority vote labels, overlooking important nuances that individuals introduce. In a recently published study, we establish a criterion for making a definitive yes/no decision on whether an LLM can replace a human annotator and provide a rigorous statistical analysis to ensure that replacement is justified (see Figure 1 for an illustration of the method) [9]. This is a first step in accomplishing the objectives of this research proposal that are outlined in the next section. However, agreement between annotators should not be regarded as the sole indicator of reliable annotators. There are several reasons for this:

Subjectivity of Interpretations: Different annotators may have varying perspectives, biases, or expertise levels, leading to differences in judgment even when following the same guidelines. This is particularly relevant for tasks involving emotions, opinions, or cultural aspects.

Task Complexity and Ambiguity: Some annotation tasks inherently involve ambiguous cases where multiple interpretations are reasonable, such as measuring the quality of a generated story or text summaries. High agreement does not necessarily imply correctness, nor does low agreement always indicate poor-quality annotations. For example, in summarization evaluation, one annotator might prioritize factual coverage of key events, while another might value readability and conciseness more. Both perspectives are valid, yet they may lead to different judgments about the same summary.

I think this should be the linchpin argument LLM As An Everything flattens the world into the LLM representation space, turning our lives into a slop-optimized existence.

Not sure what that sentence was meant to mean.

Is the last part qualifying the replacement or the limitations?

such as?

Note: this is your first

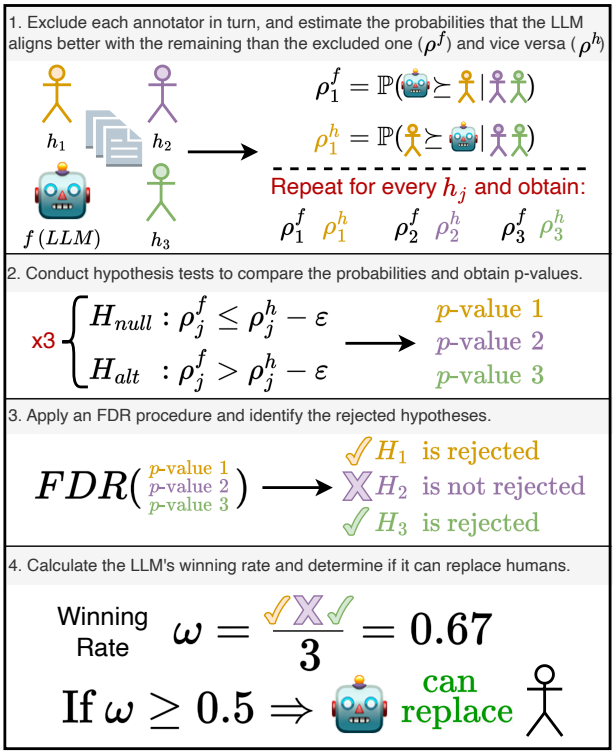

Figure 1: **An Illustration of the Alt-Test:** Given instances annotated by human annotators, we first exclude each annotator in turn to estimate the probabilities that the LLM better represents the remaining annotators and that the excluded annotator better represents them. We then test whether the LLM probability exceeds the annotator probability (considering a cost-benefit penalty ε), and apply a False Discovery Rate (FDR) controlling procedure. Then, we calculate the winning rate, ω, as the proportion of rejected hypotheses. If $\omega \geq 0.5$, we conclude that the LLM is more likely to hold an advantage over human annotators, which justifies using it.

Systematic Biases and Group Effects: Certain annotators might systematically agree due to shared biases rather than objective correctness. Conversely, disagreement among annotators with different perspectives can provide a richer understanding of the data and populations. For example, when asking whether eating cockroaches is "disgusting," annotators from cultures where insects are a common source of protein may label it as acceptable or even appealing, while annotators from cultures where insects are not considered food are more likely to rate it as highly aversive.

Subjective annotations represent a particularly interesting and challenging case for this proposal because they do not have a single, universally correct answer. Instead, they reflect personal judgments, cultural norms, and individual differences in perception. This makes them fundamentally different from objective annotations such as part-of-speech tagging or named entity recognition, where annotator agreement is more straightforward to interpret and giving a single label to each data instance makes sense. Precisely because subjective tasks involve diversity in interpretation, they expose the limitations of relying solely on annotator agreement as a measure of quality. Moreover, they offer an opportunity to explore how LLMs, with their ability to model and emulate diverse perspectives as personas, can complement or even enhance traditional annotation practices. The next subsection discusses how subjective tasks provide a natural setting to examine the potential and limitations of LLMs as annotators and their general importance to NLP.

Isn't this opposite? Maybe rephrase as 'agree/di with the statement'

Maybe talk a little bit more about your work

1.3 Evaluation of Subjective Annotations

Many aspects of human language understanding are inherently subjective and reflect diverse perspectives. Phenomena such as humor, sarcasm, offensiveness, or embarrassment are closely tied to personal experiences, cultural backgrounds, and individual sensitivities. As such, these tasks do not admit a single universally agreed upon gold label. Instead, disagreements among annotators are an intrinsic and meaningful feature rather than an error to be eliminated.

Recognizing this, the NLP community has devoted significant efforts to developing models and evaluation frameworks that better capture subjectivity. Previous research has highlighted the limitations of consensus-based evaluation and proposed alternative approaches that embrace annotator diversity. For example, Basile et al. [5] modeled individual annotators as distinct sources of personalized signals, while Davani et al. [15] introduced multi-annotator models that explicitly predict each annotator's label, thereby advancing personalization in NLP systems. Similarly, Amidei et al. [3] argued for defining acceptable bounds of annotator agreement, rather than enforcing full consensus, to account for the intrinsic variability of subjective judgments.

Other approaches emphasize the importance of considering the full distribution of annotations. Works such as Hou et al. [29], Cheng et al. [11], Uma et al. [51] argued that predictive models should approximate label distributions rather than a single output. Our previous work [22] provided methodological tools for distribution-level evaluation, while Van Der Meer et al. [52] introduced annotator-centric metrics that compare predicted and gold distributions across subjective tasks using Jensen-Shannon divergence [44].

Recent advances in calibration methods further refine model evaluation for subjective tasks. For instance, Khurana et al. [32] proposed the Crowd-Calibrator architecture, which explicitly accounts for disagreement by measuring the distance between model predictions and the crowd's label distribution. This approach enables models to abstain from making overly confident predictions when annotator disagreement is high, thereby aligning system behavior with human uncertainty.

While these contributions advance the treatment of subjectivity at the dataset level, an important gap remains: the lack of methodologies for **assessing the reliability and quality of individual annotators** within subjective frameworks. Developing such methods constitutes a central research objective of this proposal.

2 Research Objectives and Expected Significance

The overarching goal of this research is to solve the core problem that is the absence of a standardized, statistically sound methodology for validating the use of LLMs as replacements and supplements for human annotators and judges. This problem manifests in two distinct domains:

Objective Tasks: For tasks with a presumed ground truth, how can we statistically justify that an LLM annotator is a comparable or superior alternative to recruiting a human annotator, considering the benefits of cost and speed?

Subjective Tasks: For tasks where disagreement is meaningful and there is no single ground truth, how can we evaluate an LLM's quality as an annotator?

To solve these questions, I will focus on three main research objectives:

Try to differentiate more; maybe give

RO1 **Develop rigorous evaluation methods to determine when LLMs can replace or augment human annotators:** This objective focuses on extending the statistical frameworks proposed in Calderon et al. [9] to provide principled justification for the use of LLMs as annotators and judges across a variety of scenarios and tasks. This work package will include refining statistical tests for comparing LLMs and human annotators, calibrating thresholds for decision-making based on cost-benefit tradeoffs, and adapting the methodology to different annotation types (classification, regression, structured outputs and free-text). Additionally, this objective entails the design of hybrid annotation workflows in which LLMs assist annotation efforts while human experts provide supervision, adjudication, and calibration.

RO2 **Advance evaluation metrics for subjective tasks by focusing on annotator consistency, expected disagreement, and persona modeling:** The second objective addresses the unique challenges of subjective annotation tasks, where diversity of opinions is both expected and meaningful. We will redefine the notion of annotator quality for these tasks by introducing new measures of self-consistency (intra-annotator consistency) and relative-reliability (inter-annotator stable patterns of disagreement). These metrics will be applied to both human and LLM annotators, with comparisons to traditional agreement-based measures. We will also explore persona modeling techniques for LLMs, enabling them to emulate diverse human perspectives and examining them as representatives of human groups.

RO3 **Benchmark LLM annotation reliability across diverse datasets and domains, from objective classification to open-ended subjective tasks:** To ensure generalizability, this objective involves systematic benchmarking of LLM annotation performance across a broad set of domains and task types. We will curate a suite of datasets that vary along key dimensions: annotation type, task subjectivity, and annotator expertise. Each dataset will be annotated by both humans and LLMs under controlled conditions, and performance will be evaluated using the proposed statistical tests and consistency-based metrics. This benchmarking effort will produce a public repository of annotated datasets, evaluation results, and analysis protocols, enabling reproducibility and providing the community with practical guidelines on when and how LLMs can reliably serve as annotators.

2.1 Expected Significance

Mention the Hitchhiker's Guide as an example of your work that impacted in a similar manner.

The proposed research will substantially advance both the theory and practice of data annotation in NLP. By addressing the methodological and conceptual gaps raised above, the project will generate outcomes with significant scientific, practical, and societal impact.

First, the development of rigorous statistical evaluation methods (RO1) will provide the NLP community with principled tools for determining when LLMs can reliably replace or complement human annotators. This contribution will enhance the transparency and reproducibility of research that relies on LLM-generated labels, mitigating the risks associated with adopting automated annotation pipelines without sufficient validation.

Second, advancing metrics for subjective tasks (RO2) will establish a new and novel paradigm for evaluating annotators in domains where disagreement is inherently desired. By shifting the focus from consensus to consistency and reliability, this research will ensure that diversity of perspectives is preserved rather than erased in the annotation process. This has broader implications for fairness and inclusivity, as it enables models to better reflect under-represented or minority viewpoints. Moreover,

the development of meta-evaluation strategies will directly inform best practices for building more robust, fair, and accurate models, and even personalized models for various tasks.

Third, the systematic benchmarking of LLM annotation reliability across diverse datasets and domains (RO3) will provide the community with concrete empirical evidence regarding the strengths and limitations of LLMs as annotators. The resulting datasets, benchmarks, and analysis protocols will serve as critical resources for both researchers and practitioners, enabling evidence-based decisions about integrating LLMs into annotation workflows.

Taken together, these contributions will transform current practices in annotation by offering rigorous methodologies, validated metrics, and practical guidelines. Beyond NLP, the significance extends to interdisciplinary domains such as psychology, social sciences, and human-computer interaction (HCI), where annotation quality directly shapes downstream analyses and conclusions. Ultimately, this project will help ensure that annotation practices in the LLM era remain scientifically rigorous, cost-efficient, and socially responsible.

Cite personalized models literature?

And, maybe, also a roadmap for doing this for even more tasks? As in, give a framework, a 'fishing line'?

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 both objective and subjective NLP tasks.

3.1 Working Hypotheses

In this research, we hypothesize that an LLM should not significantly alter the distribution of annotations or judgments more than any individual human annotator would. Instead of measuring the LLM's agreement with an average label derived from multiple human annotations, we assess the effect of substituting each human annotator with the LLM. This method simulates the LLM in the role of an annotator by directly replacing each of the human annotators in turn and observing the consequences.

For subjective annotations, traditional agreement metrics are insufficient, and discarding annotators who deviate from the majority may eliminate meaningful minority perspectives. Building on this perspective, we define a good annotator as one who follows their "inner truth" rather than producing random noise. We introduce two criteria for assessing the quality of subjective annotators; importantly, these criteria can also serve as complements to traditional agreement-based measures in objective tasks:

1. **Self Consistency**: Does the annotator make similar judgments across similar items?
2. **Relative Consistency**: Does the annotator's bias (disagreement) with respect to the other annotators remain constant across all examples?

Do you mean the whole proposal? It makes it look like a derivative of the Calderon papern.

3.2 Methodology

3.2.1 The Alt-Test for Objective Tasks

In Calderon et al. [9], we propose using an LLM instead of human annotators when it offers a comparable alternative to recruiting an annotator. By comparing the predictions of the LLM to those of

humans, we can evaluate which more closely emulates the gold label distribution which is approximated using the collective responses of multiple annotators. Accordingly, a key consideration in our method is that the perspective of every annotator is valued. Specifically, our leave-one-out approach excludes one annotator at a time from the pool of annotators and evaluates how well the LLM's annotations align with those of the remaining annotators. The procedure is illustrated in Figure 1 and detailed below.

Notations and Definitions For a dataset of n instances $\{x_1, \ldots, x_n\}$ and m human annotators $\{h_1, \ldots, h_m\}$, we denote the annotation of the jth annotator for instance x_i as $h_j(x_i)$. The annotation predicted by the LLM is denoted as $f(x_i)$. In addition, $[-j] = \{1, \ldots, j-1, j+1, \ldots, m\}$. The set of indices of the instances annotated by h_j is denoted as $\mathbb{I}_j$. Similarly, $\mathbb{H}_i$ is the set of indices of human annotators that annotated x_i.

Instance Alignment Score We start by examining the removal of each human annotator h_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 function* between $f(x_i)$ and the annotations of $\mathbb{H}_i[-j]$. Below, we formally define three variants of S:

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

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

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

Advantage Probabilities After computing the alignment score for each data instance, we estimate the likelihood that the LLM achieves a comparable alignment with the annotators to that of the excluded annotator by calculating the percentage of instances for which the score of the LLM, $S(f, x_i, j)$, was higher or equal to the score of the jth excluded human annotator, $S(h_j, x_i, j)$. We represent this event (for x_i) using the indicator:

$$W_{i,j}^f = \begin{cases} 1, & \text{if } S(f, x_i, j) \geq S(h_j, x_i, j) \\ 0, & \text{otherwise} \end{cases}$$

Similarly, we define the indicator $W_{i,j}^h$ by reversing the inequality (to $\leq$) in the definition above, representing that the annotation of h_j for x_i is comparable to that of the LLM.

The expectation of $W_{i,j}^f$ represents the probability that the LLM annotations are as good as or better than those of h_j. We denote this probability as the *advantage over h_j probability* and estimate this probability by averaging $W_{i,j}^f$ values across all instances:

$$\rho_j^f = \hat{\mathbb{P}}(\text{LLM} \succeq h_j) = \hat{\mathbb{E}}[W_{i,j}^f] = \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} W_{i,j}^f$$

Similarly, ρ_j^h estimates the probability that h_j holds an advantage over the LLM, calculated by averaging the values of $W_{i,j}^h$.

Should the LLM Replace Annotators? If ρ_j^f is **significantly** larger than ρ_j^h it indicates that employing the LLM instead of h_j is a *justified evidence-based decision*. Notice, however, that employing an LLM is a cheaper and less labor-intensive alternative. Therefore, we introduce ε, a *cost-benefit hyperparameter* which penalizes ρ_j^h to reflect the higher cost and effort associated with human annotation. We define the following set of hypothesis testing problems to test if the LLMs' relative advantage probability is significantly larger than that of h_j:

$$\mathbf{H_{0j}}: \rho_j^f \leq \rho_j^h - \varepsilon \quad \text{vs.} \quad \mathbf{H_{1j}}: \rho_j^f > \rho_j^h - \varepsilon$$

If the p-value $< \alpha$ (typically $\alpha = 0.05$), we reject the null hypothesis, concluding that *the LLM holds a statistically significant advantage over h_j when considering the cost-benefit tradeoff.*

So far, we discussed the advantage of LLMs over a single human annotator. To generalize our conclusion to any annotator, we measure the percentage of annotators that the LLM "wins", i.e., the proportion of rejected null hypotheses. We denote this *winning rate (WR)* by ω, formally:

$$\omega = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{H_{0j} \text{ is rejected}\}$$

where $\mathbf{1}\{H_{0j} \text{ is rejected}\}$ is an indicator that receives one if the null hypothesis is rejected and zero, otherwise. If $\omega \geq 0.5$, then the LLM wins the majority of human annotators, hence *we assert that it can replace human annotators.*

Simply counting the number of rejected null hypotheses is problematic due to the accumulation of Type-I errors when performing multiple hypothesis tests, particularly when the hypotheses are dependent [20]. In our case, the dependency arises because the score of h_j relies on the annotations of the remaining $[-j]$ annotators (see how S is defined). We recommend using the Benjamini-Yekutieli procedure [6] to control the false discovery rate (FDR), as it is specifically suited for scenarios where the null hypotheses are dependent.

How to Compare LLM Judges? In many scenarios, we wish to compare different LLM judges. While it is possible to compare LLMs by their winning rate (ω), we argue this is suboptimal. First, ω does not account for the magnitude of the wins. Second, ω depends on the value of ε, and third, the range of its possible values depends on the number of human annotators, making it a coarse measure. Therefore, for comparing LLM judges, we propose the *Average Advantage Probability (AP)*:

$$\rho = \frac{1}{m} \sum_{j=1}^{m} \rho_j^f$$

We argue that ρ is a good measure for comparing LLM judges due to its desirable properties. Unlike ω, ρ spans a denser range of values and accounts for the magnitude of ρ_j^fs. Furthermore, it is more interpretable than traditional measures like F1, Cohen's κ, or correlation—it directly represents the probability that the LLM annotations are as good as or better than those of a randomly chosen annotator. This intuitive interpretation makes it accessible and meaningful for decision-makers.

It feels like a 'micro'-measure as opposed to ω's macro, is this

3.2.2 Measuring Self Consistency for Subjective Tasks

To measure the self-consistency of annotators and data quality in subjective tasks we define the following implementable definitions:

- A high-quality annotator is one who provides consistent labels for similar samples.
- A high-quality sample is one that elicits consistent labels from similar annotators, while producing divergent labels from annotators with genuinely different annotation strategies.

Given an annotation matrix A with rows representing annotators and columns representing samples, we aim to learn two quality vectors: one assigning quality scores to annotators and the other to samples. This approach is (a) generalizable, as it does not depend on the type of data, annotation, or embedding methods; and (b) flexible, since it enables us to sort annotators by their estimated quality and decide whether to exclude them from the pool of annotators based on their self-consistency on high-quality samples, rather than their agreement with the majority.

We estimate the annotator-quality vector (q_A) and the sample-quality vector(q_I) simultaneously using an iterative, SVD-based framework, inspired by recommendation systems [47]. Importantly, the original annotation matrix A remains fixed throughout the process. The algorithm begins by assigning equal trust to annotators and equal weight to samples. At each iteration, annotations are reweighted and factorized to reveal latent structures. From these, similarity graphs are built for annotators and samples. Quality is then updated by measuring how consistent each annotator is across similar samples, and how consistent each sample is across similar annotators. Consistent behavior is rewarded, inconsistency penalized. The process continues until annotator and sample quality scores stabilize. The proposed algorithm appears in Algorithm 1. We plan to develop convergence proofs and theoretical guarantees for the proposed framework.

Algorithm 1: Joint Estimation of Annotator and Sample Quality

Input: Annotation matrix $A \in \mathbb{R}^{m \times n}$, tolerance ϵ, max iterations $T_{\max}$

Output: Annotator quality q_A, sample quality q_I

Initialize: $q_A \leftarrow \frac{1}{m}\mathbf{1}_m$, $q_I \leftarrow \frac{1}{n}\mathbf{1}_n$.

for $t = 1, 2, \ldots, T_{\max}$ **do**

- $A' \leftarrow \mathrm{Diag}(q_A)\, A\, \mathrm{Diag}(q_I)$
- $(U', \Sigma, V') \leftarrow \mathrm{SVD}(A')$
- $S_A \leftarrow U'U'^{\top}, \quad S_I \leftarrow V'V'^{\top}$
- $L_A \leftarrow D_A - S_A, \quad L_I \leftarrow D_I - S_I$
- **for** *annotator* a **do**
  - $r_A(a) \leftarrow 2A_{a\cdot} L_I A_{a\cdot}^{\top}$
  - $q_A(a) \leftarrow 1/(r_A(a) + \delta)$
- **for** *sample* i **do**
  - $r_I(i) \leftarrow 2A_{\cdot i}^{\top} L_A A_{\cdot i}$
  - $q_I(i) \leftarrow 1/(r_I(i) + \delta)$
- Normalize q_A, q_I so that $\|q_A\|_1 = \|q_I\|_1 = 1$
- **if** $\|q_A^{(t)} - q_A^{(t-1)}\|_1 + \|q_I^{(t)} - q_I^{(t-1)}\|_1 < \epsilon$ **then**
  - break

return q_A, q_I

3.2.3 Measuring Relative Consistency for Subjective Tasks

We introduce a reliability measure inspired by Cohen's κ [14], adapted to capture *relative consistency* rather than raw agreement. Multiple annotators label a shared dataset, and their *relative consistency* is assessed under the assumptions that: (a) instances are independent, (b) the label space

So this is RO 2? Might wanna explicitly state

is mutually exclusive and exhaustive, (c) annotators act independently with comparable competence, and (d) no ground truth is presumed.

Our approach is probabilistic. For each annotator pair (h_i, h_j), we define a task-specific notion of *consistency* (for nominal or ordinal labels), then compute observed vs. expected frequencies under independence. Pairwise scores are aggregated into a global coefficient.

Setup Let $D = \{x_1, \ldots, x_{|D|}\}$ be the data, $\mathcal{L}$ the label set (ordinal or nominal), and $H = \{h_1, \ldots, h_n\}$ the group of annotators. For two annotators, let $D(joint)$ be the subset they both labeled and let $C_{i,j} \subseteq \mathcal{L}^2$ denote the set of pairs of labels that are considered **consistent**. The definition of a consistent pair of labels will depend on the type of annotations and task. We define the following probabilities:

Why does this depend on the annotator pair?

Marginal Probability: The probability that the annotator h_i assigns a specific label, $x \in \mathcal{L}$.

$$\hat{P}(h_i = x) = \frac{\#\{k : Label_{h_i}(x_k) = x\}}{|D(joint)|}.$$

Joint Probability: The probability that the two annotators, h_i and h_j, assign the specific set of labels x and y to the **same instance**, x_k.

$$\hat{P}(h_i = x, h_j = y) = \frac{\#\{k : Label_{h_i}(x_k) = x \wedge Label_{h_j}(x_k) = y\}}{|D(joint)|}.$$

Pairwise Consistency For a set of consistent label pairs $C_{i,j} \subseteq \mathcal{L}^2$ we define:

$$P_o = \sum_{(x,y) \in C_{i,j}} \hat{P}(h_i = x, h_j = y), \quad P_e = \sum_{(x,y) \in C_{i,j}} \hat{P}(h_i = x)\hat{P}(h_j = y).$$

Pairwise Relative Consistency:

$$RC(h_i, h_j) = \frac{P_o - P_e}{1 - P_e}.$$

$RC > 0$ indicates above-chance consistency, $RC \approx 0$ chance-level, $RC < 0$ systematic divergence. The overall relative consistency is the mean across annotator pairs:

$$RC = \tfrac{2}{n(n-1)} \sum_{i<j} RC(h_i, h_j).$$

Consistency Definitions

Ordinal: Two labels (x, y) are rank-order consistent if, say, h_i assigns higher scores than h_j both on this instance and more often across $D(joint)$.

Nominal: Consistency may be based on: (i) *Exact Match* (lenient or stricter variants), (ii) *Binary Case* (e.g., $P(h_i = x, h_j = y) > P(h_i = y, h_j = x)$), (iii) *Similarity-based*, using a label similarity function $\mathcal{S}$ with threshold α.

This framework generalizes Cohen's κ: it rewards not only exact matches but also consistent disagreement patterns, making it suitable for subjective annotation tasks.

I think this whole section needs to be clarified. Add intuition, as well

3.3 Preliminary Results

Discrete Annotation Tasks															
	WAX ($\varepsilon = 0.1$)			**LGBTeen** ($\varepsilon = 0.2$)			**MT-Bench** ($\varepsilon = 0.2$)			**Framing** ($\varepsilon = 0.15$)			**CEBaB-A** ($\varepsilon = 0.1$)		
	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ
Gemini-Flash	0.38	0.38	0.69	0.54	0.25	0.71	0.62	0.0	0.72	0.69	1.0	0.83	0.88	0.7	0.91
Gemini-Pro	0.39	0.5	**0.74**	0.47	0.0	0.67	0.62	0.0	0.76	0.79	1.0	0.91	0.91	0.9	**0.94**
GPT-4o	0.38	0.5	0.73	0.63	0.75	**0.77**	0.68	0.0	**0.77**	0.80	1.0	**0.92**	0.90	0.9	0.93
GPT-4o-mini	0.24	0.0	0.59	0.59	0.75	0.76	0.60	0.0	0.74	0.74	1.0	0.87	0.86	0.5	0.90
Llama-3.1	0.24	0.0	0.57	0.54	0.0	0.72	0.54	0.0	0.69	0.66	0.5	0.80	0.87	0.6	0.89
Mistral-v3	0.17	0.0	0.50	0.58	0.25	0.75	0.52	0.0	0.68	0.66	0.25	0.80	0.78	0.1	0.81
Continuous and Textual Annotation Tasks															
	SummEval ($\varepsilon = 0.2$)			**10K Prompts** ($\varepsilon = 0.15$)			**CEBaB-S** ($\varepsilon = 0.1$)			**Lesion** ($\varepsilon = 0.15$)			**KiloGram** ($\varepsilon = 0.1$)		
	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Sim	WR ω	AP ρ
Gemini-Flash	0.51	0.0	0.46	0.44	0.31	0.67	0.75	0.6	0.82	0.70	0.17	0.71	0.79	0.66	**0.61**
Gemini-Pro	0.47	0.0	0.44	0.33	0.08	0.63	0.78	0.8	0.87	0.73	1.0	**0.81**	0.77	0.08	0.43
GPT-4o	0.54	0.0	0.48	0.47	0.69	0.76	0.80	0.9	**0.90**	0.67	0.0	0.62	0.78	0.2	0.53
GPT-4o-mini	0.50	0.0	0.54	0.46	0.92	**0.80**	0.79	0.9	0.89	0.72	0.67	0.73	0.78	0.16	0.49
Llama-3.1	0.36	0.0	0.58	0.23	0.15	0.67	0.78	0.6	0.85	–	–	–	–	–	–
Mistral-v3	0.12	0.0	**0.62**	0.28	0.15	0.67	0.76	0.5	0.83	–	–	–	–	–	–

Table 1: **Main Results (zero-shot) — Full Datasets:** For all tasks, we report a traditional LLM-human alignment measure, such as accuracy with the majority vote (Acc) for discrete tasks, Pearson's correlation (Pears) for continuous tasks, and average similarity (Sim) for textual tasks. Additionally, we present our proposed measures: the winning rate (WR ω, the ε value is stated next to the dataset name) and the average advantage probability (AP ρ). Bold values indicate the best-performing LLM according to ρ, while a light green background highlights $\omega \geq 0.5$.

3.3.1 The Alt-Test

Table 1 presents the performance of various LLMs across discrete, continuous, and free-text tasks. We report three key measures: traditional LLM-human alignment measures (accuracy, Pearson's correlation, and similarity), the winning rate (WR, denoted as ω), and the average advantage probability (AP, denoted as ρ). For each dataset, we selected ε values based on the type of annotators: experts ($\varepsilon = 0.2$), skilled annotators ($\varepsilon = 0.15$), and crowd-workers ($\varepsilon = 0.1$). Below, we summarize our main findings. A complete presentation of our findings can be found at Calderon et al. [9].

LLMs can sometimes replace humans. Table 1 shows that many LLMs pass the alt-test across various datasets. While in two datasets (MT-Bench, and SummEval), none of the LLMs pass the test, in four (Framing, CEBAB-A, CEBaB-S and Lesion), almost all LLMs achieve $\omega \geq 0.5$. In the free-text dataset KiloGram, only Gemini-Flash passes the test. The results suggest that *in many scenarios, employing LLMs can be an alternative to recruiting additional human annotators.*

Our results also demonstrate that test success depends on the dataset and annotation aspect, with LLMs often failing to pass it. This emphasizes the relevance of the alt-test: researchers cannot simply rely on LLM annotations without justifying this choice.

Few-Shot improves LLM-human alignment. Table 1 indicates that the closed-source LLMs (GPTs and Geminis), outperform open-source LLMs.[1] However, Table 1 reports only zero-shot experiments. Thus, we also conducted experiments using three other strategies: few-shot, CoT, and ensemble. The results are presented in Table 2 and are based on 100 bootstraps of three annotators and 100 randomly sampled instances from five datasets. The reduced sample size was chosen to minimize computational costs[2] and primarily to reflect practical constraints better, as researchers are unlikely to annotate thousands of instances for testing whether the LLM is a good judge.

[1]Further experiments across varying model sizes are necessary to support broader claims about model openness.
[2]We annotated a maximum of 300 instances per dataset, which were then used for bootstrapping.

3 Annotators and 100 Instances Subsets															
	WAX ($\varepsilon = 0.1$)			**LGBTeen** ($\varepsilon = 0.2$)			**MT-Bench** ($\varepsilon = 0.2$)			**SummEval** ($\varepsilon = 0.2$)			**10K Prompts** ($\varepsilon = 0.15$)		
	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ
Gemini-Flash	0.34	0.0	0.67	0.47	0.0	0.71	0.62	0.0	0.71	0.52	0.0	0.46	0.45	0.0	0.72
+ 4-shots	0.38	0.33	0.74	0.60	1.0	**0.82**	0.59	0.0	0.72	0.65	0.67	**0.82**	0.51	1.0	**0.82**
+ CoT	0.36	0.33	0.75	0.42	0.0	0.67	0.64	0.0	0.77	0.39	0.0	0.38	0.47	0.0	0.68
Gemini-Pro	0.40	0.33	0.72	0.45	0.0	0.69	0.61	0.0	0.77	0.40	0.0	0.44	0.42	0.0	0.69
+ 4-shots	0.38	0.33	0.70	0.55	0.0	0.75	0.65	0.0	**0.79**	0.58	0.33	0.76	0.34	0.0	0.68
+ CoT	0.39	0.33	0.71	0.53	0.0	0.75	0.56	0.0	0.76	0.50	0.0	0.54	0.48	0.0	0.76
GPT-4o	0.37	0.33	0.73	0.57	0.0	0.78	0.69	0.0	0.78	0.54	0.0	0.48	0.50	0.33	0.77
+ 4-shots	0.37	0.33	0.72	0.51	0.0	0.74	0.70	0.33	**0.79**	0.60	0.67	0.76	0.44	0.0	0.74
+ CoT	0.35	0.33	0.72	0.53	0.0	0.71	0.65	0.33	**0.79**	0.59	0.0	0.66	0.46	1.0	0.77
GPT-4o-mini	0.24	0.0	0.63	0.47	0.0	0.72	0.57	0.0	0.73	0.44	0.0	0.50	0.39	0.67	0.78
+ 4-shots	0.32	0.33	0.69	0.53	0.0	0.77	0.58	0.0	0.74	0.62	0.67	0.78	0.39	0.0	0.72
+ CoT	0.39	0.0	0.71	0.52	0.0	0.72	0.58	0.0	0.72	0.58	0.0	0.58	0.34	0.33	0.75
Ens. Geminis	0.40	0.33	0.73	0.54	0.0	0.77	0.63	0.0	0.76	0.53	0.0	0.60	0.45	0.0	0.75
Ens. GPTs	0.37	0.33	0.71	0.54	0.0	0.76	0.64	0.0	0.74	0.61	0.33	0.70	0.45	0.67	0.77
Ens. All	0.44	0.33	**0.77**	0.56	0.0	0.78	0.59	0.0	0.72	0.56	0.0	0.66	0.43	0.67	0.76

Table 2: **Results – Advanced LLM Judges:** Each subset contains three annotators and one hundred instances. *Ens.* stands for "Ensemble". Please see the caption of Table 1 for information about the metrics.

As shown in Table 2, the few-shot approach (with four demonstrations) improved the performance of nearly all LLM judges. Importantly, two few-shot LLMs achieved $\omega \geq 0.5$ on SummEval, a result not observed in the zero-shot setting. This success can be attributed to the demonstrations in the prompt, which helped align the LLMs' scoring distributions more closely with the human distributions. In contrast, the CoT methodology led to a decline in performance in many cases (45%). Finally, ensembling few-shot models did not improve performance.

First appearance, I think—expand and maybe cite and define

Ok, so what's left to do on this RO?

3.3.2 Self Consistency for Subjective Tasks

We conducted a controlled experiment to test convergence of the algorithm in practice. We generated a synthetic annotation matrix $A \in \mathbb{R}^{m \times n}$ with $m = 60$ annotators and $n = 120$ items. Each item was assigned a latent "true" position $s_i \in [1, 5]$, corresponding to a score on a Likert scale. Each annotator a was modeled with three parameters: a base bias $b_a \in [1.5, 4.5]$, a sensitivity $\alpha_a \in [0.5, 1.5]$, and a noise level $\sigma_a \in [0.2, 0.8]$. Each annotation was generated as

$$A_{ai} = \text{clip}_{[1,5]}\Big(\text{round}\big(b_a + \alpha_a(s_i - 3) + \varepsilon_{ai}\big)\Big), \qquad \varepsilon_{ai} \sim \mathcal{N}(0, \sigma_a^2).$$

We then applied our algorithm with the following hyperparameters: stability constant $\delta = 10^{-4}$, maximum iterations $T_{\max} = 300$, and tolerance $\epsilon = 5 \cdot 10^{-5}$.

Metrics. At each iteration we recorded:

- $\Delta_t = \|q_A^{(t)} - q_A^{(t-1)}\|_1 + \|q_I^{(t)} - q_I^{(t-1)}\|_1$ (iterate change).
- $\text{Resid}_t = \|q_A^{(t)} - q_A^{\text{fp}}\|_1 + \|q_I^{(t)} - q_I^{\text{fp}}\|_1$ (residual after one more update).
- $\Phi_t = \sum_a \log(r_A(a) + \delta) + \sum_i \log(r_I(i) + \delta)$ (potential function).

Results. As shown in Figure 2, the algorithm consistently converged within 20–40 iterations. Both Δ_t and Resid_t decayed geometrically toward zero, reaching values below 10^{-4}. The potential function Φ_t showed monotone stabilization across iterations, confirming the presence of a Lyapunov-like descent property. Annotator and item quality vectors q_A, q_I quickly stabilized: the set of top-ranked

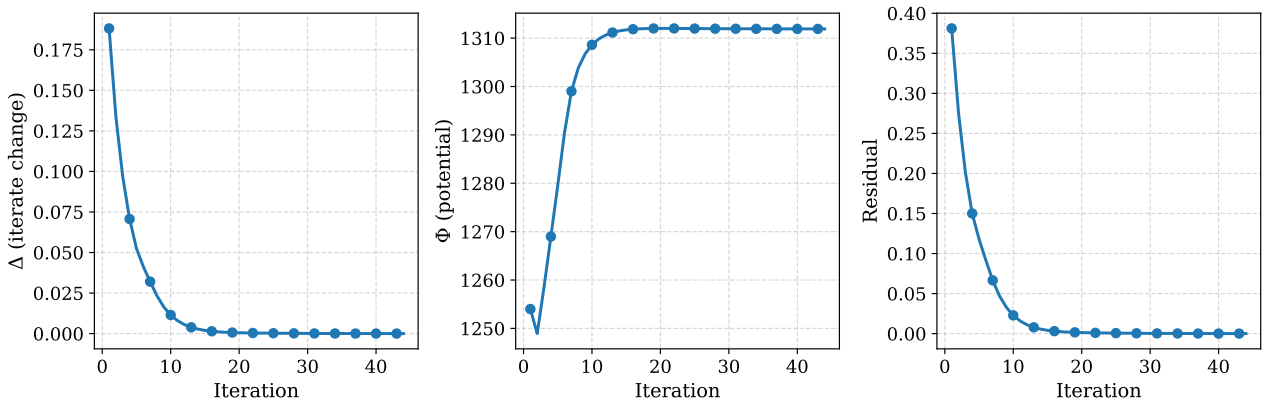

Figure 2: Convergence diagnostics of the proposed method: iterate change (Δ), potential (Φ), and residual across iterations.

annotators and items became fixed after roughly 10 iterations, while weights continued to fine-tune until convergence.

Same: what's left? (Here it's more obvious but still worth making explicit)

3.3.3 Relative Consistency for Subjective Tasks

Table 3 presents results across various simulated scenarios. RC assigns high scores in cases of structured subjectivity (*'Threshold Bias'* and *'Outlier Annotator'*) where Cohen's κ remains low or near zero. In contrast, both metrics present near zero in random annotation scenarios that do not show item-level coordination (*'Random Disagreement'*, *'Rank Bias'*).

In the *Divergent Preferences* scenario, both metrics yield negative values, with RC producing more extreme scores. This outcome reflects a key difference in how the two metrics respond to disagreement patterns: Cohen's κ given a near zero score because the annotators do not agree much, however, the label distribution is very skewed. RC on the other hand gives a very negative score since the annotators consistently act in an inconsistent way that contradicts disagreement pattern. This demonstrates RC's sensitivity to systematic inconsistencies in annotator behavior, rather than chance alignment alone. These trends are consistent across both nominal and ordinal versions of the simulated tasks.

Once more, what's left? 'future work' section needed.

4 Resources

4.1 Resource Availability

The proposed research is highly feasible given the accessibility of resources and the availability of the required expertise. The project will be carried out by the PI, one PhD student, and two MSc students, ensuring sufficient personnel to cover the theoretical, experimental, and implementation aspects of the work. The computational infrastructure available through the hosting institution already supports large-scale NLP research; however, the requested funding will be used to expand and advance these resources, enabling more efficient training and evaluation of models, large-scale statistical analyses, and systematic benchmarking of human and LLM-based annotations. The research also leverages widely accessible NLP datasets, many of which are already curated with multiple annotators and demographic information, reducing the need for costly new data collection. Where additional data is required, annotation will be supported through established institutional pipelines and access to public annotation archives. The PI has extensive experience in annotation studies, statistical evaluation, and

Be concrete.

Scenario	Cohen's κ Mean	Cohen's κ CI	Consistency (RC) Mean	Consistency (RC) CI	Description	Interpretation
Nominal Scenarios						
Random Disagreement	0	[-0.06, 0.07]	0	[-0.07, 0.08]	Annotators label randomly; no structure.	**Sanity check:** ⚠ both metrics reflect randomness.
Rank Bias	0	[-0.05, 0.06]	0	[-0.12, 0.13]	Two groups prefer opposite labels; no item-level coordination.	**Sanity check:** ⚠ shared group priors, both metrics remain low due to unaligned preferences.
Outlier Annotator	0.21	[0.20, 0.24]	0.74	[0.73, 0.75]	One annotator always flips labels from others.	**Consistent Opposer:** ✅ RC captures structure; ❌ κ penalizes.
Clustered Disagreement	0.40	[0.34, 0.45]	0.40	[-0.12, 0.85]	Annotators split into two internally consistent groups.	**Clustered Opinions:** ⚠ both metrics penalize inter-group disagreement.
Threshold Bias	0.29	[0.23, 0.36]	1.00	[1.00, 1.00]	Each annotator applies a fixed threshold to latent signal.	**Consistent Thresholds:** ✅ RC detects perfect consistency; ❌ κ sees disagreement.
Noisy Threshold Bias	0.18	[0.11, 0.25]	0.40	[0.19, 0.60]	Same as above with added label noise.	**Consistently Noisy Thresholds:** ⚠ both metrics degrade; RC remains more stable.
Divergent Preferences	0.01	[-0.04, 0.07]	-0.47	[-0.78, -0.23]	Annotators showing strong individual preference but opposite labels, resulting in high marginals but low joint agreement; random noise introduces occasional agreement.	**Adversarial-Like Behavior:** ✅ RC detects systematic bad behavior.

Table 3: **Nominal Disagreement Scenarios:** Simulated comparisons of Cohen's κ and Relative Consistency (RC) under various structured and unstructured disagreement settings. Each scenario is based on 100 annotations from 5 annotators, with 10,000 bootstrap samples used to estimate means and confidence intervals. RC better captures consistent disagreement in scenarios where κ fails.

reproducible NLP research, ensuring that both methodological development and empirical validation are well within reach. Together, the personnel, infrastructure, and accessible resources provide a strong foundation for the successful and timely execution of the proposed research.

I think this makes the project look too conservative.

4.2 Expertise and Relevant Work of the 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 [20]. 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 [22, 18, 9].

All very good but relate it to the proposed work

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 [16, 17]. Additionally, she has contributed to understanding the robustness of LLMs to prompt paraphrasing and their abilities as judges and annotators [41, 9].

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 [21, 23]. 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.

5 Expected Results and Pitfalls

We plan to conduct a series of experiments to both validate the theoretical foundations of our frameworks and demonstrate their empirical utility on real-world annotation tasks. First, we will run sanity checks to show that our self-consistency and relative-consistency measures capture intuitive properties: good annotators label similar examples consistently, and good examples receive consistent labels from similar annotators. Using datasets with multiple annotators and available demographic information, we will test whether the annotator similarity matrix recovers meaningful clusters (e.g., along demographic lines), whether similar examples are grouped together, and whether minority annotators are not unfairly penalized. Beyond these checks, we will design two core experiments. The first is **outlier detection**, where we inject different types of random annotators (uniform, label-distribution-based, demographically-biased, imitators, etc.) and measure our method's ability to identify them compared to baselines such as IAA or random detection. The second is **learning dynamics**, where we train models on the labels provided by individual annotators and test the correlation between the annotator's quality score q and model performance (both per-annotator models and unified models). These experiments will be repeated across multiple datasets to establish robustness. In addition, we will analyze existing datasets by applying our method to characterize annotator populations and example quality, and extend the analysis to LLM-generated **personas**, comparing their annotation quality across models.

We expect our experiments to confirm that the proposed frameworks satisfy both theoretical and empirical desiderata. In particular, we anticipate that the similarity-based modeling will produce meaningful annotator clusters that align with demographic or behavioral subgroups, and that example quality scores will correlate with prediction difficulty for trained models. We further expect that minority annotators will not be systematically assigned low quality scores, thereby supporting the claim that our approach captures informative minority perspectives rather than discarding them. In the outlier detection experiments, we predict that our method will successfully identify various forms of random or adversarial annotators more effectively than baseline approaches, while maintaining high retention rates for minority annotators. In the learning dynamics setting, we anticipate observing strong positive correlations between annotator quality scores and the performance of models trained on their annotations, demonstrating that our metric reflects the practical utility of using our methods to filter out bad annotators in both objective and subjective tasks. Finally, when analyzing existing datasets and LLM-generated personas, we expect to uncover new insights into the structure of annotator populations, the distribution of example difficulty, and the strengths and weaknesses of different LLMs as synthetic annotators.

At the same time, several pitfalls must be considered. First, the availability of suitable datasets with both rich annotation and demographic information may be limited, which could constrain the scope of our validation. To mitigate this, our contingency plan includes performing a small-scale, targeted data collection effort to supplement existing resources should they prove insufficient. Second, while we aim to avoid penalizing minority annotators, there is a risk that if the data are too sparse or skewed, quality scores may still inadvertently favor majority groups. We will address this by

A lot of the information here is appearing too late (like the persona idea), and the way that it's presented without citations makes it look like an afterthoughl I think you should extract a §3.4 from this (and expand it).

systematically monitoring for such effects and are prepared to implement methodological adjustments, such as stratified analysis or re-weighting schemes, to ensure fairness. Third, our outlier detection experiments rely on the injection of synthetic noise, which may not fully capture the complexities of real-world low-quality annotators. To ensure our findings generalize, we will complement these synthetic experiments with validation on real-world datasets. Fourth, model-based validation (e.g., learning dynamics) may be sensitive to the choice of model architecture, training setup, or dataset size, introducing variance unrelated to our method. To ensure the robustness of our conclusions, we will conduct a comprehensive evaluation across a diverse suite of models (including both closed and open-source architectures) and datasets. Finally, when analyzing LLM-generated annotations, interpreting results may be challenging since there is no definitive ground truth about which synthetic personas are "better," leaving some conclusions more exploratory than confirmatory. We frame this limitation as an exploratory objective to characterize the behavior of different LLM personas, providing a valuable typology of their annotation styles.

(Any plans for working on non-English data?)

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Toufique Ahmed, Premkumar T. Devanbu, Christoph Treude, and Michael Pradel. Can llms replace manual annotation of software engineering artifacts? *CoRR*, abs/2408.05534, 2024. doi: 10.48550/ARXIV.2408.05534. URL https://doi.org/10.48550/arXiv.2408.05534.

[3] Jacopo Amidei, Paul Piwek, and Alistair Willis. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, 2018.

[4] Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. Aligning human and LLM judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *CoRR*, abs/2410.00873, 2024. doi: 10.48550/ARXIV.2410.00873. URL https://doi.org/10.48550/arXiv.2410.00873.

[5] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics, 2021.

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

[7] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121, 2025.

[8] 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.

[9] Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

[10] 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.

[11] Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim O'Gorman, Nalini Singh, Andrew McCallum, and Xiang Li. Every answer matters: Evaluating commonsense with probabilistic measures. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 493–506, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.29. URL https://aclanthology.org/2024.acl-long.29/.

[12] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *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.

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[15] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[25] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[26] Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box NLP models using llm-generated counterfactuals. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=UMfcdRIotC.

[27] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[28] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[29] Zhaoyi Joey Hou, Adriana Kovashka, and Xiang Lorraine Li. Leveraging large models for evaluating novel content: A case study on advertisement creativity. *arXiv preprint arXiv:2503.00046*, 2025.

[30] Lucas Joos, Daniel A. Keim, and Maximilian T. Fischer. Cutting through the clutter: The potential of llms for efficient filtration in systematic literature reviews. *CoRR*, abs/2407.10652, 2024. doi: 10.48550/ARXIV.2407.10652. URL https://doi.org/10.48550/arXiv.2407.10652.

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

[32] Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? *arXiv preprint arXiv:2408.14141*, 2024.

[33] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

[34] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[35] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.29. URL https://aclanthology.org/2023.findings-acl.29/.

[36] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.

[37] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[38] Hsuan Wei Liao, Christopher Klugmann, Daniel Kondermann, and Rafid Mahmood. Minority reports: Balancing cost and quality in ground truth data annotation. *arXiv preprint arXiv:2504.09341*, 2025.

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[40] Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.312/.

[41] 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.

[42] Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. Are llms better than reported? detecting label errors and mitigating their effect on model performance. *arXiv preprint arXiv:2410.18889*, 2024.

[43] Arbi Haza Nasution and Aytug Onan. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language NLP tasks. *IEEE Access*, 12:71876–71900, 2024. doi: 10.1109/ACCESS.2024.3402809. URL https://doi.org/10.1109/ACCESS.2024.3402809.

[44] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*, 2020.

[45] Maja Pavlovic and Massimo Poesio. The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. *CoRR*, abs/2405.01299, 2024. doi: 10.48550/ARXIV.2405.01299. URL https://doi.org/10.48550/arXiv.2405.01299.

[46] 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.

[47] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35, 2021.

[48] Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. *CoRR*, abs/2410.20266, 2024. doi: 10.48550/ARXIV.2410.20266. URL https://doi.org/10.48550/arXiv.2410.20266.

[49] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *CoRR*, abs/2410.12784, 2024. doi: 10.48550/ARXIV.2410.12784. URL https://doi.org/10.48550/arXiv.2410.12784.

[50] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 930–957. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.54.

[51] Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470, 2021. doi: 10.1613/JAIR.1.12752. URL https://doi.org/10.1613/jair.1.12752.

[52] Michiel Van Der Meer, Neele Falk, Pradeep K Murukannaiah, and Enrico Liscio. Annotator-centric active learning for subjective nlp tasks. *arXiv preprint arXiv:2404.15720*, 2024.

[53] 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.

[54] Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *CoRR*, abs/2310.01929, 2023. doi: 10.48550/ARXIV.2310.01929. URL https://doi.org/10.48550/arXiv.2310.01929.

[55] 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.

[56] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, 11(8):240197, 2024.

[57] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Ben Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6):160:1–160:32, 2024. doi: 10.1145/3649506. URL https://doi.org/10.1145/3649506.

[58] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. *CoRR*, abs/2410.02736, 2024. doi: 10.48550/ARXIV.2410.02736. URL https://doi.org/10.48550/arXiv.2410.02736.

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

[60] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics*, 4:371–383, 2016.

[61] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Comput. Linguistics*, 50(1):237–291, 2024. doi: 10.1162/COLI_A_00502. URL https://doi.org/10.1162/coli_a_00502.