OLD EXPERIENCE HELPS: LEVERAGING SURVEY METHODOLOGY TO IMPROVE AI TEXT ANNOTATION RELIABILITY IN SOCIAL SCIENCES

Linzhuo Li
Zhejiang University
Hanzhou, Zhejiang 310058
linzhuoli@zju.edu.cn

February 28, 2025

ABSTRACT

This paper introduces a framework for assessing the reliability of Large Language Model (LLM) text annotations in social science research by adapting established survey methodology principles. Drawing parallels between survey respondent behavior and LLM outputs, the study implements three key interventions: option randomization, position randomization, and reverse validation. While traditional accuracy metrics may mask model instabilities, particularly in edge cases, our framework provides a more comprehensive reliability assessment. Using the F1000 dataset in biomedical science and three sizes of Llama models (8B, 70B, and 405B parameters), the paper demonstrates that these survey-inspired interventions can effectively identify unreliable annotations that might otherwise go undetected through accuracy metrics alone. The results show that 5-25% of LLM annotations change under these interventions, with larger models exhibiting greater stability. Notably, for rare categories approximately 50% of "correct" annotations demonstrate low reliability when subjected to this framework. The paper introduce an information-theoretic reliability score (R-score) based on Kullback-Leibler divergence that quantifies annotation confidence and distinguishes between random guessing and meaningful annotations at the case level. This approach complements existing expert validation methods by providing a scalable way to assess internal annotation reliability and offers practical guidance for prompt design and downstream analysis.

Keywords AI · Large Language Model · Text Annotation · Reliablity · Survey Research

1 Introduction

Artificial intelligence has increasingly become the fundamental infrastructure for research ([1, 2]). Particularly, Large Language Models (LLMs) emerged as powerful tools for large-scale text annotation in many research settings in social sciences ([3], [4]). Although many empirical studies demonstrate LLMs' remarkable "out-of-the-box" annotation capabilities, increasing concerns also emerge regarding the reliability of these annotations ([5], [6], [7], [8], [9], [10], [11]. This reliability question is particularly crucial for social science researchers who rely on LLM-annotated variables for downstream statistical analyses or as training data for supervised models, as reliability may substantially impact subsequent coefficient estimates, confidence intervals, or model performance. Thus, social scientists need to find ways to draw the boundary between "effective annotations" and "ineffective annotations" in their daily use.

Traditional approaches to validating annotation reliability have primarily relied on expert verification. However, this expensive and time-consuming method faces significant scaling challenges in the era of big data and LLMs. The very appeal of LLMs lies in their ability to analyze subtle dimensions (cultural, psychological, social, political, knowledge-based, etc.) across massive datasets - a scale that makes comprehensive expert verification impractical.

This raises a critical question: how can researchers develop reliability measures that match the boundless possibilities LLMs offer while maintaining rigorous standards? Is it possible to construct a reasonable boundary for assessing result

credibility before resorting to costly expert verification? This paper addresses these questions by drawing insights from survey methodology - a field with decades of experience in handling response reliability issues.

1.1 The Challenge of LLM Annotation Reliability

The dominant approach to assessing the reliability of the LLM annotation is mainly based on human expert verification. Although valuable, this "external" method adopts a somewhat agnostic perspective, disregarding the underlying mechanisms of how LLMs generate responses. This limitation becomes particularly apparent when facing the exponential growth in both data volume and annotation dimensions. Critically, even when LLM annotations align with external validation benchmarks, this consistency may mask deeper reliability concerns analogous to "careless" or "inattentive" human respondents in surveys—instances where models exploit brittle shortcuts or superficial cues rather than genuinely engaging with the task's substantive intent. Such behavior undermines construct validity([12]), particularly when annotations inform latent variable measurement or causal inference in social science research.

The challenge also highlights a mismatch between conventional validation paradigms and researchers' needs. Social scientists increasingly require case-level reliability metrics—granular indicators of whether an LLM's annotation for a specific text-problem pair reflects meaningful reasoning versus stochastic or shortcut-driven outputs. Traditional model- or variable-level evaluations, while informative for aggregate performance, fail to address the heterogeneity of LLM reliability across individual cases.

Meanwhile, the external approach also ignores recent opportunities that leverage LLMs' probabilistic nature to examine token probability distributions and measures like hallucination indices. These methods acknowledge LLMs' probabilistic output mechanism ([13], [14]) and have identified many aspects of challenges of LLM annotation reliability ([15], [16]). Notably, some studies([17]) have started to move beyond accuracy and noticed the reliability issue in a more general sense, yet still a well-established framework is currently lacking to assess the reliability problem both theoretically and empirically in a systematic way. This gap persists despite parallels to decades of survey methodology research, which has refined techniques to detect inattentive respondents, validate response consistency, and distinguish substantive engagement from brittle heuristic shortcuts.

1.2 A Survey-inspired Framework

While recent literature explores using LLMs to simulate human survey respondents ([18, 19, 17, 20]), this paper reverses the direction of knowledge transfer and argue that survey methodology's rich tradition of diagnosing and improving response reliability offers a framework for addressing LLM annotation challenges. This paper proposes such a framework that draws largely from established survey design principles to help assess LLM annotation effectiveness. The survey design methodology offers both theoretical interpretations of various respondent behaviors and corresponding strategies to evaluate response quality beyond simple accuracy metrics, addressing issues such as response consistency, attention checks, and response patterns.

The proposed framework offers three key advantages:

1. Case-Level Assessment: Unlike model-level or variable-level evaluations, this framework provides fine-grained information about LLMs' ability to annotate specific cases effectively. This allows researchers to screen for question-case pairs where LLM annotations are more reliable and incorporate reliability measures into downstream analyses.

2. Complementarity with Expert Validation: Rather than replacing expert validation, our framework provides additional information that can guide and enhance the validation process by detecting "shortcut-driven" annotations that pass external validation but lack construct validity.

3. Prompt Design Guidance: The framework can help social scientists construct more effective prompts for their annotation tasks —mirroring survey pretesting practices—to mitigate ambiguities or unintended cueing that lead LLMs toward brittle reasoning "circuit" paths.

Building upon these survey-inspired interventions, the paper further proposes an information-theoretic reliability metric (R-score) that quantifies the degree to which LLM annotations deviate from random guessing. This metric provides researchers with a practical tool to identify unreliable annotations at the case level, complementing the expert-validation approach and offering a continuous measure of annotation confidence.

2 When Survey Methodology Meets AI Annotation

2.1 Survey Strategies for the Assessment of Response Effectiveness

Despite the apparent differences between survey respondents and Large Language Models (LLMs), both can exhibit behavior that does not necessarily stem from deep engagement or true "understanding." Survey research has documented how participants sometimes take cognitive shortcuts—known as satisficing ([21], [22]). Proposed by [23] in a general sense and by [21] specifically in survery research, satisficing describes how respondents may skip or abbreviate the four cognitive processes of comprehension, retrieval, judgment, and response selection to conserve effort. Instead of formulating a fully reasoned answer, they opt for a response that seems good enough. Sometimes, satisfying can also take a more extreme form as in the case of careless responding ([24], [25], [26]): not only are participants skipping deeper thought, but they may be ignoring the survey content altogether.

Satisficing theory offers a useful theoretical framework for understanding how both human survey respondents and Large Language Models (LLMs) might produce answers that appear sufficient without necessarily reflecting deep engagement. Within the realm of survey research, satisficing often becomes more prevalent when questions are lengthy, complex, or placed toward the end of a survey, where respondent fatigue sets in. In a parallel way, LLMs can exhibit similar tendencies by relying on frequent patterns or default abilities (such as assiging high probabilities to certain tokens) instead of logically parsing the contextual aspect of a prompt. Consequently, just as survey data can be skewed by inattentive or rushed respondents, LLM-generated answers may reflect superficial cues rather than meaningful reasoning.

To address these issues, it is useful to translate three classic survey-design interventions into the context of LLM annotation. Each intervention (also known as "screener question" [27]) targets a different facet of potential satisfice-like behavior in LLMs, encouraging more thoughtful and substantively effective (regardless of accurate or not) outputs.

First, option randomization involves rearranging the possible responses or labels to which an LLM can map its answers. Surveys have long embraced this approach to mitigate response order effects, such as primacy (a tendency to favor the first-listed option) and recency (a tendency to pick the last option offered). In LLM scenarios, if choices are consistently offered in the same order (in the training data), the model may adopt a habitual preference for a particular choice. Previous work shows that LLMs can be highly sensitive to reordering of choices ([28]), LLMs can exhibit inherent selection biases toward certain option IDs ([29]). Thus, an option randomization intervention similar to survey design is needed. By randomly shifting the positions of these options, the LLM is pushed to consider their actual meaning rather than relying on their option preferences in the list.

Second, position randomization changes the very order of prompts, segments, or option texts within a single annotation or classification task. Survey researchers often shuffle question sequences to avoid complacency and to distribute respondents' cognitive effort more evenly across all items. In much the same way, LLMs reading prompts in a static, predictable order could have a pattern of mechanical responses related to the position of the content. Introducing randomness in the positioning of prompt text can help detect such tendencies.

Third, reverse validation (also called reverse coding in survey contexts) helps detect inconsistencies by presenting the LLM with inverted or negated versions of the same question. When humans respond inconsistently to positively worded and negatively worded statements of the same concept, it signals inattentiveness or insincerity. With LLMs, this intervention can reveal whether the model truly grasps the logical underpinnings of the text. For instance, after asking "What is the main type of this paper?" one might invert the question to "What is NOT the main focus of this paper?" to see whether the model remains coherent. If the LLM mismatches its answers, it suggests it may be defaulting to superficial or learned patterns rather than engaging in a deeper semantic analysis effectively.

Although the broader survey research literature includes other measures, such as open-ended questioning or extensive probing, these may have more limited utility for LLM annotation, at least in their direct form. This paper thus focuses on adopting the following survey-inspired approaches, option randomization, position randomization, and reverse validation, in line with robust experimental design traditions [30], to assess the reliability of LLM responses.

Table 1: Parallels Between Survey Responses and LLM Behavior

Survey Phenomenon	**LLM Parallel**	**Satisficing Interventions**
Response order effects (primacy, recency) in visual or oral surveys	Position/token bias when the model is repeatedly presented with the same label ordering	Option randomization to vary the sequence of answer choices
Attention checks and reverse-coded items to ensure consistency	Checking consistency across rephrased or negated prompts to detect "lazy" or contradictory LLM outputs	Reverse validation: Require logical coherence despite altered wording
Framing/context effects: subtle shifts in question wording	LLMs can exhibit sensitivity to small changes in prompt context, often yielding superficial or inconsistent answers	Position randomization to disrupt any fixed interpretation route

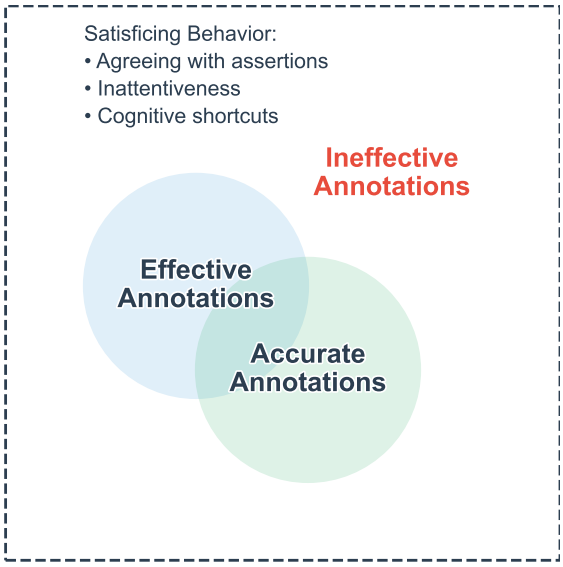

Figure 1: Effective and ineffective AI annotations

As shown in Figure 1, there is a conceptual difference between the boundary of reliability of Large Language Models (LLMs) text annotation and a common metric formerly emphasized – accuracy. While previous studies often adopt accuracy based evaluation of LLM annotation, here I aruge that their reliability cannot be assessed through accuracy metrics alone. Accuracy is typically measured by comparing LLM outputs with expert annotations. However, this comparison captures only one dimension of reliability. This situation mirrors challenges in survey methodology. Survey respondents may provide responses that appear "correct" but still ineffective. Similarly, LLM annotations require evaluation for their substantive effectiveness. The figure shows two overlapping circles representing accurate and effective annotations. Their partial overlap illustrates this distinction. Accurate annotations may match expert coding but still fail to capture the underlying construct of interest. LLMs can exhibit behaviors similar to survey respondent satisficing. These include agreeing with assertions without analysis, showing inattentiveness to context, and using cognitive shortcuts and so on. Thus, similar survey interventions need to be introduced to identify, at the case level, whether the model give an output in effectively or not.

3 Data and Methods

3.1 Data

F1000 Dataset We utilized the F1000 dataset (also known as Faculty Opinions) from previous studies in science of science. It comes from a post-publication peer-review platform in which invited scholars — practicing scientists and clinicians — select and evaluate biomedical papers they deem significant. The experts are asked to label papers with predefined tags in about five categories. For demonstration purposes, the labels used here contain three primary

contribution types that takes the majority of cases: (A). Interesting Hypothesis (7.5%) (B). Technical Advance (13.3%) and (C). New Finding (79%). These expert annotations has been shown to align with different types of novelty ([31]), thus are important in studying science and innovation.

This dataset is suitable here as it provides expert-validated classifications that requires nuanced understanding. The categories are distinct yet related, making it a moderately challenging test for LLM reliability. Notably, the class imbalance mirrors real-world scientific output - most papers make empirical findings over theoretical or methodological contributions. This distribution creates natural test conditions for evaluating LLM reliability across frequent and rare categories. After preprocessing and cleaning, a total of 816 biomedical papers with expert annotations are included.

Microsoft Academic Graph (MAG) Dataset To investigate how annotations might affect downstream tasks such as regression, this study uses a simple example of predicting a paper's citation impact (with in 3 years of publication) based on the paper's contribution types using linear regression models. To do so, the F1000 data is merged with Microsoft Academic Graph Dataset, which has publicly available Microsoft Academic Graph, to get the citation count of 816 papers. The merge was done by matching their MAG paper ids with their PMID in the PubMed dataset. Citation counts follow a heavy-tailed distribution (mean=142, SD=213, max=2,184), typical of scientific impact patterns. We log-transform citations after adding 1 to handle zeros. We also added year and team size as control variables.

3.2 LLM Models

Model Selection and Architecture We employ three variants of the LLaMA-3.1 Instruct series (8B, 70B, and 405B parameters) to systematically examine how model annotation reliability changes under survey-inspired interventions. This progression captures the full spectrum from lightweight to state-of-the-art LLMs (open sourced), allowing us to test whether larger models exhibit greater robustness to survey-inspired interventions. All models use the standard dense transformer architecture along with supervised fine-tuning and direct preference optimization after pretraining [32].

Inference Configuration To ensure comparability across model sizes, we maintain identical generation parameters: temperature=0 for controlled randomness, top-p=0.7 sampling, and maximum output length=1 token). For probability distribution analysis, we extract logits directly from the final unembedding layer by using TogetherAI api with the parameter "logprobs" equals True . This setup allows us to precisely track how intervention-induced perturbations affect the models' internal confidence metrics at the precise token of interest.

Rationale for Multi-Scale Analysis The tripartite model selection directly informs key findings in later analyses: 1) The 8B model serves as a baseline for "commodity" LLMs accessible on consumer hardware, showing high intervention sensitivity (Figure 5 - 7) due to limited contextual reasoning capacity. 2) The 70B variant represents current practical limits of dense models, demonstrating partial robustness to position randomization but remaining vulnerable to reverse validation (Figure 5 - 7). 3) The 405B model tests whether large models with extreme scale can overcome satisficing tendencies - our results suggest even this frontier model retains non-trivial sensitivity to option ordering (Figure 5 - 7), indicating limitations in current LLM paradigms.

3.3 Annotation Implementation

To setup, we design prompts for AI annotation in a multiple choces format– a fundamental and widely used format in AI tasks ([33], [34], [35], [36]), and is suggested for good performance for prompting answers ([37]), and as LLM to predict a single option token. This setting allows us to trace and evaluate the probablity distribution of the single token prediction easily, and allows for consistent comparison across models.

3.3.1 Prompt Design

Basic Prompt Template

Given the following scientific paper abstract:
`[Abstract Text]`
What is the main contribution type of this paper?

A. Interesting Hypothesis

B. Technical Advance

C. New Finding

Please respond with only the option of your choice.

3.4 Survey-Inspired Interventions

For each paper in the dataset, we implement three types of interventions. Below are the example for each intervention.

1. **Option Randomization**
   - Addresses token bias by varying the order of response options

   Example: Option Randomization

   Original:
   What is the main contribution type of this paper?
   $\mathcal{A}$. New method $\mathcal{B}$. New finding $\mathcal{C}$. New theory

   Randomized Variants:

   1) $\mathcal{A}$. New finding $\mathcal{B}$. New theory $\mathcal{C}$. New method
   2) $\mathcal{A}$. New theory $\mathcal{B}$. New method $\mathcal{C}$. New finding

2. **Position Randomization**
   - Mitigates the impact of prompt position structure on responses

   Example: Position Randomization

   Original:
   What is the main contribution type of this paper?
   $\mathcal{A}$. New method $\mathcal{B}$. New finding $\mathcal{C}$. New theory

   Position Variants:

   1) $\mathcal{B}$. New finding $\mathcal{C}$. New theory $\mathcal{A}$. New method
   2) $\mathcal{C}$. New theory $\mathcal{A}$. New method $\mathcal{B}$. New finding

3. **Reverse Validation**
   - Creating reverse-coded items

   Example: Reverse Validation

   Original:
   What is the main contribution type of this paper?
   $\mathcal{A}$. New method $\mathcal{B}$. New finding $\mathcal{C}$. New theory

   Reverse:
   What items are **NOT** the main contribution of this paper?
   $\mathcal{A}$. New finding or New theory
   $\mathcal{B}$. New method or New theory
   $\mathcal{C}$. New method or New finding

3.5 Reliability Metric Design

Beyond implementing interventions, we also need to quantify the inherent uncertainty in LLM annotations. While LLMs always provide a definitive answer when forced to choose, this masks their internal uncertainty. The intervention results (as shown in the next section) suggest that when models are genuinely uncertain about the correct classification, their outputs become highly sensitive to prompt variations—exactly what the survey-inspired interventions are designed to detect.

3.5.1 Independent Probability Assessment

To measure model uncertainty accurately, we need to access the probability distribution over possible answers. However, directly presenting all options simultaneously in a prompt introduces a problematic asymmetry due to the causal attention mechanism of transformer-based LLMs. Information from earlier options accumulates and influences the processing of later options, creating position-dependent biases (see Figure 2). Thus, the model isn't actually evaluating the options independently.

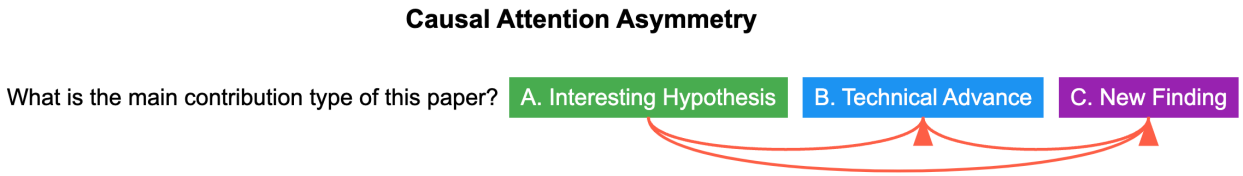

Figure 2: The Causal Asymmetry of Information Accumulation in LLM

To address this issue, the paper propose independently evaluating each category through separate binary queries:

Independent Probability Assessment

Example query for each category:
Given the following scientific paper abstract:
`[Abstract Text]`
Is the main contribution of this paper "Technical Advance"?
Please answer only Yes or No.

By repeating this process for each category ("Interesting Hypothesis", "Technical Advance", and "New Finding"), we can obtain independent probability assessments $p(\text{Yes}|\text{Category})$ for each category. This approach can be understood as the LLM-version of projecting meanings onto specific dimensions ([38, 39, 40]) —thus representing a "geometry of thinking" where each query projects the input text onto a distinct semantic axis, albeit using natural language. This controls for information asymmetry, allowing us to capture the model's genuine preference distribution.

This metric can be viewed as the "dual" of survey intervention. While survey intervention introduces asymmetry into prompts to test reliability, this reliability score does the opposite—it controls for these asymmetries to uncover the model's inherent preferences. In essence, survey intervention deliberately perturbs the system, while our reliability measure filters out these perturbations to reveal the underlying signal.

3.5.2 Information-Theoretic Reliability Score

Drawing from information theory, the paper proposes a reliability score (R-score) based on Kullback-Leibler divergence that quantifies how much an LLM's probability distribution differs from a uniform random distribution:

R-score Calculation

$$R = D_{KL}(P||U) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{1/k}\right) \tag{1}$$

where:

- $P = \{p_1, p_2, ..., p_k\}$ is the normalized probability distribution across k annotation options
- $U = \{1/k, 1/k, ..., 1/k\}$ is the uniform distribution representing random guessing
- Higher R-scores indicate greater divergence from random guessing, suggesting more reliable annotations

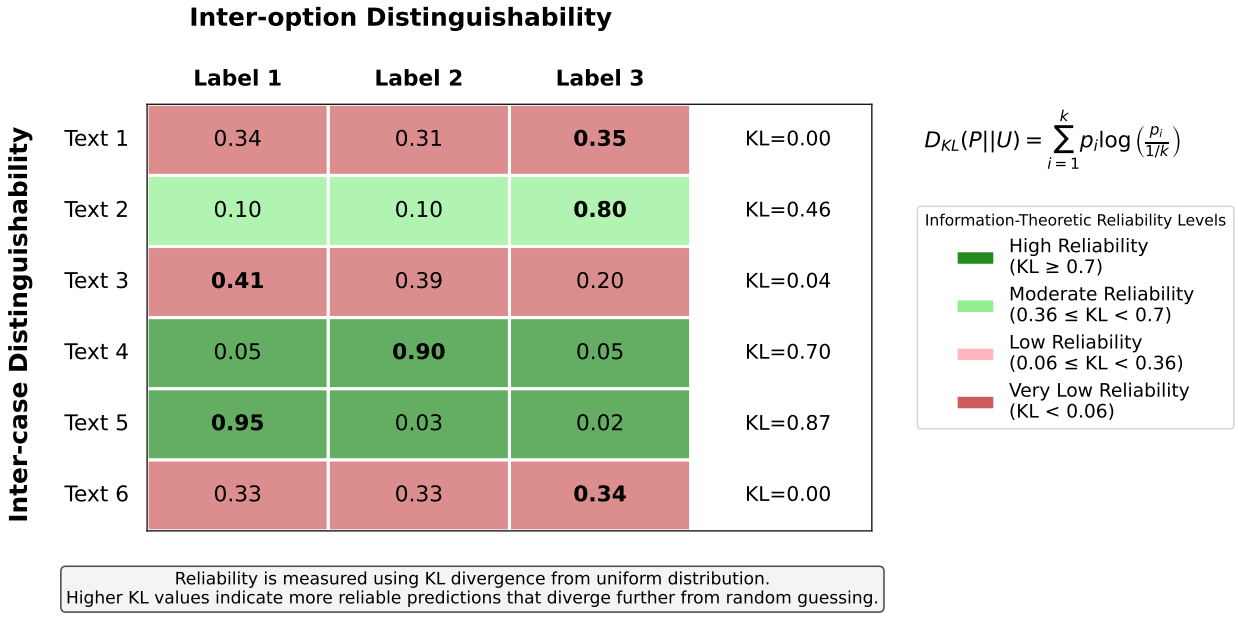

Figure 3: **Information-Theoretic Reliability Score (R-score).**

While Figure 1 illustrates the conceptual difference between accuracy and effectiveness, we need a concrete metric to operationalize this distinction. Figure 3 introduces an information-theoretic reliability score (R-score).

The reliability of a large language model (LLM) annotation can be quantified by measuring how distinctly it can differentiate between multiple classification options. Figure 3 gives an example matrix. The matrix shows probability distributions for six example texts, with each row representing a text and each column representing a classification option. Specifically, we want LLM to assign a probability for each case under each option, independently, and then normalized these probabilities. Then, the reliability is calculated using Kullback-Leibler (KL) divergence from a uniform distribution: $R = D_{KL}(P||U) = \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{1/k}\right)$, where p represents the normalized probability distribution across k annotation options and U is a uniform distribution ($1/k$ for each option, indicating random guessing). Color coding reflects reliability levels: dark green indicates high reliability (KL $\geq$ 0.7, strong differentiation between options), light green indicates moderate reliability (0.36 $\leq$ KL $<$ 0.7), light red indicates low reliability (0.06 $\leq$ KL $<$ 0.36), and dark red indicates very low reliability (KL $<$ 0.06, near-uniform distribution suggesting random guessing). For example, Text 2 shows high reliability (KL = 0.87) with a clear preference for Label 1 (0.95), while Text 1 shows very low reliability (KL = 0.00) with near-equal probabilities across all options, indicating the model cannot meaningfully distinguish between classification choices. This information-theoretic approach thus provides a principled metric for assessing when LLM annotations should be trusted for downstream analyses, at the case level.

Based on empirical analysis with three-option classification tasks, we establish four reliability thresholds:

Reliability Thresholds

- **High Reliability** ($R \geq 0.7$): Corresponds approximately to a distribution of [0.9, 0.05, 0.05], indicating strong confidence in the top prediction
- **Moderate Reliability** ($0.36 \leq R < 0.7$): Corresponds approximately to a distribution of [0.75, 0.125, 0.125], showing clear preference for the top option
- **Low Reliability** ($0.06 \leq R < 0.36$): Corresponds approximately to a distribution of [0.5, 0.25, 0.25], where the top probability equals the sum of remaining options
- **Very Low Reliability** ($R < 0.06$): Distribution approaches uniformity, suggesting the model is essentially guessing

The thresholds of R-score were empirically selected based on intuitive probability distributions for a three-option classification scenario. A distribution close to uniform (low R-score) suggests the model cannot meaningfully differentiate between categories, while a highly skewed distribution (high R-score) indicates strong preference for a particular category. The threshold of KL = 0.06 corresponds approximately to a distribution of [0.5, 0.25, 0.25], where the top probability is equal to the sum of remaining options. KL = 0.36 corresponds to a distribution of [0.75, 0.125, 0.125], where the top probability is three times of the rest others, demonstrating a clearer model preference. KL = 0.7 corresponds to a distribution of [0.9, 0.05, 0.05], representing a case where the model shows very strong confidence (90% of the time) in its top prediction. In other words, if the model repeatedly annotates this case randomly, the result is consistent with the "top" label in 90% cases.

The R-score complements our intervention-based approach by providing a continuous measure of annotation confidence. While interventions reveal sensitivity to prompt variations, the R-score quantifies the inherent uncertainty in the model's predictions.

4 Results

Figure 4 provides a comparative view of two metrics—entropy (panel a) and Kullback-Leibler (KL) divergence (panel b)—to show how survey-inspired interventions affect LLM annotation consistency across different model sizes. In panel (a), higher entropy values indicate that the model's probability distribution over the "A/B/C" labels is more uncertain (less confident), whereas lower entropy values signal more concentrated (confident) predictions. It is clear that, smaller models have higher uncertainty in general, and the variance of uncertainty is also larger in 8B and 70B than in 405 models. Interestingly, across all three LLM variants (8B, 70B, and 405B parameters), introducing intervention of reverse validation (red) leads to visibly much higher median entropy in most instances than in randomization, position randomization. This suggests that the models become much less certain when the question format is reversed than when option labels or option positions are altered.

Panel (b) further quantifies the degree of deviation from the original prompt by measuring KL divergence. Larger KL divergence values indicate that the model's predicted distributions shift more substantially under the interventions. Notably, reverse validation (red) also exhibits relatively robust divergence shifts for all model sizes, implying that negating or inverting the question meaning exposes latent inconsistencies in the model's reasoning.

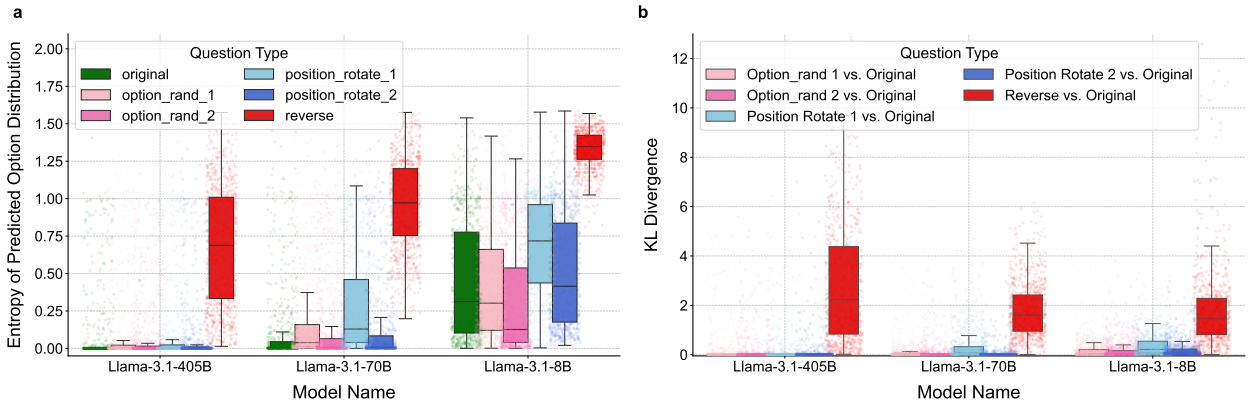

Figure 4: **Measuring annotation consistency and variance under survey-inspired interventions. a,** Entropy of predicted multinomial probability distributions across different question formats. Higher entropy values (measured in bits) indicate more uniform distributions across options, while lower values suggest more concentrated predictions. Each box represents the distribution of entropy values for 816 papers, with individual points showing specific responses. The original format (green) serves as the baseline, with variations shown in different colors (n = 816 per format). **b,** Kullback-Leibler (KL) divergence analysis comparing modified formats against the original question format. KL divergence quantifies the information loss when using modified formats compared to the original, with higher values indicating greater deviation from original responses. Each comparison includes 816 paired observations across 3 different models. Option randomization (pink shades) and position rotation (blue shades) show [specific pattern] compared to complete reversal (red), suggesting [specific insight about format effects]. Both panels demonstrate [overall insight about model behavior across formats]. Boxes show median and interquartile ranges; whiskers extend to 1.5× the interquartile range.

Figure 5 reports the "flip rate," i.e., the percentage of instances in which an LLM's top predicted label changes after being subjected to each of the three survey-inspired interventions. Despite potential expectations that larger models (405B parameters) might be more robust, the figure shows that all models—including the largest—are prone to flips at non-trivial rates. Specifically, the 8B model demonstrates a wide flip rate range (10.5–25.4%), reflecting heightened susceptibility to reordering of answer options, to rotated prompt structures, or to reversed question logic. Even the 405B model exhibits a flip rate averaging above 5%, suggesting that, while it is relatively more stable than its smaller counterparts, it is not immune to intervention-induced variability.

These results show that relying solely on standard accuracy measures may obscure the underlying instability of LLM annotations, especially for smaller-sized models. The high flip rates observed in the 8B model point toward more pronounced reliance on "shortcut" or pattern-based reasoning—akin to survey respondents who guess based on question ordering rather than content. Conversely, while the 70B and 405B models show somewhat reduced flip rates, the changes are still non-negligible. Thus, even advanced models can lapse into "satisficing"-like behaviors under small prompt adjustments.

Figure 6 zooms in on how flip rates vary by paper category—Interesting Hypothesis (N=62), Technical Advance (N=109), and New Finding (N=645)—across all three LLMs. A key takeaway is that the "Interesting Hypothesis" category shows the highest flip rates in most interventions for all model sizes. This suggests that rarer or more conceptually demanding classes may induce greater model uncertainty, making them more susceptible to small perturbations in option ordering, prompt position, or question framing. In other words, the models struggle disproportionately with classifying the less common or more abstract paper types, indicating a particular instability that could significantly affect downstream analyses focused on such minority categories. This instability in underrepresented cases will not be identified if researchers soling relying on the external metrics for model evaluation.

For "Technical Advance," the flip rates also remain elevated, especially for the 8B model, but somewhat moderate for the 70B and 405B models. "New Finding," by contrast, shows lower but still non-zero flip rates; this is likely due to its dominance in the training distribution and, hence, the model's learned preference for that label. Nonetheless, the models still register flips in the "New Finding" category under certain interventions (especially reverse validation).

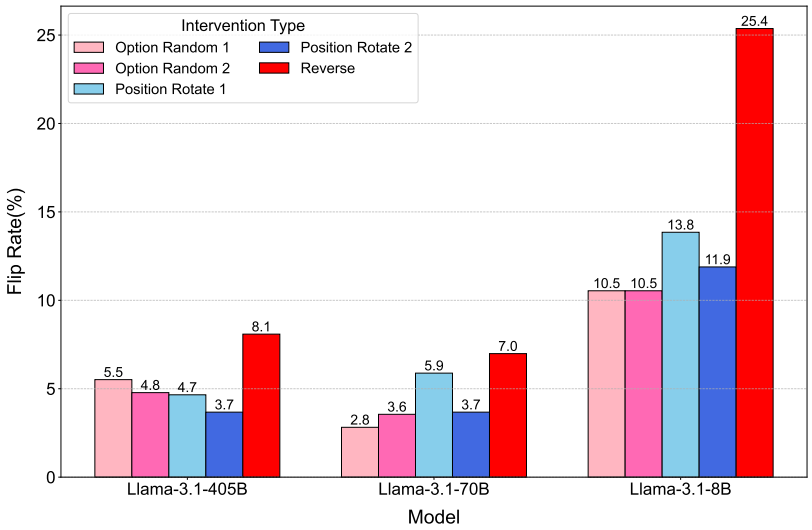

Figure 5: **Impact of survey-inspired interventions on LLM Annotation**. The figure presents the flip rates of LLM answers under three survey-methodology interventions: option randomization (varying the order of choices, shown in pink shades), position randomization (altering prompt structure, shown in blue shades), and reverse validation (using inversely coded questions, shown in red). An answer is considered to **flip** if the predicted class (with largest probability among the options) under a certain intervention is different from that under the original prompt. These interventions, adapted from classical survey design principles addressing satisficing behavior, were tested across three Llama model variants (8B, 70B, and 405B parameters). The results show that larger models are not necessarily more robust to these interventions, with Llama-3.1-8B showing notably high sensitivity (10.5-25.4%) compared to its larger counterparts. This suggests that, at least for the task of paper contribution annotation, model responses still rely to a non-negligible extent on superficial patterns rather than deep semantic understanding, analogous to satisficing behavior in human survey responses.

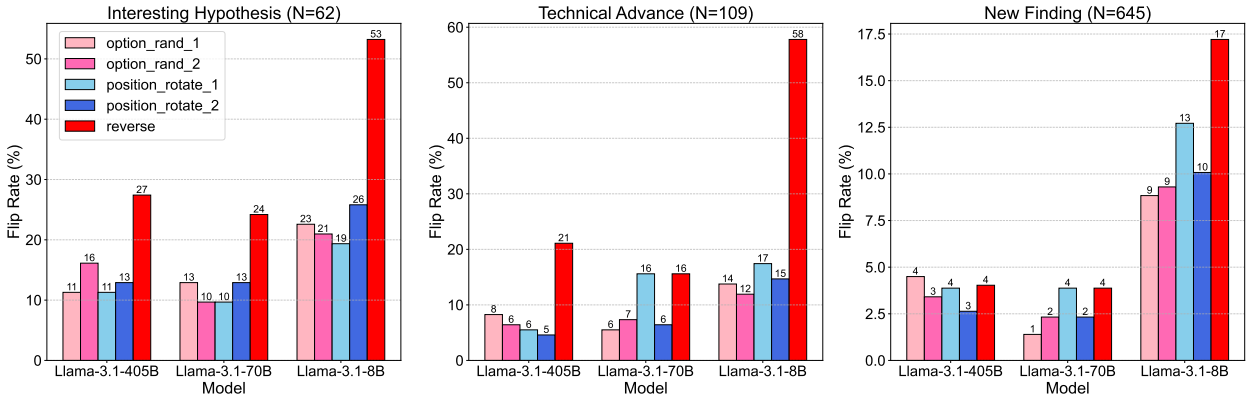

Figure 6: **Impact of survey-inspired interventions on LLM Annotation accross different paper categories**. The figure presents flip rates for three contribution types of 816 papers in the dataset: Interesting Hypothesis (N=62), Technical Advance (N=109), and New Finding (N=645). Each panel shows how different intervention types (option randomization in pink shades, position rotation in blue shades, and reverse validation in red) affect the three Llama models' response consistency. The flip rates vary substantially across paper categories, suggesting that the models' reliability is content-dependent. The Interesting hypothesis category generally exhibits higher flip rates across all interventions and model sizes, indicating particular challenges in maintaining consistent classifications for less frequent categories. This pattern persists regardless of model size, with even the largest 405B parameter model showing significant response instability under various intervention conditions.

Figure 7: **Consistency-accuracy relationships across different scientific claim types and model sizes**. Each matrix shows the relationship between consistency (whether the LLM annotation flips under certain intervention) and accuracy (Δacc) for different intervention types (y-axis) and model sizes (x-axis) across three contribution categories of scientific papers: Interesting Hypothesis, Technical Advance, and New Finding. Δacc is calculated as accuracy(flipped) - accuracy(original), where positive values (red) indicate that flipped versions achieve higher accuracy than original versions, and negative values (green) indicate higher accuracy on original versions. The intensity of color represents the magnitude of the effect. Results show distinct patterns: Interesting Hypotheses generally exhibit positive Δacc (maximum +0.55), Technical Advances show predominantly negative Δacc except for specific interventions (ranging from -0.58 to +0.98), and New Findings demonstrate consistent negative Δacc (minimum -0.86), suggesting that model behavior varies systematically across different types of scientific papers. Small categories suffer more significantly problems.

Figure 7 explores how "internal consistency" (whether an annotation flips under an intervention) aligns—or fails to align—with "external accuracy" (agreement with expert ground truth). Each cell reports Δacc, defined as accuracy(flip) minus accuracy(no-flip). By design, a positive Δacc (shown in red) means that those annotations which flipped under an intervention ironically ended up being more accurate than those that stayed the same. Conversely, negative Δacc (green) indicates that flips tend to be bad indicators for correctness in those cases—i.e., changing an answer correlates with lower accuracy relative to not flipping. The result shows that only the "New Finding" category shows alignment between external accuracy and internal consistancy for all interventions. In the other categories, all three models show misalignment in certain situations. This finding carries significant implications: although conventional wisdom suggests that higher internal consistency should correspond to higher accuracy and thus better for use, these positive Δacc values in minor categories (especially in "Interesting Hypothesis") shows that this assumption does not always hold. The areas of discrepancy consistute a "non-sense region" in which the model is unreliable, and in which simple reliance on external validation alone can be misleading.

To show how unreliable annotation might affect downstream regression tasks, a series of regression analyses are done. For simplicity, the regression task is chosen to be the relationship between paper classification (Technical Advance vs. New Finding) and the paper's citation impact. The dependent variable is the logarithm of citations received within three years of publication, and all specifications include controls for publication year and team size fixed effects. To assess the robustness of language model classifications, we implement various prompt interventions and analyze their effects on the downstream regression results.

The regression results presented in Table 2 reveal substantial impacts of various interventions on downstream classification tasks across different model sizes. The findings demonstrate that interventions frequently alter both the magnitude and statistical significance of the relationship between technical advances and citation counts.

For the 8B model, while the original classification shows no significant relationship between technical advances and citations ($\beta = -0.060$, $p = 0.615$), multiple interventions lead to statistically significant effects. Notably, the reverse intervention produces the largest effect ($\beta = -1.206$, $p < 0.01$), followed by option randomization ($\beta = -0.322$, $p < 0.05$) and position rotation ($\beta = -0.255$, $p < 0.05$). The emergence of significance across multiple intervention types suggests that the original classification's null result may not be robust to alternative prompting strategies.

Table 2: Impact of Different Interventions on Downstream Regression Tasks

			Original Classification		Intervention Classification	
Intervention Type	N	R^2	Coef.	p-value	Coef.	p-value
Panel A: 8B Model						
Option Random 1	615	0.171	-0.060	0.615	-0.322	0.010**
Option Random 2	615	0.171	-0.060	0.615	-0.071	0.594
Position Rotate 1	615	0.171	-0.060	0.615	-0.255	0.023**
Position Rotate 2	615	0.171	-0.060	0.615	0.045	0.739
Reverse	615	0.171	-0.060	0.615	-1.206	0.005***
Panel B: 70B Model						
Option Random 1	612	0.178	-0.446	0.009***	-0.268	0.076*
Option Random 2	612	0.178	-0.446	0.009***	-0.396	0.010***
Position Rotate 1	612	0.178	-0.446	0.009***	-0.341	0.026**
Position Rotate 2	612	0.178	-0.446	0.009***	-0.322	0.041**
Reverse	612	0.178	-0.446	0.009***	-0.071	0.690
Panel C: 405B Model						
Option Random 1	619	0.169	-0.189	0.196	-0.127	0.346
Option Random 2	619	0.169	-0.189	0.196	-0.268	0.060*
Option Random 1	619	0.169	-0.189	0.196	-0.185	0.202
Option Random 2	619	0.169	-0.189	0.196	-0.100	0.520
Reverse	619	0.169	-0.189	0.196	0.038	0.793

Notes: Dependent variable is the log of journal citations within 3 years of publication. All regressions include year and team size fixed effects.
Coefficients show the effect of Technical Advance (as compared to New Finding) on citation counts. Highlighted rows indicate significant changes in statistical significance level of coefficients.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. Heteroskedasticity-robust standard errors (HC1).

The 70B model exhibits the most consistent original effect ($\beta = -0.446$, $p < 0.01$), indicating that technical advances receive fewer citations compared to new findings. However, the magnitude and significance of this relationship varies substantially under different interventions. While option randomization and position rotation maintain statistical significance, their coefficients show reduced magnitudes (ranging from -0.268 to -0.396). Most strikingly, the reverse intervention completely eliminates the significant negative effect ($\beta = -0.071$, $p = 0.690$).

The 405B model's results demonstrate more modest intervention effects, though option randomization produces a marginally significant negative effect ($\beta = -0.268$, $p < 0.10$) where none existed in the original classification ($\beta = -0.189$, $p = 0.196$). The reverse intervention actually changes the direction of the coefficient, albeit non-significantly ($\beta = 0.038$, $p = 0.793$).

The prevalence of highlighted rows across all model sizes (8 out of 15 specifications) indicates that the results of the downstream regression can be very sensitive to prompt interventions. 1 out of 8 highlighted rows have changed from significant results to non-significant ones. 4 out of 8 highlighted rows have changed from non-significant to significant ones. This sensitivity suggests that survery-inspired interventions are necessary when using large language models annotated classifications that feed into statistical analyses, as seemingly minor changes in prompt design can lead to substantially different empirical conclusions.

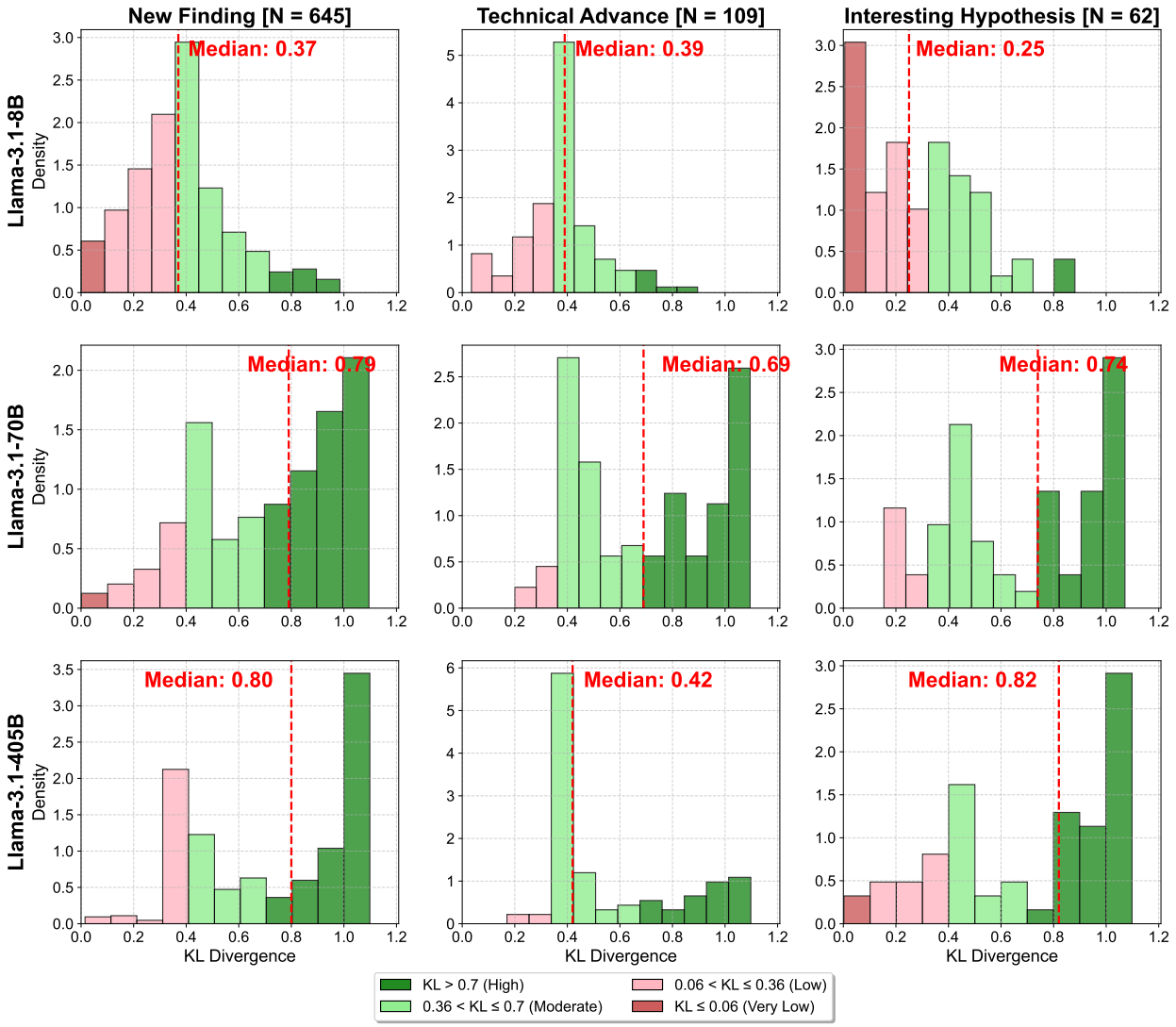

Figure 8: **Distribution of Reliable Annotations by Model and Category**.

Figure 8 presents the distribution of Reliability scores (R-score) for three sizes of the Llama-3.1 model family (8B, 70B, and 405B parameters) across three annotation categories (New Finding, Technical Advance, and Interesting Hypothesis). The R-score measures how much a model's probability distribution diverges from a uniform distribution using KL divergence, with higher values indicating greater reliability. Color-coding corresponds to reliability thresholds: dark green ($KL > 0.7$, high reliability), light green ($0.36 < KL \leq 0.7$, moderate reliability), light red ($0.06 < KL \leq 0.36$, low reliability), and dark red ($KL \leq 0.06$, very low reliability). While larger models generally demonstrate higher overall reliability (median KL of 0.37, 0.79, and 0.80 for 8B, 70B, and 405B respectively in the New Finding category), there is notable variation in reliability across different annotation categories within the same model. For instance, the 405B model shows high reliability for New Finding (median KL = 0.80) and Interesting Hypothesis (median KL = 0.82) but only moderate reliability for Technical Advance (median KL = 0.42). This variability highlights an important consideration: even when a model demonstrates high overall reliability, its performance may still be notably unreliable for specific categories or individual cases, necessitating reliability assessment at the case level rather than relying solely on model-wide metrics.

Table 3 and Table 4 further summarizes both correlations between model size and reliability, as well as category and reliablity in assessing scientific contributions of biomedical papers. The largest model, Llama-3.1-405B, demonstrated the lowest overall unreliability rate (5.0%), while the smallest model, Llama-3.1-8B, showed significantly higher unreliability (45.8%). Interestingly and importantly, all models exhibited the highest unreliability when evaluating "Interesting Hypothesis" claims (29.6% overall), compared to "New Finding" (20.2%) and "Technical Advance"

Table 3: Percentage of Unreliable Cases (KL $\leq$ 0.36) by Model and Category

Model	Interesting Hypothesis	New Finding	Technical Advance
Llama-3.1-405B	16.1% (10/62)	4.2% (27/645)	3.7% (4/109)
Llama-3.1-70B	12.9% (8/62)	10.1% (65/645)	5.5% (6/109)
Llama-3.1-8B	59.7% (37/62)	46.2% (298/645)	35.8% (39/109)
Overall by Tag	29.6% (55/186)	20.2% (390/1935)	15.0% (49/327)

Table 4: Overall Unreliability by Model

Model	Unreliability Rate
Llama-3.1-405B	5.0% (41/816)
Llama-3.1-70B	9.7% (79/816)
Llama-3.1-8B	45.8% (374/816)

(15.0%) categories. This suggests that less frequent categories present greater challenges for LLMs to evaluate reliably, regardless of model scale. The gap does NOT always become narrower in large models (as in the case of Interesting Hypothesis from 70B model to 405B model). This underscores the importance of distinguishing reliable annotation from unreliable ones, since in more general situations the underlying labels are not known as here.

5 Discussion

High accuracy scores often give a sense of confidence in model outputs. Yet, these scores can hide subtle instabilities when models handle edge cases or require deep reasoning. This paper's results demonstrate that accuracy metrics by external validation alone may not reveal fragile reasoning paths of LLM annotation. In the era of AI, adapting survey-inspired interventions for LLM annotations helps surface these hidden vulnerabilities. Social science researchers thus can use the findings to create a reliability boundary when deploying LLMs in scientific applications.

Given survey-inspired intervention results, this paper propose a information-theoretic reliability score that is simple enough to implement in any text annotation tasks without opening the black box of LLM. Rather than providing a coarse-grained metric for the model in general, one advantage of this score is to distinguish at the case level whether a LLM is capable of confidently distinguish among the categories in label tasks and consequently sensitive to altered prompt designs, as revealed by survey-inspired prompt interventions.

The study also empirically shows that one important aspect of annotation reliability is model choices. smaller models (e.g., 8B parameters) are found to exhibit substantially higher flip rates under survey-inspired interventions compared to larger counterparts. This pattern underscores the importance of reasoning capacity in annotation reliability—a capability that scales with model size. However, larger models are not completely immune. They also exhibit non-negligible rates of flips and inconsistencies. Thus, when designing large-scale studies, researchers should cautiously weigh the trade-offs between model cost and reliability. Even top-tier models can produce shortcuts that distort downstream analyses.

Another key takeaway is that reliability is not evenly distributed across classes in text data. Regions with fewer training examples, such as rarer categories, are found to be especially prone to brittle reasoning. The results show that for less common categories in science papers, many seemingly correct annotations fail under small prompt perturbations. The significance is such that paradoxically, rare categories may appear more accurate in conventional validation if models exploit dataset-specific shortcuts. Given the long-tailed distribution of categories in social data, this indicates that when using LLM for annotation, there is always a "non-sense region" where task is beyond the model's reasoning capability. Researchers need devote extra attention to this region, by creating a task-specific boundary with the help of survey interventions, and incorporate the unreliability metrics into downstream statistical analysis to mitigate this issue.

Practically, reverse validation appears to give the strongest signal of potential unreliability among the three types of interventions discussed here. It consistently triggers high flip rates and large shifts in probability distributions. Thus, researchers may want to prioritize reverse-coded checks in classification tasks that requires more rigidity or rely on more subtle distinctions, given the time and budget constraint.

While option randomization, position randomization, and reverse validation prove effective, the proposed intervention framework is not exhaustive. Other approaches may be adapted from survey methodology principles. Deeper probing of LLM sensitivities could involve more comprehensive rephrasings, semantic inversions, and multi-step consistency checks. Expanding this toolkit can strengthen reliability assessments before resorting to expensive expert labeling.

Finally, survey-inspired interventions and expert validation should be viewed as complementary rather than competing approaches. By calculating reliability scores, researchers can implement stratified validation strategies where high flip-rate or low entropy cases receive priority expert review while conserving resources on more reliable annotations. This targeted sampling approach can be valuable particularly for (hidden) rare categories and edge cases. The resulting feedback loop not only validates outputs but improves the entire annotation pipeline, as expert corrections reveal patterns of model failure that inform prompt refinement or model selection. When reporting findings, researchers can enhance methodological transparency by documenting both reliability metrics and expert agreement rates. A practical workflow might include initial LLM annotation, reliability assessment through interventions, establishing validation thresholds based on reliability distributions, concentrated expert review of below-threshold cases plus random sampling of high-reliability instances, and incorporating correction rates into final uncertainty estimates. This hybridized approach maintains the rigor of expert validation while maximizing the efficiency advantages of LLMs.

References

[1] Laura K Nelson. Ai and the transformation of social science research. *Science*, 379(6634):728–731, 2023.

[2] Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

[3] Weixin Liu, Xiaogang Shen, and Stan Matwin. Large language models for data annotation: A survey. *arXiv preprint arXiv:2310.03791*, 2023.

[4] Yiming Li, Jiaxing Xu, Dongyan Zhang, and Junru Zhao. Cipta: Contrastive-based iterative prompt-tuning using text annotation from large language models. *arXiv preprint arXiv:2310.08397*, 2023.

[5] Emily M Bender and Timnit Gebru. Using proprietary language models in academic research requires explicit justification. *Nature Machine Intelligence*, 5(7):633–634, 2023.

[6] Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. Towards coding social science datasets with language models. *arXiv preprint arXiv:2306.02177*, 2023.

[7] Zhexin Zhang, Leixin Wu, Zikang Hou, Jinghui Wang, et al. Safetybench: Evaluating the safety of large language models. *arXiv preprint arXiv:2309.07045*, 2023.

[8] Petter Törnberg. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*, 2024.

[9] Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. Can large language models transform computational social science? *Science*, 381(6662):1148–1152, 2023.

[10] Claudia Wagner and Markus Strohmaier. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *arXiv preprint arXiv:2309.11564*, 2023.

[11] Yiming Wang, Mingyu Ding, Zhen Zhao, Dongyan Zhang, and Jiaxing Xu. On the limitations of large language models (llms): False attribution. *arXiv preprint arXiv:2309.12156*, 2023.

[12] Milton E Strauss and Gregory T Smith. Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, 5(1):1–25, 2009.

[13] Yiming Wang, Mingyu Ding, Zhen Zhao, Junru Zhao, Hang Xu, Dongyan Zhang, and Jiaxing Xu. Strengthened symbol binding makes large language models reliable multiple-choice selectors. *arXiv preprint arXiv:2309.12156*, 2023.

[14] Kamila Misiejuk, Rogers Kaliisa, and Jennifer Scianna. Augmenting assessment with ai coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, 6:100216, 2024.

[15] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

[16] Gracjan Góral and Emilia Wisnios. When all options are wrong: Evaluating large language model robustness with incorrect multiple-choice options. *arXiv e-prints*, pages arXiv–2409, 2024.

[17] James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[18] Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. Vox populi, vox ai? using language models to estimate german public opinion. *arXiv preprint arXiv:2407.08563*, 2024.

[19] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[20] Austin C Kozlowski, Hyunku Kwon, and James A Evans. In silico sociology: forecasting covid-19 polarization with large language models. *arXiv preprint arXiv:2407.11190*, 2024.

[21] Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

[22] Scott Barge and Hunter Gehlbach. Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, 53:182–200, 2012.

[23] Herbert Alexander Simon. *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. New York: Wiley, 1957.

[24] Adam W Meade and S Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[25] John A Johnson. Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, 39(1):103–129, 2005.

[26] Mary K Ward and Adam W Meade. Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74(1):577–596, 2023.

[27] Adam J Berinsky, Michele F Margolis, and Michael W Sances. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3):739–753, 2014.

[28] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[29] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

[30] Paul M Sniderman and Douglas B Grob. Innovations in experimental design in attitude surveys. *Annual review of Sociology*, 22(1):377–399, 1996.

[31] Feng Shi and James Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641, 2023.

[32] Meta AI. Introducing llama 3.1: Our most capable models to date, 2024. Accessed: 2025/01/28.

[33] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[34] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[35] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[36] Yue Zhang and Felix Naumann. Vox populi, vox ai? using language models to estimate german public opinion. *arXiv preprint arXiv:2310.15619*, 2023.

[37] Joshua Robinson. Leveraging large language models trained on code for symbol binding. Master's thesis, Brigham Young University, 2022.

[38] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[39] Pedro L Rodriguez and Arthur Spirling. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):101–115, 2022.

[40] Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119, 2022.