To Err Is Human; To Annotate, SILICON? Reducing Measurement Error in LLM Annotation

Xiang Cheng
Robert H. Smith School of Business, University of Maryland
xccheng@umd.edu

Raveesh Mayya, João Sedoc
Leonard N. Stern School of Business, New York University
{raveesh, jsedoc}@stern.nyu.edu

Updated: 1 December 2025.
(For the Latest Version: Click Here)

ABSTRACT

Unstructured text data annotation is foundational to management research and Large Language Models (LLMs) promise a cost-effective and scalable alternative to human annotation. The validity of insights drawn from LLM-annotated data critically depends on minimizing the discrepancy between LLM-assigned labels and the unobserved ground truth, as well as ensuring long-term reproducibility of results. This study addresses gaps in the literature on LLM annotation by decomposing measurement error in LLM-based text annotation into four distinct sources: (1) guideline-induced error from inconsistent annotation criteria, (2) baseline-induced error from unreliable human reference standards, (3) prompt-induced error from suboptimal meta-instruction formatting, and (4) model-induced error from architectural differences across LLMs. We develop the SILICON workflow (Systematic Inference with LLMs for Information Classification and Notation) to systematically reduce measurement error from these four sources. Empirical validation across seven management research cases shows iteratively refined guidelines substantially increase the LLM-human agreement compared to one-shot guidelines; expert-generated baselines exhibit higher inter-annotator agreement as well as are less prone to producing misleading LLM-human agreement estimates compared to crowdsourced baselines; placing guidelines in the system prompt reduces prompt-induced error; and model performance varies substantially across tasks. To further reduce error, we introduce a cost-effective multi-LLM labeling method, where only low-confidence items receive additional labels from alternative models. Beyond error reduction, we address closed source model retirement cycles by introducing an intuitive regression-based methodology to establish robust reproducibility protocols. Our evidence indicates that reducing each error source is necessary, and that SILICON supports reproducible, rigorous annotation in management research. An open-source toolkit to operationalize SILICON will be released upon publication.

Key words: Large Language Models, Measurement Error, Annotation, Robust Reproducibility

1. Introduction

Annotation of unstructured text data, using either humans or artificial intelligence tools, is an integral part of management research (Ghose and Ipeirotis 2011, Humphreys and Wang 2018, Loughran and Mcdonald 2014). In particular, machine learning and Natural Language Processing

tools have been widely employed to extract insights from unstructured text data and explain underlying mechanisms via sentiment and emotional content analysis (Chakraborty et al. 2022, Liu et al. 2019, Melumad et al. 2019, Oh et al. 2023, Yang et al. 2019), product attribute classification (Banerjee et al. 2021, Kwark et al. 2021), advertisement analysis (Lee et al. 2018, Shi et al. 2022), managerial response analysis (Deng and Ravichandran 2023), consumer review analysis (Hong et al. 2021, Mayya et al. 2021), and product categorization (Lee and Hosanagar 2021).

Yet annotation errors, either by humans or automated systems, can fundamentally distort research conclusions and business decisions. Consider how Saha et al. (2023) found that fear speech is subtle and often appears non-toxic to automated toxicity or hate-speech systems, which can slip past platform moderation. Sap et al. (2019) revealed that toxicity detection systems are twice as likely to incorrectly flag communications from racial minorities as offensive, creating systematic bias in organizational text processing.[1] Given how annotation errors can have real-world consequences, minimizing measurement error is the most critical requirement for any annotation workflow.

Large Language Models (LLMs) promise to revolutionize text data annotation offering time-efficiency and cost-effectiveness compared to human annotators. Naturally, researchers are exploring LLMs for text annotation tasks (Huang et al. 2024, Pangakis et al. 2023, Yeverechyahu et al. 2024, Zhang et al. 2024). Organizations are adopting LLMs for annotation tasks, with studies showing ChatGPT can match or exceed crowdsourced worker performance in certain contexts (Gilardi et al. 2023). However, this rapid adoption has proceeded without systematic approaches to minimize measurement error, which is the discrepancy between assigned labels and unobserved ground truth. The quality of text annotation processes fundamentally depends on how closely assigned labels align with this ground truth (Yang et al. 2018), yet extant LLM annotation pipelines overlook multiple sources of error that can compound to produce misleading results.

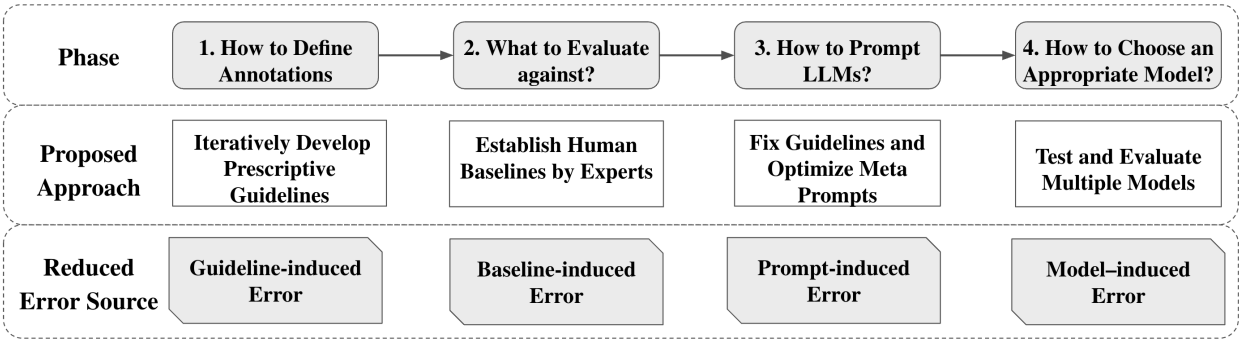
Despite growing research interest in LLMs for text annotation tasks (Tan et al. 2024), existing studies seldom systematically identify and address the distinct sources of measurement error. Extant work validates LLM outputs against human-generated labels (Pangakis et al. 2023) and offers general implementation guidelines (Törnberg 2024), but these efforts examine isolated components of the annotation process rather than systematically decomposing measurement error. Even structured approaches like de Kok (2025) and Carlson and Burbano (2025) overlook how to systematically assess model confidence for determining when multiple labels are needed (Ipeirotis et al. 2014), and fail to address reproducibility concerns as proprietary models undergo retirement

[1] These errors extend beyond research settings: in 2023, multiple universities reported cases where "AI-detection" classifiers incorrectly flagged legitimate student work as AI-generated, resulting in false academic misconduct accusations. For instance: see `https://www.independent.co.uk/news/world/americas/chatgpt-ai-plagiarism-texas-a-and-m-b2341238.html`.

cycles. Together, these gaps hinder the development of annotation pipelines that ensure quality and long-term scientific validity.

We address this critical gap by formalizing an analytical decomposition of measurement error in LLM annotation. Specifically, we partition measurement error into four sources: (1) guideline-induced error from inconsistent annotation criteria, (2) baseline-induced error from unreliable human reference standards, (3) prompt-induced error from suboptimal meta-instruction formatting, and (4) model-induced error from architectural differences across LLMs. We then propose the SILICON workflow (**S**ystematic **I**nference with **LL**Ms for **I**nformation **C**lassification and **N**otation), as illustrated in Figure 1, which provides intervention strategies for each source of error. Beyond these systematic interventions, we also present a cost-effective multi-LLM labeling methodology, where only low-confidence items receive additional labels from different models, motivated by similar benefits in human annotation (Ipeirotis et al. 2014). Finally, we introduce a regression-based statistical equivalence test to ensure robust reproducibility across multiple LLMs.

Figure 1 Proposed Four-Phase Process to Reduce Measurement Error in LLM Annotation



To empirically evaluate the effectiveness of SILICON workflow in reducing measurement error, we test multiple state-of-the-art LLMs across nine representative text classification tasks from management research. These tasks span binary classification, multi-label classification, and multi-class classification tasks. Specific applications include business proposal evaluation (Obermeier and Mayya 2024), review attribute detection (Liu et al. 2019), customer support dialog analysis, affective content evaluation (Cheng et al. 2024), language toxicity detection (Saha et al. 2023), criticism stance detection (Peng et al. 2022), and sentiment analysis (Sen et al. 2020).[2]

Our empirical analysis quantifies measurement error reduction across all four sources. First, iteratively refined guidelines reduce measurement error by 0.1–0.5 agreement level (measured by Cohen's

[2] We used seven case studies (seven datasets). Two of them include multiple tasks. Specifically, customer support dialog analysis consists of two subtasks: intent classification and breakdown analysis. Affective content evaluation supports two tasks: connectedness presence (binary classification) and connectedness classification (multiclass classification).

Kappa between LLM and human baselines) compared to one-shot guidelines across six tasks for which we developed the guidelines. Second, expert-generated baselines exhibit 0.1–0.4 higher inter-annotator agreement than crowdsourced baselines, indicating higher reliability of expert baselines. Third, system prompt placement generally reduces prompt-induced error relative to user prompt placement. Finally, model performance varies substantially across tasks, with no single model dominating all contexts, which necessitates task-specific selection to minimize model-induced error.

Beyond measurement error decomposition, we address two critical concerns: model confidence calibration and robust reproducibility. To address model confidence, we use First-Second Distance metrics (Lyu et al. 2025) to demonstrate a cost-effective approach to determine confidence thresholds for triggering multiple label generation. We document that optimal thresholds vary by model type and task complexity. For robust reproducibility, our regression-based statistical equivalence tests reveal that high-performing open-parameter models (e.g., GPT-OSS-120B) can substitute for proprietary models (e.g., GPT-4o) across most tasks, enabling long-term research sustainability despite regular model retirement cycles of frontier models.

Our work contributes to management research by providing the first systematic decomposition and quantification of measurement-error sources in LLM annotation. Whereas existing approaches focus primarily on isolated performance benchmarking, SILICON embeds error-minimization at every stage while addressing two critical gaps: determining when model confidence warrants multiple labels and ensuring long-term reproducibility despite model retirement cycles. We introduce a cost-effective approach using First-Second Distance metrics to empirically calibrate confidence thresholds for triggering additional model votes, and develop regression-based statistical equivalence testing to identify open-parameter models that can substitute for proprietary ones. Practically, SILICON offers a resource-efficient approach to LLM annotation. By evaluating model performance and measurement error on representative samples rather than entire datasets, researchers can achieve statistical validity with relatively modest sample sizes (typically 100-200 documents). These findings remain applicable as LLM capabilities evolve because they address fundamental sources of measurement error rather than model-specific implementations. To facilitate adoption, we will release an open-source toolkit that implements SILICON for practice and research.

2. Conceptual Background and Related Literature

2.1. Science of Annotation: Prescriptive Annotation Process and Reproducibility

In management research using text analysis, the effectiveness of text annotation processes fundamentally depends on establishing consistent interpretation across annotators, thereby minimizing measurement error that stems from heterogeneous mental models. Annotation tasks face challenges when annotator perspectives, influenced by their socio-economic backgrounds, lead to divergent

interpretations. For example, toxicity detection labels can vary significantly based on annotators' racial identities (Sap et al. 2019). Researchers have addressed this inherent subjectivity by choosing between *descriptive* approaches that capture diverse perspectives or *prescriptive* approaches that enforce consistent interpretation (Rottger et al. 2022). While neither approach is inherently superior, prescriptive approaches are particularly valuable in management research, where researchers aim to measure constructs using precise, theoretically grounded, unified definitions.

Prescriptive guidelines become critical when transitioning from human to LLM annotators due to fundamental differences in how the two process information. Human annotators can seek clarification during ambiguous cases and rely on contextual understanding beyond the written guidelines (Salehi et al. 2017). In contrast, LLMs operate exclusively on textual input without the ability to request clarification or access implicit domain knowledge (Pryzant et al. 2023). This creates a critical asymmetry: whereas human annotation benefits from concise guidelines to maintain attention and reduce cognitive load (Snow et al. 2008), LLMs perform optimally with detailed, exhaustive instructions that anticipate edge cases and provide explicit decision rules. Especially since LLMs cannot seek clarification during the annotation process, a prescriptive approach becomes critical to prevent any guideline ambiguity translating directly into increased measurement error.

Researchers have established that the prescriptive approach to any guideline requires an "iterative" refinement. An iterative guideline refinement through multiple rounds of documented discussions among annotators can achieve substantial agreement on the intended interpretation, since discussions resolve edge-cases (Lee et al. 2023). This iterative refinement process becomes even more critical when LLMs serve as annotators because, unlike human annotators, models cannot adapt their interpretation mid-task through discussion.

While these principles are established for human annotation processes, their application to LLM-based annotation introduces new challenges, including model selection and the need for carefully crafted "meta" system prompts that help models understand tasks and return results in structured, usable formats. The literature remains sparse on the establishment of well-established principles of reducing measurement error through annotation guidelines when using LLMs for annotation.

2.2. LLMs for Text Annotation

Given that LLMs are trained on vast volumes of human-generated text, researchers have explored their capacity to replicate or match human judgment across various domains (e.g., Chen and Chan 2024, Goli and Singh 2024, Ye et al. 2024). Early studies document that LLMs, such as GPT-3.5, could replace human coders for psychological construct detection (Rathje et al. 2024), policy interpretation (Leek et al. 2024), or act as experimental subjects or agents (Doshi et al. 2024, Manning et al. 2024, Xie et al. 2024).

For text annotation specifically, where the objective is to produce consistent labels for clearly specified tasks rather than to mimic human cognition, LLMs have demonstrated strong potential (Tan et al. 2024). For instance, Gilardi et al. (2023) found that zero-shot ChatGPT labeling achieved higher accuracy in political science classification tasks than crowdsourced workers. Mousavi et al. (2024) demonstrated that GPT-4 matched or surpassed traditional lexicon-based, supervised learning, and embedding approaches in psychometric text analysis. Indeed, LLMs can serve as cost-effective alternatives to human annotation for well-defined classification tasks.

Several studies have emerged to guide LLM annotation pipelines. Pangakis et al. (2023) highlights the importance of validating LLM annotation results and propose several model selection procedures. Törnberg (2024) offer implementation guidelines focusing on prompt engineering and evaluation metrics. de Kok (2025) presents a framework for GPT-3.5 and GPT-4 deployment for accounting annotation tasks. Closely related, Carlson and Burbano (2025) develop a five-stage implementation framework for LLM annotation with a focus on comparisons across prompts and models.

Our work builds on the extant literature and contributes in three ways that prior studies have not addressed. First, rather than another performance demonstration of LLM annotation, we target the critical problem of reducing measurement error, formally decompose its sources, and show empirically how structured procedures at each step of the pipeline reduce error. Second, given the realistic threat of model retirement and the need for researchers to demonstrate reproducibility across models, we propose a statistical procedure to test and establish robust reproducibility. Third, while the human annotation literature highlights the benefits of obtaining additional labels, current studies have yet to cover this in LLM annotation; we introduce a cost-effective multi-LLM approach that adds complementary labels from other models to improve label quality. We summarize our contributions and differentiation from the literature in Table 1.

2.3. Robust Reproducibility and Model Confidence

Reproducibility means that independent researchers obtain same results using the same methodology (Fišar et al. 2024). The rapid evolution of LLMs introduces two previously unaddressed methodological challenges: ensuring robust reproducibility as proprietary models face retirement cycles, and determining when to generate multiple labels per instance based on model confidence.

Robust reproducibility is the ability to reproduce annotations when original model becomes unavailable. Proprietary models frequently become unavailable or deprecated, threatening the long-term reproducibility of research findings (Palmer et al. 2023). Unlike traditional statistical methods that remain stable across decades, LLM behaviors have been shown to vary drastically between adjacent versions (Chen et al. 2023). Since existing reproducibility frameworks do not address these LLM-specific challenges, LLM-based analyses face fundamental reproducibility constraints

Table 1 Comparison of Methodological Components in Prior Work

Paper	Are Guidelines Systematically Developed	Who Develops Annotation Baselines	Meta Prompt Optimization	Model Testing⋆	Multi-LLM Labeling
Gilardi et al. (2023)	Not mentioned	Crowdsourced workers	No	Single	No
Rathje et al. (2024)	Not mentioned	Not mentioned	Yes	Single	No
Pangakis et al. (2023)	Yes†	Labels from past literature	No	Single	No
de Kok (2025)	No‡	Expert	Yes	Single	No
Mousavi et al. (2024)	Uses Extant Literature	Labels from past literature	Yes	Multiple	No
Carlson and Burbano (2025)	Not mentioned	Crowdsourced workers and experts	Yes	Multiple	No
Our Approach	**Yes**	**Comparing experts and Crowds**	**Placement¶, persona, CoT**	**Multiple; robust reproducibility**	**Cost-effective multi-LLM labeling**

Notes. ⋆ In model testing, "single" implies models from same series; e.g., OpenAI's GPT-3.5 and GPT-4.
¶"Placement" refers to whether instructions are included in the user prompt or system prompt.
†In Pangakis et al. (2023), the annotation guidelines were updated while optimizing the prompts. This LLM performance-oriented guideline development may be model specific and may not be "prescriptive."
‡de Kok (2025) iteratively develops the annotation guidelines, but we are unclear what these iterations do.

when the underlying models become inaccessible. Furthermore, open-parameter models (such as GPT-OSS-120B) and proprietary models potentially present a tradeoff between immediate quality and long-term availability. Robustly reproducible annotation protocols must balance this tradeoff by selecting the best current model while establishing pathways for future reproducibility.

Assessing *model confidence* in LLM annotation presents additional methodological challenges that intersect with measurement error and cost considerations. While Ipeirotis et al. (2014) demonstrate that generating multiple labels per instance improves annotation quality when human annotator confidence is low, applying this approach to LLMs requires new approaches for measuring model confidence and determining optimal thresholds. LLMs cannot express uncertainty explicitly like human annotators can, requiring post-hoc confidence estimation methods such as First-Second Distance metrics to identify when multiple model votes are warranted.

Existing LLM annotation studies lack systematic approaches to model-confidence based strategies and statistical equivalence testing between models. The absence of cost-effective methods for determining confidence thresholds and regression-based approaches for assessing model substitutability creates methodological gaps that limit the annotation efficiency and long-term scientific reliability of LLM-based annotation studies.

3. Methodology

3.1. Decomposing Measurement Error in LLM Annotation

We cannot observe ground truth labels directly, yet we need to evaluate LLM performance. This creates a fundamental measurement challenge: improvements in observed agreement (LLM vs. human reference) may not reflect true improvements (LLM vs. ground truth). To address this question, this section analytically decomposes measurement error and derives empirical predictions.

To formalize this problem, we denote a *configuration* as $c = (G, H, P, M)$, where G represents annotation guidelines, $H \in \{\text{experts}, \text{crowds}\}$ indicates the human source of the reference baselines, P represents the meta-prompt, and M represents the LLM model. Let x be an item to annotate, drawn from the distribution $x \sim \mathcal{D}$. In practice, we work with a finite sample $D = \{x_i\}_{i=1}^{n}$ from this distribution. For an item x, the ground truth (unobserved) is $y(x)$, the human reference label is $r(x)$, and the LLM label under configuration c is $\hat{y}^{(c)}(x)$. We take all expectations and probabilities with respect to $x \sim \mathcal{D}$, and compute empirical estimates using the finite sample D.

Our objective is to minimize the expected LLM annotation loss against the ground truth by systematically evaluating how each configuration component affects performance, formalized as:

$$\min_{c} \mathbb{E}\Big[\mathcal{L}\big(\hat{y}^{(c)}(x),\, y(x)\big)\Big], \tag{1}$$

where $\mathcal{L}$ represents a task-appropriate loss function. In classification tasks with 0–1 loss, this minimization is equivalent to maximizing accuracy. We approach this optimization through component-wise analysis, examining how variations in guidelines (G), human baselines (H), meta-prompts (P) and models (M) individually contribute to measurement error reduction.

To connect this objective to measurable outcomes, we define two types of agreement: (i) the *true agreement* against the unobserved truth, $A_T(c) \equiv \Pr\big(\hat{y}^{(c)} = y\big)$; and (ii) the *reference agreement* (or *observed agreement*) against the human reference baselines, $A_R(c) \equiv \Pr\big(\hat{y}^{(c)} = r\big)$.[3] Since the best case scenario (perfect annotation) is when LLM annotations perfectly match the unobserved truth, the highest possible value of $A_T(c)$ is 1. Hence, we define *measurement error* as the shortfall from perfect agreement:

$$\varepsilon_c \;\equiv\; 1 - A_T(c). \tag{2}$$

To operationalize how configuration changes translate into measurable performance differences, we introduce a central unit of analysis, the *contrast*, which isolates the effect of modifying a single component of c, defined below:

[3] In our analytical derivations, we use accuracy (proportion of matches) as our agreement measure because it maps clearly to the standard label-noise model detailed below. In empirical validation, we report Cohen's Kappa for interpretability. The qualitative conclusions hold for any agreement statistic that increases monotonically as task loss on the pairwise local comparisons in our study.

Key Definitions

A *contrast* is a pairwise comparison between two LLM-annotation configurations that are identical except for one component in $\{G, H, P, M\}$.
A *reference agreement contrast* is a contrast measured on reference agreement A_R.

Intuitively, a positive reference agreement contrast indicates the existence of a "better" configuration; we now derive the precise condition under which such a positive contrast corresponds to a reduction in measurement error.

3.1.1. Relation between reference agreement contrast and measurement error reduction. Measurement error concerns true agreement (i.e., agreement with the unobserved ground truth), but we only observe agreement with a reference produced by a particular human source. We therefore need to relate the observable A_R to the target A_T. To this end, we adopt two assumptions. The first is a standard assumption of symmetric K-class random classification noise (Natarajan et al. 2013).

ASSUMPTION 1 **(Symmetric label noise)**. *Let $K \geq 2$ be the number of classes and let $e \in [0, 1)$ be the (class-independent) error rate of the human reference. For every item with truth $y \in \{1, \ldots, K\}$,*

$$\Pr(r = y \mid y) = 1 - e, \qquad \Pr(r = \ell, y \neq \ell \mid y) = \frac{e}{K-1} \quad \textit{for each } \ell \in \{1, \ldots, K\} \setminus \{y\}.$$

We also assume the independence between reference baselines and prompt/model configurations:

ASSUMPTION 2 **(Prompt/Model-independent reference)**. *For fixed items x and (G, H), the human reference labels do not depend on which meta-prompt or model (P, M) is being evaluated:*

$$r \perp (P, M) \mid (G, H).$$

This assumption holds by design in our proposed workflow: the guideline G is developed independently of any specific (P, M); annotators use G to label a fixed D; and they never see which LLM produced $\hat{y}^{(c)}$. Importantly, $\hat{y}^{(c)}$ and r may be correlated conditional on the true value y, resulting in the LLM and the reference to share error modes (i.e., both could err identically).

This creates a challenge: reference agreement A_R can improve for two reasons: (1) the LLM actually gets closer to truth, or (2) the LLM starts making the same mistakes as the human reference. To address this challenge, we formalize this distinction by introducing the concept of co-labeling covariance. Specifically, we formalize the relationship between reference agreement, true agreement, and co-labeling covariance in the following lemma (proof in Appendix A.4):

LEMMA 1. *Under Assumptions 1 and 2, for any configuration c,*

$$A_R(c) \;=\; aA_T(c) + b + J(c),$$

where $a = (1-e) - \frac{e}{K-1}$, $b = \frac{e}{K-1}$, $J(c) = \mathbb{E}_y\Big[\sum_{\ell=1}^{K}\Big(\Pr(\hat{y}^{(c)} = \ell, r = \ell \mid y) - \Pr(\hat{y}^{(c)} = \ell \mid y)\Pr(r = \ell \mid y)\Big)\Big]$. *This implies that any rise in reference agreement* $A_R(c)$ *equals a linear contribution from true accuracy* $A_T(c)$ *(through a and b) plus a covariance term* $J(c)$.

Here, $J(c)$ measures the co-labeling covariance between the LLM and human reference. It captures how often the LLM and the reference assign the *same* label (either both correct or both incorrect with the same wrong label), beyond what would occur under conditional independence given the truth. Positive $J(c)$ is problematic when the LLM and the reference make the *same* errors (assigning the same wrong class), inflating $A_R(c)$ without improving $A_T(c)$. These common misclassifications occur when systematic, ambiguous guidelines cause *both* the LLM and the reference to the same wrong class, which occurs when tasks are inherently complex or the guidelines are not prescriptive.

Because a, b depend only on (G, H) and on the item distribution (and not on (P, M)), for any contrast c_1, c_0 that keeps (G, H) fixed,

$$A_R(c_1) - A_R(c_0) \;=\; a\big(A_T(c_1) - A_T(c_0)\big) \;+\; \big(J(c_1) - J(c_0)\big). \tag{3}$$

Equation (3) shows that the *reference agreement contrast* decomposes into the truth-level improvement plus the change in the co-labeling covariance J. Rearranging Equation (3) isolates the corresponding change in measurement error, which we formalize in the following proposition:

PROPOSITION 1. *Under Assumptions 1 and 2, the difference in measurement error between the contrast* c_1 *and* c_0 *is:*

$$\varepsilon_{c_1} - \varepsilon_{c_0} \;=\; -\big(A_T(c_1) - A_T(c_0)\big) \;=\; -\tfrac{1}{a}\Big(\big(A_R(c_1) - A_R(c_0)\big) - \big(J(c_1) - J(c_0)\big)\Big). \tag{4}$$

Hence, holding the human reference fixed, a positive reference agreement contrast $\big(A_R(c_1) - A_R(c_0)\big)$ reduces *measurement error if it exceeds the increase in co-labeling covariance* $\big(J(c_1) - J(c_0)\big)$.

Proposition 1 reveals that observed improvements translate to measurement error reduction when the change in co-labeling covariance is sufficiently small. However, determining whether co-labeling covariance is sufficiently small requires empirical validation of two critical aspects. First, we need to empirically compare expert versus crowd baselines to identify which procedure produces more stable identification (which we do in Section 5.2). Second, when reference agreement $A_R(c)$ is low, we need to establish empirical cutoffs beyond which reference agreement contrasts reliably indicate measurement error reduction. Our analysis in Section 5.2 also identifies a conservative threshold ($\kappa \geq 0.5$) where J-induced confounding becomes minimal, providing guidance for when LLM annotation achieves sufficient quality for research applications.

3.1.2. Designing contrasts. We detail how we design the contrasts to study the impact of each element of $c = (G, H, P, M)$. Each component requires a different comparison protocol: changing guidelines (G) or baselines (H) affects the human reference itself, whereas changing prompts (P) or models (M) changes only the generated output being evaluated.

(i) Guideline changes via anchored evaluation. We select a preferred guideline $G^\star$ as the canonical construct definition and build the human reference under $G^\star$ (holding H fixed). We then compare labels produced under competing guidelines against this anchored reference baseline:

$$\widehat{A}_R^{[G^\star]}(G, P, M; H) \;\equiv\; \Pr\big(\hat{y}^{(G,P,M)} = r^{(G^\star,H)}\big).$$

We define the guideline contrast as

$$\Delta^{(G)} \;\equiv\; \widehat{A}_R^{[G^\star]}(G_1, P, M; H) \;-\; \widehat{A}_R^{[G^\star]}(G_0, P, M; H). \tag{5}$$

This anchoring is *asymmetric* by design: it evaluates which guideline drives the LLM to better align with the construct as instantiated by $G^\star$. Using a single anchor ensures stable target when comparing guidelines. In practice, we designate the iteratively refined guidelines as $G^\star$, treating it as the canonical construct definition. This choice is motivated by the annotation science literature, which establishes that iterative refinement converges toward more consistent interpretations (Lee et al. 2023). Given this anchor, the contrast then evaluates whether one-shot guidelines can achieve comparable alignment with this established standard. Interpretation follows (3), where improvements in $\widehat{A}_R^{[G^\star]}$ reflect truth-level gains *net of* any change in co-labeling covariance with the anchored reference. We expect that refined guidelines (G_1) are clearer about the annotation criteria compared to the base guidelines (G_0), which plausibly decreases the probability of LLM and human reference both being wrong, so $J(c_1) - J(c_0) \leq 0$ is reasonable. In this case, $\Delta^{(G)}$ is conservative as it underestimates the improvement in A_T.

(ii) Human procedure selection via internal consistency. For a fixed $G^\star$, let $\widehat{\mathrm{IAA}}(r^{(G^\star,H)})$ denote within-procedure inter-annotator agreement (IAA) computed on D (e.g., Krippendorff's α or mean Cohen's κ over annotator pairs). We define

$$\Delta_{\mathrm{IAA}}^{(H)} \;\equiv\; \widehat{\mathrm{IAA}}\big(r^{(G^\star,H_1)}\big) \;-\; \widehat{\mathrm{IAA}}\big(r^{(G^\star,H_0)}\big). \tag{6}$$

Under standard reliability assumptions (symmetric K-class noise and conditionally independent annotator errors within a procedure), higher IAA implies a lower per-annotator error rate e, which yields a more accurate reference aggregate (Krippendorff 2004).[4] Under Assumption 1, this yields

[4] Since $a = (1 - e) - e/(K - 1)$ from Assumption 1, lower e yields a larger identification slope, making observed LLM-reference contrasts more faithful to truth-level contrasts.

a larger identification slope for H and a reference more closely aligned with the unobserved truth. Intuitively, when annotators using $(G^\star, H)$ agree more with one another, their majority-vote aggregation better approximates y, making reference agreement contrasts better approximate reduction in measurement error.

IAA is not sufficient to guarantee validity of downstream comparisons. IAA does not capture whether the model and the reference tend to make the same mistakes on the same items, which is captured by co-labeling covariance J. Even with good IAA, a large positive "both-incorrect" component can inflate reference agreement without any improvement against the ground truth. This scenario is especially plausible for crowd references for two reasons. First, for complex tasks, many crowd workers and current LLMs may lack domain knowledge that experts apply. Second, when guidelines are under-specified, experts draw on tacit annotation criteria developed through guideline development process that are not fully captured in the written guidelines. By contrast, crowdworkers and LLMs, prompted only with written guidelines, do not apply these tacit rules and consequently may make similar classification errors.

We therefore use $\Delta^{(H)}_{\text{IAA}}$ as the primary criterion to choose $H^\star$: pick the human procedure with the higher IAA. However, co-labeling covariance J remains unobserved. To assess whether expert baselines provide a more stable identification than crowd baselines, we perform a sensitivity analysis in Section 5.2: we examine how introducing crowd labels into expert baselines affects reference agreement contrasts.

(iii) Prompt and model changes under identification. With $(G^\star, H^\star)$ fixed, we have the following two contrasts:

$$\Delta^{(P)} \equiv \widehat{A}_R(G^\star, P_1, M; H^\star) \;-\; \widehat{A}_R(G^\star, P_0, M; H^\star), \tag{7}$$

$$\Delta^{(M)} \equiv \widehat{A}_R(G^\star, P, M_1; H^\star) \;-\; \widehat{A}_R(G^\star, P, M_0; H^\star). \tag{8}$$

According to Proposition 1, a positive contrast implies improvement in true agreement when $A_R(c_1) - A_R(c_0) \;>\; J(c_1) - J(c_0)$. In other words, with the reference held fixed, a superior P or M (as indicated by a higher A_R) implies a better alignment with y and a reduced measurement error when the change in co-labeling covariance is sufficiently small. We argue that the co-labeling covariance being sufficiently small is plausible for two reasons. First, with $G^\star$ and $H^\star$ fixed, any residual shift in J arises solely from the LLM side and should be small. Second, when prompts and models improve true agreement, they are more likely to reduce "both incorrect" component of J, making $J(c_1) - J(c_0) \leq 0$ more plausible than > 0. Hence the observed A_R gains are conservative as they tend to *understate* the improvement in A_T.

3.1.3. Sequential optimization strategy. Empirically, we adopt a sequential approach on a fixed item set D, starting from a baseline configuration (M_0, P_0), where M_0 is a widely-adopted model (e.g., GPT-4o) and P_0 embeds guidelines in a standard prompt structure. First, holding (H, P, M) fixed, we compare candidate guidelines using the anchored-contrast in Equation (5) and select $G^\star$. Second, we compare human procedures via $\Delta^{(H)}_{\text{IAA}}$ in Equation (6) to select $H^\star$. As the final two steps, with $(G^\star, H^\star)$ fixed, we evaluate prompts and models using the reference agreement contrasts in Equations (7)–(8), respectively.[5] We adopt this sequential design for two practical reasons. First, it surfaces robust, model-agnostic principles about implementation details (e.g., always placing guideline content in the system role, as empirically shown in Table 2). Second, exhaustive joint optimization over all (G, H, P, M) combinations is methodologically cumbersome, computationally prohibitive and, therefore, practically unusable. A stable reference $(G^\star, H^\star)$ ensures the identification in Equation (4) applies cleanly when interpreting contrasts.

3.2. Measurement Error Reduction Mechanisms

Our design approach, which we call SILICON (**S**ystematic **I**nference with **L**LMs for **I**nformation **C**lassification and **N**otation) systematically isolates and quantifies the four sources of measurement error identified in Section 3.1. The design follows three core principles.

1. We isolate individual error sources through controlled comparisons that hold other factors constant, enabling attribution of performance differences to specific methodological choices.
2. We maintain verbatim guideline integrity across all treatments to ensure that observed differences reflect genuine methodological improvements rather than instruction variations.
3. We establish statistical equivalence criteria for robust reproducibility, addressing the model retirement problem that threatens long-term research validity.

Appendix Figure C1 illustrates our four-stage protocol, where each stage targets a specific measurement error source. The architecture enables both sequential implementation for optimal error reduction and modular application based on existing research infrastructure. Details of each component's theoretical motivation, protocol design, and measurement strategy are below.

3.2.1. Preliminary: measuring agreement. Evaluating LLMs against human baselines requires a performance metric that captures annotation agreement and supports IAA analysis. Several established metrics exist for evaluating IAA, such as Cohen's Kappa (Cohen 1968), Fleiss' Kappa (Fleiss 1971), and Krippendorff's Alpha (Hayes and Krippendorff 2007). The choice of metric often depends on the characteristics of the annotation task and no single metric fits all contexts.

[5] This ordering reflects two dependencies: (i) guideline selection must precede baseline creation since annotators apply the guidelines, and (ii) baseline establishment must precede prompt/model optimization to ensure a stable evaluation reference for identification.

We consistently use Cohen's Kappa (κ), which measures the level of agreement between two annotators while adjusting for chance agreement (Cohen 1968). Formally, it is given by: $\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o represents the relative observed agreement among raters, and P_e denotes the hypothetical probability of agreement by chance. We choose this metric for three main reasons. First, Cohen's Kappa is designed for pairwise agreement analysis, which aligns with our objective of comparing a single LLM against the human baseline. Second, it can be extended to scenarios involving multiple annotators through calculation and averaging of pairwise Kappa scores. This is useful during the phases of annotation guideline development and human baseline establishment, where evaluation typically involves more than two human annotators. Third, it can be adapted to multi-label classification tasks, where items may belong to multiple categories simultaneously.[6]

Ideally, the agreement metric should be as high as possible, but practical constraints often lead researchers to predetermine a threshold as an acceptable target. Here, we do not argue for the universal applicability of any specific threshold, as such thresholds reflect a trade-off between rigor and feasibility and are domain-specific (Paun et al. 2022). A commonly referenced threshold suggests that values above 0.4 indicate "moderate agreement" (Landis and Koch 1977). We set a target Cohen's Kappa threshold of 0.5, which incorporates a conservative buffer of 0.1 to account for potential variability and uncertainty in determining moderate agreement. As discussed in Section 3.1, this threshold also serves as an empirical cutoff where co-labeling covariance effects are low.

3.2.2. Guideline-induced error reduction. Ambiguous annotation criteria create systematic measurement error as annotators apply inconsistent interpretations to edge cases. For LLM annotation, this problem becomes more acute because models cannot seek clarification during annotation, making explicit guidelines the sole mechanism for consistent interpretation (Törnberg 2024). Extant literature acknowledges the importance of detailed guidelines for human annotation tasks (Sheng et al. 2008, Lee et al. 2023) and shows that iterative refinement improves consistency (Hovy and Lavid 2010). Yet many LLM annotation studies adopt existing guidelines directly or draft new ones in a one-shot fashion (Gilardi et al. 2023, Rathje et al. 2024), risking guideline-induced error.
Protocol design. We compare iteratively refined guidelines against one-shot guidelines adapted from existing research. The iterative refinement follows a structured protocol based on established annotation science (Lee et al. 2023), where multiple annotators independently apply preliminary definitions, discuss disagreements through documented conversations (e.g., recorded Zoom call), and refine guidelines until achieving predetermined IAA thresholds. Figure 2 presents our guideline development protocol. In brief, annotators independently annotate initial samples, measure IAA

[6] By incorporating context-dependent weighting matrices during the calculation, the weighted Cohen's Kappa metric enables assessment of agreement across multi-label classification tasks. We provide detailed specifications of this adaptation in Appendix A.1.

against predefined thresholds, conduct documented discussions focusing on disagreements, and iterate until achieving consistent agreement on new samples. This process continues until annotators achieve the agreement threshold on first pass with a fresh subsample of data.

Figure 2 Iterative Process to Create Annotation Guidelines

Iterative Annotation Process Flow

1. RAs independently annotate an initial sample dataset based on preliminary definitions
2. Measure IAA and compare it with the predefined threshold
3. If IAA threshold is not met:
   - Conduct documented discussions focusing on disagreements
   - Reshuffle and re-annotate the *same* sample independently
   - Repeat until threshold is met
4. Once threshold is met, proceed to annotate a *new* sample
5. Iteration concludes when RAs achieve the IAA threshold on first pass with a new sample
6. RAs independently draft annotation guidelines and then collaboratively merge them.

Contrast of Interest: Guideline Effect

With (H, P, M) fixed and evaluation anchored at $G^\star$, replacing one-shot guidelines G_0 by iteratively refined ones G_1 increases agreement against the anchored reference (Equation (5)):

$$\Delta^{(G)} > 0.$$

Measurement strategy. We operationalize guideline-induced error reduction through Cohen's κ comparisons between iteratively refined and one-shot guidelines, holding constant the evaluation baseline,[7] meta-prompt structure, and model selection.

3.2.3. Baseline-induced error reduction. Human annotation baselines serve as the reference standard against which LLM annotations are evaluated, making baseline quality crucial for valid performance assessment. The choice between expert and crowdsourced annotators involves fundamental trade-offs. Domain experts bring specialized knowledge and can apply nuanced judgment (Snow et al. 2008), but are costly to recruit and may introduce systematic biases when their shared training leads to correlated errors (Aroyo and Welty 2015). Crowdsourced workers offer scalability and diverse perspectives (Ipeirotis et al. 2014), but typically receive minimal training and lack deep understanding of annotation guidelines (Snow et al. 2008, Zhang et al. 2023), creating measurement error in the evaluation process itself.

Protocol design. A key challenge in getting expert-generated baselines is the infeasibility of recruiting domain experts in many scenarios, due to time and cost considerations. The SILICON

[7] We do not differentiate between expert- and crowd-based implementations for this step; the development of expert baselines is detailed in Section 3.2.3.

workflow addresses this challenge by creating task-specific "annotation experts" through its iterative and collaborative prescriptive guideline development process. The task-specific "annotation experts" are individuals who participated in the iterative guideline development process and demonstrated consistent understanding of task definitions through high IAA. Having internalized consistent interpretations through the collaborative development process, they reduce evaluation-driven measurement error by minimizing disagreements about ambiguous cases that could confound LLM performance assessment. The protocol compares expert-generated baselines against crowdsourced baselines using identical annotation guidelines and evaluation protocols.

Contrast of Interest: Baseline Effect

Under fixed $(G^\star, M, P)$, annotation expert procedures (H_1) exhibit higher within-procedure agreement than crowdsourced worker procedures (H_0) (Equation (6)):

$$\Delta_{\mathrm{IAA}}^{(H)} > 0.$$

Measurement strategy. We assess baseline quality through internal consistency (mean κ among annotators). Sample sizes ranging from 120 to 200 items provide adequate statistical power while maintaining practical feasibility (Lee et al. 2023).

3.2.4. Prompt-induced error reduction. Meta-prompt structure affects LLM annotation performance even when guidelines remain verbatim constant, creating prompt-induced measurement error. To ensure fair comparisons between LLMs and human annotators, annotation guidelines should be incorporated verbatim into LLM prompts, with optimization focused solely on meta-prompt elements. Prior work suggests structure and reasoning strategies can influence outputs (Törnberg 2024, Wei et al. 2022).

Protocol design. We implement controlled meta-prompt optimization while preserving guideline integrity. Optimization focuses on three elements: (1) strategic positioning of annotation guidelines within prompt structure (system prompt versus user/human prompt), (2) adoption of task-relevant personas to align model perspective with domain expectations (Hu and Collier 2024), and (3) incorporation of reasoning strategies such as chain-of-thought prompting (Wei et al. 2022).

The experimental protocol maintains strict separation between guideline content and meta-prompt elements. Guidelines are incorporated verbatim into all prompt variations, ensuring that performance differences reflect meta-prompt optimization rather than instruction changes. This design principle enables fair comparison between human and LLM annotation performance.

Contrast of Interest: Prompt Effect

For fixed (M, G, H), an improved meta-prompt reduces measurement error and raises reference agreement (Equation (7)):

$$\Delta^{(P)} > 0.$$

Measurement strategy. We quantify prompt-induced error reduction by comparing optimized meta-prompts against baseline prompt structure, holding guidelines, human baselines, and model selection constant. The contrast identifies corresponding improvements with respect to truth by Equation (4).

3.2.5. Model-induced error assessment. Different LLMs exhibit varying annotation performance due to architectural differences, training data variations, and alignment strategies (Mousavi et al. 2024). Model-induced error arises from this heterogeneity. Our approach addresses a critical question: *do annotation results remain stable across different models?* If outcomes are consistent across multiple high-performing models, then downstream analyses do not depend on idiosyncrasies of single model implementations. This establishes flexibility in model selection based on accessibility, cost, or computational requirements while maintaining annotation quality.

Protocol design. We evaluate multiple state-of-the-art LLMs using identical guidelines, baselines, and meta-prompts to isolate model-specific performance differences. The evaluation includes both proprietary models (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro) and open-parameter ones (GPT-OSS-120B, LLaMA 3.3 70B, DeepSeek R1) to address reproducibility concerns (see Section 3.3).

Contrast of Interest: Model Effect

For fixed (P, G, H), model changes yield non-zero contrast (Equation (8)):

$$\Delta^{(M)} \neq 0.$$

Measurement strategy using statistical equivalence testing. We employ a regression-based method to test for statistical equivalence[8] in model outputs. Since evaluation samples are randomly drawn from the population of interest, models producing statistically indistinguishable results on representative samples maintain similar performance when scaled to larger datasets.

Each model represents a distinct treatment, with item-treatment level annotations as the unit of analysis. The dependent variable indicates agreement between model annotation and human ground truth (binary). We perform logistic regression with treatment indicators, clustering standard errors at the item level to account for repeated-measures structure. The regression equation is presented

[8] Statistical equivalence is defined as producing annotation outcomes that do not differ significantly in agreement metrics relative to human baselines.

in Section 5.5, with more details provided in Appendix A.2. When observed contrasts are positive and the change in co-labeling covariance is sufficiently small, Equation (4) implies strictly reduced measurement error. Beyond identifying which models perform best, Section 3.3.2 addresses when model confidence warrants generating multiple labels per instance rather than relying on a single model's output.

3.3. Robust Reproducibility Protocol

3.3.1. Model confidence assessment. Model confidence serves as a proxy for annotation reliability, enabling cost-sensitive strategies where uncertain predictions receive additional scrutiny. Unlike human annotators who can express uncertainty explicitly, LLMs require post-hoc confidence estimation methods. High-confidence predictions typically correlate with higher accuracy, making confidence assessment crucial for identifying when additional model votes or human review may be warranted (Lyu et al. 2025)

We evaluate intrinsic model reliability using the First-Second Distance (FSD) method (Lyu et al. 2025). FSD measures the discrepancy between the most common and second most common outputs across multiple generations from the same model. For each item, we generate five independent responses using temperature = 1, then calculate the difference between the most frequent response and the second most frequent response.[9] A larger FSD indicates higher confidence, suggesting the model's responses concentrate on a single answer across multiple runs.

3.3.2. Multi-model annotation strategy. Following Ipeirotis et al. (2014), adding multiple annotators is most valuable when existing labels show uncertainty. Running multiple models on every item yields high-quality labels but can be prohibitively expensive. Conversely, relying on a single model is cost-effective but error-prone. The optimal strategy balances cost against quality by strategically deploying multiple models based on confidence indicators.

Protocol design. The threshold selection strategy depends on three factors: budget constraints, quality requirements, and task-specific error costs. Researchers should first establish whether their focal model is a cost-effective option (typically open-parameter) or a high-performing proprietary model, as this determines the appropriate threshold range.

We recommend empirical calibration using the expert-annotated baseline as the validation sample: compute FSD distributions for your focal model, test 2-3 threshold values, and select the threshold that achieves acceptable quality improvements within budget constraints. The optimal threshold balances marginal quality gains against incremental annotation costs.

Measurement strategy. We evaluate threshold performance by varying τ across [0,1] interval and measuring both quality improvements and cost implications. Quality is assessed through Cohen's

[9] Here we obtain FSD scores by sampling; Appendix A.3 details how to compute FSD from token-level log probabilities.

Kappa against human ground truth, comparing single-model performance against multi-model performance triggered at different thresholds. Cost is operationalized as the fraction of items requiring additional model labels $q(\tau)$. The analysis identifies threshold ranges where marginal quality gains justify incremental costs.

3.3.3. Open-parameter model substitution for robust reproducibility. Proprietary model retirement threatens research reproducibility, as findings based on deprecated models cannot be replicated by future researchers. Open-parameter models provide permanent accessibility but may exhibit different performance characteristics than proprietary alternatives. Establishing statistical equivalence between open-parameter and proprietary models enables reproducible annotation pipelines that maintain quality while ensuring long-term availability.

Protocol design. We test whether high-performing open-parameter models can substitute proprietary models without significant quality loss. Using our regression-based equivalence testing framework, we compare open-parameter models (GPT-OSS-120B, LLaMA 3.3 70B, DeepSeek R1) against proprietary baselines (GPT-4o, Claude 3.5 Sonnet) across all validation tasks.

Measurement strategy. Statistical equivalence is established when the confidence intervals for the performance difference encompass zero, thus indicating that there is no significant quality gap. Models achieving equivalence provide reproducible alternatives to proprietary options, enabling long-term research sustainability despite model retirement cycles.

3.4. Toolkit and Implementation

We have created an open-source toolkit that operationalizes SILICON, which we will release upon publication to respect the double-blind process. The toolkit includes utilities for agreement calculation for guideline refinement and baseline establishment; automated pipelines to compare different meta-prompt settings and multiple models; modules for model-confidence estimation and multi-LLM annotation; and statistical tests for robust reproducibility and equivalence.

4. Empirical Validation Design

We validate the SILICON workflow across seven annotation cases that are representative scenarios in management research. Our evaluation design employs two categories of validation approaches based on the level of methodological control achievable for each error source.

4.1. Primary Cases with Full Methodological Control.

The first category comprised four cases: business proposal evaluation, review attribute detection, dialog intent classification and breakdown analysis, and connectedness presence detection and classification. For these scenarios, we implemented the complete SILICON workflow, developing the annotation guidelines iteratively and establishing annotation-expert-validated human baselines

with three research assistants per task. The second category comprised three supplementary scenarios: language toxicity detection, criticism stance detection, and sentiment analysis. For this category, we adopted annotation guidelines and human baselines directly from published papers. Whereas this approach offers valuable comparative insights, it presents limitations-particularly our inability to verify the original IAA or whether the baselines were derived from expert or crowdsourced annotations. Consequently, we tread these scenarios as supplementary evidence, drawing our primary conclusions from the first four scenarios. Appendix B presents a summary of original research questions, annotation approaches, SILICON process implementation, and key findings for each case, and Online Appendix presents corresponding prompts.[10]

4.1.1. Business proposal evaluation. Business documentation analysis, particularly of proposals, crowdfunding initiatives, and corporate statements, plays a critical role in conveying business-related information and is essential for analyzing organizational decision making and strategic behavior (Cao et al. 2019). In this case study, we focus on the classification of decentralized autonomous organization (DAO) business proposals, building on the work of Obermeier and Mayya (2024). This classification task involves categorizing user "proposals" based on their primary function or intention. Each proposal was categorized into one of six categories: Organizational, Business Model, Marketing, Functionality, Security, or Other. The categorization process aims to reveal patterns in how DAOs delegate and structure decision rights across their platforms.

4.1.2. Review attribute detection. Online reviews contain critical information that help the understanding of consumer purchase decisions. We obtained the annotated dataset from Liu et al. (2019), where authors used a large scale deep learning approach to identify the presence of six attribute in the product reviews, including price, performance feature, relatability/durability, conformance, aesthetics, and perceived quality. This is a multi-label classification task which provides insights into how different product attributes influence consumer decision-making.

4.1.3. Dialog intent classification and breakdown analysis. Dialog analysis has been critical in various domains, particularly in understanding and improving human-computer interaction. We focus on two critical aspects of dialog analysis: dialog intent classification (Stolcke et al. 2000) and breakdown analysis (Higashinaka et al. 2016), using a sample of customer service dialog data obtained from a Hugging Face dataset.[11] The first task, dialog intent classification, requires human annotators to classify the intention of each utterance into one or multiple categories (i.e., multi-label classification): factual questions, yes/no questions, task commands, invalid

[10] Link to Online Appendix: `https://drive.google.com/file/d/14sFKxeSANFX3Xl-umVxGNR_6qKZ5_M5Y/view?usp=sharing` (This file is anonymous and non-tracking).

[11] https://huggingface.co/datasets/bitext/Bitext-customer-support-llm-chatbot-training-dataset

commands, appreciations, complaints, comments, statement non-opinions, positive answers, and negative answers. The second task, the dialog breakdown analysis, focuses on identifying whether an utterance in the conversation leads to a breakdown, classifying each utterance as: (1) not a breakdown, (2) a possible breakdown, or (3) confirmed breakdown.

4.1.4. Affective content evaluation (connectedness presence and classification). Affective content in marketing—such as in advertisements, product packaging, and product descriptions—plays a significant role in shaping consumer decision-making (Paharia et al. 2011). Evaluating affective content is therefore critical for understanding consumer behavior. In this study, we focus on one specific dimension of affective content: the sense of connectedness. We use a dataset from Cheng et al. (2024), which investigates the presence and nature of connectedness in online product descriptions. The first task, connectedness presence detection, is a binary classification task where human annotators assess whether a product description conveys a sense of connectedness or not. The second task, connectedness classification, is conducted only when connectedness is present. This is a multi-label classification task where each product description may be assigned one or more of the following categories: Emotional Appeal, Storytelling, Relatable Scenarios, Empathy, and Cultural or Community Connection.

4.2. Complementary Cases with Limited Methodological Control.

4.2.1. Language toxicity detection. Detecting toxic content in textual data plays a critical role in management research, offering insights into a range of organizational and social dynamics. Toxic interactions, such as hate speech or fear-based messaging, can significantly impact online communities (Matook et al. 2022, Sibai et al. 2024), and organizational behavior (Rosette et al. 2013). For this task, we utilize the pre-existing annotation guidelines, human baseline, and raw data publicly available from Saha et al. (2023). The annotated items are social media posts from Gab. The task is to mark each post as a) fear speech, b) hate speech, c) normal, or d) both fear speech and hate speech (if the post contains elements of both).

4.2.2. Criticism stance detection. We selected this case because the core task—criticism stance detection—can be generalized to address a broad range of organizational and social challenges. Specifically, we use data from Peng et al. (2022), which focuses on detecting the criticism stance of tweets about certain academic papers, framing it as a binary classification task.

4.2.3. Sentiment analysis. Sentiment analysis plays a pivotal role in management research, with applications across various contexts such as social media platforms (Zhang et al. 2016, Oh et al. 2023, Tang et al. 2024), e-commerce platforms (Homburg et al. 2015), and firm-public communications (Choudhury et al. 2019). In this study, we assess whether text annotations generated by LLMs can produce similarly well-aligned results. For this task, we utilize annotation guidelines, human baselines, and raw data from Sen et al. (2020).

4.3. Prompt and Model Specifics

The prompt strategies included a base prompt, a prompt with persona settings, and a prompt with chain-of-thought, all unified in their placement within the system role rather than the user role (for reasons that will appear in Section 5.3). These prompts were initially tested using GPT-4o to establish a benchmark. Based on the results from GPT-4o, we identified the optimal prompt by balancing performance and cost considerations. This optimal prompt was then applied across several models: GPT-4o, o3-mini, Claude 3.5 Sonnet, Gemini 1.5 Pro, LLaMA 3.3 70B, DeepSeek-R1, and GPT-OSS-120B.[12] Acknowledging that the best-performing prompt might vary between models, we further examined this variability in the business proposal evaluation task. Specifically, we tested all three prompt strategies, altering their placement (system role versus user role) to generate six distinct prompts. These prompts were applied to a selective subset of four models. We set the temperature at 1 across all models, aligning with the default value for GPT-4o.

5. Empirical Validation Results

We present results in the same sequence as described in Section 3.1.3: guideline contrast, reference baseline contrast, meta prompt contrast, and model contrast. We then evaluate robust reproducibility protocols and multi-LLM annotation strategies. Our empirical validation demonstrates systematic measurement error reduction across all four sources, with each intervention targeting distinct aspects of the annotation process.

5.1. Reducing Guideline-induced Error

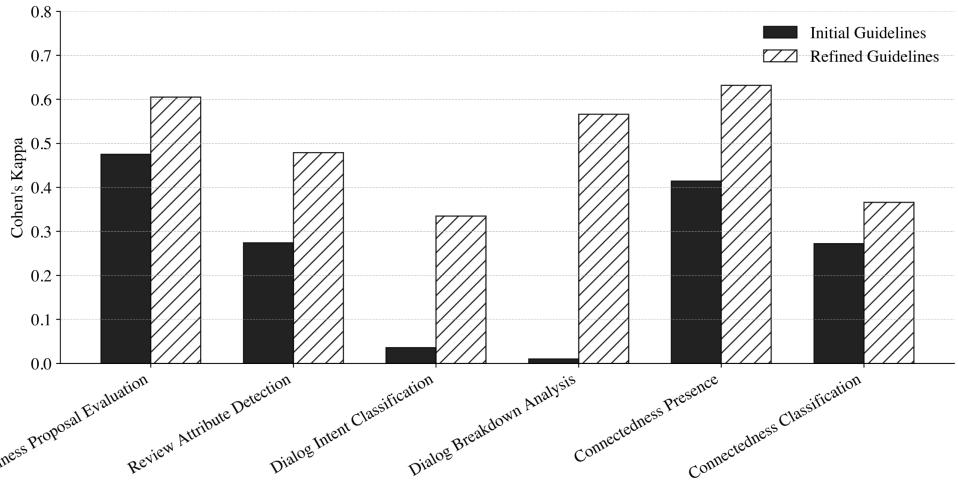
Ambiguous or incomplete annotation guidelines can create systematic measurement error when annotators apply inconsistent interpretations. We examine whether iterative guideline refinement reduces this source of error by comparing LLM annotation performance using one-shot (pre-existing) guidelines versus iteratively refined ones developed through our structured protocol. Figure 3 presents Cohen's Kappa values between LLM annotations and expert-generated ground truth across six tasks. LLM performance was consistently lower when using one-shot guidelines compared to iteratively refined guidelines. The improvement ranged from 0.1–0.5 Kappa points, with the largest gains observed in dialog intent classification and breakdown analysis tasks, where initial agreement levels below 0.1 (indicating no meaningful agreement) improved to fair agreement level after guideline refinement.

These findings provide strong support for the argument that holding human baseline procedures (H), meta-prompts (P), and models (M) constant while anchoring evaluation at refined guidelines ($G^* = G_1$), the transition from G_0 to G_1 increases reference agreement ($\Delta^{(G)} > 0$). Appendix Figure

[12] We used the following versions of models: GPT-4o-2024-08-06, o3-mini-2025-01-31, Claude-3-5-Sonnet-20240620, Gemini-1.5-pro, Llama-3.3-70B-Instruct, DeepSeek-R1, and GPT-OSS-120B.

C3 illustrates the gains from iterative refinement process, showing that IAA among annotation experts rises consistently over multiple revision rounds. Achieving high IAA typically requires several iterations for complex tasks, supporting our workflow that prescriptive guidelines must be systematically developed rather than created in isolation.

Figure 3 LLM Performance based on One-shot vs. Iteratively Refined Annotation Guidelines



Notes. The y-axis shows the κ between LLM annotation and human annotation baselines (i.e., the majority votes from experts). The LLM used is GPT-4o. LLM performance based on one-shot annotation guidelines is always worse than that based on iteratively refined annotation guidelines.

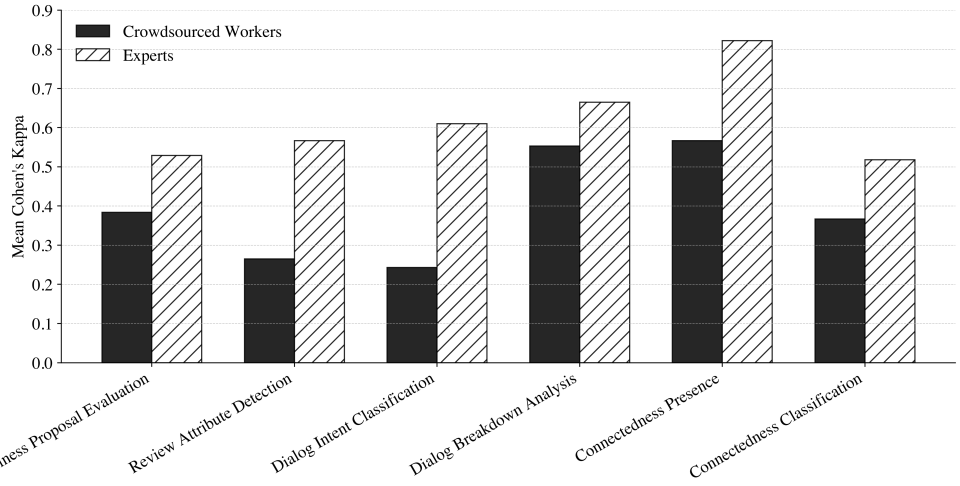
5.2. Reducing Baseline-induced Error

Unreliable human reference standards introduce systematic baseline-induced error when the evaluation standard itself diverges from ground truth. We test whether expert-generated baselines reduce this source of error by comparing the internal consistency of annotation experts versus crowdsourced workers using identical annotation guidelines.

For each of the four primary cases (comprising six sub-tasks), we constructed two distinct annotation baselines using our finalized annotation guidelines. The expert baseline was obtained from three undergraduate research assistants who participated in the iterative guideline development process, while the crowdsourced baseline was obtained from three different research assistants who received the guidelines without participating in their development. Both groups annotated 200 to 300 randomly chosen documents using identical protocols. Figure 4 presents the IAA comparisons between the two baseline procedures. The annotation experts consistently achieved higher consensus compared to crowdsourced annotators across all six tasks. For instance, in the dialog intent classification task, the mean Cohen's κ among expert annotators was 0.61 (indicating substantial agreement), whereas crowdsourced annotators achieved only 0.23, reflecting weaker alignment and less consistent understanding of the annotation criteria.

These findings provide strong support for the argument that under fixed guidelines (G^*) and models/prompts (M,P), expert annotation procedures (H_1) exhibit higher within-procedure agreement than crowdsourced procedures (H_0) yielding ($\Delta^{(H)}_{IAA} > 0$). The superior consistency occurs because the iterative refinement process allows experts to develop shared mental models of the task, facilitating more consistent interpretations of complex phenomena (Ruggeri et al. 2024). This higher baseline reliability provides a more demanding and stable evaluation standard to assess LLM annotation quality.

Figure 4 Inter-Annotator Agreement Comparison: Expert- and Crowdsourced Worker-labeled Baseline

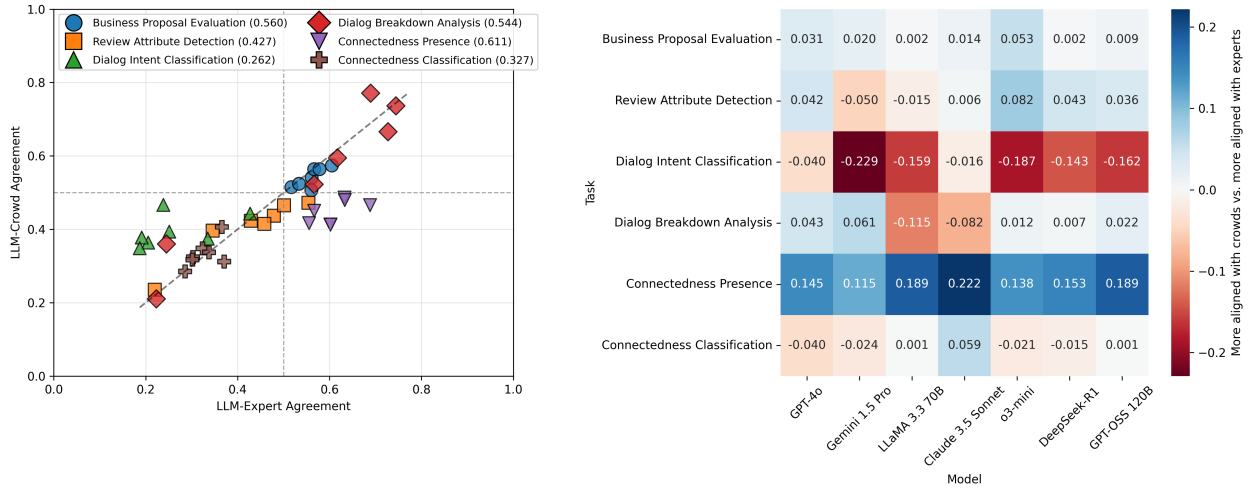


Notes. The y-axis shows mean κ among three annotators (either annotation experts or crowdsourced workers). Expert-labeled annotations consistently yield higher agreement across all tasks.

We next compare how well LLM annotation align with annotation-experts versus crowdsourced baselines. Figure 5 presents this comparison through two complementary views. The left panel plots each model-task combination as a point with coordinates ($x = \kappa(\text{LLM}, \text{expert})$, $y = \kappa(\text{LLM}, \text{crowd})$). The diagonal indicates equal agreement; points *below* the diagonal indicate stronger alignment with experts, and points *above* it indicate stronger alignment with crowds.

Two insightful patterns emerge from this scatter plot. First, alignment is not uniformly expert-leaning: for instance, in *Dialog Intent Classification* (green triangles), most points lie above the diagonal, implying that LLMs agree more with the crowd than with experts. Within our workflow, this pattern is consistent with positive co-labeling covariance J (Lemma 1): LLMs and crowd workers tend to make the same mistakes, inflating reference agreement A_R without improving true agreement A_T. Second, when agreement reaches acceptable level ($\kappa \geq 0.5$ marked by dashed line), nearly all points fall on or below the diagonal. In other words, whenever agreement with LLMs is "good," it aligns more closely with experts than crowdsourced workers.

Figure 5 LLM Agreement with Expert vs. LLM Agreement with Crowds



Notes. In the left panel, each point represents a model-task pair with coordinates ($x = \kappa(\text{LLM}, \text{expert})$, $y = \kappa(\text{LLM}, \text{crowd})$). Points below the diagonal in the left panel indicate stronger LLM alignment with experts; points above indicate stronger alignment with crowds. Values after each task name are the mean expert agreement for the focal task across models. In the right panel, each cell reports $\kappa(\text{LLM}, \text{expert}) - \kappa(\text{LLM}, \text{crowd})$ by task and model. Blue cells indicate higher agreement with experts relative to crowds for the corresponding (task, model) cell.

The right heatmap displays $\kappa(\text{LLM}, \text{expert}) - \kappa(\text{LLM}, \text{crowd})$ for each task-model combination, revealing task-specific patterns. The heatmap provides similar results: *Connectedness Presence* is uniformly positive (+0.14 to +0.22), implying that, regardless of model choice, expert reference baselines provide a higher-quality reference for that construct. By contrast, *Dialog Intent Classification* is uniformly negative (−0.16 to −0.23), indicating crowd-leaning alignment consistent with positive J between LLMs and crowd annotators (shared confusions). The remaining tasks cluster near zero, suggesting small baseline effects. Together, these panels demonstrate that expert references yield stronger alignment precisely in the higher-agreement regime, supporting the use of expert baselines for downstream prompt and model comparisons.

Furthermore, Appendix Figure D1 shows how measured LLM–human agreement changes when expert-only labels are replaced with mixes that include crowd labels. The gap grows quickly for Dialog Intent Classification and Connectedness Presence, so conclusions would shift under crowd baselines, while Business Proposal Evaluation and Review Attribute Detection are largely flat, so conclusions are stable. These patterns show how baseline choice can influence reference agreement through co-labeling covariance and motivate our use of high-IAA expert baselines.

5.3. Reducing Prompt-induced Error

Meta-prompt structure can introduce systematic measurement error even when annotation guidelines remain constant. We test whether strategic meta-prompt optimization reduces this source of error by examining three prompting strategies across multiple models while maintaining verbatim guideline integrity. Building on the methodological principle of fixing the annotation guidelines

and optimizing only the meta-prompt, we explored three prompting strategies: base, persona, and chain-of-thought (CoT). We also examined their placement in either the system role or user role, resulting in six distinct prompt configurations. We deployed these variants across four state-of-the-art LLMs in the business proposal evaluation task to isolate prompt-induced effects.

Table 2 presents LLM-human agreement under different prompt configurations. The best-performing prompts across all models consistently involved system-role placement of the guidelines. However, our regression-based statistical analysis does not provide sufficient evidence to reject the null hypothesis of performance equivalence between system and user role configurations. This provides mixed support for the impact of meta prompt on measurement error: while we observe consistent trends favoring system role placement (which we advocate as a default strategy), the improvements do not reach statistical significance ($\Delta^{(P)} \approx 0$). Furthermore, the relative performance of the base, persona, and CoT prompt types show considerable heterogeneity across models. For instance, when guidelines were placed in the system role, GPT-4o achieved optimal performance with the base prompt, while Claude 3.5 Sonnet peaked with the chain-of-thought prompting. This model-specific sensitivity suggests that meta-prompt optimization benefits vary by architecture.

Practically, the findings suggest that system role placement offers practical advantages for researchers operating under time or budget constraints. Furthermore, competitive performance can be obtained with minimal meta-prompt engineering. The prompt-induced error reduction may be smaller than anticipated, mainly because prescriptive guidelines limit the marginal gains from such meta-prompt optimization.

Table 2 LLM-Human Agreement under Different Prompts (Business Proposal Evaluation Task)

Model	Guidelines in System Role			Guidelines in User Role		
	Base	Persona	CoT	Base	Persona	CoT
GPT-4o	0.605	0.582	**0.610**	0.582	0.592	0.608
Claude 3.5 Sonnet	0.578	0.503	**0.616**	0.540	0.577	0.580
Gemini 1.5 Pro	**0.561**	0.512	0.487	0.547	0.544	0.515
LLaMA 3.3 70B	0.531	**0.589**	0.527	0.569	0.555	0.541

Notes. The three prompting strategies (base prompt, with personas, and chain-of-thought) and their placement (system role versus user role) generates six distinct prompts. The table presents Cohen's Kappa scores between LLM annotations and human ground truth.

5.4. Reducing Model-choice–induced Error

Even with guidelines, baselines, and prompts held constant, different LLM architectures introduce model-induced error through varying performance on identical annotation tasks. We test whether models produce significantly different annotation outcomes by evaluating seven state-of-the-art LLMs across nine representative management research tasks.

5.4.1. Cross-model differences warrant model testing and selection. Table 3 demonstrates the absence of a universally superior model, with performance leadership varying substantially across annotation contexts, providing strong support for the argument that for fixed (P, G, H), model changes yield significant contrast ($\Delta_S^{(M)} \neq 0$).

Among the tested models, no single model dominates across contexts. Claude 3.5 Sonnet excels in tasks such as dialog intent classification ($\kappa = 0.427$), connectedness presence ($\kappa = 0.688$), and connectedness classification tasks ($\kappa = 0.371$), and whereas DeepSeek-R1 performs best in dialog breakdown analysis ($\kappa = 0.744$), criticism stance detection ($\kappa = 0.580$), and sentiment analysis ($\kappa = 0.935$). However, these two models show dramatic drops on other tasks. For instance, GPT-4o achieves the highest performance in business proposal evaluation ($\kappa = 0.605$), and language toxicity detection ($\kappa = 0.471$) , while o3-mini leads in review attribute detection ($\kappa = 0.554$). One interesting observation is that reasoning-focused models, such as o3-mini and DeepSeek-R1, display greater performance variability across tasks. For instance, o3-mini attains the high agreement level in review attribute detection, while its performance in dialog breakdown analysis drops sharply. This volatility reflects optimization of the reasoning models for analytical cognition (conscious deliberation) rather than intuitive cognition (i.e., grounded, situational pattern recognition) (Patterson and Eggleston 2017) that annotation tasks require, suggesting completion-optimized models may be more suitable for consistent annotation performance.

Overall, these findings establish the necessity of task-specific model evaluation and highlight two critical challenges: determining when low model confidence warrants additional votes from multiple models, and ensuring robust reproducibility as high-performing proprietary models face retirement cycles. We address these challenges in the following sections.

Table 3 Cross-Model Performance Comparison: Human-LLM Agreement

Task	GPT-4o	Gemini 1.5 Pro	LLaMA 3.3 70B	Claude 3.5 Sonnet	OpenAI o3-mini	DeepSeek-R1	GPT-OSS-120B
Business Proposal Evaluation	**0.605**	0.561	0.531	0.578	0.560	0.517	0.533
Review Attribute Detection	0.479	0.346	0.22	0.429	**0.554**	0.458	0.501
Dialog Intent Classification	0.335	0.238	0.205	**0.427**	0.191	0.251	0.187
Dialog Breakdown Analysis	0.566	0.727	0.245	0.689	0.223	**0.744**	0.617
Connectedness Presence	0.632	0.566	0.602	**0.688**	0.555	0.633	0.602
Connectedness Classification	0.366	0.324	0.338	**0.371**	0.303	0.302	0.286
Language Toxicity Detection	**0.471**	0.407	0.366	0.390	0.307	0.340	0.389
Criticism Stance Detection	0.540	0.550	0.500	0.570	0.510	**0.580**	0.520
Sentiment Analysis	0.918	0.885	0.900	0.822	0.901	**0.935**	0.900

Notes. The values represent Cohen's Kappa scores between LLM annotations and human ground truth. Performance varies by task type, and no single model is uniformly strongest across contexts. Binary and multi-class tasks generally yield higher and more stable agreement, while multi-label tasks exhibit lower and more variable agreement.

5.4.2. LLM suitability is task-dependent. The empirical evidence from our comprehensive evaluation presents a subtle picture of LLM viability for text annotation tasks, instead of a binary verdict on LLM suitability. A clear picture emerges: there is a clear relationship between task structure and annotation performance (measurement error).

Annotation tasks exist on a continuum of complexity that directly corresponds to LLM performance reliability. For binary classification tasks, such as sentiment analysis, connectedness presence, and criticism stance detection, LLMs perform reliably well. Across almost all models, Cohen's Kappa scores against human annotations exceed 0.5, with the best-performing models in each task approaching or surpassing 0.6. In multi-class classification tasks where each document receives a single label from mutually exclusive options, some models perform well, but variability across models is large. For business proposal evaluation, better-performing models reach $\kappa \approx 0.5$ to 0.6, whereas dialog breakdown analysis spans roughly $\kappa \approx 0.2$ to 0.7 across models. This dispersion indicates that current LLMs can address moderately complex categorization, yet reliability is highly task and model dependent, so validation on the focal task is essential. Finally, LLMs face significant challenges in multi-label classification tasks, where instances can belong to multiple categories simultaneously. Examples include dialog intent classification, review attribute detection, and connectedness classification. Performance in these tasks drops considerably. Dialog intent classification ($\kappa \approx 0.2$), review attribute detection ($\kappa \approx 0.4$), and connectedness classification ($\kappa \approx 0.3$) all demonstrate substantially lower agreement with expert annotations. This low performance suggests a substantial limitation in LLMs' ability to handle nuanced, overlapping, and multidimensional labeling schemes.

These results suggest that researchers should adopt a decision framework based on task structure when considering LLM implementation for annotation tasks. Binary and simple multi-class classification tasks represent "low-hanging fruit" where LLMs can reliably scale human annotation efforts. In contrast, complex multi-label classification tasks remain better suited to human annotation or may require more sophisticated annotation architectures that combine human expertise with LLM capabilities. Ultimately, our findings caution against viewing LLMs as universal replacements for human annotation.[13]

5.5. Establishing Robust Reproducibility Through Statistical Equivalence

To robustly establish statistical equivalence among models, we employ the regression-based method for all six tasks in primary cases. Denoting the tested LLMs by j (1, 2, ..., J) and the item (each

[13] In addition, we applied four more advanced models (Claude 3.7 Sonnet, Gemini 2.5 Pro, GPT-4.5, and GPT-4.1) as well as one less advanced model (LLaMA 3.3 70B), to evaluate their performance. Models from the same series exhibit some improvement across versions (e.g., Claude 3.7 Sonnet compared to Claude 3.5 Sonnet), though the performance gains are generally modest. The main conclusions remain robust across all models. Detailed results are reported in Table D1.

annotation document, such as each business proposal document) by i (1, 2, ..., I), we formally present the logistic regression model below:

$$\text{logit}(P(Match_{ij}=1)) = \alpha_0 + \alpha_1 \mathbb{1}(j=1) + \alpha_2 \mathbb{1}(j=2) + \cdots + \alpha_{J-1} \mathbb{1}(j=J-1) + \epsilon_{ij}, \quad (9)$$

where the dependent variable ($Match_{ij}$) is an indicator variable equal to 1 if the model j's label for item i matches the human reference baselines, and 0 otherwise. Standard errors are clustered at the item level (at i level) to account for the repeated-measures structure. Coefficients are interpreted as log-odds differences relative to the omitted baseline model. This methodology enables systematic assessment of whether models produce statistically equivalent annotation outcomes, addressing the critical challenge of ensuring research reproducibility as proprietary models face retirement cycles. By establishing formal equivalence criteria, researchers can identify open-parameter alternatives that maintain annotation quality while providing long-term accessibility.

We present the statistical equivalence results in Figure 6, which displays coefficient estimates and 95% confidence intervals for all six primary tasks. The omitted baseline is GPT-4o,[14] and the coefficients represents the difference in log-odds of matching the expert ground truth relative to GPT-4o on the same items within that task. Intervals that cross zero indicate statistical equivalence to GPT-4o for that task; intervals entirely below zero indicate significantly worse performance, and entirely above zero indicate significantly better performance.

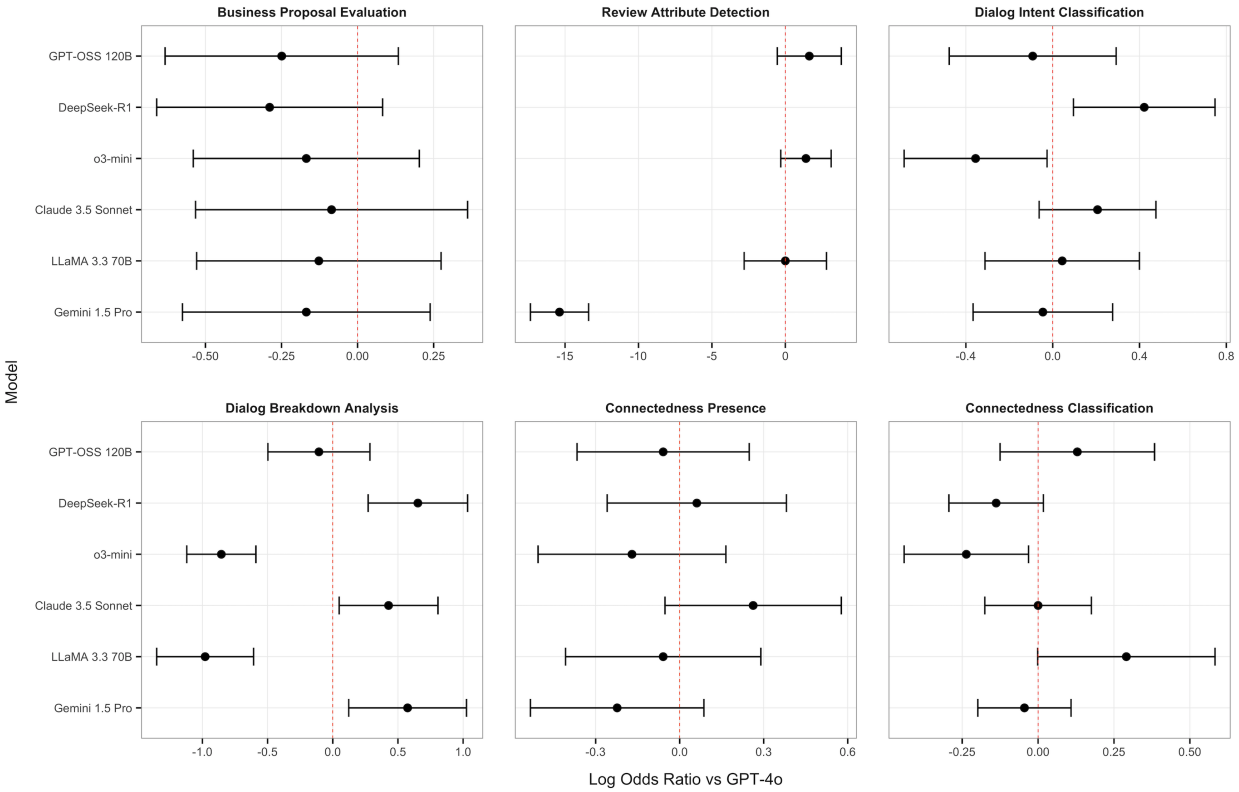
Two patterns emerge across tasks. First, on tasks where most models achieve moderate-to-high performance, statistical equivalence across models is common. For instance, in Business Proposal Evaluation and Connectedness Presence, all tested models are statistically equivalent to GPT-4o, suggesting that reproducibility holds when baseline performance levels are adequate across models. In contrast, on the more complicated multi-label classification tasks, equivalence breaks down and the panels show dispersion where several models' intervals sit cleanly above or below zero. In Review Attribute Detection, Gemini 1.5 Pro falls below, which implies it is significantly worse than GPT-4o model in terms of human-LLM agreement. Similarly, in Dialog Breakdown Analysis, Claude 3.5 outperform GPT-4o significantly, whereas LLaMA-70B and o3-mini are worse than GPT-4o significantly.

The second pattern is about robust reproducibility. Modern open-parameter models are able to demonstrate statistical equivalence to proprietary alternatives across tasks. GPT-OSS-120B achieves statistical equivalence to GPT-4o in all six tasks, providing a permanently accessible alternative that maintains annotation quality. While some proprietary models may achieve marginally better performance on specific tasks, this represents a fundamental tradeoff between immediate

[14] GPT-4o serves as the baseline given its position as the most widely deployed large language model globally at the time of this study. Its potential future retirement motivate our search for statistically equivalent open alternatives.

performance gains and robust reproducibility. The statistical equivalence of open-parameter models ensures long-term research sustainability regardless of commercial model availability, addressing the critical challenge that proprietary model retirement poses to scientific reproducibility.

Figure 6 Regression-based Performance Comparison Across Models



Notes. Regression-based comparison of model performance relative to GPT-4o (red dashed line, reference). Each point represents the estimated log-odds ratio of accuracy for a given model, with horizontal bars indicating 95% confidence intervals. Points to the right of zero suggest higher relative performance compared to GPT-4o, while points to the left indicate lower relative performance. Missing estimates occur when no exact matches exist between model annotations and the reference baseline, which is rare but possible for multi-label classification tasks.

5.6. Multi-LLM Annotation Performance

Building on our confidence assessment methodology from 3.3.2, we implement a cost-effective routing procedure that strategically deploys multiple models based on First-Second Distance (FSD) confidence scores. Using GPT-OSS-120B as the focal model M_1 across the subset where we have obtained the human ground truth, we compute FSD_i from five repeated generations for each item and vary τ from 0 to 1 with an interval of 0.1.[15] The idea is that items with confidence below the threshold ($\text{FSD}_i < \tau$) are routed to auxiliary models.[16]

[15] For $\tau = 0$, performance equals the main model alone.

[16] As τ increases, a larger fraction $q(\tau)$ of low-confidence items is triaged to M_2 and M_3, and at $\tau = 1$, all items are labeled by the three-model majority votes.

Figure 7 plots Cohen's Kappa against human ground truth as a function of τ across six tasks when augmenting GPT-OSS-120B with GPT-4o and Claude 3.5 Sonnet as auxiliaries. It demonstrates that adding labels from good auxiliary models improves agreement, and the gain in agreement comes from items where the focal model is less certain. As the FSD threshold τ increases from 0, routing a larger share of low-FSD items to the auxiliaries raises Cohen's Kappa in a smooth, monotone way when errors concentrate in the low-confidence region. For instance, in Review Attribute Detection task, Cohen's Kappa climbs sharply as more ambiguous items are triaged and then levels off once the uncertain tail has been covered. Crucially, the curves flatten in the middle of the range, which means there is no need to label the full dataset with all models: across tasks the marginal benefit largely saturates once τ enters the range of 0.3–0.7, and pushing toward $\tau = 1$ (an always-ensemble regime) adds cost with little additional improvement. When the focal model is already confident and accurate, extra votes add little gains; for instance, the line in Business Proposal Evaluation and Connectedness Presence tasks remain high and flat because M_1's labels already align with human baselines.

We conclude with two actionable insights based on the multi-LLM annotation approach: (i) introducing additional labels from strong, complementary models reliably improves LLM–human agreement on items the focal model is less confident about; and (ii) a mid-range FSD threshold, $\tau \in [0.3, 0.7]$, captures essentially all attainable gains. There is no need to set $\tau = 1$ (i.e., no need to label the entire dataset with all models).

6. Discussion and Conclusion

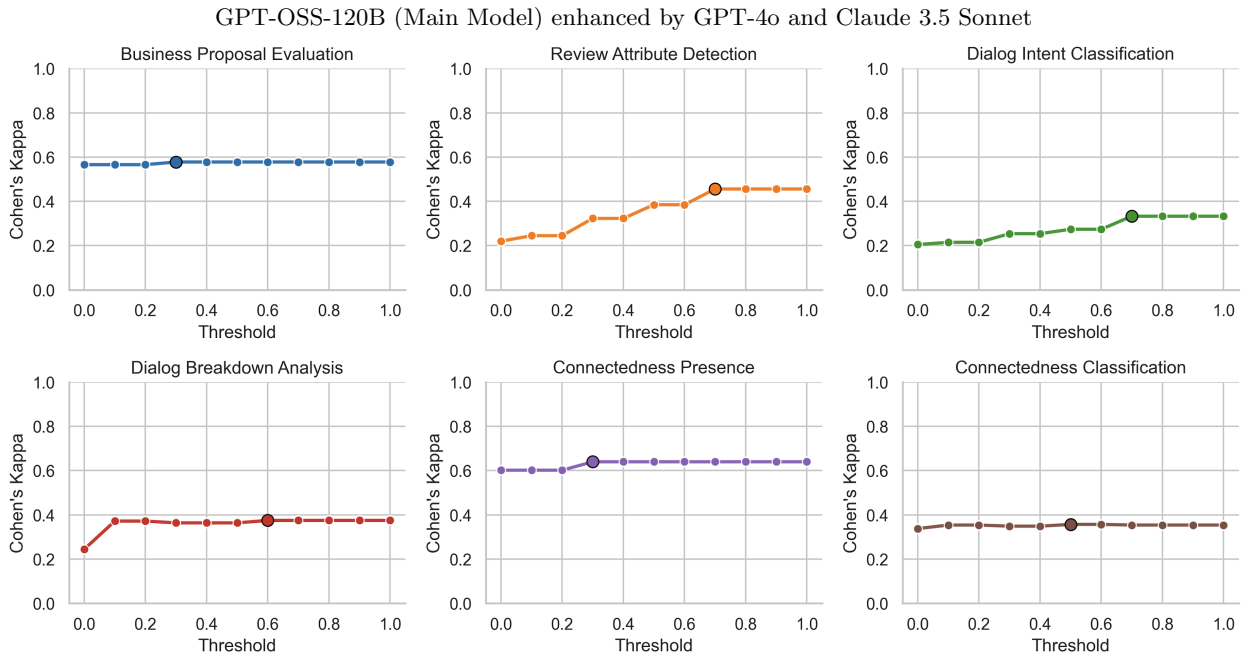
This paper offers a systematic workflow to reduce measurement error in LLM annotation. By decomposing error into guideline-, baseline-, prompt-, and model–induced sources and addressing each with targeted interventions, the SILICON workflow delivers reproducible, higher-agreement labels across diverse research tasks. Beyond raw agreement gains, SILICON adds two elements frequently missing from prior work: (i) a confidence-aware routing procedure that decides when additional model votes are warranted and (ii) a regression-based equivalence design that enables robust reproducibility even as proprietary models change.

While our primary focus has been on management research, the SILICON workflow is generalizable to other disciplines where prescriptive guidelines are preferred over descriptive guidelines. For instance, fields requiring precise ground truth, such as ethics, morality, or safety, may necessitate highly prescriptive annotation guidelines.

6.1. Managerial Implications

Our findings have immediate relevance for organizations that rely on text classification for operational decisions. In practice, organizations run recurring text-classification workloads across diverse

Figure 7 Performance of Multi-LLM Annotation across Confidence Threshold



Notes. Cohen's Kappa vs. FSD routing threshold τ across six tasks when GPT-OSS-120B (focal model) triages items with $\text{FSD}_i < \tau$ to GPT-4o and Claude 3.5 Sonnet for majority voting (at $\tau = 0$ only the focal model is used; larger τ routes a larger share of low-confidence items). The curves represents agreement under increasing triage intensity, with the highlighted marker denotes a representative mid-range threshold $\tau \in [0.3, 0.7]$ where a flattening curve signifies diminishing marginal benefit from routing more items.

functions: customer service departments analyze support tickets and chat transcripts to identify emerging issues, human resources screens resumes and employee feedback for cultural fit and satisfaction indicators, marketing teams mine social media and product reviews for brand perception, and compliance units monitor internal communications for regulatory risks. SILICON provides these organizations with a systematic approach to deploy LLM-based annotation that balances quality, cost, and reproducibility.

The workflow's parsimonious design makes it particularly suitable for organizational adoption. Rather than requiring annotation of entire datasets, SILICON's validation approach uses representative samples of 100-200 documents to assess annotation quality, a scale that fits within typical pilot project budgets. For instance, a firm seeking to classify customer complaints can first apply SILICON to iteratively refine guidelines with domain experts from customer service, establish baselines using these experts rather than generic crowdworkers, and test multiple LLMs on a sample before committing to full-scale deployment. Our results showing that open-parameter models like GPT-OSS-120B achieve statistical equivalence to proprietary models in most tasks offers organizations a path to avoid vendor lock-in while maintaining annotation quality. Furthermore, our confidence-based routing procedure (Section 3.3.2) enables cost-effective quality control: organizations can use inexpensive models for high-confidence cases while selectively routing ambiguous

instances to stronger models or human review, optimizing the cost-quality tradeoff. By decomposing measurement error into interpretable sources, SILICON also provides actionable diagnostics: if annotation quality is poor, organizations can identify whether the issue stems from ambiguous guidelines, inadequate baselines, suboptimal prompts, or model limitations, enabling targeted improvements rather than costly trial-and-error.

6.2. Limitations

There are several limitations that warrant future research. First, our study adopts a prescriptive approach to address annotation tasks. The workflow assumes that high inter-annotator agreement is a desirable standard, which may not apply in research contexts that intentionally seek to capture variability in human perception or interpretation. In such settings, enforcing LLM alignment with a single "ground truth" may obscure meaningful heterogeneity. Furthermore, reasoning models may be especially desirable in such settings, compared to chat completion models. Future work could explore how LLMs can support, rather than suppress, annotation plurality, particularly when diversity of opinion is part of the theoretical construct being studied. Second, while we emphasize expert-developed baselines, we do not account for the individual-level characteristics of annotators. Prior research suggests that LLMs can be tuned to emulate specific annotator styles or preferences (Hashemi et al. 2024), opening the door to more personalized or demographic-aware annotation pipelines. Incorporating these perspectives remains an important direction for future methodological development.

Competing Interests Policy Disclosure Statement

The authors declare that there is no funding to disclose. The research assistants for the tasks as well as the LLM models were funded by the research budgets that authors' department allocated.

References

Aroyo L, Welty C (2015) Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36(1):15–24.

Banerjee S, Dellarocas C, Zervas G (2021) Interacting User-Generated Content Technologies: How Questions and Answers Affect Consumer Reviews. *Journal of Marketing Research* 58(4):742–761.

Cao J, Liang H, Zhan X (2019) Peer Effects of Corporate Social Responsibility. *Management Science* 65(12):5487–5503.

Carlson NA, Burbano V (2025) The Use of LLMs to Annotate Data in Management Research: Foundational Guidelines and Warnings. *Strategic Management Journal* forthcoming.

Chakraborty I, Kim M, Sudhir K (2022) Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research* 59(3):600–622.

Chen L, Zaharia M, Zou J (2023) How is ChatGPT's behavior changing over time?

Chen Z, Chan J (2024) Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. *Management Science* 70(12):9101–9117.

Cheng X, Mayya R, Shi L, Ye S (2024) Support or Setback? Unpacking the Impact of the Small Business Badge .

Choudhury P, Wang D, Carlson NA, Khanna T (2019) Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal* 40(11):1705–1732.

Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4):213–220.

de Kok T (2025) ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science* 71(9):7888–7906.

Deng C, Ravichandran T (2023) Managerial Response to Online Positive Reviews: Helpful or Harmful? *Information Systems Research* 35(4):1802–1823.

Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2024) Generative artificial intelligence and evaluating strategic decisions. *Strategic Management Journal* smj.3677.

Fišar M, Greiner B, Huber C, Katok E, Ozkes AI, and the Management Science Reproducibility Collaboration (2024) Reproducibility in Management Science. *Management Science* 70(3):1343–1356.

Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Ghose A, Ipeirotis PG (2011) Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23(10):1498–1512.

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120(30):e2305016120.

Goli A, Singh A (2024) Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science* 43(4):709–722.

Hashemi H, Eisner J, Rosset C, Van Durme B, Kedzie C (2024) LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. Ku LW, Martins A, Srikumar V, eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834 (Bangkok, Thailand: Association for Computational Linguistics).

Hayes AF, Krippendorff K (2007) Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1(1):77–89.

Higashinaka R, Funakoshi K, Kobayashi Y, Inaba M (2016) The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3146–3150 (Portorož, Slovenia: European Language Resources Association (ELRA)).

Homburg C, Ehm L, Artz M (2015) Measuring and Managing Consumer Sentiment in an Online Community Environment. *Journal of Marketing Research* 52(5):629–641.

Hong Y, Peng J, Burtch G, Huang N (2021) Just DM Me (Politely): Direct Messaging, Politeness, and Hiring Outcomes in Online Labor Markets. *Information Systems Research* 32(3):786–800.

Hovy E, Lavid J (2010) Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22(1):13–36.

Hu T, Collier N (2024) Quantifying the Persona Effect in LLM Simulations. Ku LW, Martins A, Srikumar V, eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10289–10307 (Bangkok, Thailand: Association for Computational Linguistics).

Huang A, Huang N, Hong Y (2024) Algorithm-enabled Workflow Automation in Open-source Development: Evidence from GitHub Actions. *SSRN Electronic Journal* .

Humphreys A, Wang RJH (2018) Automated Text Analysis for Consumer Research. *Journal of Consumer Research* 44(6):1274–1306.

Ipeirotis PG, Provost F, Sheng VS, Wang J (2014) Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28(2):402–441.

Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44:223–270.

Krippendorff K (2004) Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30(3):411–433.

Kwark Y, Lee GM, Pavlou PA, Qiu L (2021) On the Spillover Effects of Online Product Reviews on Purchases: Evidence from Clickstream Data. *Information Systems Research* 32(3):895–913.

Landis JR, Koch GG (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159.

Lee D, Hosanagar K (2021) How Do Product Attributes and Reviews Moderate the Impact of Recommender Systems Through Purchase Stages? *Management Science* 67(1):524–546.

Lee D, Hosanagar K, Nair HS (2018) Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science* 64(11):5105–5131.

Lee S, DeLucia A, Nangia N, Ganedi P, Guan R, Li R, Ngaw B, Singhal A, Vaidya S, Yuan Z, Zhang L, Sedoc J (2023) Common Law Annotations: Investigating the Stability of Dialog System Output Annotations. Rogers A, Boyd-Graber J, Okazaki N, eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 12315–12349 (Toronto, Canada: Association for Computational Linguistics).

Leek LC, Bischl S, Freier M (2024) Introducing Textual Measures of Central Bank Policy-Linkages Using ChatGPT.

Liu X, Lee D, Srinivasan K (2019) Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning. *Journal of Marketing Research* 56(6):918–943.

Loughran T, Mcdonald B (2014) Measuring Readability in Financial Disclosures. *The Journal of Finance* 69(4):1643–1671.

Lyu Q, Shridhar K, Malaviya C, Zhang L, Elazar Y, Tandon N, Apidianaki M, Sachan M, Callison-Burch C (2025) Calibrating Large Language Models with Sample Consistency. *Proceedings of the AAAI Conference on Artificial Intelligence* 39(18):19260–19268.

Manning BS, Zhu K, Horton JJ (2024) Automated Social Science: Language Models as Scientist and Subjects.

Matook S, Dennis AR, Wang YM (2022) User Comments in Social Media Firestorms: A Mixed-Method Study of Purpose, Tone, and Motivation. *Journal of Management Information Systems* 39(3):673–705.

Mayya R, Ye S, Viswanathan S, Agarwal R (2021) Who Forgoes Screening in Online Markets and Why? Evidence from Airbnb. *MIS Quarterly* 45(4):1745–1776.

Melumad S, Inman JJ, Pham MT (2019) Selectively Emotional: How Smartphone Use Changes User-Generated Content. *Journal of Marketing Research* 56(2):259–275.

Mousavi R, Kitchens B, Oliver A, Abbasi A (2024) From Lexicons to Large Language Models: A Holistic Evaluation of Psychometric Text Analysis in Social Science Research.

Natarajan N, Dhillon IS, Ravikumar PK, Tewari A (2013) Learning with Noisy Labels. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds., *Advances in Neural Information Processing Systems*, volume 26 (Curran Associates, Inc.).

Obermeier D, Mayya R (2024) Decentralized Governance as an Open Platform Strategy: An Empirical Study of Devolving Controls through DAOs .

Oh H, Goh KY, Phan TQ (2023) Are You What You Tweet? The Impact of Sentiment on Digital News Consumption and Social Media Sharing. *Information Systems Research* 34(1):111–136.

Paharia N, Keinan A, Avery J, Schor JB (2011) The Underdog Effect: The Marketing of Disadvantage and Determination through Brand Biography. *Journal of Consumer Research* 37(5):775–790.

Palmer A, Smith NA, Spirling A (2023) Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* 4(1):2–3.

Pangakis N, Wolken S, Fasching N (2023) Automated Annotation with Generative AI Requires Validation .

Passonneau R (2006) Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. Calzolari N, Choukri K, Gangemi A, Maegaard B, Mariani J, Odijk J, Tapias D, eds., *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)* (Genoa, Italy: European Language Resources Association (ELRA)).

Patterson RE, Eggleston RG (2017) Intuitive Cognition. *Journal of Cognitive Engineering and Decision Making* 11(1):5–22.

Paun S, Artstein R, Poesio M (2022) *Statistical Methods for Annotation Analysis*. Synthesis Lectures on Human Language Technologies (Cham: Springer International Publishing), ISBN 978-3-031-03753-5 978-3-031-03763-4.

Peng H, Romero DM, Horvat EA (2022) Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences* 119(25):e2119086119.

Pryzant R, Iter D, Li J, Lee Y, Zhu C, Zeng M (2023) Automatic Prompt Optimization with "Gradient Descent" and Beam Search. Bouamor H, Pino J, Bali K, eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7957–7968 (Singapore: Association for Computational Linguistics).

Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson C, Van Bavel JJ (2024) GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* 121(34):121 (34) e2308950121.

Rosette AS, Carton AM, Bowes-Sperry L, Hewlin PF (2013) Why Do Racial Slurs Remain Prevalent in the Workplace? Integrating Theory on Intergroup Behavior. *Organization Science* 24(5):1402–1421.

Rottger P, Vidgen B, Hovy D, Pierrehumbert J (2022) Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. Carpuat M, de Marneffe MC, Meza Ruiz IV, eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–190 (Seattle, United States: Association for Computational Linguistics).

Ruggeri F, Misino E, Muti A, Korre K, Torroni P, Barrón-Cedeño A (2024) Let Guidelines Guide You: A Prescriptive Guideline-Centered Data Annotation Methodology.

Saha P, Garimella K, Kalyan NK, Pandey SK, Meher PM, Mathew B, Mukherjee A (2023) On the Rise of Fear Speech in Online Social Media. *Proceedings of the National Academy of Sciences* 120(11):e2212270120.

Salehi N, Teevan J, Iqbal S, Kamar E (2017) Communicating Context to the Crowd for Complex Writing Tasks. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1890–1901, CSCW '17 (New York, NY, USA: Association for Computing Machinery), ISBN 978-1-4503-4335-0.

Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The Risk of Racial Bias in Hate Speech Detection. Korhonen A, Traum D, Màrquez L, eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678 (Florence, Italy: Association for Computational Linguistics).

Sen C, Hartvigsen T, Yin B, Kong X, Rundensteiner E (2020) Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4596–4608 (Online: Association for Computational Linguistics).

Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622 (Las Vegas Nevada USA: ACM), ISBN 978-1-60558-193-4.

Shi ZJ, Liu X, Srinivasan K (2022) Hype News Diffusion and Risk of Misinformation: The Oz Effect in Health Care. *Journal of Marketing Research* 59(2):327–352.

Sibai O, Luedicke MK, De Valck K (2024) Why Online Consumption Communities Brutalize. *Journal of Consumer Research* ucae022.

Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Lapata M, Ng HT, eds., *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263 (Honolulu, Hawaii: Association for Computational Linguistics).

Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Van Ess-Dykema C, Meteer M (2000) Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3):339–374.

Tan Z, Li D, Wang S, Beigi A, Jiang B, Bhattacharjee A, Karami M, Li J, Cheng L, Liu H (2024) Large Language Models for Data Annotation and Synthesis: A Survey. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 930–957 (Miami, Florida, USA: Association for Computational Linguistics).

Tang C, Li SK, Ding Y, Gopal RD, Zhang G (2024) Racial Discrimination and Anti-discrimination: The COVID-19 Pandemic's Impact on Chinese Restaurants in North America. *Information Systems Research* 35(3):1274–1295.

Törnberg P (2024) Best Practices for Text Annotation with Large Language Models.

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837, NIPS '22 (Red Hook, NY, USA: Curran Associates Inc.), ISBN 978-1-71387-108-8.

Xie C, Chen C, Jia F, Ye Z, Lai S, Shu K, Gu J, Bibi A, Hu Z, Jurgens D, Evans J, Torr P, Ghanem B, Li G (2024) Can Large Language Model Agents Simulate Human Trust Behavior? (38th Conference on Neural Information Processing Systems).

Yang M, Adomavicius G, Burtch G, Ren Y (2018) Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining. *Information Systems Research* 29(1):4–24.

Yang M, Ren Y, Adomavicius G (2019) Understanding User-Generated Content and Customer Engagement on Facebook Business Pages. *Information Systems Research* 30(3):839–855.

Ye Z, Yoganarasimhan H, Zheng Y (2024) LOLA: LLM-Assisted Online Learning Algorithm for Content Experiments.

Yeverechyahu D, Mayya R, Oestreicher-Singer G (2024) The Impact of Large Language Models on Open-Source Innovation: Evidence from GitHub Copilot. *SSRN Electronic Journal* .

Zhang C, D'Haro LF, Chen Y, Zhang M, Li H (2024) A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators.

Zhang K, Bhattacharyya S, Ram S (2016) Large-Scale Network Analysis for Online Social Brand Advertising. *MIS Quarterly* 40(4):849–868.

Zhang L, Mille S, Hou Y, Deutsch D, Clark E, Liu Y, Mahamood S, Gehrmann S, Clinciu M, Chandu KR, Sedoc J (2023) A Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14944–14982 (Toronto, Canada: Association for Computational Linguistics).

Appendix A: Methodological Details: Agreement, Robust Reproducibility, and Confidence

A.1. Measurement of Agreement Rate: Cohen's Kappa

To measure agreement, one simple way is counting the raw number of matching annotations. However, the raw agreement fails to account for agreements that might occur by chance. This can lead to misleading conclusions, especially in cases where the likelihood of random agreement is high (Lee et al. 2023). To overcome this limitation, we employed Cohen's Kappa (κ), which measures the level of agreement between two annotators while adjusting for chance agreement (Cohen 1968). The formula for Cohen's Kappa is given by:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{10}$$

where P_o is the relative observed agreement among raters, and P_e is the hypothetical probability of agreement by chance. Specifically, $\kappa = 1$ indicates perfect agreement. $\kappa = 0$ indicates agreement no better than chance. $\kappa < 0$ suggests less agreement than would be expected by chance. A κ of 0.4 to 0.6 is commonly regarded as a threshold for sufficient inter-annotator agreement (Landis and Koch 1977).

In our multi-label classification task, annotators could assign multiple distinct labels (referred to as "units") to a single item (e.g., a social media post). We define the overall label(s) assigned to an item as a "set". For instance, in the toxicity detection task, there are in total three units, "fearspeech", "hatespeech", and "normal", and a set could be "fearspeech, hatespeech". It is apparent that the agreement level between "fearspeech, hatespeech" and "fearspeech" should be higher than that between "hatespeech" and "fearspeech". To account for this, we use Weighted Cohen's Kappa (Cohen 1968), which is given by

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}, \tag{11}$$

where w_{ij} is the weight matrix, x_{ij} is the observed matrix, and m_{ij} is the expected matrix.

We follow Passonneau (2006) to derive the weight matrix W for our calculations. The weight w ranges from 0 (identical sets) to 1 (disjoint sets). The weight w between two sets P and Q is defined as:

$$w = 1 - J \cdot M, \tag{12}$$

where J is the Jaccard metric (Jaccard 1908) and M represents monotonicity. Specifically, J measures the size difference between two sets, independently of their structural relationship. It is calculated as the ratio of the cardinality of the intersection to the cardinality of the union of the two sets. J ranges from 0 (disjoint sets) to 1 (identical sets). The M term captures the structural relationship between sets: if two sets Q and P are identical, M is 1; if one set is a subset of the other, M is 2/3; if the intersection and the two set differences are all non-null, then M is 1/3; if the sets are disjoint, M is 0. Altogether, the weight w reflects both the size difference and the structural relationship between the two sets. For instance, in the language toxicity detection task, the weight between sets "fearspeech" and "hatespeech" is 1, and the weight between "fearspeech" and "fearspeech, hatespeech" is 2/3.

A.2. A Regression-based Approach to Compare LLM Annotation Performance

Consider the scenario where we use the same prompt across multiple models applied to the same text sample.[17] Our goal is to statistically compare model performances and identify models with equivalent results. Specifically, we aim to determine whether a model's performance metric differs significantly from others or if certain models yield statistically indistinguishable performance metrics.

In this setting, there are two sources of uncertainty:

1. Sampling uncertainty: This arises from inferring population parameters (e.g., differences in LLM performance across models) based on the human baseline sample, which is randomly drawn from the human baseline.

2. Stochastic output uncertainty: This stems from the non-deterministic nature of LLM outputs. The literature (e.g., Pangakis et al. 2023) documents this variability, and common approaches to account for this uncertainty include running the model multiple times and calculating the consistency of its annotation results.

Our focus is on addressing the first source of uncertainty. Instead of computing overall performance metrics for each model and comparing them with pairwise tests or bootstraps, we propose a logistic regression-based approach.

Suppose we use J models to annotate a sample with I items (e.g.., a sample consisting of I business proposals). We treat each model as a distinct "treatment" and aim to assess the "treatment effect" of one model compared to a baseline model. Denoting the model by j (1, 2, ..., J) and the item by i (1, 2, ..., I), the unit of analysis is the item-model pair, ij. The dependent Variable ($Matched_{ij}$) is an indicator variable equal to 1 if the model j's label for item i matches the human ground truth, and 0 otherwise.[18]

We then perform a logistic regression of the dependent variable $Matched_{ij}$ on all treatment dummies, excluding one due to perfect collinearity. The regression equation is:

$$\text{logit}(P(Match_{ij}=1)) = \alpha_0 + \alpha_1 \mathbb{1}(j=1) + \alpha_2 \mathbb{1}(j=2) + ... + \alpha_{J-1}\mathbb{1}(j=J-1) + \epsilon_{ij}. \tag{13}$$

Here, the J-th treatment is omitted as the baseline. The coefficients are interpreted as follows: The coefficients are interpreted in terms of log-odds:

$$\begin{aligned}
\alpha_0 &= \log\left(\frac{P(Matched_{ij}=1|j=J)}{1-P(Matched_{ij}=1|j=J)}\right), \\
\alpha_1 &= \log\left(\frac{P(Matched_{ij}=1|j=1)}{1-P(Matched_{ij}=1|j=1)}\right) - \alpha_0, \\
\alpha_2 &= \log\left(\frac{P(Matched_{ij}=1|j=2)}{1-P(Matched_{ij}=1|j=2)}\right) - \alpha_0, \\
&\dots \\
\alpha_{J-1} &= \log\left(\frac{P(Matched_{ij}=1|j=J-1)}{1-P(Matched_{ij}=1|j=J-1)}\right) - \alpha_0.
\end{aligned} \tag{14}$$

[17] This approach in essence is about comparing LLM annotation performance across different treatments. Treatments could be different models or different prompts, and the same logic outlined here applies to both scenarios.

[18] This binary "matched" measure is straightforward and easily interpretable at the item level. When aggregated across items, it provides the model's accuracy. However, a slight discrepancy exists since in our case studies, we use Cohen's Kappa to account for chance agreement. This adjustment is made because Cohen's Kappa is a chance-corrected measure calculated over all items, which makes it less intuitive at the per-item level. Nonetheless, applying a Kappa-like transformation to derive the dependent variable would still align with the same regression logic.

Crucially, we cluster standard errors at the item level to account for the fact that each item i is "tested" multiple times—once per model. By design, the regression estimates represent differences in log-odds of matching the human ground truth between each model and the baseline model. Hypotheses can be tested within this logistic framework, such as (1) joint test: H_a: $\alpha_1 = \alpha_2 = \cdots = \alpha_{J-1} = 0$, using a likelihood ratio test, and (2) individual tests: H_{b1}: $\alpha_1 = 0$, H_{b2}: $\alpha_2 = 0$, ..., $H_{b(J-1)}$: $\alpha_{J-1} = 0$.

Furthermore, under this framework, the second source of uncertainty (stochasticity in LLM outputs) can also be addressed with additional assumptions. Specifically, we consider that LLMs aim to produce an "intended output"[19] but sometimes generate a different "actual output" due to non-determinism. From a regression perspective, this discrepancy can be treated as a form of measurement error. Assuming the measurement error is independently and identically distributed (i.i.d.), the regression estimates remain unbiased when the sample size is sufficiently large.

A.3. Computing *FSD* and Two Practical Approaches

FSD evaluates LLM confidence (or consistency) by measuring the discrepancy in agreement between the most common and second most common answers from multiple sampled outputs. A larger FSD indicates higher confidence, indicating that the model's responses are more concentrated on a single answer over multiple runs. We summarize two ways to obtain the required distribution in practice.

1. **Log-probability approach (when available).** If token log-probabilities are exposed (common for open-parameter models and some proprietary APIs), infer the top two option probabilities from the first disambiguating token of each option and compute FSD from those. There are some limitations to this approach: (i) some APIs do not expose log-probs; (ii) with very large or combinatorial label spaces (e.g., multi-label), mapping options to first tokens becomes cumbersome.
2. **Sampling route (model-agnostic).** Sample multiple independent outputs from the focal model for each item, form the empirical distribution over options, and compute FSD directly from observed frequencies. This resolves both limitations above at the expense of additional cost.

A.4. Proofs

Proof of Lemma 1. Write $\pi_k = \Pr(y = k)$ and, for each configuration c, define

$$p^{(c)}_{\ell|k} := \Pr(\hat{y}^{(c)} = \ell \mid y = k), \qquad q^{(c)}_{\ell|k} := \Pr(r = \ell \mid y = k, c).$$

By Assumption 2, $r \perp c \mid (x, G, H)$. Hence for any k, ℓ,

$$\begin{aligned} q^{(c)}_{\ell|k} &= \Pr(r = \ell \mid y = k, c) = \mathbb{E}[\Pr(r = \ell \mid y = k, x, G, H, c) \mid y = k] \\ &= \mathbb{E}[\Pr(r = \ell \mid y = k, x, G, H) \mid y = k] =: q_{\ell|k}, \end{aligned}$$

which shows that the conditional reference probabilities $q_{\ell|k}$ (and thus the reference error rate $e = \Pr(r \neq y)$) are *invariant across c*.

[19] One could argue that the intended output is the one with the highest probability of being generated by the LLM model when the temperature is set to 0. However, we take the stance that the intended output may not necessarily align with the output generated under this condition. This difference is not the focus of the study, and we leave its discussion for future work.

Now we decompose $A_R(c)$. For any c,

$$A_R(c) = \sum_{k=1}^{K} \pi_k \Pr(\hat{y}^{(c)} = r \mid y = k) = \sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} \Pr(\hat{y}^{(c)} = \ell,\, r = \ell \mid y = k).$$

Add and subtract $p^{(c)}_{\ell|k}\, q_{\ell|k}$ inside the inner sum:

$$\Pr(\hat{y}^{(c)} = r \mid y = k) = \sum_{\ell=1}^{K} p^{(c)}_{\ell|k}\, q_{\ell|k} + \sum_{\ell=1}^{K} \Big(\Pr(\hat{y}^{(c)} = \ell, r = \ell \mid y = k) - p^{(c)}_{\ell|k}\, q_{\ell|k} \Big). \tag{15}$$

By Assumption 1, $q_{k|k} = 1 - e$ and $q_{\ell|k} = e/(K-1)$ for $\ell \neq k$. Therefore,

$$\sum_{\ell=1}^{K} p^{(c)}_{\ell|k}\, q_{\ell|k} = (1-e)\, p^{(c)}_{k|k} + \frac{e}{K-1} \sum_{\ell \neq k} p^{(c)}_{\ell|k} = \Big((1-e) - \frac{e}{K-1} \Big)\, p^{(c)}_{k|k} + \frac{e}{K-1}. \tag{16}$$

Substituting Equation (16) back to Equation (15), we have:

$$\begin{aligned} A_R(c) &= \sum_{k=1}^{K} \pi_k \Big[\Big((1-e) - \frac{e}{K-1} \Big)\, p^{(c)}_{k|k} + \frac{e}{K-1} \Big] + \sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} \Big(\Pr(\hat{y}^{(c)} = \ell, r = \ell \mid y = k) - p^{(c)}_{\ell|k} q_{\ell|k} \Big) \\ &= a \sum_{k=1}^{K} \pi_k p^{(c)}_{k|k} + b + \underbrace{\sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} \Big(\Pr(\hat{y}^{(c)} = \ell, r = \ell \mid y = k) - p^{(c)}_{\ell|k} q_{\ell|k} \Big)}_{=:\, J(c)}, \end{aligned}$$

with $a = (1-e) - \frac{e}{K-1} > 0$ and $b = \frac{e}{K-1}$. Recognizing $\sum_k \pi_k p^{(c)}_{k|k} = \Pr(\hat{y}^{(c)} = y) = A_T(c)$ completes the identity

$$A_R(c) = a\, A_T(c) + b + J(c).$$

Note that the term $J(c)$ captures conditional dependence between $\hat{y}^{(c)}$ and r given y; if they were conditionally independent, $J(c)$ would be 0 and the mapping would reduce to the familiar linear form. □

Appendix B: Case Study Details

We present the research contexts, original annotation approach, our implementation, and key findings of the seven cases studies in Table B1.

Table B1 Summary of Case Studies

Task	Research Context	Original Text Annotation Approach	SILICON Implementation	Key Findings
Business Proposal Evaluation	Understanding DAO governance through systematic analysis of business proposal content and decision delegation patterns	Manual classification by RAs into six predefined categories	Developed expert-validated guidelines with 3 RAs; established baseline using 121 proposals; tested multiple LLM configurations	High-performing LLMs achieved moderate to high agreement with expert baseline
Review Attribute Detection	Quantifying consumers' purchasing decisions through systematic analysis of product review content	Training a deep learning model using 5,000 reviews annotated by Amazon Mechanical Turk workers	Developed expert-validated guidelines with 3 RAs; established baseline using 180 reviews; tested multiple LLM prompt and model configurations	1. Moderate agreement achieved with expert baseline across selected models 2. Significant performance variation observed between different LLMs for this complex multi-label task
Dialog Intent Classification and Breakdown Analysis	Evaluating human-computer interactions in customer support through systematic dialog analysis	None. This is a novel evaluation of human-computer interactions in customer support through systematic dialog analysis	Developed expert-validated guidelines with 3 RAs; established baseline using 195 conversational turns; tested multiple LLM prompt and model configurations	1. Intent classification showed consistently low agreement between LLMs and expert baseline 2. Breakdown analysis demonstrated moderate to high agreement levels across tested models
Affective Content Evaluation	Understanding how a sense of connectedness affects consumer willingness to pay and product sales performance	Ongoing work involving manual annotations	Developed expert-validated guidelines with three research assistants; established a baseline using 120 product descriptions; tested multiple LLM prompting and model configurations	1. Moderate to high agreement with expert baseline achieved for connectedness presence detection 2. Low agreement observed for connectedness classification across detailed categories
Language Toxicity Detection	Demonstrating fear speech's prevalence, influence, and subtlety compared to hate speech on social media platforms	Manual classification into one of four types using a combination of experts and Amazon Mechanical Turk workers	Used the annotation guidelines and a random sample of the human baseline from the original paper, which involves 160 posts; explored multiple prompts and models	Agreement level between LLMs and crowdsourced baseline is moderate to low
Criticism Stance Detection	Assessing online attention and criticism toward retracted scientific papers over time	Manual classification of Twitter posts to identify expressions of criticism toward specific scientific papers, using trained annotators	Used the annotation guidelines and a random sample of the human baseline from the original paper consisting of 200 posts; explored multiple prompts and models	Agreement level between LLMs and crowdsourced baseline is moderate
Sentiment Analysis	Comparing human and computational attention in text classification tasks, focusing on sentiment analysis	Manual labeling of sentiment value and human attention maps of Yelp customer reviews using Amazon Mechanical Turk workers	Used the annotation guidelines and a random sample of the human baseline from the original paper consisting of 120 reviews; explored multiple models	Agreement level between LLMs and crowdsourced baseline is very high across all models tested

Appendix C: Workflow Details

Figure C1 presents SILICON workflow in detail. The workflow organizes how we (i) develop reliable human annotation guidelines, (ii) establish human baselines, and (iii) subsequently optimize meta prompts and evaluate multiple LLMs.

Figure C1 SILICON: A Systematic Workflow for LLM-based Text Annotation

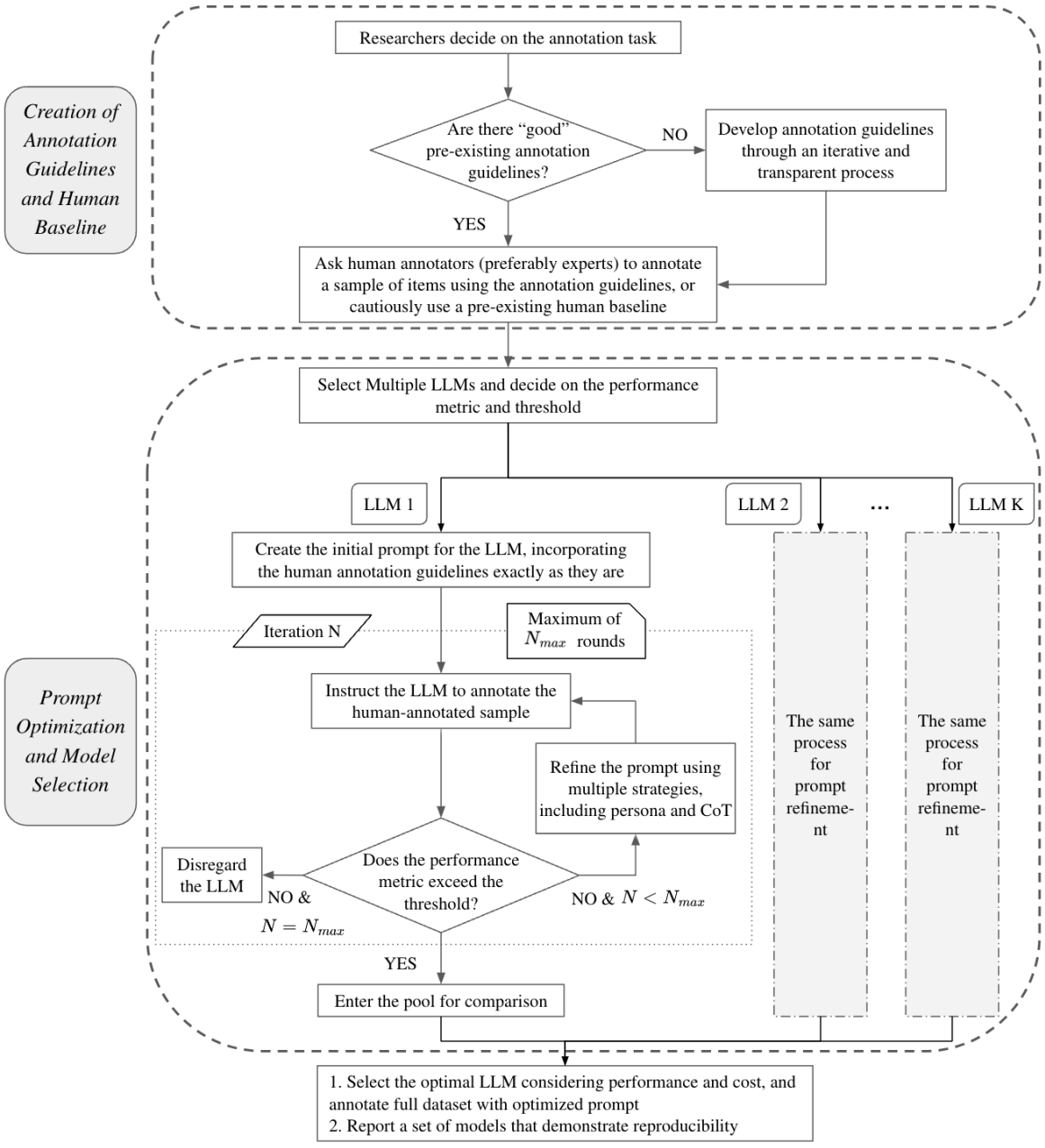
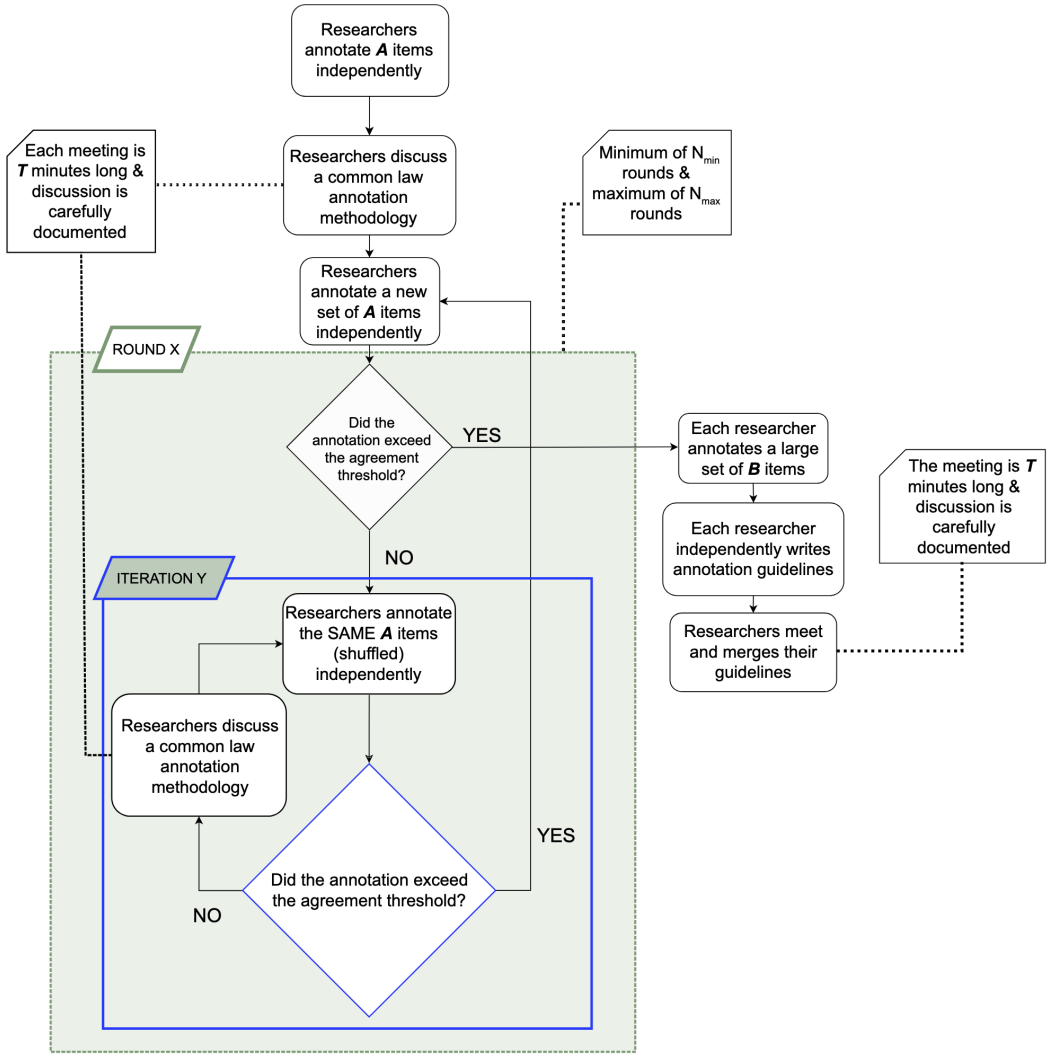


Figure C2 shows the detailed workflow for developing the annotation guidelines that we adopted. This workflow was originally proposed by Lee et al. (2023) and we have integrated it into the SILICON workflow.

In the figure, the term "researcher" denotes expert annotators, which may include research assistants or the researchers themselves.

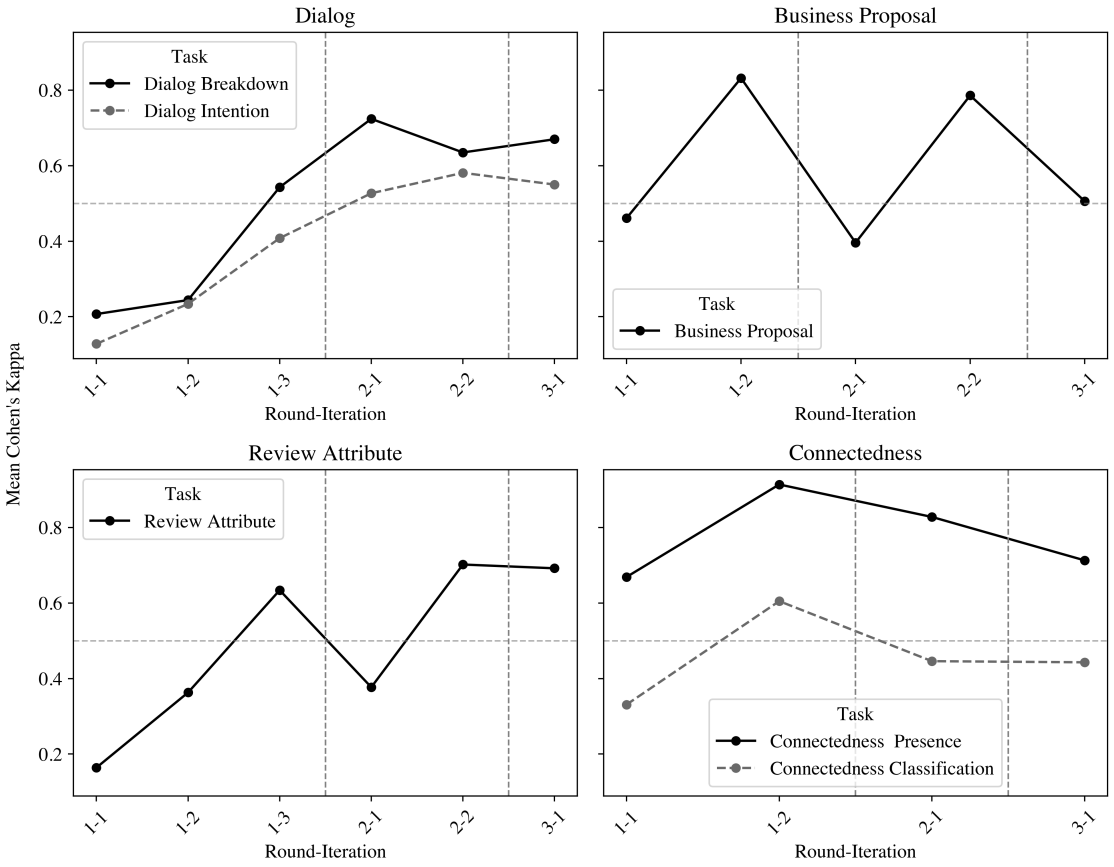
Figure C2 A Detailed Workflow of Developing Annotation Guidelines from Lee et al. (2023)



We emphasize two practical takeaways from applying this workflow across tasks: (1) pre-existing guidelines often fail to produce reliable human baselines due to inconsistent interpretations, and (2) achieving high IAA typically requires multiple rounds of revision. Figure C3 shows the evolution of agreement levels across multiple rounds. Most tasks start with very poor agreement levels in the first iteration. Over a few iterations, we observe substantial improvements, which highlights the importance of iteratively refining annotation guidelines to resolve ambiguities and ensure consistent interpretation, especially for complex tasks.

To illustrate the process in practice, we provide an example. Specifically, we detail the creation of guidelines for the dialog analysis task, as presented in Table C1. We recruit three undergraduate RAs to undertake this iterative process. Initially, they receive an overview of the task, including the contextual background and

Figure C3 Iterative Processes of Annotation Guideline Development



Notes. The y-axis shows the IAA among three annotation experts. Achieving a high level of IAA (i.e., κ) often requires multiple iterations of refinement of annotation guidelines.

labels for the two classification tasks. The RAs independently label a small sample and then discuss their results via Zoom, emphasizing points of disagreement and refining label definitions (Iteration 1). Following this initial meeting, the RAs annotate the same shuffled sample twice more, meeting after each annotation to discuss their results (Iterations 2 and 3). After reaching an agreement rate threshold in their third iteration, they move to a new sample, repeating the annotation and meeting process twice (Iterations 4 and 5). In their next sample, they achieve the agreement threshold on their first attempt (Iteration 6), concluding the iteration process. Finally, the RAs meet to consolidate their independent annotation guidelines and annotate a larger sample to establish a human baseline. Throughout the process, their mean κ for the intent classification task improves from 0.128 to 0.61, and their agreement for breakdown analysis increases from 0.207 to 0.665. The annotation guidelines and human annotation baseline are further used in the next phase of SILICON: optimizing prompts for LLMs and testing multiple LLMs' performance.

Table C1 Development of Annotation Guidelines for Dialog Analysis Task

Round	Iteration	Number of conversations	Number of utterances	Intent Anaylsis Mean κ	Breakdown Analysis Mean κ
1	1	5	81	0.128	0.207
1	2	5	81	0.234	0.244
1	3	5	81	0.408	0.543
2	1	6	101	0.527	0.724
2	2	6	101	0.581	0.635
3	1	6	100	0.550	0.670
–	Large sample (the human baseline)	18	195	0.61	0.665

The processes for business proposal evaluation, review attribute detection, and affective content evaluation follow the same approach, wherein we recruit three RAs for each task to go through the iterative process. The details are summarized in Tables C2 through C4.

Table C2 Development of Annotation Guidelines for Business Proposal Evaluation

Round	Iteration	Number of proposals	Mean κ
1	1	50	0.461
1	2	50	0.832
2	1	49	0.396
2	2	49	0.786
3	1	50	0.506
–	Large sample (the human baseline)	109	0.529

Table C3 Development of Annotation Guidelines for Review Attribute Detection Task

Round	Iteration	Number of Reviews	Mean κ
1	1	67	0.164
1	2	67	0.363
1	3	67	0.634
2	1	65	0.377
2	2	65	0.702
3	1	65	0.692
–	Large sample (the human baseline)	180	0.567

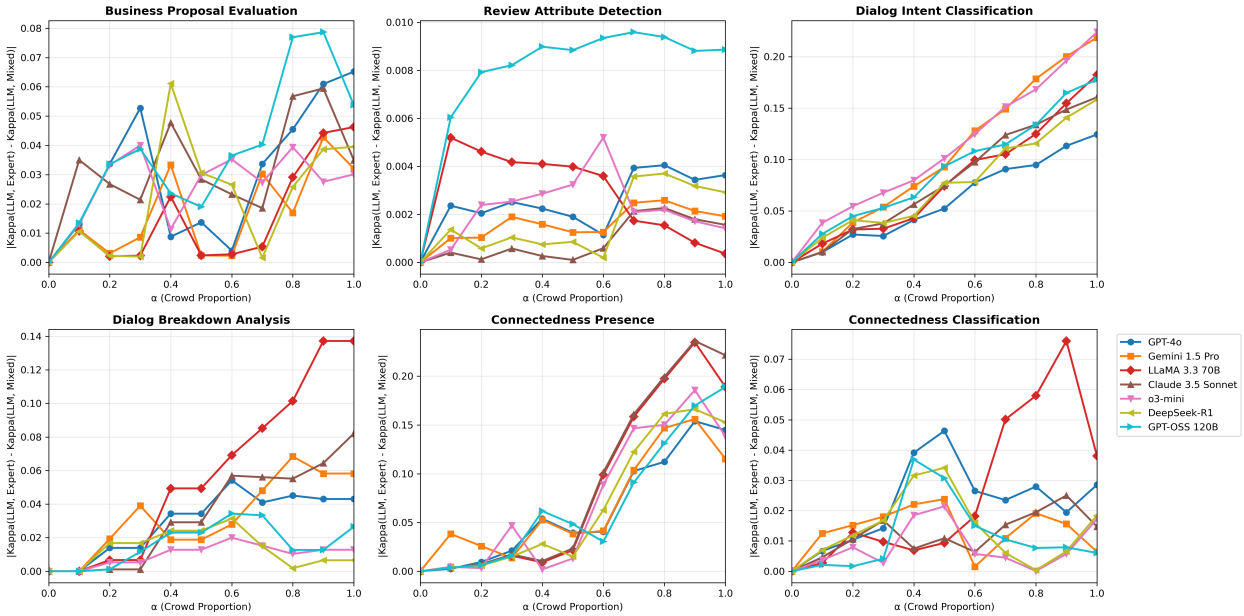
Table C4 Development of Annotation Guidelines for Affective Content Evaluation Task

Round	Iteration	Number of product pages	Connectedness Presence Mean κ	Connectedness Classification Mean κ
1	1	50	0.618	–
1	2	50	0.669	0.331
2	1	50	0.914	0.605
2	2	50	0.828	0.446
3	1	50	0.713	0.443
–	Large sample (the human baseline)	120	0.822	0.518

Appendix D: Additional Empirical Validation Results

To quantify how baseline choice affects identification, we progressively mix crowd labels into the expert baseline and track $|\kappa(\text{LLM}, \text{expert}) - \kappa(\text{LLM}, \text{mixed})|$ as the crowd proportion α increases; results are reported in Figure D1. Sensitivity is strongly task dependent. *Dialog Intent Classification* exhibits a near-linear, steep rise up to around 0.20 at $\alpha = 1$, indicating that substituting crowds for experts as reference baselines substantially changes the observed contrasts. This is precisely the scenario where increases in J can spuriously boost A_R under the crowd references. *Connectedness Presence* also shows pronounced growth (especially for $\alpha \geq 0.6$), while *Business Proposal Evaluation* and *Review Attribute Detection* are relatively flat (≤ 0.08), suggesting limited baseline-induced error for these tasks. Non-monotonicity in *Connectedness Classification* points to heterogeneous, non-systematic crowd deviations. These patterns are consistent with our methodological approach: when $(G^\star, H^\star)$ establishes a high-IAA, lower-error reference, pairwise ordered contrasts $\Delta^{(P)}$ and $\Delta^{(M)}$ are less confounded by co-labeling covariance. Observed improvements therefore more tightly track reductions in true measurement error, as formalized in Equation (4).

Figure D1 Sensitivity of LLM–Expert Agreement to Crowd–Expert Label Mixing



Notes. For each task and model, the y-axis reports the absolute change in Cohen's κ when moving from expert labels to a mixed label set, $|\,\kappa(\text{LLM}, \text{expert}) - \kappa(\text{LLM}, (1-\alpha)\cdot\text{expert} + \alpha\cdot\text{crowd})\,|$, as the crowd proportion α increases. The slope and shape of each curve, i.e., how quickly the gap grows, reveal robustness to crowd labels, where flatter curves indicates robustness and steep curves indicate κ sensitive to expert-only annotations).

In addition to the seven models consistently evaluated in the main analysis, we also assess four additional models: four more advanced ones—Claude 3.7 Sonnet, Gemini 2.5 Pro, GPT-4.5, and GPT-4.1—and one less advanced model, LLaMA 3 70B. Their performance is reported in Table D1. Overall, we observe only modest performance improvements within model families across versions. In terms of the best-performing model for each task, we find that the four additional models emerge as top performers in several cases: Claude 3.7 Sonnet performs best on dialog breakdown analysis and connectedness classification; GPT-4.1 performs best

in detecting language toxicity; and Gemini 2.5 Pro achieves the highest agreement level in criticism stance detection. The improvement in criticism stance detection by Gemini 2.5 Pro is substantial, with Kappa value increasing from around 0.5 (all other models) to 0.76. We suspect this gain may be attributed to memorization rather than genuine advances in inference capability, as the underlying dataset (Peng et al. 2022) is publicly available and may have been included in the model's training data.

Table D1 Cross-Model Performance Comparison (with Additional Models)

Task	GPT-4o	GPT-4.5	GPT-4.1	Gemini 1.5 Pro	Gemini 2.5 Pro	LLaMA 3 70B	LLaMA 3.3 70B	Claude 3.5 Sonnet	Claude 3.7 Sonnet	OpenAI o3-mini	DeepSeek-R1	GPT-OSS-120B
Business Proposal Evaluation	**0.605**	0.55	0.592	0.561	0.562	0.485	0.531	0.578	0.566	0.560	0.517	0.533
Review Attribute Detection	0.479	0.422	0.466	0.346	0.382	0.152	0.220	0.429	0.388	**0.554**	0.458	0.501
Dialog Intent Classification	0.335	0.372	0.399	0.238	0.375	0.193	0.205	**0.427**	0.364	0.191	0.251	0.187
Dialog Breakdown Analysis	0.566	0.717	0.717	0.727	0.685	0.252	0.245	0.689	**0.796**	0.223	0.744	0.617
Connectedness Presence	0.632	0.678	0.625	0.566	0.566	0.232	0.602	**0.688**	0.681	0.555	0.633	0.602
Connectedness Classification	0.366	0.354	0.355	0.324	0.303	0.177	0.338	0.371	**0.372**	0.303	0.302	0.286
Language Toxicity Detection	0.471	0.497	**0.516**	0.407	0.463	0.479	0.366	0.390	0.436	0.307	0.340	0.389
Criticism Stance Detection	0.540	0.540	0.470	0.550	**0.760**	0.500	0.500	0.570	0.540	0.510	0.580	0.520
Sentiment Analysis	0.918	0.905	0.918	0.885	0.902	0.901	0.900	0.822	0.847	0.901	**0.935**	0.900

Notes. The values represent Cohen's Kappa scores between LLM annotations and human ground truth.