# Subjective Annotation Studies: Valid Inference and Annotation Design*

Ziyu Jiang
Department of Economics, University College London

26 October 2026

Click here for the latest version

**Abstract**

We develop a statistical framework for annotation studies of inherently subjective features of unstructured data. In the absence of a conventional ground-truth measurement, we define document-level measurements relative to a reference population of annotators whose responses are relevant to the underlying economic question. Our goal is valid inference for economic parameters which can be written as weighted averages of these measurements and smooth functions of such weighted averages. We develop an annotation design and establish asymptotically valid inference for estimators that combine machine-generated measurements for the full corpus with human annotations of sampled documents. Our analysis accounts for uncertainty from document sampling, limited annotations per document, and annotator reuse. Moreover, the resulting variance formula allows us to derive the optimal allocation of a limited annotation budget for weighted corpus averages. The resulting optimal rule always prioritises increasing a document's sampling probability before increasing its number of annotations above the required minimum. It assigns more resources to documents with larger absolute weights in the target, less accurate machine predictions, or more variable annotator responses. We illustrate the method with simulations calibrated to Wikipedia toxicity annotations and an empirical application based on Shapiro and Wilson (2022).

*I am grateful to Tim Christensen, Áureo de Paula, Stephen Hansen, and Andrei Zeleneev for their guidance, encouragement, and support throughout this project. I also thank Andrew Chesher, Ben Deaner, Raffaella Giacomini, Dennis Kristensen, Daniel Lewis, Guanyi Wang, Martin Weidner, and Tian Xie, as well as participants at the 2026 ESIF Economics and AI+ML Meeting, for helpful comments and suggestions. All remaining errors are my own. Email: uctpzji@ucl.ac.uk.

# 1 Introduction

Economists increasingly use unstructured data, such as text, audio, images, and video, to construct interpretable numerical measures that conventional datasets do not provide (Gentzkow et al., 2019; Ash and Hansen, 2023). Human annotations are often crucial in this process. For some measurement tasks, these annotations can reveal a verifiably correct answer (the ground truth), but collecting them for an entire dataset is costly. Machine learning offers a scalable alternative. For example, Hansen et al. (2023) use a language model to identify job vacancies that allow remote work. Using these data, Christensen and Hansen (2026) show how limited human annotation can be used to correct regression bias caused by classification errors. More generally, recent work in econometrics and statistics shows how limited human validation data can be combined with machine-generated measurements to obtain valid inference on economic parameters (Ludwig et al., 2025; Egami et al., 2024; Angelopoulos et al., 2023). There is, however, a wide class of applications in which a verifiable ground truth is conceptually difficult to establish. Dupas et al. (2026), for example, investigate whether women presenting research in economics seminars encounter more negative interactions than men. What constitutes a negative interaction is inherently subjective and can mean different things to different people. In these settings, human annotations no longer directly supply a benchmark for validating machine outputs and correcting downstream estimates. Researchers must therefore establish what the measurements should represent and how to conduct valid inference on the economic parameters constructed from them.

Such settings are not uncommon in empirical work. Researchers have constructed measures of the ideological slant of news coverage (Gentzkow and Shapiro, 2010), the emotional tone conveyed by central-bank communication (Gorodnichenko et al., 2023), and the perceived trustworthiness of borrowers from their photographs (Duarte et al., 2012). Recent work in machine learning has argued that disagreement in subjective annotations can be systematic and informative and that the resulting labels depend on whose responses are represented (Plank, 2022; Gordon et al., 2022). We give this insight a statistical formulation. In subjective annotation settings, researchers can specify a population of annotators whose judgments are substantively relevant to the underlying economic question. We term this the reference population. For each unit of unstructured data, which we refer to generically as a "document," researchers can still define meaningful measurements by summarizing the distribution of responses to that document in the reference population, for example through its mean or variance, even in the absence of a conventional ground truth.

This paper develops theory for annotation studies that enable the estimation of economic parameters constructed from these measurements across a fixed corpus of documents. When combined with the practical constraints of annotation studies, this setting presents a distinct set of challenges for both design and inference. On the inference front, cost constraints often mean that only a sample of documents can be annotated, and measurements for those documents must be estimated from a limited number of annotator responses. In practice, annotators are commonly assigned multiple documents. Their personal tendencies and preferences can therefore induce dependence among the estimated measurements of the documents they assess. The resulting

correlation pattern depends on both the sampled documents and the assignment of annotators, posing a nonstandard inference problem. On the design front, obtaining more annotations can improve the precision of measurement for a given document but leaves fewer resources to annotate others, making the optimal allocation of annotation resources nontrivial. We specify a class of annotation designs that serves two purposes. It first provides sufficient structure to develop valid inference procedures for established estimators that combine machine predictions with limited human annotations. Once the uncertainty arising from the annotation experiment has been characterized, the flexibility within this class allows us to use the resulting asymptotic variance to derive optimal allocations of a limited annotation budget to estimate the economic parameters as precisely as possible.

The choice of reference population determines the economic interpretation of the resulting measurement and is therefore crucial to this framework. Researchers may choose a reference population for its members' domain knowledge and, more importantly, for the direct relevance of their perceptions to the economic question. Economics researchers, for example, have the knowledge needed to interpret the content and context of interactions in economics seminars, while their perceptions of the seminar environment may influence their willingness to attend and participate in future seminars. These considerations make economics researchers a natural reference population for the seminar example discussed above.

Our document-level measurements capture both average judgments and disagreement within the reference population. Under a fixed annotation protocol, our primary measurement for document $j$ is the population mean response, $\mu_j$. Disagreement among annotators can also be economically informative. For example, differences in responses to written descriptions of macroeconomic shocks reveal differences in how people understand the economy (Andre et al., 2022). To capture such disagreement, we also develop theory for measurements based on pairs of independently drawn responses to the same document. These include response variance and the probability that two categorical responses differ. The same extension also provides unbiased estimates of the squared means $\mu_j^2$, allowing us to extend the analysis to regressions that use $\mu_j$ as an explanatory variable.

We then study estimation and inference for estimands that can be written as weighted averages of these document-level measurements and smooth functions of those weighted averages. We take the $J$-document corpus and its observed characteristics as given, with weights determined by these characteristics and the underlying economic question. This class and its applications include corpus means (Gentzkow and Shapiro, 2010), comparisons across subcorpora (Dupas et al., 2026), and regression coefficients when the document-level measurement is the outcome (Shapiro and Wilson, 2022). The analysis extends naturally to fixed-dimensional vectors of such averages and smooth transformations of these vectors. These extensions cover regression coefficients when a subjective measurement is used as a regressor (Möbius and Rosenblat, 2006) and parameters from some nonlinear specifications (Kessler et al., 2019). We return to these examples in Section 2, where we show in detail how their estimands fit within our framework.

To estimate these corpus-level parameters within a limited annotation budget, researchers must decide which documents to annotate, how many annotations each receives, and how annotators

are assigned to documents. We specify a class of designs under which researchers choose in advance each document's sampling probability and the number of annotations it will receive if sampled. Documents are then selected using Poisson sampling (Särndal et al., 1992), with each document sampled independently according to its predetermined probability. We impose a fixed cap on each annotator's workload, as in many large-scale annotation studies (Agirre et al., 2013; Derczynski et al., 2016; Kenyon-Dean et al., 2018). To limit the recruitment burden, we use the smallest feasible workforce under this cap and fill all but possibly one annotator to capacity. We then match annotators to documents uniformly at random among assignments that satisfy the prescribed annotation counts and annotator workloads, with no annotator assigned to the same document twice. Within this class, we seek the sampling probabilities and annotation counts that yield the most precise estimates of the corpus-level parameters under a limited annotation budget. To do so, we must first specify the estimator, characterize the uncertainty arising from the annotation study, and establish valid inference.

Our estimator builds on established methods for combining machine predictions with observed measurements, but human annotations in our setting do not directly reveal the document-level measurements. Human annotations are costly to collect, whereas machine-generated outputs can be obtained inexpensively for the full corpus. We adopt an established solution used in econometrics (Graham, 2011; Carlson and Dell, 2025; Egami et al., 2024) and statistics (Särndal et al., 1992; Angelopoulos et al., 2023; Zrnic and Candès, 2024): we first compute the relevant corpus averages using machine predictions for all documents, then add a correction based on human measurements from sampled documents. This correction averages the differences between those measurements and the corresponding predictions, accounting for unequal sampling probabilities. In our setting, the human measurements entering this correction must themselves be estimated from a limited number of annotator responses. Reusing annotators can therefore correlate the resulting estimation errors across documents, with the correlation pattern depending on both the sampled documents and the assignment of annotators.

This complicates statistical inference, which must account for the random selection of documents, the random allocation of their evaluations across annotators, and the dependence created when annotators evaluate multiple documents. A Gaussian approximation for the estimator may be straightforward to establish conditional on the documents selected for annotation and their assignment to annotators. To quantify uncertainty about the corpus-level parameter, however, we must also account for variation across possible document samples and annotator assignments. This requires averaging the conditional distributions over these possibilities, and the resulting limit need not be Gaussian under an arbitrary allocation mechanism. We show that under our proposed annotation design, the estimator has a Gaussian limit under standard regularity conditions. We also construct a consistent estimator of the resulting asymptotic variance, allowing researchers to use standard Gaussian inference procedures. This analysis also gives practitioners flexibility in how annotators are assigned to documents. If they prefer or are compelled to use a different allocation mechanism, the same variance estimator supports standard inference, as long as the regularity conditions for the Gaussian limit can be established.

Having characterized the asymptotic variance, we can now use it to guide how annotation re-

sources should be allocated. Choosing a sampling design to improve precision under a budget is a classical problem in survey sampling (Stuart, 1954; Särndal et al., 1992). More recent work on active statistical inference and stratified prediction-powered inference also started studying how the allocation of labels can improve precision (Zrnic and Candès, 2024; Fisch et al., 2024). Our reference-population setting introduces a distinct allocation decision because each document-level measurement must itself be estimated from annotator responses. Collecting more evaluations thus improves the precision of that measurement but leaves fewer resources for annotating other documents. We therefore jointly choose document inclusion probabilities and annotation depths to minimize the asymptotic variance of the resulting estimator under an annotation budget.

We derive the optimal allocation of annotation resources for estimating weighted averages of the document-level means. We refer to the weights as target weights, as they determine how each document-level mean enters the target parameter. Importantly, within our design class, the first-order variance contribution from annotator reuse is common across allocation policies, so annotator reuse affects attainable precision but not how resources should be distributed across documents. We also show that, for any fixed expected expenditure on a document, it is weakly preferable to increase its inclusion probability before increasing the labels it receives. The optimal allocation therefore follows a coverage-before-depth rule, increasing labels per document only after the document is sampled with certainty. Profiling out this choice yields a convex problem with a generalized Neyman solution (Neyman, 1934). At a specified annotation depth, the solution assigns expected expenditure according to the document's absolute target weight and the square root of a document-specific difficulty measure. This measure depends on how much annotators' responses vary for a document, how far the machine prediction is from the reference-population mean, and how many evaluations the document receives if sampled. Thus, more resources are directed toward documents with larger absolute target weights, noisier human responses, or less accurate machine predictions.

In practice, the machine-prediction errors and annotator-response variance needed by the optimal allocation rules are unknown before annotation. A pilot with repeated annotations, or comparable annotated data, can be used to predict them from observed document characteristics. When reliable information for ranking document difficulty is unavailable, we propose a transparent default that treats documents as equally difficult in expectation for planning purposes and divides the budget using only the fixed weights in the weighted corpus average. When these weights are not too concentrated and the budget is small relative to the corpus, expected expenditure on each document is proportional to the absolute value of its weight, and every sampled document receives the minimum annotation depth. Two examples of this allocation rule are particularly interpretable. For a corpus mean, this rule samples all documents with the same probability. For a contrast between two group means, it assigns half of the expected budget to each group and therefore samples documents in the smaller group with higher probability.

To illustrate the performance of our estimator in a setting grounded in real annotation data, we conduct a calibrated simulation using the Wikipedia Talk Labels: Toxicity corpus (Wulczyn et al., 2017; Thain et al., 2017). We use the observed comments and human judgments to

estimate comment-level toxicity probabilities and calibrate a synthetic population of annotators. We apply a pretrained Cardiff offensive-language classifier to each comment to generate the machine predictions used by the estimator (Barbieri et al., 2020). For the corpus mean, nominal 95% confidence intervals achieve 94.3–94.9% coverage, compared with 63.8–66.0% when each comment's average label is treated as error-free. In a stress test with approximately 30 or 60 annotators, coverage ranges from 93.3% to 94.6%. For the difference in mean toxicity between unregistered and registered contributors, allocation based on target weights reduces empirical variance by 32.8% relative to uniform sampling. The full-information optimal allocation and its pilot-estimated version reduce variance by 41.5% and 36.8%, respectively.

We apply our framework to the setting studied by Shapiro and Wilson (2022), who relate sentiment in Federal Open Market Committee (FOMC) deliberations to economic conditions to investigate central-bank objectives. They measure sentiment using a dictionary-based score: the share of negative words minus the share of positive words. They then fit a quadratic relationship between this score and inflation to estimate the inflation rate associated with the least negative sentiment (the inflation target). We conduct an annotation study with Prolific participants selected for relevant educational, professional, or investment experience, replacing dictionary-based sentiment with the mean assessment of this reference population. We combine these human annotations with either predictions from GPT-6 Astra or a rescaled version of the authors' original dictionary score. The resulting turning-point estimates are 1.68% and 1.79%, respectively. Both are close to the authors' original estimate of approximately 1.69%, supporting the effectiveness of their dictionary-based measure for this economic task.

The remainder of the paper is organized as follows. Section 2 defines the reference-population measurements and corpus-level targets, and Section 3 specifies the annotation design. Section 4 develops the estimator and establishes valid inference under annotator reuse. Section 5 derives the optimal allocation of annotation resources and discusses its practical implementation. Section 6 evaluates inference and allocation rules through a simulation calibrated to Wikipedia toxicity data. Section 7 applies the framework to sentiment in FOMC deliberations and the estimation of the Federal Reserve's inflation target. Section 8 concludes. The appendices collect proofs and supporting results.

# 2 Model

## 2.1 Annotation outcomes

Let $W_{1:J} := (W_1, \ldots, W_J)$ denote a fixed but large corpus of $J$ documents. In the main text, we treat corpus construction as completed before the annotation study begins, and $W_{1:J}$ is held fixed throughout the main-text analysis. The parameters below therefore concern this fixed corpus and require no probability model for document generation. When a superpopulation model for document generation can be credibly imposed, it can be combined with the main analysis. Online Appendix OA.4 provides such an example for the model-based extensions.

Let $P_\theta$ denote the distribution of annotator characteristics in a prespecified reference population. Potential annotators are indexed by $i \in \mathbb{N}$, and $\theta_i$ collects all latent characteristics that may

persist across evaluations, such as background, prior beliefs, and interpretive tendencies. For each annotator–document pair, let $\varepsilon_{ij}$ represent influences on annotator $i$'s response to document $j$ that are not captured by the document $W_j$ or the annotator's persistent characteristics $\theta_i$. Let $Y_{ij}^*$ denote the response that annotator $i$ would provide if assigned to evaluate document $j$. Notice that this potential response is defined whether or not that assignment is made. We assume the responses are bounded, that is $|Y_{ij}^*| \leq C_Y$ for all $i, j$, where $C_Y < \infty$ is a fixed constant independent of $i$, $j$, and $J$. We write $\mathcal{Y} \subseteq [-C_Y, C_Y]$ for the set of possible response values. We impose the following assumptions.

**(M1) Stable potential responses.** There is a measurable function $h$, common to all pairs and otherwise unrestricted, such that

$$Y_{ij}^* = h(W_j, \theta_i, \varepsilon_{ij}), \qquad i \geq 1, \quad j \leq J.$$

The response is invariant to the realized document sample, annotator assignment, evaluation order, and workload.

**(M2) Pair-specific shocks.** The shocks $\{\varepsilon_{ij} : i \geq 1, j \leq J\}$ are mutually i.i.d. and independent of $\{\theta_i : i \geq 1\}$.

**(M3) Reference-population sampling.** The annotator types $\{\theta_i\}_{i \geq 1}$ are i.i.d. draws from $P_\theta$.

Assumption **(M1)** ensures that the potential-response notation is well defined: the response of annotator $i$ to document $j$ is a fixed function of the pair alone, invariant to the realized annotation design. It rules out fatigue, order, and workload effects, under which an annotator's response to a document would depend on the rest of their assignment.

Assumption **(M2)** makes $\varepsilon_{ij}$ a genuinely pair-specific shock. It rules out response dependence across annotator–document pairs beyond that induced by persistent annotator characteristics, including communication or herding among annotators. Responses from the same annotator may nevertheless remain dependent across documents through $\theta_i$. Importantly, $\theta_i$ need not represent only an additive annotator effect: it may encode the annotator's entire response rule, allowing different annotator types to map the same document into very different response distributions. The common function $h$ therefore does not impose a common interpretation rule across annotators.

A well-designed annotation exercise should seek to limit violations of Assumptions **(M1)**–**(M2)**. Preventing communication among annotators can limit dependence across annotators, while limiting the number of documents per session and requiring regular breaks can help reduce fatigue. Attention checks can also help detect inattention during the study. These practices are increasingly practical to implement as researchers gain access to tools for managing annotation workloads, scheduling breaks, and monitoring attention both in online and offline settings.

Assumption **(M3)** permits the reference population and its eligibility rules to reflect the application and corpus domain. Such choices define the population being averaged over and therefore form part of the measurement itself. The design determines the required workforce size. Given

this number, annotators must be recruited to represent the announced reference population. The choice of whom to recruit should not depend on the sampled documents or any final-study annotation outcomes.

## 2.2 Document-level measurements, characteristics, and pre-trained predictions

Let $\theta$ denote a generic draw from $P_\theta$, and let $\varepsilon$ denote an independent draw from the distribution of the pair-specific shocks. Our primary document-level measurement is the reference-population mean response

$$\mu_j := \mathbb{E}_{\theta,\varepsilon}[h(W_j, \theta, \varepsilon)]\,.$$

The expectation averages over annotator characteristics and pair-specific shocks, with the document $W_j$ held fixed.

To interpret $\mu_j$, consider Dupas et al. (2026), who study, in one specification, the number of seminar interruptions that fellow economists rate as negative in tenor. In our notation, let $W_j$ contain the interactions in seminar $j$, and let $Y_{ij}^*$ be the number that economist $i$ would rate as negative. Their in-person seminar sample illustrates how such a measure can be constructed by asking economists to assess the interactions directly.

If annotation resources were unlimited, the researcher could draw $M$ economists independently from a prespecified population of economics researchers, ask each to evaluate seminar $j$ once, and average their responses. Under Assumptions **(M1)**–**(M3)**,

$$\frac{1}{M}\sum_{i=1}^{M} Y_{ij}^* \xrightarrow{\text{a.s.}} \mu_j \qquad \text{as } M \to \infty.$$

The mean provides a natural measure of response level and is the primary document level measurement we consider. However, in some applications, heterogeneity in reference-population responses is itself substantively important. To some extent, the mean framework itself can accommodate questions about response heterogeneity by changing the outcome being averaged. For example, if $Y_{ij}^*$ is an ordered categorical response, the researcher may instead take the outcome to be the indicator that the response lies at or below a prespecified cutoff $c$. The resulting document-level mean is

$$\mu_j(c) := \mathbb{E}\big[\mathbf{1}\{Y_{ij}^* \le c\}\big] = \Pr(Y_{ij}^* \le c),$$

the probability that a randomly drawn member of the reference population returns a response at or below that cutoff. Different cutoffs reveal different features of the response distribution.

In many applications, however, researchers may prefer a measurement that directly captures the extent of disagreement or contestedness. Let $Y_{1j}^*$ and $Y_{2j}^*$ denote the responses of two independently drawn annotators, including independent pair-specific shocks. A measure of

response dispersion can be defined as

$$\nu_j := \frac{1}{2}\mathbb{E}\big[(Y_{1j}^* - Y_{2j}^*)^2\big] = \mathrm{Var}(Y_{ij}^*).$$

Pairs of responses are also useful when $\mu_j$ enters a linear regression as a regressor. Such regressions require corpus averages of $\mu_j^2$. However, simply squaring a document's average annotation overestimates $\mu_j^2$ on average by the variance of that average. The resulting measurement-error bias need not disappear as the corpus grows if each document receives only a few evaluations. Products of responses from two independently drawn annotators instead have expectation $\mu_j^2$, allowing us to estimate the required corpus moment consistently.

More generally, we extend our analysis to pairwise response measurements of the form

$$\vartheta_j(\ell) := \mathbb{E}\big[\ell_j(Y_{1j}^*, Y_{2j}^*)\big]\,,$$

where $\ell_j$ is a prespecified symmetric comparison. Their natural sample analogues compare pairs of annotations within a document and are order-two $U$-statistics.

Before annotation begins, the researcher may already have some information for every document in the corpus. Because $W_j$ is an unstructured object, it enters the econometric analysis only after being mapped into numerical variables or scores. Let $X_j \in \mathbb{R}^d$, for fixed $d$, collect prespecified document characteristics and numerical representations of its content, such as its source and date, characteristics of its author or speaker, pretrained text, image, or audio embeddings. Researchers may use these characteristics as regressors in downstream analyses or to guide the choice of document sampling probabilities and the number of evaluations per sampled document.

Before documents are selected for human annotation, the researcher may also construct a scalar prediction $g_j$ of $\mu_j$ from $W_j$, $X_j$, or both. The prediction procedure may be an off-the-shelf classifier, a model trained on auxiliary data, or an LLM. In the LLM case, the prespecified procedure includes the model version, prompt, conversion of the response into a numerical score, number of calls, and any rule used to combine outputs across calls.

Once the procedure has been applied to every document, the resulting vector $g_{1:J}$ is held fixed: it is not regenerated, re-trained, or revised after the annotation sample is drawn or in response to the final-study annotations. We call $g_j$ a *frozen prediction* precisely in this timing sense. If the procedure is stochastic, the subsequent analysis conditions on the realized predictions. Thus, both a single realized call and a prespecified aggregation of repeated calls are permitted.[1] The frozen prediction may inform the annotation design and enter the prediction-assisted estimator. It need not be unbiased, calibrated, or consistent. As we show below, its predictive quality affects efficiency rather than validity. However, we do require it to be bounded, given the original annotation outcome is bounded.

**(R1) Bounded frozen predictions.** We assume that $|g_j| \le \bar{g}$ for every document $j$, where $\bar{g} < \infty$ is a fixed constant independent of $j$ and $J$.

[1] Generative-model outputs may vary across calls. Han (2025) discusses the use of separate LLM sessions, while Gao et al. (2026) treat a pretrained LLM and its prompting protocol as a fixed prediction rule.

## 2.3 Economic parameters generated by corpus moments

The preceding subsection defines document-level measurements, including the reference-population mean $\mu_j$ and the pairwise measurements $\vartheta_j(\ell)$. Economic applications, however, generally concern corpus-level parameters. We consider parameters built from weighted averages of these document-level measurements of the form

$$\Phi_J(a) = \frac{1}{J}\sum_{j=1}^{J} a_j \mu_j, \qquad \Phi_J^{\ell}(a) = \frac{1}{J}\sum_{j=1}^{J} a_j \vartheta_j(\ell),$$

where the weights $a_j$ may depend on observed corpus characteristics, such as the $X_j$'s, but not on the potential responses $Y_{ij}^*$.

Let $\boldsymbol{\Phi}_J$ denote a fixed-dimensional vector whose components take the form $\Phi_J(a)$ or $\Phi_J^{\ell}(a)$. We study corpus-level parameters that can be written as

$$\boldsymbol{\beta}_J = H_J(\boldsymbol{\Phi}_J), \tag{2.1}$$

where $H_J$ is a known, corpus-specific map that is continuously differentiable in a neighborhood of $\boldsymbol{\Phi}_J$. The map may depend on fixed observed attributes of the entire corpus.

The following applications fall in this class.

**Reader-perceived media slant (Gentzkow and Shapiro, 2010).** The authors study political slant in U.S. newspapers and construct an automated measure based on partisan phrase use. In our framework, $W_j$ is the text of article $j$, and $Y_{ij}^*$ is reader $i$'s assessment of the article's political position. A natural reference population is U.S. newspaper readers. Thus $\mu_j$ is the mean perceived political position of article $j$ in this population. For a newspaper $n$ with a corpus $C_n$ of $J_n$ articles, the newspaper-level target is

$$\beta_{J_n,n} = \frac{1}{J_n}\sum_{j\in C_n} \mu_j = \Phi_{J_n}(\mathbf{1}).$$

An article-level adaptation of the authors' phrase-scoring rule could provide a natural $g_j$.

**Negative interruptions in economics seminars (Dupas et al., 2026).** The authors study whether female presenters are treated more harshly than male presenters in economics seminars. They combine human coding with audio-processing algorithms. In our framework, $W_j$ is the available audio or audiovisual record of seminar $j$, $X_j$ includes seminar duration and presenter characteristics, and $D_j$ indicates a female presenter. The authors recruited human coders from the economics research community, mainly graduate students. We therefore take the reference population to be economics researchers. Let $Y_{ij}^*$ be the number of audience interactions in recording $j$ that researcher $i$ rates as negative in tenor. Thus $\mu_j$ is the mean number of audience interactions rated as negative in seminar $j$, averaging over the population of economics

researchers. With $J_1 := \sum_{j=1}^{J} D_j$ and $J_0 := J - J_1$, the female–male contrast is

$$\beta_J = \frac{1}{J_1} \sum_{j:D_j=1} \mu_j - \frac{1}{J_0} \sum_{j:D_j=0} \mu_j = \Phi_J(a^\Delta), \qquad a_j^\Delta = J \left\{ \frac{D_j}{J_1} - \frac{1 - D_j}{J_0} \right\}.$$

This unadjusted contrast is a weighted corpus moment. If $\widetilde{D}_j$ residualizes presenter gender on prespecified seminar-series effects and fixed controls, the corresponding finite-corpus regression coefficient is also a weighted mean, with $a_j = J\widetilde{D}_j / \sum_{k=1}^{J} \widetilde{D}_k^2$. A natural $g_j$ is the seminar-level count of interruptions classified by the authors' pretrained audio model as having negative tone.

More generally, linear regressions with $\mu_j$ as the outcome and observed corpus characteristics as regressors also fall into this class. For the seminar example, let $\widetilde{X}_j = (1, D_j, X_j')'$, where $X_j$ contains the prespecified controls. Provided the regressor second-moment matrix is nonsingular, the least-squares coefficients are

$$\boldsymbol{\beta}_J = \arg\min_{\boldsymbol{b}} \frac{1}{J} \sum_{j=1}^{J} (\mu_j - \widetilde{X}_j' \boldsymbol{b})^2 = \left( \frac{1}{J} \sum_{j=1}^{J} \widetilde{X}_j \widetilde{X}_j' \right)^{-1} \left( \frac{1}{J} \sum_{j=1}^{J} \widetilde{X}_j \mu_j \right).$$

**Perceived attractiveness as a regressor (Möbius and Rosenblat, 2006).** The authors study the beauty premium in an experimental labor market. They have photographs of all participants rated on a five-point attractiveness scale by 50 high-school students from Tucumán and use the resulting beauty measure in regressions for worker confidence and productivity and for employers' wage assessments. To illustrate how our framework could be applied to a worker-level regression, let the corpus consist of photographs of the workers included in that regression. Then $W_j$ is worker $j$'s photograph and $Y_{ij}^*$ is the attractiveness rating that annotator $i$ would assign to it. One natural reference population, matching the original annotator pool, is high-school students in Tucumán. Thus $\mu_j = \mathbb{E}[Y_{ij}^*]$ is worker $j$'s mean perceived attractiveness in this population.

Let $R_j$ be a fixed worker outcome, such as reported confidence or realized productivity, let $X_j$ contain the remaining observed regressors, and write $\widetilde{X}_j = (1, X_j', \mu_j)'$. The regression target under our framework is

$$\boldsymbol{\beta}_J = \arg\min_{\boldsymbol{b}} \frac{1}{J} \sum_{j=1}^{J} (R_j - \widetilde{X}_j' \boldsymbol{b})^2.$$

The document-level measurements enter these coefficients through corpus averages of $\mu_j$, $X_j\mu_j$, $R_j\mu_j$, and $\mu_j^2$; all remaining quantities are observed for the full corpus. The first three are weighted averages of $\mu_j$. The last can be expressed using a pairwise measurement, since the product of responses from two independently drawn annotators has expectation $\mu_j^2$. If the regressor second-moment matrix is nonsingular, the coefficients are a smooth function of these corpus averages, as in equation (2.1).

A natural $g_j$ is a frozen vision-model prediction of the reference-population mean rating under the same rating protocol. Its square, $g_j^2$, can serve as an auxiliary prediction for $\mu_j^2$.

**Employer evaluations of résumés (Kessler et al., 2019).** The authors ask employers to rate hypothetical résumés on an ordered hiring-interest scale; their appendix also dichotomizes the rating at several cutoffs and estimates logit specifications for the resulting counterfactual-callback outcomes. Let $W_j$ be the complete résumé, let $Y_{ij}^*$ be employer $i$'s hiring-interest rating, and, for a prespecified cutoff $c$, define

$$B_{ij}^*(c) = \mathbf{1}\{Y_{ij}^* \geq c\}, \qquad \mu_j(c) = \mathbb{E}[B_{ij}^*(c)].$$

The reference population can be defined as employers recruiting Penn seniors in the relevant field, with separate populations for STEM and business/humanities/social-science recruiting when their evaluation criteria differ. Let $Z_j$ collect the fixed résumé characteristics in the economic specification. Let $\Lambda(t) := \{1 + \exp(-t)\}^{-1}$ denote the logistic link. We define the logistic regression coefficients by

$$\boldsymbol{\beta}_J(c) = \arg\min_{\boldsymbol{b}} \frac{1}{J} \sum_{j=1}^{J} \big[ -\mu_j(c) \log \Lambda(Z_j'\boldsymbol{b}) - \{1 - \mu_j(c)\} \log\{1 - \Lambda(Z_j'\boldsymbol{b})\} \big]. \tag{2.2}$$

Define

$$\boldsymbol{\Phi}_J(c) = \frac{1}{J} \sum_{j=1}^{J} Z_j \mu_j(c).$$

Each coordinate of $\boldsymbol{\Phi}_J(c)$ is a weighted first-order corpus moment. For a finite, locally unique minimizer of (2.2), the score equation is

$$\boldsymbol{\Phi}_J(c) - \frac{1}{J} \sum_{j=1}^{J} Z_j \Lambda\big(Z_j'\boldsymbol{\beta}_J(c)\big) = 0.$$

If its Jacobian with respect to $\boldsymbol{\beta}$ is nonsingular, the implicit-function theorem gives a locally continuously differentiable map $H_J$ such that

$$\boldsymbol{\beta}_J(c) = H_J\{\boldsymbol{\Phi}_J(c)\}.$$

Although these applications report different economic parameters, their dependence on the reference-population response distributions is captured by a fixed-dimensional collection of weighted first-order and order-two corpus moments. These moments are therefore the common population objects developed in the sections that follow.

# 3 Annotation design

Machine predictions can be generated inexpensively for the full corpus, whereas human evaluations are costly and time-consuming. Recruiting fresh annotators for every annotation may also be impractical. Under a limited annotation budget, the researcher must therefore choose which documents to annotate, how many evaluations each receives, and how annotators are assigned to documents.

### 3.1 Document sampling and the annotation budget

Throughout the analysis, annotation expenditure is measured by the number of requested labels. The document characteristics $X_{1:J}$, machine predictions $g_{1:J}$, and target weights $a_{1:J}$ are fixed. Using this information, the researcher specifies in advance the probability that each document $j$ is sampled (its inclusion probability, $\pi_j \in (0, 1]$). The researcher also specifies how many evaluations it receives if sampled (its annotation depth, $m_j \in \{1, \ldots, \bar{m}\}$), where $\bar{m}$ is a fixed finite integer. These are the two document-level design choices. The inclusion probabilities govern the extensive margin: how much of the corpus is expected to be covered and which documents are more likely to be annotated. The annotation depths govern the intensive margin: how precisely each sampled document's measurement can be estimated.

Let $I_j = 1$ if document $j$ is sampled and $I_j = 0$ otherwise. Let $\mathcal{S}_J := \{j : I_j = 1\}$ denote the set of sampled documents.

**(D1) Poisson document sampling.** The indicators $\{I_j\}_{j=1}^{J}$ are mutually independent, with

$$I_j \sim \text{Bernoulli}(\pi_j), \qquad j = 1, \ldots, J.$$

Poisson sampling with unequal inclusion probabilities is standard in survey sampling (Särndal et al., 1992; Tillé, 2006). It is particularly useful here because it gives the researcher direct control over each document's marginal inclusion probability while preserving independence across document-selection decisions. This independence simplifies the variance analysis. The expected budget constraint $\sum_j \pi_j m_j = \bar{B}$ is also additive across documents, making it easier to optimize inclusion probabilities and annotation depths for later analysis.

This convenience comes with a limitation: the realized annotation expenditure $B$ is random. A constraint on expected expenditure does not guarantee that realized expenditure stays within that budget. Such situations also arise in optimal experimental design and treatment-policy choice (Hahn et al., 2011; Sun, 2026). In our framework, a strict upper bound can nevertheless be enforced with asymptotically negligible distortion. Let $B_{\max}$ be the largest feasible number of labels and choose the expected budget below this cap so that, for example,

$$B_{\max} - \bar{B} \geq \sqrt{\bar{B} \log \bar{B}}.$$

Under the bounded-depth condition, Lemma 5 implies

$$\Pr(B > B_{\max}) \longrightarrow 0.$$

If the cap must never be exceeded, samples with $B > B_{\max}$ can be rejected and redrawn before annotation begins. The accepted-sample law has total variation distance $\Pr(B > B_{\max})$ from the original Poisson law, so the rejection step does not alter the asymptotic conclusions.

When the annotation budget must be fixed exactly, Online Appendix OA.2 develops a stratified simple-random-sampling alternative for these applications. A fixed number $n_h$ of documents is sampled without replacement from each prespecified stratum $h$ of size $J_h$, and every selected

document in that stratum receives a common annotation depth $m_h$. Consequently,

$$B = \sum_{h=1}^{H} n_h m_h$$

is fixed by design. We refer to this as the stratified fixed-budget design, Design **(SFB)**. It sacrifices fully document-specific control of $\pi_j$ and $m_j$ in exchange for stratum-level control and an exact finite-sample annotation budget. Poisson sampling will remain the main sampling regime in this paper, while Design **(SFB)** provides a fixed-budget alternative for applications that require strict budget control.

### 3.2 Annotator allocation

After document sampling, the researcher must assign the $B$ requested evaluations to annotators. Each selected document $j$ is evaluated by $m_j$ distinct annotators. We allow annotators to evaluate multiple documents, but impose a fixed workload cap $K$.

Recruiting and managing annotators is itself costly, so it is desirable to use recruited annotators up to their workload cap. For $B > 0$, define

$$n_0 := \left\lceil \frac{B}{K} \right\rceil, \qquad r_B := B - K(n_0 - 1) \in \{1, \dots, K\}.$$

Our capacity-filled default uses $n_0$ annotators: $n_0 - 1$ receive $K$ documents and the remaining annotator receives $r_B$. Thus $n_0$ is the smallest workforce compatible with the cap, and all but possibly one active annotator are used to capacity. The cap $K$ limits the number of documents evaluated by the same annotator and hence the amount of dependence created by annotator reuse.

Once the document and annotator workloads have been specified, the remaining design choice is how to match documents to annotators. We represent such a matching by a simple bipartite graph $A_J$, whose two vertex sets are the selected documents and the active annotators. An edge $(j, i)$ means that annotator $i$ evaluates document $j$. The document degrees therefore record the requested annotation depths, while the annotator degrees record annotator workloads. We impose simplicity on the graph, that is, the same annotator does not evaluate a document more than once. Let $M_j$ denote the set of annotators assigned to evaluate document $j$, with $M_j = \varnothing$ when $I_j = 0$.

In the baseline design, annotators are ex ante symmetric: they are recruited from the same reference population, and the allocation rule does not use annotator types or potential responses. Conditional on the prescribed document and annotator workloads, there is, therefore, no reason to prefer one feasible matching over another. We implement this symmetry by drawing the allocation uniformly from all feasible matchings.

Formally, let

$$\boldsymbol{d}_J := (m_j)_{j \in \mathcal{S}_J}, \qquad \boldsymbol{k}_J^{\mathrm{cf}} := (\underbrace{K, \dots, K}_{n_0 - 1}, r_B)$$

denote the document- and annotator-degree sequences. Both have degree sum $B$. For degree vectors $\boldsymbol{d}$ and $\boldsymbol{k}$, let

$$\mathfrak{A}_J(\boldsymbol{d}, \boldsymbol{k})$$

denote the class of labelled simple bipartite graphs on document vertices $\mathcal{S}_J$ and annotator vertices $\{1, \ldots, n_0\}$ having those degree sequences.

**(D2) Random-annotator allocation.** If $B > 0$ and the prescribed workloads are feasible,[2] then, conditional on $\mathcal{S}_J$,

$$A_J \sim \mathrm{Unif}\Big(\mathfrak{A}_J(\boldsymbol{d}_J, \boldsymbol{k}_J^{\mathrm{cf}})\Big)\,.$$

The allocation can be generated using established sampling methods. A simple approach uniformly matches the requested evaluations to annotator slots with the prescribed workloads, repeating the entire draw if any annotator is assigned the same document twice. This produces an exact uniform allocation (Blanchet and Stauffer, 2013). When rejection becomes computationally costly, established swap algorithms provide an alternative: they approximate the uniform distribution by repeatedly proposing exchanges of assignments between annotators while preserving workloads and preventing repeated annotator–document assignments (Erdős et al., 2018). Appendix E.1 describes this implementation and shows that its approximation error can be made negligible for our asymptotic results.

Let $\mathcal{G}_J$ denote the information contained in the realized document sample and annotator allocation, and let $\mathcal{L}_J$ denote the information contained in all annotator types and annotator–document response shocks:

$$\begin{aligned}\mathcal{G}_J &:= \sigma(\mathcal{S}_J, A_J),\\ \mathcal{L}_J &:= \sigma(\{\theta_i\}_{i\geq 1}, \{\varepsilon_{ij} : i \geq 1,\ j \leq J\})\,.\end{aligned}$$

**(D3) Design exogeneity.** The sigma-fields $\mathcal{G}_J$ and $\mathcal{L}_J$ are independent.

Assumption **(D3)** permits the design to use document-side information fixed before the current annotation study, but rules out allocation based on annotator types, potential responses, or pair-specific shocks from the study.

**Allocation under the exact-budget design** Under Design **(SFB)**, the same capacity-filled allocation is conducted separately within each stratum. Stratum $h$ contributes the fixed workload $B_h = n_h m_h$; conditional on its document sample, the allocation is drawn uniformly from the feasible graphs with document degree $m_h$ and the corresponding capacity-filled annotator degrees. Annotator pools are disjoint and graph draws are independent across strata. Thus each stratum fills its assigned budget, all but possibly one active annotator per stratum are used to capacity, and no annotator is shared across strata.

[2] If $B > 0$ and these workloads are infeasible, we recruit $B$ annotators and assign each requested evaluation to a different annotator. The probability that this alternative is needed tends to zero as $\bar{B} \to \infty$; see Appendix C.3. If $B = 0$, no annotators are recruited and $A_J$ is the empty graph.

# 4 Estimation and inference

Subsection 2.3 expressed the economic parameters of interest as smooth functions of weighted averages at the corpus level. The main text develops estimation and inference for the building block of such parameters:

$$\Phi_J(a) := \frac{1}{J}\sum_{j=1}^{J} a_j \mu_j,$$

a weighted average of the primary document-level measurements $\mu_j$. Online Appendix OA.1 develops the corresponding results for two-response document measurements, while Online Appendix OA.3 combines a fixed number of mean and two-response moments and extends their joint limit theory to smooth and implicitly defined economic parameters using delta-method and implicit-function arguments.

The framework targets applications with reasonably large annotation budgets, typically comprising hundreds or thousands of labels, but much larger corpora containing thousands, tens of thousands, or more documents. Formally, we use the framework of finite-population asymptotics (Abadie et al., 2020, Section 3.1), under which

$$J \to \infty \qquad \text{and} \qquad \bar{B} \to \infty.$$

No limit is imposed on $\bar{B}/J$. In particular, the analysis permits the sparse regime $\bar{B}/J \to 0$, at any rate consistent with $\bar{B} \to \infty$, as well as regimes in which the annotation budget remains nonnegligible relative to the corpus size.

The researcher observes $X_j$, $g_j$, and $a_j$ for every document in the corpus, but observes human evaluations only for documents selected for annotation. This section investigates the commonly used prediction-assisted estimator for $\Phi_J(a)$ and its statistical properties.

## 4.1 Prediction-assisted estimation and unbiasedness

For each selected document $j \in \mathcal{S}_J$, let

$$\bar{Y}_j := \frac{1}{m_j}\sum_{i \in M_j} Y_{ij}^*$$

denote its average observed evaluation. Set $\bar{Y}_j = 0$ when $I_j = 0$. Assumptions **(D3)** and **(M3)** imply

$$\mathbb{E}\big[\bar{Y}_j \mid \mathcal{G}_J\big] = \mu_j \qquad \text{on } \{I_j = 1\}.$$

Thus, conditional on the realized sample and annotator allocation, $\bar{Y}_j$ is an unbiased measurement of $\mu_j$.

We estimate $\Phi_J(a) = J^{-1}\sum_{j=1}^{J} a_j \mu_j$ following a standard approach in survey sampling (Cassel et al., 1976; Särndal et al., 1992) that is also used in econometrics (Graham, 2011; Carlson and

Dell, 2025; Egami et al., 2024):

$$\widehat{\Phi}_J(a) = \frac{1}{J}\sum_{j=1}^{J} a_j g_j + \frac{1}{J}\sum_{j\in\mathcal{S}_J} \frac{a_j}{\pi_j}\big(\bar{Y}_j - g_j\big). \tag{4.1}$$

The first term applies the machine prediction to every document in the corpus. The second uses the annotated sample to estimate the corresponding full-corpus weighted prediction error, with inverse-probability weighting correcting for document selection. The predictions therefore provide an initial full-corpus estimate, while the human evaluations estimate the correction required to recover the target. If $g_j = 0$ for every $j$, equation (4.1) reduces to the ordinary Horvitz–Thompson estimator based on the observed document means.

**Proposition 1** (Unbiasedness)**.** *Suppose Assumptions **(M1)**–**(M3)** and **(D3)** hold, and the allocation is simple, so that each selected document is evaluated by $m_j$ distinct annotators. Under any document-sampling design with known positive first-order inclusion probabilities*

$$\Pr(I_j = 1) = \pi_j > 0, \qquad j = 1, \ldots, J,$$

*for every finite $J$,*

$$\mathbb{E}\Big[\widehat{\Phi}_J(a)\Big] = \Phi_J(a).$$

*Proof.* Design exogeneity and reference-population sampling give

$$\mathbb{E}\big[\bar{Y}_j - g_j \,\big|\, I_j = 1\big] = \mu_j - g_j.$$

Consequently,

$$\mathbb{E}\left[\frac{I_j}{\pi_j} a_j(\bar{Y}_j - g_j)\right] = \frac{a_j}{\pi_j}\Pr(I_j = 1)(\mu_j - g_j) = a_j(\mu_j - g_j).$$

Summing across documents and adding $J^{-1}\sum_{j=1}^{J} a_j g_j$ proves the result. □

The proof uses only the first-order inclusion probabilities and does not require independent document selection. Under the baseline Poisson design, these probabilities are the $\pi_j$ specified in Assumption **(D1)**. Under Design **(SFB)**, a document in stratum $h$ has inclusion probability $\pi_h = n_h/J_h$. Since $\rho_h = J_h/J$,

$$\frac{1}{J\pi_h} = \frac{\rho_h}{n_h},$$

so equation (4.1) reduces exactly to the stratified version in Online Appendix OA.2. Notice that the sampling law here affects only the variance and subsequent inference, but not our derivation of unbiasedness.

Equation (4.1) is also the fixed-corpus, unequal-probability analogue of the mean estimator in prediction-powered inference (Angelopoulos et al., 2023). The channel through which predictions can improve efficiency is the same. As discussed by Angelopoulos et al. (2023), predictions available for the full corpus allow the annotated sample to estimate only the prediction error. When $g_j$ tracks $\mu_j$ well in the design-weighted sense, the residual correction is less variable,

reducing document-sampling uncertainty. The predictions need not be unbiased, calibrated, or consistent; poor predictions may indeed increase this component of uncertainty, but cannot introduce bias.

The inferential challenge here, however, is substantially different. In standard model-assisted survey sampling and prediction-powered inference, the outcome is observed once a unit enters the labeled sample. Here, selecting document $j$ does not reveal $\mu_j$: its correction must itself be estimated from finitely many annotator evaluations, and annotator reuse can correlate these measurement errors across documents. Valid inference must therefore account jointly for document sampling, finite annotation depth, and the dependence induced by the allocation graph.

### 4.2 Asymptotic normality under annotator reuse

The asymptotic argument begins with a decomposition of the estimation error. Let

$$r_j := \mu_j - g_j$$

denote the prediction residual for document $j$. Adding and subtracting the sampled residual correction gives

$$\widehat{\Phi}_J(a) - \Phi_J(a) = \underbrace{\frac{1}{J}\sum_{j=1}^{J}\left(\frac{I_j}{\pi_j} - 1\right) a_j r_j}_{T_J^s(a)} + \underbrace{\frac{1}{J}\sum_{j\in\mathcal{S}_J}\frac{a_j}{\pi_j}\big(\bar{Y}_j - \mu_j\big)}_{T_J^m(a)}. \tag{4.2}$$

The first term is the usual Horvitz–Thompson error applied to the prediction residual. As in model-assisted survey sampling and prediction-powered inference, it arises because the residual correction is observed only for the sampled documents. The second term is annotation measurement error and marks the main departure from those settings. Even for a sampled document, $\mu_j$ is not directly observed: the collected labels are finitely many draws from the reference population whose mean is $\mu_j$. We analyze the two components separately before combining them.

As with other estimators built from inverse-probability weights, asymptotic normality requires some control over extreme design weights. In our setting, the inclusion probability $\pi_j$ and annotation depth $m_j$ determine expected expenditure $\pi_j m_j$ on document $j$. Dividing the expected annotation budget equally across documents would assign $\bar{B}/J$ labels to each document in expectation. Define

$$\lambda_j := \frac{J\pi_j m_j}{\bar{B}},$$

which compares document $j$'s expected expenditure with the expenditure under this equal allocation.

**(R2) Balanced spending.** For some constant $c_\lambda \in (0, 1]$, independent of $J$,

$$\lambda_j \geq c_\lambda, \qquad j = 1, \ldots, J. \tag{4.3}$$

Assumption **(R2)** imposes a one-sided balance requirement: expected spending on each document must be at least a fixed fraction $c_\lambda$ of the expenditure under equal allocation. Individual expected expenditures may be small, provided they satisfy this relative lower bound. The condition permits heterogeneous inclusion probabilities and annotation depths, including sparse sampling with $\bar{B}/J \to 0$, provided that $\bar{B} \to \infty$. Together with the maintained boundedness conditions and fixed annotator capacity, this lower bound controls the inverse-probability weights and ensures that no individual document or annotator dominates the normalized estimation error. These properties support the central limit theorem established below.

### 4.2.1 Document-sampling uncertainty

We first consider the sampling term, $T_J^s(a)$. Under Poisson sampling, the scaled variance of the first term in (4.2) is

$$v_{r,J}(a) := \bar{B}\,\mathrm{Var}\{T_J^s(a)\} = \frac{\bar{B}}{J^2}\sum_{j=1}^{J}\frac{1-\pi_j}{\pi_j}a_j^2 r_j^2. \tag{4.4}$$

The document-sampling CLT requires that no document receive enough inverse-probability weight to affect the limit on its own. Under uniformly bounded target weights, $|a_j| \le \bar{a}$, and annotation depths, $m_j \le \bar{m}$, balanced spending gives

$$\delta_J^{\mathrm{lev}}(a) := \max_{1\le j\le J}\frac{\sqrt{\bar{B}}}{J\pi_j}|a_j| = \max_j \frac{m_j|a_j|}{\lambda_j\sqrt{\bar{B}}} \le \frac{\bar{m}\bar{a}}{c_\lambda\sqrt{\bar{B}}} \longrightarrow 0. \tag{4.5}$$

The quantity $\delta_J^{\mathrm{lev}}(a)$ measures the largest document-level leverage after the $\sqrt{\bar{B}}$ normalization. With uniformly bounded prediction residuals, (4.5) implies the Lindeberg condition for the document-sampling term: no individual document makes a first-order contribution to its asymptotic distribution.

**Lemma 1** (Document-sampling central limit theorem)**.** *Suppose documents are selected independently by Poisson sampling, $\bar{B} \to \infty$, Assumption **(R2)** holds, and the target weights, annotation depths, and prediction residuals are uniformly bounded. If*

$$v_{r,J}(a) \longrightarrow v_r(a), \tag{4.6}$$

*then*

$$\sqrt{\bar{B}}\,T_J^s(a) \xrightarrow{d} \mathcal{N}(0, v_r(a))\,.$$

Equation (4.6) is Condition **(R3)**. Conditions of this form, requiring a normalized design variance or covariance matrix to converge, are standard in finite-population design-based asymptotics; see, for example, Francisco and Fuller (1991, Condition 3), Deville and Särndal (1992, p. 379), and Boistard et al. (2017, Condition **(HT2)**). In our setting, the condition requires only that the design-weighted second moment of the prediction residuals converge to a finite limit $v_r(a) \ge 0$. The condition follows, for example, from an ordinary law of large numbers under a well-behaved document superpopulation. It can equally hold for a deterministic sequence of fixed corpora, for example when the corpus has finitely many document strata whose

proportions and within-stratum weighted second moments stabilize. The lemma is the standard triangular-array Lindeberg–Feller result; its proof is given in Appendix C.1.

### 4.2.2 Measurement uncertainty and the high-level limit

The measurement term requires more care. Reusing an annotator can induce correlation across the documents evaluated by that annotator. Following the usual remedy in clustered-data settings (Hansen and Lee, 2019), we rewrite the measurement term as a sum of annotator contributions. Let $\mathcal{J}_i$ denote the documents assigned to annotator $i$, and define

$$\zeta_i(a) := \frac{\sqrt{\bar{B}}}{J} \sum_{j \in \mathcal{J}_i} \frac{a_j}{\pi_j m_j} (Y_{ij}^* - \mu_j), \qquad \sqrt{\bar{B}}\, T_J^m(a) = \sum_i \zeta_i(a).$$

Conditional on $\mathcal{G}_J$, the annotator contributions are independent and have mean zero. Balanced spending, together with bounded annotator capacity and the maintained boundedness conditions, ensures that no annotator dominates the sum. Appendix C.1 therefore gives a conditional Gaussian approximation for the complete measurement term.

However, this is not quite enough. The remaining issue is most directly illustrated by the conditional variance. Define

$$\mu_j^\theta(\theta) := \mathbb{E}\big[Y_{ij}^* \mid \theta_i = \theta\big], \qquad \tau_{jk} := \mathrm{Cov}_\theta\Big\{\mu_j^\theta(\theta), \mu_k^\theta(\theta)\Big\},$$

and let $N_{jk} := |M_j \cap M_k|$. As shown in the appendix, the graph-specific measurement variance is

$$\begin{aligned} v_{\varepsilon,J}(a) &:= \bar{B}\, \mathrm{Var}\{T_J^m(a) \mid \mathcal{G}_J\} \\ &= \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{I_j a_j^2 \nu_j}{\pi_j^2 m_j} + \frac{\bar{B}}{J^2} \sum_{j \neq k} \frac{I_j I_k N_{jk}}{\pi_j \pi_k m_j m_k} a_j a_k \tau_{jk}, \end{aligned} \tag{4.7}$$

where the second sum is over ordered document pairs. The first component captures variation across fresh evaluations of the same document. The second is the overlap component generated by annotator reuse. In general, this variance—especially its overlap component—depends on the realized sample and allocation graph through quantities such as $N_{jk}$.

Conditional on $\mathcal{G}_J$, the annotator-level argument gives a centered Gaussian approximation for $\sqrt{\bar{B}}\, T_J^m(a)$ with variance $v_{\varepsilon,J}(a)$. The sampling term $T_J^s(a)$, however, is then a fixed and generally nonzero shift. Integrating over $\mathcal{G}_J$ therefore averages Gaussian approximations with different locations and variances. The resulting limit need not be Gaussian, even though the sampling and measurement terms can be shown to be uncorrelated.

This resembles the random-environment problem in Andrews (2005), where random conditional variances generate mixed-normal limits. The additional complication here is that the environment also contains the document-selection realization and hence fixes the separate sampling shift. Standardizing by the graph-specific measurement variance alone therefore does not remove the random location.

Martingale central limit theory suggests a way forward. A standard condition in this literature is that the conditional variance converges to a nonrandom limit (Hall and Heyde, 1980, Chapter 3). We impose a similar variance-stabilization condition **(R4)**:

$$v_{\varepsilon,J}(a) \xrightarrow{p} v_{\varepsilon}(a), \tag{4.8}$$

where $v_{\varepsilon}(a)$ is a finite, nonrandom, nonnegative constant. To understand its role here, recall that the normalized measurement term is approximately Gaussian conditional on the document sample and annotator assignment, with conditional mean zero. Since a Gaussian distribution is determined by its mean and variance, this approximation depends on the realized sample and assignment only through its conditional variance. Variance stabilization makes the approximating distribution asymptotically the same across realized document samples and annotator assignments, so the measurement term becomes asymptotically independent of both. Since the document-sampling term is determined by the selected documents, the two terms have independent Gaussian limits. Their sum therefore has a Gaussian limit with variance $v_r(a) + v_{\varepsilon}(a)$. Appendix C.1 formalizes this argument.

**Theorem 1** (Asymptotic normality under annotator reuse)**.** *Maintain Assumptions **(M1)**–**(M3)**, **(D1)**, **(D3)**, and **(R1)**. Suppose $J \to \infty$, $\bar{B} \to \infty$, the target weights and annotation depths are uniformly bounded, and the allocation is simple with every annotator assigned at most a fixed number $K$ of documents. Suppose also that Assumption **(R2)** holds. If* (4.6) *and* (4.8) *hold and*

$$v_r(a) + v_{\varepsilon}(a) > 0,$$

*then*

$$\sqrt{\bar{B}} \left\{ \widehat{\Phi}_J(a) - \Phi_J(a) \right\} \xrightarrow{d} \mathcal{N}(0, v_r(a) + v_{\varepsilon}(a)) \,. \tag{4.9}$$

*Because $B/\bar{B} \xrightarrow{p} 1$, the same conclusion holds with $\sqrt{B}$ in place of $\sqrt{\bar{B}}$.*

Although similar, Condition **(R4)** is fundamentally different from Condition **(R3)**. Condition **(R3)** concerns convergence of the deterministic document-sampling variance along the fixed corpus sequence, whereas Condition **(R4)** requires the random conditional measurement variance to converge to a nonrandom limit despite variation in the sampled documents and annotator assignments. Its validity therefore depends on the allocation rule as well as document sampling. At this stage, we impose this variance behavior as a high-level condition. The next subsection shows that, under additional regularity conditions, our uniform allocation scheme ensures variance stabilization.

### 4.2.3 Variance stabilization under uniform allocation

We begin with two document-side stability conditions analogous to Condition **(R3)**. Suppose, for $v_{\text{diag}}(a; \pi, m) \geq 0$ and $\tau(a) \geq 0$, that

$$\frac{1}{J} \sum_{j=1}^{J} \frac{a_j^2 \nu_j}{\lambda_j} \longrightarrow v_{\text{diag}}(a; \pi, m), \qquad \frac{1}{J^2} \sum_{j \neq k} a_j a_k \tau_{jk} \longrightarrow \tau(a), \tag{4.10}$$

where the second sum is over ordered document pairs. As with Condition **(R3)**, these limits may hold along a regular deterministic sequence of fixed corpora or follow from ordinary and U-statistic laws of large numbers under a well-behaved optional document superpopulation. Although the finite-$J$ off-diagonal average can be negative, any limit $\tau(a)$ is nonnegative: the covariance matrix $(\tau_{jk})_{j,k\leq J}$ is positive semidefinite, and its omitted diagonal contribution is $O(J^{-1})$ under the maintained bounds.

Conditions (4.10) alone do not ensure variance stabilization because they do not determine which document pairs share annotators. Even with bounded annotator loads, an allocation rule can systematically concentrate overlaps on particular document classes and make the conditional measurement variance depend persistently on the realized design. Appendix C.2 gives such a counterexample.

Uniform sampling over all feasible bipartite graphs provides this much-needed allocation regularity. The proof relies on two averaging arguments. First, capacity filling fixes the overall amount of annotator reuse, while uniform allocation randomizes which document pairs share annotators. On the high-probability event that the capacity-filled allocation is feasible, the fixed-degree graph result of McKay (2010, Theorem 2.2) gives, uniformly for selected $j \neq k$,

$$\mathbb{E}[N_{jk} \mid \mathcal{S}_J] = (K-1)\frac{m_j m_k}{B} + O\Big(\frac{m_j m_k}{B^2}\Big). \tag{4.11}$$

Thus pairs receiving more annotations are proportionally more likely to share an annotator. The proof also shows that the variance contribution from annotator reuse concentrates around the value implied by the expected numbers of shared annotators. Intuitively, with bounded document depths and annotator workloads, the allocation behaves locally like uniform random matching of document slots to annotator slots. As the annotation budget grows, the assignments of any two annotators become approximately independent, while balanced spending ensures that each annotator's contribution becomes small. The sum of these small, approximately independent contributions therefore concentrates, yielding the aforementioned result.

Second, the factors $m_j m_k$ in (4.11) cancel the corresponding annotation-depth adjustments in (4.7). The remaining average over selected document pairs converges under Poisson sampling to the corpus-wide limit $\tau(a)$, while the diagonal component converges to $v_{\text{diag}}(a;\pi,m)$ following standard Horvitz–Thompson concentration arguments.

**Theorem 2** (Variance stabilization under uniform allocation)**.** *Maintain Assumptions **(M1)**–**(M3)**, **(D1)**–**(D3)**, and **(R1)**. Suppose $J \to \infty$, $\bar{B} \to \infty$, the target weights and annotation depths are uniformly bounded, annotator capacity $K$ is fixed, and Assumption **(R2)** holds. Suppose also that* (4.10) *holds. Then*

$$v_{\varepsilon,J}(a) \xrightarrow{p} v_{\text{diag}}(a;\pi,m) + (K-1)\tau(a). \tag{4.12}$$

Combining the theorem with the high-level Gaussian result gives the following operational conclusion.

**Corollary 1** (Operational Gaussian limit)**.** *Suppose Condition **(R3)** and the conditions of*

*Theorem 2 hold. The conclusion of Theorem 1 then holds with*

$$v_\varepsilon(a) = v_{\text{diag}}(a; \pi, m) + (K-1)\tau(a),$$

*provided*

$$v_r(a) + v_{\text{diag}}(a; \pi, m) + (K-1)\tau(a) > 0.$$

It should be noted that uniform allocation is one transparent sufficient mechanism, not a necessary condition. Other allocation schemes can be treated case by case when their worker reuse structure and the corresponding, potentially allocation-dependent, fixed-array limits jointly imply Condition **(R4)**. Under an optional document-superpopulation model, those limits may often be verified using ordinary or U-statistic laws of large numbers, but regularity of the allocation mechanism must still be established case by case.

### 4.3 Variance estimation and feasible inference

Throughout this subsection, we restrict attention to designs with $m_j \geq 2$ for every document (we explain this restriction later in the subsection). If the document-level means $\{\mu_j\}_{j=1}^J$ were known, the prediction residuals $r_j = \mu_j - g_j$ would also be known, and the document-sampling variance $v_{r,J}(a)$ could be calculated directly from (4.4). Conditional on the sampled documents and annotator assignment, the measurement variance $v_{\varepsilon,J}(a)$ could be estimated by centering responses using the known document means and summing squared annotator contributions. This annotator-level "cluster robust" type calculation would account for annotator reuse. The difficulty is that the document means needed for these calculations are unknown and must themselves be estimated from a small number of responses.

Although $\bar{Y}_j$ is unbiased for $\mu_j$, directly plugging it into squared residuals and covariance products can introduce bias. Because the number of labels per document remains bounded, this bias may not disappear as more documents are annotated. To address this problem, we follow the leave-out logic of Kline et al. (2020) and construct unbiased estimators of the moments needed for variance estimation. The remaining mean-zero estimation errors can then average out under the maintained regularity conditions.

#### 4.3.1 Document-sampling variance

For the document-sampling component, the finite-label bias takes a particularly simple form. For a selected document,

$$\mathbb{E}\big[(\bar{Y}_j - g_j)^2 \,\big|\, \mathcal{G}_J\big] = r_j^2 + \frac{\nu_j}{m_j}. \tag{4.13}$$

The additional term is the finite-label variance of the observed document mean. Because $m_j$ remains bounded, this term need not vanish as the annotation budget grows, and its weighted aggregate can generate first-order bias in the estimated document-sampling variance.

This expression suggests a direct correction. For $j \in \mathcal{S}_J$, let

$$\widehat{\sigma}_j^2 := \frac{1}{m_j - 1} \sum_{i \in M_j} \big(Y_{ij}^* - \bar{Y}_j\big)^2$$

denote the usual within-document sample variance, and define

$$\widehat{r}_j^2 := (\bar{Y}_j - g_j)^2 - \frac{\widehat{\sigma}_j^2}{m_j}. \tag{4.14}$$

Since $\widehat{\sigma}_j^2$ is unbiased for $\nu_j$, $\widehat{r}_j^2$ is unbiased for $r_j^2$. Equivalently, $\widehat{r}_j^2$ can be written as the average product of each annotator's label and the mean of the remaining annotators' labels, after both have been centered at $g_j$. Thus, (4.14) is a document-level leave-out correction of the type used to estimate quadratic objects in Kline et al. (2020).

Substituting the corrected residual square into the Horvitz–Thompson variance formula gives

$$\widehat{v}_{r,J}(a) := \frac{\bar{B}}{J^2} \sum_{j \in \mathcal{S}_J} \frac{1 - \pi_j}{\pi_j^2} a_j^2 \widehat{r}_j^2. \tag{4.15}$$

### 4.3.2 Measurement variance

The diagonal term in the measurement variance is already handled by $\widehat{\sigma}_j^2$. It remains to estimate the overlap contribution $N_{jk}\tau_{jk}$ for document pairs that share annotators. For an annotator $i$ shared by documents $j$ and $k$, the corresponding oracle covariance product is

$$\big(Y_{ij}^* - \mu_j\big)\big(Y_{ik}^* - \mu_k\big),$$

whose conditional expectation is $\tau_{jk}$. Because $\mu_j$ and $\mu_k$ are unknown, centering instead by the full document means would use annotator $i$'s own labels in both centering terms. We remove this self-centering by defining

$$\bar{Y}_{j,-i} := \frac{1}{m_j - 1} \sum_{\ell \in M_j \setminus \{i\}} Y_{\ell j}^*.$$

When documents $j$ and $k$ share more than one annotator, however, their leave-out means can still contain the other $N_{jk} - 1$ shared annotators. The following factor removes the resulting additional covariance:

$$\kappa_{jk}^{\mathrm{loo}} := \begin{cases} \dfrac{(m_j - 1)(m_k - 1)}{(m_j - 1)(m_k - 1) + N_{jk} - 1}, & N_{jk} \geq 1, \\ 0, & N_{jk} = 0. \end{cases}$$

For distinct selected documents $j$ and $k$, define

$$\widehat{\Gamma}_{jk} := \kappa_{jk}^{\mathrm{loo}} \sum_{i \in M_j \cap M_k} \big(Y_{ij}^* - \bar{Y}_{j,-i}\big)\big(Y_{ik}^* - \bar{Y}_{k,-i}\big).$$

With this adjustment, as shown in the appendix,

$$\mathbb{E}[\widehat{\Gamma}_{jk} \mid \mathcal{G}_J] = N_{jk}\tau_{jk}.$$

Combining the diagonal and overlap terms gives the measurement-variance estimator

$$\widehat{v}_{\varepsilon,J}(a) := \frac{\bar{B}}{J^2}\left[\sum_{j\in\mathcal{S}_J} \frac{a_j^2\widehat{\sigma}_j^2}{\pi_j^2 m_j} + \sum_{\substack{j,k\in\mathcal{S}_J \\ j\neq k}} \frac{a_j a_k \widehat{\Gamma}_{jk}}{\pi_j\pi_k m_j m_k}\right], \tag{4.16}$$

where the second sum is over ordered document pairs, as in (4.7). The estimator of the total asymptotic variance is

$$\widehat{v}_J(a) := \widehat{v}_{r,J}(a) + \widehat{v}_{\varepsilon,J}(a). \tag{4.17}$$

Under the maintained response, design-exogeneity, Poisson-sampling, and simple-allocation assumptions, the two corrections are unbiased when $m_j \geq 2$:

$$\mathbb{E}[\widehat{v}_{r,J}(a)] = v_{r,J}(a), \qquad \mathbb{E}[\widehat{v}_{\varepsilon,J}(a) \,|\, \mathcal{G}_J] = v_{\varepsilon,J}(a). \tag{4.18}$$

Define

$$\bar{v}_{\varepsilon,J}(a) := \mathbb{E}[v_{\varepsilon,J}(a)].$$

Applying the law of iterated expectations to average over the document sample and annotator assignment therefore gives

$$\mathbb{E}[\widehat{v}_J(a)] = v_{r,J}(a) + \bar{v}_{\varepsilon,J}(a) = \bar{B}\ \mathrm{Var}\Big\{\widehat{\Phi}_J(a) - \Phi_J(a)\Big\}. \tag{4.19}$$

Unbiasedness alone does not imply consistency. The intuition behind concentration is as follows. Conditional on the document sample and annotator assignment, responses from different annotators are independent. Bounded workloads and annotation depths ensure that changing one annotator's responses affects only a limited number of terms in the variance estimators. Balanced spending makes these individual contributions small as the budget grows. The Efron–Stein inequality translates this limited influence into an $O(\bar{B}^{-1})$ mean-square bound for the error due to annotator responses. For the sampling term, independent Poisson sampling controls the remaining error due to document selection in the sampling-variance estimator. Appendix D gives the formal bounds.

Neither finite-sample unbiasedness nor this concentration argument requires uniform allocation; graph simplicity and bounded annotator load are the only allocation-specific requirements. Uniform allocation enters through variance stabilization, which makes $v_{\varepsilon,J}(a)$ converge to the nonrandom limit used for studentization. Consequently, $\widehat{v}_J(a)$ consistently estimates the asymptotic variance in Corollary 1. This retains the crucial modular structure of the results: if researchers desire, or are compelled, to use a different type of allocation mechanism, then, as long as variance stabilization can be established, the variance estimator facilitates feasible

inference.

Because the leave-out corrections involve subtraction, the raw estimator $\widehat{v}_J(a)$ is not constrained to be nonnegative in finite samples. This is a familiar possibility for dependence-robust covariance estimators involving subtractive corrections, such as multiway-clustered covariance estimators (Cameron et al., 2011). For any deterministic sequence $\eta_J > 0$ with $\eta_J \to 0$, define

$$\widehat{v}_J^+(a) := \max\{\widehat{v}_J(a), \eta_J\}.$$

This floor is only a numerical finite-sample safeguard.

**Theorem 3** (Feasible inference)**.** *Suppose the conditions of Corollary 1 hold and $m_j \geq 2$ for every document. Then*

$$\widehat{v}_{r,J}(a) - v_{r,J}(a) \xrightarrow{p} 0, \qquad \widehat{v}_{\varepsilon,J}(a) - v_{\varepsilon,J}(a) \xrightarrow{p} 0,$$

*and hence*

$$\widehat{v}_J(a) \xrightarrow{p} v_r(a) + v_{\text{diag}}(a; \pi, m) + (K-1)\tau(a).$$

*If*

$$v_r(a) + v_{\text{diag}}(a; \pi, m) + (K-1)\tau(a) > 0,$$

*then*

$$\frac{\sqrt{\bar{B}}\,\{\widehat{\Phi}_J(a) - \Phi_J(a)\}}{\sqrt{\widehat{v}_J^+(a)}} \Rightarrow \mathcal{N}(0,1).$$

*The same conclusion holds with $\sqrt{B}$ in place of $\sqrt{\bar{B}}$.*

*Proof.* Proposition 4 gives componentwise consistency under the stated conditions. Corollary 1, variance stabilization, and Slutsky's theorem then give the studentized limit; the replacement of $\sqrt{\bar{B}}$ by $\sqrt{B}$ follows from Lemma 5. □

Relative to Corollary 1, feasible inference requires only that each document receive at least two labels, so that the within-document variance and leave-out means are well defined. Strict positivity is the usual nondegeneracy condition: it holds whenever document sampling or annotation contributes nonvanishing first-order uncertainty in the target direction. Under this condition, the probability that the numerical safeguard binds converges to zero.

## 4.4 Important extensions

### 4.4.1 Pairwise response measurements

Online Appendix OA.1 extends the analysis to the order-two measurements $\vartheta_j(\ell)$ introduced in Subsection 2.2. For a selected document, the natural estimator is the within-document $U$-statistic

$$\widehat{\vartheta}_j = \binom{m_j}{2}^{-1} \sum_{\substack{i,r \in M_j \\ i<r}} \ell_j(Y_{ij}^*, Y_{rj}^*).$$

Substituting $\widehat{\vartheta}_j$ for the document mean in the same prediction-assisted Horvitz–Thompson construction gives an unbiased estimator of the corresponding corpus target. Its document-sampling component is handled exactly as in the mean case.

The new issue is again in the measurement term. The Hoeffding decomposition separates the error in $\widehat{\vartheta}_j$ into a linear component indexed by individual annotators and a canonical component indexed by pairs of annotators. The linear component has the same structure as that of the mean, so the same overlap-averaging argument applies.

The canonical component requires additional care. Because annotation depth remains bounded, its contribution to the within-document measurement variance cannot be discarded. Canonical degeneracy means that each pairwise term has mean zero conditional on either annotator's responses. Together with independence across annotators, this implies that, conditional on the sample and assignment, canonical terms from different documents have zero covariance unless the same pair of annotators evaluates both. Under our allocation scheme, such overlaps are sufficiently rare that their combined contribution to the normalized measurement variance is asymptotically negligible.

One additional complication is that the measurement error can no longer be written as a sum of independent annotator contributions, so the annotator-block central limit theorem used for the mean is unavailable. Nevertheless, balanced spending and bounded document and annotator degrees permit an application of the dependency-graph central limit theorem of Janson (1988, Theorem 2), which provides a conditional Gaussian approximation for the measurement term. With variance stabilization established, the measurement and document-sampling components are then combined using the same conditioning argument as in the mean case.

The appendix consequently establishes Gaussian and feasible inference for bounded symmetric pairwise kernels, including response variance and categorical disagreement. Point estimation and the central limit theorem require $m_j \geq 2$. The feasible variance estimator uses products formed from disjoint annotator pairs and therefore requires $m_j \geq 4$.

### 4.4.2 Stratified exact-budget inference

Online Appendix OA.2 develops the corresponding inference theory under Design **(SFB)**, which fixes the number of sampled documents and their common annotation depth within each stratum, so that total expenditure is fixed by design.

The estimator retains the same decomposition into document-sampling and measurement components. The sampling component is now a stratified finite-population average, so the Poisson-sampling central limit theorem is replaced by a stratumwise finite-population central limit theorem under the usual Hájek no-dominance condition (Li and Ding, 2017, Theorem 1). Independence across the fixed number of strata then combines the stratum-specific limits. Conditional on the samples and allocation graphs, the measurement component remains a sum of independent annotator contributions, so the conditional annotator-block argument used for the mean continues to apply.

The remaining step is variance stabilization. Within each stratum, simple random sampling

makes the sampled document and document-pair moments concentrate around their finite-population counterparts. Common annotation depth makes selected documents exchangeable under uniform allocation, which spreads the capacity-fixed overlap opportunities symmetrically across document pairs. Separate annotator pools eliminate cross-stratum measurement dependence. These properties make the allocation-dependent measurement variance concentrate around its deterministic expectation. The sampling and measurement components can then be combined using the same conditioning argument as under the baseline design.

The appendix also constructs a consistent feasible variance estimator using the same leave-out moments, together with a within-stratum correction for the finite-label noise in observed prediction residuals.

### 4.4.3 Joint and smooth-functional inference

Appendix D.3 establishes joint inference for a fixed-dimensional vector of weighted corpus means. Online Appendix OA.3 extends these results to vectors containing both mean and order-two corpus moments. In essence, a mean coordinate based on a bounded single-response transform $h_j$ can be represented exactly by the symmetric kernel $\{h_j(y_1) + h_j(y_2)\}/2$, allowing mean and genuinely pairwise coordinates, including their mixed covariances, to be treated within one construction. Every fixed linear combination of the moment vector is then a scalar target covered by the preceding results. Applying the scalar result to every fixed linear combination and invoking the Cramér–Wold device, we obtain a joint Gaussian limit. We then estimate the covariance between two coordinates by subtracting their individual variance estimates from the estimated variance of their sum and dividing the result by two. Standard delta-method and implicit-function arguments consequently provide feasible inference.

## 5 Choosing document inclusion probabilities and annotation depths

We now use the preceding variance characterization to choose the document inclusion probabilities $\{\pi_j\}_{j=1}^J$ and annotation depths $\{m_j\}_{j=1}^J$ subject to the expected label budget

$$\sum_{j=1}^{J} \pi_j m_j = \bar{B}.$$

The analysis is restricted to the scalar weighted mean target

$$\Phi_J(a) = \frac{1}{J} \sum_{j=1}^{J} a_j \mu_j$$

under the baseline Poisson document-sampling design. We leave the optimal allocation for pairwise response measurements and stratified exact-budget designs to future research.

All comparisons are restricted to the admissible design class maintained for inference. In particular, design choices are made before the final annotation outcomes are observed, documents are

selected independently, annotator allocation follows the uniform capacity-filled mechanism with fixed capacity $K$, and annotation depths are uniformly bounded. We also maintain balanced spending,

$$c_\lambda \le \frac{J\pi_j m_j}{\bar{B}}, \qquad j = 1, \dots, J,$$

together with $\bar{B} \to \infty$ and the corpus-array convergence and nondegeneracy conditions required by the preceding inference results. For feasible mean inference, we impose $2 \le m_{\min} \le m_j \le \bar{m}$. Documents with zero target weight can be removed from the design frame, with the remaining weights rescaled so that the target is unchanged.

### 5.1 Oracle benchmark

To expose the allocation trade-offs, suppose initially that the prediction residual

$$r_j := \mu_j - g_j$$

and the fresh-annotator response variance $\nu_j$ are known. For designs satisfying the conditions stated above, our variance-stabilization result gives

$$v_{r,J}(a) + v_{\varepsilon,J}(a) = \frac{\bar{B}}{J^2} \sum_{j=1}^{J} a_j^2 \left\{ \frac{1-\pi_j}{\pi_j} r_j^2 + \frac{\nu_j}{\pi_j m_j} \right\} + v_{\mathrm{ov},J}(a) + o_p(1), \tag{5.1}$$

$$v_{\mathrm{ov},J}(a) := \frac{K-1}{J^2} \sum_{j \ne k} a_j a_k \tau_{jk}.$$

The variance-stabilization result shows that, within our design class and for fixed $K$, the annotator-overlap contribution to the asymptotic variance does not depend on the document inclusion probabilities or annotation depths. Minimizing the asymptotic variance is therefore equivalent to minimizing the remaining terms in (5.1), with the annotator-overlap contribution omitted.

We approximate the allocation problem by solving its continuous relaxation, treating annotation depths $m_j$ as continuous variables between $m_{\min}$ and $\bar{m}$. Let

$$s_j := \pi_j m_j$$

denote the expected expenditure on document $j$. At a fixed value of $s_j$, that document's contribution to the allocation criterion is

$$a_j^2 \left\{ \frac{\nu_j}{s_j} + r_j^2 \left( \frac{1}{\pi_j} - 1 \right) \right\}. \tag{5.2}$$

The measurement component depends only on total expected expenditure, whereas the document-sampling component decreases as $\pi_j$ increases. For a fixed expected expenditure $s_j$, increasing the inclusion probability while reducing annotation depth strictly reduces the document's contribution to the variance whenever $r_j \ne 0$. An optimal implementation of $s_j$ therefore expands

coverage before increasing depth:

$$\pi_j^\star(s_j) = \min\left\{1, \frac{s_j}{m_{\min}}\right\}, \qquad m_j^\star(s_j) = \frac{s_j}{\pi_j^\star(s_j)}. \tag{5.3}$$

Thus depth remains at $m_{\min}$ until document $j$ is included with probability one; only then does additional expenditure increase its depth. It should be noted, at this point, this is just a within-document statement. It does not require every document to reach full coverage before depth increases for any document.

Using the coverage-first rule, we profile out $\pi_j$ and $m_j$, leaving an optimization problem in expected expenditures $s_j$ alone. Substituting (5.3) into (5.2) gives the contribution $a_j^2\phi_j(s_j)$, where

$$\phi_j(s) = \begin{cases} \dfrac{\nu_j + m_{\min} r_j^2}{s} - r_j^2, & s \le m_{\min}, \\ \dfrac{\nu_j}{s}, & s > m_{\min}. \end{cases} \tag{5.4}$$

The remaining choice is how to divide expected expenditure across documents. The allocation problem therefore becomes

$$\begin{aligned} \min_{\{s_j\}} \quad & \sum_{j=1}^{J} a_j^2 \phi_j(s_j) \\ \text{subject to} \quad & \sum_{j=1}^{J} s_j = \bar{B},\ c_\lambda \frac{\bar{B}}{J} \le s_j \le \bar{m}, \qquad j = 1, \ldots, J. \end{aligned} \tag{5.5}$$

The spending bounds preserve the balanced-spending restriction and the maximum annotation depth. The objective is convex, including where its two branches meet at $m_{\min}$. Its Karush–Kuhn–Tucker (KKT) conditions characterize the optimal allocation and yield the rule in the following proposition.

**Proposition 2** (Structure of the first-order oracle allocation)**.** *Suppose* $0 < \bar{B} \le J\bar{m}$ *and* $0 < c_\lambda \le 1$. *An optimal solution to* (5.5) *can be implemented using the coverage-first rule in* (5.3). *There is a common multiplier* $\xi \ge 0$ *on the budget constraint. When* $\xi > 0$, *documents whose optimal expenditure is neither at a spending bound nor at the full-coverage threshold* $m_{\min}$ *satisfy*

$$s_j^\star = \begin{cases} |a_j| \sqrt{\dfrac{\nu_j + m_{\min} r_j^2}{\xi}}, & s_j^\star < m_{\min}, \\ |a_j| \sqrt{\dfrac{\nu_j}{\xi}}, & s_j^\star > m_{\min}. \end{cases}$$

*The multiplier and any expenditures at the threshold or spending bounds are jointly chosen so that* $\sum_j s_j^\star = \bar{B}$.

Appendix F gives the complete constrained characterization, including the profiled convex program, documents placed exactly at full coverage, and binding spend bounds.

The proposition gives a generalized Neyman allocation with two economically distinct regimes.

Before full coverage, expected expenditure is proportional to

$$|a_j|\sqrt{\nu_j + m_{\min} r_j^2},$$

because additional expenditure reduces uncertainty from both document selection and finite annotation. Documents receive more resources when they matter more for the target, when the machine prediction is less accurate, or when annotator responses are more variable. After full coverage, sampling uncertainty has disappeared, so expenditure is proportional to

$$|a_j|\sqrt{\nu_j},$$

and only annotation uncertainty determines the return to additional depth.

### 5.2 Practical implementation

The oracle quantities $r_j^2$ and $\nu_j$ are unobserved before annotation. A pilot with repeated annotations, or other comparable annotated data, can be used to estimate these quantities for pilot documents and fit prediction rules based on document characteristics observed throughout the corpus. This parallels two-stage experimental designs that use first-wave outcomes to choose second-wave assignment probabilities or strata (Hahn et al., 2011; Tabord-Meehan, 2023). Applying the fitted rules to the full corpus provides predicted difficulty measures for the allocation problem. Negative fitted variances or squared residuals, if any, are truncated at zero. The researcher then solves the oracle program using these predictions, subject to the same budget, depth, and balanced-spending restrictions imposed above.

Without reliable outcome information for ranking document difficulty, a simple default is to assign every document the same expected response variance and the same expected squared prediction error. Let $\mathcal{I}_J$ denote the information available when the design is chosen, and let $\mathbb{E}_0$ denote the researcher's planning expectation. Our planning convention is

$$\mathbb{E}_0[\nu_j \mid \mathcal{I}_J] = \nu_J^0, \qquad \mathbb{E}_0[r_j^2 \mid \mathcal{I}_J] = q_J^0, \qquad j = 1, \ldots, J,$$

where $\nu_J^0$ and $q_J^0$ are finite, nonnegative values that are not both zero.[3] It should be stressed that these planning expectations do not assert that the actual document-specific variances or squared prediction errors are equal. They are used only to choose the allocation and are not assumptions required for inference.

Under this convention, suppose that the target weights are not too uneven and that the labeling budget is small relative to the corpus, specifically,

$$c_\lambda \leq \frac{J|a_j|}{\sum_{k=1}^J |a_k|} \quad \text{for every } j, \qquad \frac{\bar{B}}{J m_{\min}} \frac{J \max_j |a_j|}{\sum_{k=1}^J |a_k|} \leq 1. \tag{5.6}$$

Under this condition, the allocation never assigns more than the minimum admissible annotation

[3] Appendix F.1 establishes the allocation result under the weaker condition that $\mathbb{E}_0[\nu_j + m_{\min} r_j^2 \mid \mathcal{I}_J]$ has the same positive finite value across documents. We state the separate conditions for ease of interpretation.

depth to any document. The resulting rule is

$$s_j^0 = \bar{B}\frac{|a_j|}{\sum_{k=1}^{J}|a_k|}, \qquad m_j^0 = m_{\min}, \qquad \pi_j^0 = \frac{s_j^0}{m_{\min}}. \tag{5.7}$$

Thus, every sampled document receives $m_{\min}$ annotations, while documents with larger absolute target weights are sampled with higher probability. Condition (5.6) is fully determined by the target weights, the budget, and the corpus size, and can therefore be checked before annotation begins. Appendix F.1 gives the complete result for this common-expected-difficulty default.

Two especially interpretable examples are the following. For the corpus mean, $a_j = 1$, so the rule spends equally across documents. For a difference between two group means, the absolute target weights sum to the same amount within each group. Each group therefore receives one half of the expected budget, with greater expenditure per document in the smaller group. In both cases, expenditure on a document is first used to increase its inclusion probability while keeping $m_j = m_{\min}$. Annotation depth increases only after that document's inclusion probability reaches one.

Finally, we emphasize again that inaccurate or misspecified document difficulty models may reduce efficiency by leading to suboptimal inclusion probabilities and annotation depths. They do not invalidate inference when the resulting design remains within the maintained admissible class.

# 6 Calibrated Simulation: Measuring Toxicity

This section uses an empirically anchored calibrated simulation to examine three features of the proposed framework: inference under annotator reuse, finite-sample performance with few sampling units and annotators, and the efficiency gains from targeted annotation design. We use the Wikipedia Talk Labels: Toxicity corpus, which contains 159,686 comments and 1,598,289 binary toxicity judgments from 4,301 annotators (Wulczyn et al., 2017; Thain et al., 2017). The corpus links annotator-identified judgments to comment text and metadata recording contributor registration status, posting year, and talk-page namespace. These data are used to construct comment-level toxicity probabilities and calibrate the synthetic annotator population. For each comment, we also attach a fixed machine prediction $g_j$, defined as the offensive-class probability from the pretrained `cardiffnlp/twitter-roberta-base-offensive` model (Barbieri et al., 2020). Importantly, the Cardiff prediction is used directly off-the-shelf and not re-trained or cross-fitted within the simulation, satisfying our assumption on the frozen predictor.

## 6.1 Calibrated simulation population

We construct an empirically anchored, semi-synthetic population from the Wikipedia Detox corpus. For each comment, we retain its estimated toxicity probability $\mu_j$, registration status, and fixed Cardiff prediction.

Annotator responses follow a random-coefficients logit model:

$$Y_{ij} \mid u_i, \beta_i \sim \text{Bernoulli}(p_{ij}), \tag{6.1}$$

$$\text{logit}(p_{ij}) = \alpha_j + u_i + x_j^\top \beta_i. \tag{6.2}$$

The vector $x_j$ contains five comment characteristics. Three are indicators supplied with the dataset: whether the comment appears on a user-talk page, belongs to the blocked-user sample, or was posted by an unregistered contributor. We add two measures constructed from the comment text: Length is measured by $-\log(1 + \text{word count})$, so larger values indicate shorter comments. Abusive-word presence is an indicator for whether the comment contains at least one word classified as abusive in the published dictionary of Wiegand et al. (2018).

We specify the annotator intercept and slopes as

$$u_i = \sigma Z_i + \delta(A_i - q), \qquad \beta_{ik} = cZ_i + \tau V_{ik}, \quad k = 1, \ldots, 5,$$

where $Z_i, V_{i1}, \ldots, V_{i5}$ are independent standard normal draws and $A_i$ is an independent Bernoulli draw with probability $q$.

The annotator intercept $u_i$ captures differences in the overall tendency to label comments toxic, while the five slopes in $\beta_i$ allow annotators to respond differently to comment characteristics. Each annotator's intercept and slopes are drawn jointly once and retained for every comment assigned to that annotator.

To capture the skewness and upper tail of annotators' average differences from other reviewers, we allow a share $q$ of annotators to belong to a class whose average intercept is higher by $\delta$. Within each class, $\sigma$ controls variation in annotators' overall tendency to label comments toxic.

The random slopes capture differences in responses to particular types of comments that cannot be represented by an annotator intercept alone. Through the shared draw $Z_i$, the parameter $c$ allows annotators' responses to comment characteristics to be associated with their overall tendency to label comments toxic. The separate draws $V_{ik}$, scaled by $\tau$, allow further variation in each slope.

For each candidate pair $(q, \delta)$, we fit $(\sigma, c, \tau)$ to covariance comparisons across different groups of comments reviewed by the same annotator. We compare the resulting specifications using the standard deviation, skewness and excess kurtosis of annotators' average differences from other reviewers. Throughout calibration, the comment intercepts $\alpha_j$ are recomputed so that each comment's response probability, averaged over annotators, remains $\mu_j$. Appendix OA.5 reports the parameter values, the calibration procedure and the comparison of observed and simulated statistics.

## 6.2 Worker reuse and efficiency gains from annotation design

We consider two corpus-level parameters. The first is the finite-corpus mean toxicity probability,

$$\theta_M = \frac{1}{J}\sum_{j=1}^{J}\mu_j.$$

Corpus-wide averages of this kind parallel text-based measures such as the economic policy uncertainty index of Baker et al. (2016), which aggregates newspaper content into a period-level measure.

The second parameter is the difference in mean toxicity between comments posted by unregistered and registered contributors,

$$\theta_\Delta = \frac{1}{J_U}\sum_{j:G_j=1}\mu_j - \frac{1}{J_R}\sum_{j:G_j=0}\mu_j,$$

where $G_j = 1$ denotes an unregistered contributor and $J_U$ and $J_R$ are the corresponding group sizes. An analogous text-based group contrast appears in Wu (2020), who compares mean counts of professional and personal terms in EJMR posts about women with those in posts about men. In the present application, the registration status contrast measures whether toxic content is disproportionately concentrated among unregistered contributions, which may inform the allocation of moderation resources.

In the coverage comparison, each sampled comment receives $m_j = 2$ labels from distinct annotators. Comments are sampled uniformly for the corpus mean and in proportion to the absolute target weights for the registration status contrast, thereby dividing the expected annotation budget equally between the two groups as prescribed in Subsection 5.2. We consider expected budgets of $\bar{B} = 7{,}500$ and 15,000, holding annotator capacity fixed at $K = 75$. These designs correspond to approximately 100 and 200 annotators, respectively.

Table 6.1 compares the proposed method with a naive implementation. Both use the same Cardiff-assisted point estimate. The proposed method accounts for document sampling, finite annotation depth, and covariance between comments labeled by the same annotator. The naive implementation first averages the annotations for each comment and then treats the resulting comment-level outcome as fixed and error-free.

Coverage from the proposed method is close to the nominal level in both designs, whereas the naive implementation substantially undercovers, particularly for the corpus mean. For the mean, same-annotator covariances accumulate because all target weights have the same sign. The opposing weights in the registration status contrast partially offset these contributions. This explains the disparity between the undercoverage observed in the two cases. Holding annotator capacity fixed at $K = 75$, doubling both the expected annotation budget and the approximate workforce does not resolve the undercoverage of the naive implementation.

Figure 6.1 reinforces the point made earlier. With the expected annotation budget fixed at $\bar{B} = 15{,}000$, the figure shows that concentrating annotations among fewer annotators sharply

Table 6.1: Coverage under annotator overlap

| | | Coverage (%) | | |
|---|---|---|---|---|
| $(\bar{B}, K)$ | Target | Proposed method | Naive implementation | Proposed/Naive variance |
| (7,500, 75) | Corpus mean | 94.3 | 66.0 | 4.43 |
| (7,500, 75) | Registration status contrast | 94.5 | 87.4 | 1.66 |
| (15,000, 75) | Corpus mean | 94.9 | 63.8 | 4.53 |
| (15,000, 75) | Registration status contrast | 94.7 | 87.2 | 1.71 |

*Notes:* Results are based on 5,000 replications and nominal 95% intervals. Each sampled comment receives two labels from distinct annotators, and $K$ is the maximum number of comments assigned to one annotator. The final column is the ratio of the average proposed variance estimate to the average naive variance estimate. Both implementations use the same point estimate.

Table 6.2: Efficiency of alternative annotation designs

| | Variance ($\times 10^5$) | | Relative |
|---|---|---|---|
| Design | Empirical | Estimated | efficiency |
| Uniform | 8.170 | 7.879 | 1.000 |
| Target-weighted | 5.487 | 5.410 | 1.489 |
| Full oracle | 4.781 | 4.709 | 1.709 |
| Pilot-estimated | 5.164 | 5.113 | 1.582 |

*Notes:* Results are based on 5,000 replications, with $\bar{B} = 15{,}000$ and $K = 75$. Estimated variance is the average proposed variance estimate. Relative efficiency is the empirical variance under uniform sampling divided by that under the indicated design. The pilot-estimated allocation uses 1,000 historical comments with two observed labels each. These comments are excluded from the target corpus and all four designs, and their 2,000 labels are additional to the stated budget. The pilot and resulting allocation are held fixed across replications.

reduces coverage under the naive implementation, while coverage under the proposed method remains substantially more stable.

We next examine the efficiency gains from the annotation designs developed in Section 5. Table 6.2 compares four allocation designs for the registration status contrast, holding the expected main-stage budget at $\bar{B} = 15{,}000$ and annotator capacity at $K = 75$. All four designs use the same predictor-assisted estimator.

The full oracle reduces empirical variance by 41.5% relative to uniform sampling. Although it optimizes both sampling probabilities and annotation depths, it selects $m_j = m_{\min} = 2$ for every comment. The entire improvement therefore comes from reallocating sampling probabilities across comments, illustrating the coverage-before-depth rule in (5.3).

The feasible allocations introduced in Subsection 5.2 recover much of this gain. The target-weighted rule requires no pilot annotations and reduces variance by 32.8%, capturing approximately four-fifths of the oracle gain. For the pilot-estimated allocation, we use ordinary least squares to regress $(Y_{1j} - Y_{2j})^2/2$ and $(Y_{1j} - g_j)(Y_{2j} - g_j)$ separately on an intercept, the five comment characteristics in $x_j$, and $g_j$, where $Y_{1j}$ and $Y_{2j}$ are the two pilot ratings. The fitted values are designed to estimate $\nu_j$ and $r_j^2$. We set negative fitted values to zero and use them in the allocation rule. This allocation reduces variance by 36.8%, capturing 88.7% of the oracle gain. Across all four designs, average estimated variance closely tracks empirical variance, again

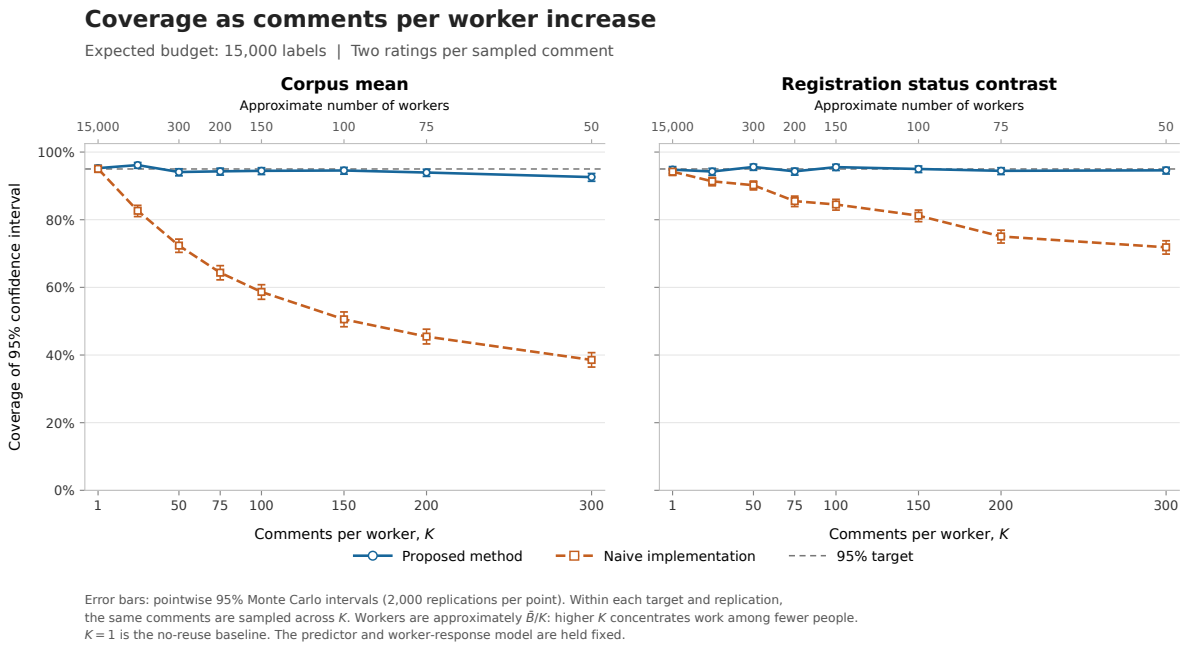



Figure 6.1: Coverage as worker reuse increases at a fixed annotation budget

showing the robustness of our inference results to allocation schemes.

## 6.3 Finite-sample performance and prediction bias

Motivated by the seminar-level counts of negative interactions in Dupas et al. (2026) and similar applications, we next examine performance when both the number of sampling units and the number of annotators are small. We construct an experiment in which negative comments can be concentrated in some units, reflecting the possibility that a seminar contains several negative exchanges while another contains few. This creates variation in the composition of the annotation tasks assigned to workers.

We aggregate the 159,686 Wikipedia Detox comments into a fixed population of 1,004 packages of 159–160 comments, each containing contributions from either registered or unregistered authors. Within each registration group, we first group comments according to whether they contain abusive vocabulary, then randomly redistribute approximately 20% of the comments from each package among packages in the same registration group.

The package is both the sampling and annotation unit, and an annotator's outcome is the number of its comments they label toxic. Each sampled package is rated by two distinct annotators, and each annotator rates at most $K = 6$ packages. We consider expected package-rating budgets of $\bar{B} = 180$ and 360 per target, corresponding to 90 and 180 sampled packages in expectation and approximately 30 and 60 annotators. The package mean count target uses uniform Poisson sampling, whereas the unregistered-minus-registered count contrast uses Poisson sampling proportional to inverse group size. The predicted count $g_P$ for package $P$ is the sum of the original Cardiff probabilities for its constituent comments.

Table 6.3 reports results from 5,000 Monte Carlo replications. With approximately 30 anno-

Table 6.3: Finite-sample stress test with package-level count outcomes

| $\bar{B}$ | Approx. annotators | Target | Coverage (%) | |
|---|---|---|---|---|
| | | | Proposed method | Naive implementation |
| 180 | 30 | Package mean | 93.8 | 72.2 |
| 180 | 30 | Registration status contrast | 93.3 | 86.8 |
| 360 | 60 | Package mean | 94.6 | 71.0 |
| 360 | 60 | Registration status contrast | 94.3 | 85.0 |

*Notes:* 5,000 Monte Carlo replications and nominal 95% confidence intervals. Both variance calculations use the same point estimate. The naive implementation treats the average of the two package ratings as noise-free.

Table 6.4: Prediction bias and efficiency with count outcomes

| $\lambda$ | Machine-only bias | Proposed bias (MC standard error) | Proposed coverage (%) | Variance reduction (%) |
|---|---|---|---|---|
| 0.00 | -5.276 | 0.026 (0.031) | 94.3 | 38.1 |
| 0.50 | -16.804 | 0.018 (0.033) | 93.8 | 27.8 |
| 0.75 | -22.569 | 0.015 (0.037) | 94.3 | 9.2 |

*Notes:* 5,000 Monte Carlo replications. Bias is measured in counts; parentheses report Monte Carlo standard errors. Coverage refers to nominal 95% confidence intervals. Variance reduction is relative to the difference of human sample group means using the same annotations.

tators, the proposed method achieves coverage of 93.8% for the package mean and 93.3% for the registration status contrast. Coverage rises to 94.6% and 94.3%, respectively, with approximately 60 annotators. Coverage moves closer to nominal with the larger workforce.

Treating the average package label as noise-free produces much greater undercoverage: coverage is approximately 71–72% for the package mean and 85–87% for the registration status contrast. Doubling the annotation budget does not resolve this problem.

We now illustrate the risk of using machine predictions alone. A black-box model trained in another setting may be informative without reproducing the average judgments of the reference population. For packages from unregistered authors, we replace $g_P$ by $g_P(1-\lambda U_P)$, where $U_P$ is drawn independently from Uniform$(0,1)$ for each package and held fixed across noise levels and Monte Carlo replications. Each predicted count is therefore reduced by a random proportion between zero and $\lambda$. We consider $\lambda \in \{0, 0.5, 0.75\}$, leaving predictions for registered authors unaltered, the underlying population and the annotation design unchanged. The case $\lambda = 0$ uses the same predictions as the preceding stress test.

Table 6.4 reports results for the registration status contrast at $\bar{B} = 360$. The true contrast is 29.20 comments. As $\lambda$ increases, the machine-only estimate understates this contrast by 5.28, 16.80 and 22.57 comments. The proposed estimator's empirical bias remains close to zero, with coverage between 93.8% and 94.3% for nominal 95% confidence intervals. Its variance reduction relative to the difference of human sample group means declines from 38.1% to 27.8% and 9.2%. Thus, efficiency gains remain possible despite substantial machine-only bias, although they diminish as predictions deteriorate.

# 7 Empirical Application: Taking the Fed at its Word

In this section, we apply our method to the setting studied by Shapiro and Wilson (2022), who relate the sentiment of Federal Open Market Committee (FOMC) deliberations to economic conditions to investigate central-bank objectives. The corpus contains 1,676 speaker-meeting documents from 42 speakers at 97 meetings during 2000–2011. The authors measure sentiment using a dictionary-based net-negativity score: the share of negative words minus the share of positive words, classified using the Loughran–McDonald dictionaries. Their quadratic specification with the full control set can be written as

$$N_{st} = \alpha_s + \theta\pi_t + \delta\pi_t^2 + z_t'\gamma + \varepsilon_{st},$$

where $N_{st}$ is the net negativity of speaker $s$ at meeting $t$, $\alpha_s$ denotes a speaker fixed effect, and $\pi_t$ is the average of current-quarter and one-quarter-ahead Greenbook forecasts of core PCE inflation. The vector $z_t$ collects the unemployment gap and real GDP growth, their squares, and financial controls: the preceding three-month S&P 500 return, the previous month's Shiller price–earnings ratio, VIX and short-term Treasury-yield volatility. Under the authors' interpretation of negativity as a proxy for the central bank's loss, positive curvature implies a loss-minimizing inflation rate $\pi^* = -\theta/(2\delta)$.

A limitation of dictionary-based sentiment measures is that word counts may miss the context in which language is used. For example, a discussion of possible risks need not express the speaker's own negative assessment. Human readers can interpret these distinctions when assessing sentiment. We define the reference population as English-fluent adult Prolific participants who satisfy at least one of three conditions: (i) their highest degree is in accounting, economics or finance; (ii) their primary employment is in finance; or (iii) they have previously invested in the stock market. Index each speaker–meeting document by a single index $j$, and let $Y_{ij}^* \in \{1, 2, 3, 4, 5\}$ denote the rating that reader $i$ would assign to document $j$, ranging from very positive (1) to very negative (5), with neutral sentiment coded as 3. We replace the dictionary score with the reference-population mean of the human response, using the linear-regression target described in Section 2.3.

The study was originally planned as a single collection with the larger combined budget. Because the full intended funding was unavailable at the outset, data collection instead proceeded in two waves with separate annotation budgets. The expected budgets were $\bar{B}_1 = 800$ and $\bar{B}_2 = 400$ individual document ratings, with realised budgets of $B_1 = 842$ and $B_2 = 384$, respectively. Each sampled document received two ratings, and each reader was assigned at most eight documents. Within each wave, documents were assigned to readers using the uniform allocation scheme in Section 3. Further details are provided in Online Appendix OA.6.

## 7.1 Results

In our preferred specification, we collapse ratings 1 and 2 to 1, retain rating 3, and collapse ratings 4 and 5 to 5. Annotator feedback indicated difficulty distinguishing degrees of positive or negative sentiment, suggesting that differences between ratings 1 and 2, or between ratings

4 and 5, may partly reflect response noise. Results using the original five-point scale, together with alternative nonlinear specifications, are retained as robustness checks in the appendix. Let $C$ denote this recoding. For each document $j$, we define

$$\mu_j = \mathbb{E}\big[C(Y_{ij}^*)\big],$$

where the expectation is taken over readers drawn from the reference population. Thus, $\mu_j$ is the expected recoded rating of document $j$.

We use two auxiliary predictors. The first is constructed using the GPT-6 Astra model, which is prompted to predict how eligible readers would rate each document. The model returns a probability for each of the five possible responses. We form its prediction by averaging the response values using these probabilities, applying the same recoding to the response values when using the recoded outcome. The second predictor is based on the original authors' dictionary score. To put this score on the 1–5 rating scale, we order all $J = 1,676$ documents from the lowest to the highest dictionary score and assign each document its position in this ordering. Documents with the same score receive the average of the positions they occupy. If $r_j$ denotes this position, the transformed score is

$$g_j^{\text{dict}} = 1 + 4\frac{r_j - 0.5}{J}.$$

Thus, documents with higher dictionary negativity receive higher predicted ratings, with the lowest and highest scores mapped close to 1 and 5. Fractional values are retained in both linear specifications.

Table 7.1: Inflation estimates using human-assessed sentiment

| Auxiliary predictor | Inflation $\theta$ | Inflation squared $\delta$ | Assisted turning point | 90% delta-method interval | Predictor-only turning point |
|---|---|---|---|---|---|
| GPT-6 Astra | −1.313 | 0.391 | 1.678 | [1.322, 2.033] | 1.326 |
| | (0.610) | (0.172) | | | |
| Authors' score | −1.261 | 0.353 | 1.788 | [1.259, 2.317] | 1.727 |
| | (0.705) | (0.193) | | | |

*Notes:* The outcome is the recoded 1–3–5 rating. Coefficient estimates are prediction-assisted, with standard errors in parentheses. Both specifications include the economic controls and speaker fixed effects described above. Standard errors account for document sampling, annotation variability and reader overlap. Turning points are measured in inflation percentage units. Predictor-only turning points use the same regression specification on full-corpus predictions, without human correction. The authors' score is rescaled using full-corpus midpoint ranks. Intervals are nominal pointwise 90% delta-method confidence intervals.

Table 7.1 reports the prediction-assisted estimates. The two estimators imply minimum-negativity inflation rates of 1.678% and 1.788%, both close to the 1.691% estimate in the full-controls quadratic specification of Shapiro and Wilson (2022, Table 1, column (4)). Both yield negative linear and positive quadratic inflation coefficients, individually significant at the nominal 10% level. Both turning-point confidence intervals include 2%. Although our point estimates are close to those reported by the authors, our evidence is less conclusive about whether the inflation target lies below 2%. This may partly reflect the limited size of our human-annotation

sample.

The GPT-6 Astra predictor alone implies a turning point of 1.326%. Incorporating human ratings raises the estimate by 0.352 percentage points to 1.678%, although the predictor-only estimate lies within the 90% confidence interval for the prediction-assisted turning point.

Relative to assistance using the authors' transformed dictionary score, GPT-6 Astra reduces the estimated standard errors of the linear and quadratic inflation coefficients by 13.4% and 11.0%, respectively. This pattern is consistent with a precision benefit from a predictor that accounts for linguistic context. The relatively modest gains in coefficient precision also suggest that the dictionary score remains a competitive auxiliary predictor in this application.

The overlap correction is small for the two inflation coefficients in the GPT-6 Astra-assisted specification, accounting for 1.5% and 2.4% of their total estimated variances, respectively. However, this is not the case for every coefficient. Two notable examples are the unemployment gap and the Shiller price–earnings ratio, for which overlap accounts for 22.4% and 35.2% of total estimated variance.

Table 7.2: Contribution of reader overlap to estimated uncertainty

| Coefficient | SE without overlap | SE with overlap | Overlap share of total variance | SE increase |
|---|---|---|---|---|
| Inflation | 0.6057 | 0.6103 | 1.5% | 0.8% |
| Inflation squared | 0.1700 | 0.1720 | 2.4% | 1.2% |
| Unemployment gap | 0.1575 | 0.1788 | 22.4% | 13.5% |
| Shiller P/E | 0.0177 | 0.0219 | 35.2% | 24.2% |

*Notes:* GPT-6 Astra auxiliary predictor and the recoded 1–3–5 outcome, with the same controls and speaker effects as Table 7.1. Let $\widehat{V}$ denote total estimated variance and $\widehat{O}$ its overlap component. The overlap share is $100\widehat{O}/\widehat{V}$. Standard errors with and without overlap are $\sqrt{\widehat{V}}$ and $\sqrt{\widehat{V}-\widehat{O}}$, respectively. The SE increase is $100\{\sqrt{\widehat{V}/(\widehat{V}-\widehat{O})}-1\}$.

# 8 Conclusion

This paper develops a statistical framework for annotation studies in which the features being measured are inherently subjective. Researchers can define meaningful measurements by specifying a reference population whose judgments are relevant to the economic question and summarizing its responses to each document. We develop inference procedures for economic parameters constructed from these measurements across a fixed corpus, using estimators that combine machine predictions with limited human annotations. These procedures account for document sampling, uncertainty in the estimated document-level measurements, and the dependence created when annotators evaluate multiple documents.

The resulting asymptotic variance also provides a basis for allocating a limited annotation budget. For weighted averages of document-level means, we derive the optimal allocation within our proposed class of designs. This allocation prioritizes increasing a document's sampling probability before increasing its number of annotations above the required minimum. It directs more resources toward documents with larger absolute weights in the target parameter,

more variable annotator responses, or less accurate machine predictions. Altogether, the paper provides a framework for subjective annotation studies: it allows researchers to specify whose judgments their measurements represent, account for the uncertainty involved in estimating those measurements, and allocate annotation resources according to the economic parameter they seek to estimate.

## Declaration of AI use

I used Claude and ChatGPT to assist with language editing, LaTeX formatting, programming, and checking mathematical arguments. I take full responsibility for the content of this paper.

## References


Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020): "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 88, 265–296.

Agirre, E., D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo (2013): "*SEM 2013 shared task: Semantic Textual Similarity," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, Atlanta, Georgia, USA: Association for Computational Linguistics, 32–43.

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022): "Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples," *The Review of Economic Studies*, 89, 2958–2991.

Andrews, D. W. K. (2005): "Cross-Section Regression with Common Shocks," *Econometrica*, 73, 1551–1585.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-Powered Inference," *Science*, 382, 669–674.

Ash, E. and S. Hansen (2023): "Text Algorithms in Economics," *Annual Review of Economics*, 15, 659–688.

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131, 1593–1636.

Barbieri, F., J. Camacho-Collados, L. Neves, and L. Espinosa Anke (2020): "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, 1644–1650.

Billingsley, P. (1995): *Probability and Measure*, New York: John Wiley & Sons, 3 ed.

Blanchet, J. and A. Stauffer (2013): "Characterizing Optimal Sampling of Binary Contingency Tables via the Configuration Model," *Random Structures & Algorithms*, 42, 159–184.

Boistard, H., H. P. Lopuhaä, and A. Ruiz-Gazen (2017): "Functional Central Limit Theorems for Single-Stage Sampling Designs," *The Annals of Statistics*, 45, 1728–1758.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference with Multiway Clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Carlson, J. and M. Dell (2025): "A Unifying Framework for Robust and Efficient Inference with Unstructured Data," arXiv preprint arXiv:2505.00282.

Carstens, C. J. and P. Kleer (2018): "Speeding up Switch Markov Chains for Sampling Bipartite Graphs with Given Degree Sequence," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, vol. 116 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 36:1–36:18.

Cassel, C.-M., C.-E. Särndal, and J. H. Wretman (1976): "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615–620.

Christensen, T. and S. Hansen (2026): "Performing Valid Inference with AI/ML-Generated Covariates: A Guide for Empirical Practice," *AEA Papers and Proceedings*, 116, 92–97.

Derczynski, L., K. Bontcheva, and I. Roberts (2016): "Broad Twitter Corpus: A Diverse Named Entity Recognition Resource," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, 1169–1179.

Deville, J.-C. and C.-E. Särndal (1992): "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Duarte, J., S. Siegel, and L. Young (2012): "Trust and Credit: The Role of Appearance in Peer-to-Peer Lending," *The Review of Financial Studies*, 25, 2455–2484.

Dupas, P., A. Handlan, A. S. Modestino, M. Niederle, M. Seré, H. Sheng, J. Wolfers, and Seminar Dynamics Collective (2026): "Gender Differences in Economics Seminars," *American Economic Review*, 116, 749–789.

Durrett, R. (2019): *Probability: Theory and Examples*, Cambridge: Cambridge University Press, 5 ed.

Egami, N., M. Hinck, B. M. Stewart, and H. Wei (2024): "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses," Working paper, November 17.

Erdős, P. L., T. R. Mezei, I. Miklós, and D. Soltész (2018): "Efficiently Sampling the Realizations of Bounded, Irregular Degree Sequences of Bipartite and Directed Graphs," *PLOS ONE*, 13, e0201995.

Fisch, A., J. Maynez, R. A. Hofer, B. Dhingra, A. Globerson, and W. W. Cohen (2024): "Stratified Prediction-Powered Inference for Effective Hybrid Evaluation of Language Models," in *Advances in Neural Information Processing Systems*, vol. 37.

Francisco, C. A. and W. A. Fuller (1991): "Quantile Estimation with a Complex Survey Design," *The Annals of Statistics*, 19, 454–469.

Gao, W., S. Han, and A. Liang (2026): "How Well Do LLMs Predict Human Behavior? A Measure of Their Pretrained Knowledge," .

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574.

Gentzkow, M. and J. M. Shapiro (2010): "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica*, 78, 35–71.

Gordon, M. L., M. S. Lam, J. S. Park, K. Patel, J. T. Hancock, T. Hashimoto, and M. S. Bernstein (2022): "Jury Learning: Integrating Dissenting Voices into Machine Learning Models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 1–19.

Gorodnichenko, Y., T. Pham, and O. Talavera (2023): "The Voice of Monetary Policy," *American Economic Review*, 113, 548–584.

Graham, B. S. (2011): "Efficiency Bounds for Missing Data Models With Semiparametric Restrictions," *Econometrica*, 79, 437–452.

Hahn, J., K. Hirano, and D. Karlan (2011): "Adaptive Experimental Design Using the Propensity Score," *Journal of Business & Economic Statistics*, 29, 96–108.

Hall, P. and C. C. Heyde (1980): *Martingale Limit Theory and Its Application*, New York: Academic Press.

Han, S. (2025): "Mining Causality: AI-Assisted Search for Instrumental Variables," .

Hansen, B. E. and S. Lee (2019): "Asymptotic Theory for Clustered Samples," *Journal of Econometrics*, 210, 268–290.

Hansen, S., P. J. Lambert, N. Bloom, S. J. Davis, R. Sadun, and B. Taska (2023): "Remote Work across Jobs, Companies, and Space," NBER Working Paper 31007, National Bureau of Economic Research.

Janson, S. (1988): "Normal Convergence by Higher Semiinvariants with Applications to Sums of Dependent Random Variables and Random Graphs," *The Annals of Probability*, 16, 305–312.

Kenyon-Dean, K., E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, R. Sarrazingendron, R. Verma, and D. Ruths (2018): "Sentiment Analysis: It's Complicated!" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 1886–1895.

Kessler, J. B., C. Low, and C. D. Sullivan (2019): "Incentivized Resume Rating: Eliciting Employer Preferences without Deception," *American Economic Review*, 109, 3713–3744.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-Out Estimation of Variance Components," *Econometrica*, 88, 1859–1898.

Levin, D. A. and Y. Peres (2017): *Markov Chains and Mixing Times*, Providence, RI: American Mathematical Society, 2 ed., with contributions by Elizabeth L. Wilmer.

Li, X. and P. Ding (2017): "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, 112, 1759–1769.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025): "Large Language Models: An Applied Econometric Framework," Working Paper 33344, National Bureau of Economic Research.

McKay, B. D. (2010): "Subgraphs of Random Graphs with Specified Degrees," in *Proceedings of the International Congress of Mathematicians 2010*, New Delhi: Hindustan Book Agency, vol. IV, 2489–2501.

Möbius, M. M. and T. S. Rosenblat (2006): "Why Beauty Matters," *American Economic Review*, 96, 222–235.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4, chap. 36, 2111–2245.

Neyman, J. (1934): "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–625.

Patriksson, M. (2008): "A Survey on the Continuous Nonlinear Resource Allocation Problem," *European Journal of Operational Research*, 185, 1–46.

Plank, B. (2022): "The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 10671–10682.

Rockafellar, R. T. (1970): *Convex Analysis*, vol. 28 of *Princeton Mathematical Series*, Princeton, NJ: Princeton University Press.

Särndal, C.-E., B. Swensson, and J. Wretman (1992): *Model Assisted Survey Sampling*, New York: Springer.

Shapiro, A. H. and D. J. Wilson (2022): "Taking the Fed at Its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis," *The Review of Economic Studies*, 89, 2768–2805.

Stuart, A. (1954): "A Simple Presentation of Optimum Sampling Results," *Journal of the Royal Statistical Society: Series B (Methodological)*, 16, 239–241.

Sun, L. (2026): "Empirical Welfare Maximization with Constraints," *Journal of Econometrics*, 253, 106169.

Tabord-Meehan, M. (2023): "Stratification Trees for Adaptive Randomisation in Randomised Controlled Trials," *The Review of Economic Studies*, 90, 2646–2673.

Thain, N., L. Dixon, and E. Wulczyn (2017): "Wikipedia Talk Labels: Toxicity," figshare dataset, version 2.

Tillé, Y. (2006): *Sampling Algorithms*, New York: Springer.

van der Klauw, T., M. E. T. Gerards, and J. L. Hurink (2017): "Resource Allocation Problems in Decentralized Energy Management," *OR Spectrum*, 39, 749–773.

Wiegand, M., J. Ruppenhofer, A. Schmidt, and C. Greenberg (2018): "Inducing a Lexicon of Abusive Words: A Feature-Based Approach," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 1046–1056.

Wu, A. H. (2020): "Gender Bias among Professionals: An Identity-Based Interpretation," *The Review of Economics and Statistics*, 102, 867–880.

Wulczyn, E., N. Thain, and L. Dixon (2017): "Ex Machina: Personal Attacks Seen at Scale," in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 1391–1399.

Zrnic, T. and E. J. Candès (2024): "Active Statistical Inference," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, vol. 235 of *Proceedings of Machine Learning Research*, 62993–63010.

# Appendix

## A Conditioning and probability conventions

We adopt a fixed-corpus asymptotic framework, analogous to the fixed-design and finite-population conventions in Kline et al. (2020, Section 1) and Abadie et al. (2020, Section 3.1). For each $J$, the corpus $W_{J,1:J}$ and all quantities determined before document sampling, including $\{X_{Jj}, a_{Jj}, g_{Jj}, \pi_{Jj}, m_{Jj}\}_{j=1}^{J}$, are treated as nonrandom. For different $J$, they form a deterministic sequence of finite arrays satisfying the assumptions made in the paper. For tractability, we suppress dependence on $J$ in the notation when there is no ambiguity. Quantities derived from the fixed corpus and the maintained annotator-response law, such as $\mu_j$, $\nu_j$, and $\tau_{jk}$, are also nonrandom.

For each $J$, let $(\Omega_J, \mathscr{F}_J, \mathbb{P}_J)$ support all final-study random quantities, including document selection, annotator allocation, and annotation outcomes. Let $\mathbb{E}_J$ and $\mathrm{Var}_J$ denote expectation and variance under $\mathbb{P}_J$. Unless stated otherwise, stochastic orders, probability limits, and weak-convergence statements refer to the corresponding row laws as $J \to \infty$. For this construction, no coupling across rows needs to be assumed.

Equivalently, all probabilities and expectations can be understood as conditional on the corpus and on the pre-study information used to construct the fixed quantities above. In particular, the predictions $\{g_j\}_{j=1}^{J}$ and the design rules may be estimated or constructed using auxiliary or pilot data available before the final study. Throughout the analysis, we condition on this information and treat the resulting values and rules as nonrandom components of the row-$J$ array.

## B Key decomposition and budget control

We maintain Assumptions **(M1)**–**(M3)** and **(D3)**. Document inclusion probabilities satisfy $\Pr(I_j = 1) = \pi_j > 0$, and each selected document $j$ is assigned to $m_j$ distinct annotators. Let $n$ denote the number of active annotators, indexed by $1, \ldots, n$, with $n = 0$ if $B = 0$.

**Lemma 2** (Decomposition and annotator representation)**.** *For $T_J^s(a)$, $T_J^m(a)$, and $\zeta_i(a)$ defined in Subsection 4.2, we have*

$$\widehat{\Phi}_J(a) - \Phi_J(a) = T_J^s(a) + T_J^m(a), \tag{B.1}$$

$$\sqrt{\bar{B}}\, T_J^m(a) = \sum_{i=1}^{n} \zeta_i(a). \tag{B.2}$$

*Conditional on $\mathcal{G}_J$, the annotator response vectors $\{(Y_{ij}^*)_{j=1}^{J}\}_{i=1}^{n}$ are mutually independent, with $\mathbb{E}[Y_{ij}^* \mid \mathcal{G}_J] = \mu_j$. Consequently, the contributions $\{\zeta_i(a)\}_{i=1}^{n}$ are conditionally independent and satisfy*

$$\mathbb{E}[\zeta_i(a) \mid \mathcal{G}_J] = 0, \qquad \mathbb{E}[T_J^m(a) \mid \mathcal{G}_J] = 0. \tag{B.3}$$

*Proof.* Using $r_j = \mu_j - g_j$, we obtain

$$\begin{aligned}
\widehat{\Phi}_J(a) - \Phi_J(a) &= \frac{1}{J}\sum_{j=1}^{J} a_j(g_j - \mu_j) + \frac{1}{J}\sum_{j\in\mathcal{S}_J} \frac{a_j}{\pi_j}(\bar{Y}_j - g_j) \\
&= -\frac{1}{J}\sum_{j=1}^{J} a_j r_j + \frac{1}{J}\sum_{j\in\mathcal{S}_J} \frac{a_j}{\pi_j}\big[r_j + (\bar{Y}_j - \mu_j)\big] \\
&= \frac{1}{J}\sum_{j=1}^{J} \left(\frac{I_j}{\pi_j} - 1\right) a_j r_j + \frac{1}{J}\sum_{j\in\mathcal{S}_J} \frac{a_j}{\pi_j}(\bar{Y}_j - \mu_j) \\
&= T_J^s(a) + T_J^m(a).
\end{aligned}$$

For the annotator representation, the definition of $\bar{Y}_j$ and interchanging the two finite sums give

$$\begin{aligned}
\sqrt{\bar{B}}\, T_J^m(a) &= \frac{\sqrt{\bar{B}}}{J}\sum_{j\in\mathcal{S}_J}\sum_{i\in M_j} \frac{a_j}{\pi_j m_j}(Y_{ij}^* - \mu_j) \\
&= \frac{\sqrt{\bar{B}}}{J}\sum_{i=1}^{n}\sum_{j\in\mathcal{J}_i} \frac{a_j}{\pi_j m_j}(Y_{ij}^* - \mu_j) \\
&= \sum_{i=1}^{n} \zeta_i(a).
\end{aligned}$$

To establish the conditional claims, write $L_i = (\theta_i, (\varepsilon_{ij})_{j\le J})$. Assumptions **(M3)** and **(M2)** make the blocks $\{L_i\}_{i\ge 1}$ mutually independent, while Assumption **(D3)** makes $\mathcal{G}_J$ independent of their joint sigma-field. Thus, for every positive integer $q$ and every bounded measurable $f$,

$$\mathbb{E}[f(L_1, \ldots, L_q) \mid \mathcal{G}_J] = \mathbb{E}[f(L_1, \ldots, L_q)]. \tag{B.4}$$

By Assumption **(M1)**, each annotator's response vector is a measurable function of that annotator's block and the fixed corpus. The response vectors therefore remain independent conditional on $\mathcal{G}_J$, with conditional means $\mu_j$. Since $\mathcal{G}_J$ fixes the number of active annotators, their assignments, and the coefficients in $\zeta_i(a)$, conditional independence and centering of the annotator contributions follow. Equation (B.2) then gives $\mathbb{E}[T_J^m(a) \mid \mathcal{G}_J] = 0$. □

**Lemma 3** (Conditional measurement variance)**.** *The conditional measurement variance is*

$$\begin{aligned}
v_{\varepsilon,J}(a) &:= \bar{B}\ \mathrm{Var}(T_J^m(a) \mid \mathcal{G}_J) \\
&= \frac{\bar{B}}{J^2}\sum_{j=1}^{J} \frac{I_j a_j^2 \nu_j}{\pi_j^2 m_j} + \frac{\bar{B}}{J^2}\sum_{\substack{1\le j,k\le J \\ j\ne k}} \frac{I_j I_k N_{jk} a_j a_k \tau_{jk}}{\pi_j \pi_k m_j m_k},
\end{aligned} \tag{B.5}$$

*where the second sum is over ordered document pairs. Equivalently,*

$$v_{\varepsilon,J}(a) = \sum_{i=1}^{n} \mathrm{Var}(\zeta_i(a) \mid \mathcal{G}_J). \tag{B.6}$$

*Proof.* Each selected document $j$ receives $m_j$ labels from distinct annotators. By (B.4), these labels are conditionally independent, each with variance $\nu_j$. Hence

$$\mathrm{Var}(\bar{Y}_j - \mu_j \mid \mathcal{G}_J) = \frac{\nu_j}{m_j}.$$

For distinct documents $j \neq k$, design exogeneity and Assumptions **(M1)**–**(M2)** give

$$\mathrm{Cov}(Y^*_{ij}, Y^*_{\ell k} \mid \mathcal{G}_J) = \mathbf{1}\{i = \ell\}\tau_{jk}.$$

Responses from different annotators are independent. Note:

$$\begin{aligned}
\mathrm{Cov}(Y^*_{ij}, Y^*_{ik} \mid \mathcal{G}_J) &= \mathrm{Cov}(Y^*_{ij}, Y^*_{ik}) \\
&= \underbrace{\mathbb{E}\big[\mathrm{Cov}(Y^*_{ij}, Y^*_{ik} \mid \theta_i)\big]}_{=0} + \mathrm{Cov}\big(\mathbb{E}[Y^*_{ij} \mid \theta_i], \mathbb{E}[Y^*_{ik} \mid \theta_i]\big) \\
&= \tau_{jk}, \qquad j \neq k.
\end{aligned}$$

Thus only the $N_{jk}$ shared annotators contribute to the covariance between the means of two selected documents:

$$\mathrm{Cov}(\bar{Y}_j - \mu_j, \bar{Y}_k - \mu_k \mid \mathcal{G}_J) = \frac{N_{jk}}{m_j m_k}\tau_{jk}.$$

Using the preceding variance and covariance identities, we obtain

$$\begin{aligned}
\bar{B}\ \mathrm{Var}(T^m_J(a) \mid \mathcal{G}_J) &= \frac{\bar{B}}{J^2} \sum_{j,k \in \mathcal{S}_J} \frac{a_j a_k}{\pi_j \pi_k} \mathrm{Cov}(\bar{Y}_j - \mu_j, \bar{Y}_k - \mu_k \mid \mathcal{G}_J) \\
&= \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{I_j a_j^2 \nu_j}{\pi_j^2 m_j} + \frac{\bar{B}}{J^2} \sum_{\substack{1 \leq j,k \leq J \\ j \neq k}} \frac{I_j I_k N_{jk} a_j a_k \tau_{jk}}{\pi_j \pi_k m_j m_k},
\end{aligned}$$

which proves (B.5). The annotator representation and conditional independence in Lemma 2 give (B.6). □

**Lemma 4** (Orthogonality and unconditional variance)**.** *The sampling and measurement components satisfy*

$$\mathrm{Cov}(T^s_J(a), T^m_J(a)) = 0. \tag{B.7}$$

*The total variance satisfies*

$$\bar{B}\ \mathrm{Var}(\widehat{\Phi}_J(a) - \Phi_J(a)) = v_{r,J}(a) + \bar{v}_{\varepsilon,J}(a), \tag{B.8}$$

*where* $v_{r,J}(a) := \bar{B}\ \mathrm{Var}(T^s_J(a))$ *and* $\bar{v}_{\varepsilon,J}(a) := \mathbb{E}[v_{\varepsilon,J}(a)]$. *Under Assumption* ***(D1)*** *(Poisson sampling),*

$$v_{r,J}(a) = \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{1 - \pi_j}{\pi_j} a_j^2 r_j^2. \tag{B.9}$$

*Proof.* By Lemma 2, $\mathbb{E}[T_J^m(a) \mid \mathcal{G}_J] = 0$. Since $T_J^s(a)$ is $\mathcal{G}_J$-measurable,

$$\begin{aligned}\operatorname{Cov}(T_J^s(a), T_J^m(a)) &= \mathbb{E}[T_J^s(a) T_J^m(a)] \\ &= \mathbb{E}[T_J^s(a) \mathbb{E}[T_J^m(a) \mid \mathcal{G}_J]] = 0.\end{aligned}$$

This proves (B.7).

Conditional centering and the law of total variance give

$$\begin{aligned}\bar{B} \operatorname{Var}(T_J^m(a)) &= \mathbb{E}\big[\bar{B} \operatorname{Var}(T_J^m(a) \mid \mathcal{G}_J)\big] \\ &= \bar{v}_{\varepsilon,J}(a).\end{aligned}$$

Combining this identity with the decomposition and orthogonality proves (B.8).

Finally, under Assumption **(D1)**, the indicators $I_j$ are independent Bernoulli variables with variances $\pi_j(1-\pi_j)$. Therefore,

$$v_{r,J}(a) = \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{a_j^2 r_j^2}{\pi_j^2} \operatorname{Var}(I_j) = \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{1-\pi_j}{\pi_j} a_j^2 r_j^2,$$

which proves (B.9). □

**Lemma 5** (Bernstein control of the realized label count)**.** *Suppose Assumption **(D1)** holds and $1 \le m_j \le \bar{m} < \infty$ for every $j$, where $\bar{m}$ does not depend on $J$. Then $\mathbb{E}[B] = \bar{B}$ and*

$$V_{B,J} := \operatorname{Var}(B) = \sum_{j=1}^{J} m_j^2 \pi_j (1-\pi_j) \le \bar{m} \bar{B}. \tag{B.10}$$

*For every $t > 0$,*

$$\mathbb{P}(|B - \bar{B}| \ge t) \le 2 \exp\left\{ -\frac{t^2}{2(V_{B,J} + \bar{m}t/3)} \right\} \le 2 \exp\left\{ -\frac{t^2}{2\bar{m}(\bar{B} + t/3)} \right\}. \tag{B.11}$$

*Consequently, for $0 < \alpha < 1$, let $L_\alpha := \log(2/\alpha)$. With probability at least $1-\alpha$,*

$$|B - \bar{B}| \le \sqrt{2 V_{B,J} L_\alpha} + \frac{2\bar{m}}{3} L_\alpha \le \sqrt{2\bar{m}\bar{B} L_\alpha} + \frac{2\bar{m}}{3} L_\alpha. \tag{B.12}$$

*If $\bar{B} \to \infty$, then*

$$\frac{B}{\bar{B}} \xrightarrow{p} 1. \tag{B.13}$$

*In particular, $\sqrt{B/\bar{B}} \xrightarrow{p} 1$, $\mathbb{P}(B = 0) \to 0$, and $\mathbb{P}(\bar{B}/2 \le B \le 2\bar{B}) \to 1$.*

*Proof.* The mean and variance identities follow from $B = \sum_j m_j I_j$, and $m_j^2 \le \bar{m} m_j$ gives $V_{B,J} \le \bar{m} \sum_j \pi_j m_j = \bar{m}\bar{B}$. The variables $m_j(I_j - \pi_j)$ are independent, centered, bounded in absolute value by $\bar{m}$, and have total variance $V_{B,J}$. Bernstein's inequality applied to these variables and their negatives gives (B.11); see Boucheron et al. (2013, Equation (2.10)).

For $t_\alpha := \sqrt{2V_{B,J}L_\alpha} + (2\bar{m}/3)L_\alpha$,

$$t_\alpha^2 - 2L_\alpha\big(V_{B,J} + \bar{m}t_\alpha/3\big) = \frac{2\bar{m}L_\alpha}{3}\sqrt{2V_{B,J}L_\alpha} \geq 0.$$

This proves the first inequality in (B.12); the variance bound proves the second. Finally, setting $t = \epsilon\bar{B}$ in (B.11) gives

$$\mathbb{P}\left(\left|\frac{B}{\bar{B}} - 1\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{\epsilon^2\bar{B}}{2\bar{m}(1+\epsilon/3)}\right\} \longrightarrow 0.$$

The remaining conclusions follow by continuity and elementary event inclusion. □

# C Proofs for asymptotic normality

## C.1 The high-level central limit theorem

### C.1.1 Document-sampling component

We first establish the Gaussian limit for the document-sampling component.

*Proof of Lemma 1.* Define

$$Z_{Jj}^s := \frac{\sqrt{\bar{B}}}{J}\left(\frac{I_j}{\pi_j} - 1\right)a_j r_j, \qquad j = 1, \ldots, J.$$

Under Assumption **(D1)**, these variables are independent and centered. By (B.9) and Condition **(R3)**,

$$\sqrt{\bar{B}}\,T_J^s(a) = \sum_{j=1}^{J} Z_{Jj}^s, \qquad \sum_{j=1}^{J}\mathrm{Var}(Z_{Jj}^s) = v_{r,J}(a) \longrightarrow v_r(a).$$

Boundedness of the response scale and Assumption **(R1)** imply that, for some finite $C_r$,

$$\max_{j\leq J}|r_j| \leq C_r.$$

Consequently,

$$\max_{j\leq J}|Z_{Jj}^s| \leq C_r \max_{j\leq J}\frac{\sqrt{\bar{B}}}{J\pi_j}|a_j| = C_r\delta_J^{\mathrm{lev}}(a) \longrightarrow 0$$

by (4.5).

Suppose first that $v_r(a) > 0$, and put $s_{s,J}^2 := v_{r,J}(a)$. Then $s_{s,J} \to \sqrt{v_r(a)} > 0$, so

$$\max_{j\leq J}\frac{|Z_{Jj}^s|}{s_{s,J}} \longrightarrow 0.$$

Hence, for every $\epsilon > 0$,

$$\frac{1}{s_{s,J}^2}\sum_{j=1}^{J}\mathbb{E}\big[(Z_{Jj}^s)^2\mathbf{1}\{|Z_{Jj}^s| > \epsilon s_{s,J}\}\big] = 0$$

for all sufficiently large $J$. The Lindeberg theorem (Billingsley, 1995, Theorem 27.2) therefore

gives

$$\frac{\sum_{j=1}^{J} Z_{Jj}^{s}}{s_{s,J}} \Rightarrow \mathcal{N}(0,1).$$

Slutsky's theorem then yields

$$\sqrt{\bar{B}}\, T_J^s(a) \Rightarrow \mathcal{N}\big(0, v_r(a)\big).$$

If $v_r(a) = 0$, the same conclusion follows directly from Chebyshev's inequality, since

$$\mathrm{Var}\Big(\sqrt{\bar{B}}\, T_J^s(a)\Big) = v_{r,J}(a) \longrightarrow 0.$$

The exact variance formula was established in Lemma 4. □

The next lemma gives the conditional Gaussian approximation needed for the measurement term. Its approximating variance remains specific to the realized sample and allocation, so Condition **(R4)** is not imposed.

### C.1.2 Measurement component

We first bound the conditional Gaussian approximation error in terms of the third absolute moments of the annotator contributions. Under the regularity conditions of Theorem 1, we can control their sum using the same constants for every admissible graph and every $J$. We can then show that the resulting error bound vanishes in probability as $\bar{B} \to \infty$.

**Lemma 6** (Gaussian approximation conditional on the realized design)**.** *Suppose the conditions of Lemmas 2 and 3 hold. Define*

$$\Lambda_J(a) := \sum_{i=1}^{n} \mathbb{E}\big[|\zeta_i(a)|^3 \mid \mathcal{G}_J\big]\,.$$

*Then, for every fixed $t \in \mathbb{R}$, the following inequality holds almost surely:*

$$\left|\mathbb{E}\Big[\exp\Big\{it\sqrt{\bar{B}}\, T_J^m(a)\Big\} \,\Big|\, \mathcal{G}_J\Big] - \exp\Big\{-\frac{t^2}{2} v_{\varepsilon,J}(a)\Big\}\right| \le \frac{|t|^3}{6}\Lambda_J(a) + \frac{t^4}{8}\Lambda_J(a)^{4/3}. \tag{C.1}$$

*Consequently, if $\Lambda_J(a) \xrightarrow{p} 0$,*

$$\mathbb{E}\Big[\exp\Big\{it\sqrt{\bar{B}}\, T_J^m(a)\Big\} \,\Big|\, \mathcal{G}_J\Big] = \exp\Big\{-\frac{t^2}{2} v_{\varepsilon,J}(a)\Big\} + o_p(1). \tag{C.2}$$

*Proof.* Conditional on $\mathcal{G}_J$, the number $n$ of active annotators and their assignments are fixed, and

$$\sqrt{\bar{B}}\, T_J^m(a) = \sum_{i=1}^{n} \zeta_i(a)$$

is a sum of independent centered variables. Write

$$v_{Ji}(a) := \mathbb{E}[\zeta_i(a)^2 \mid \mathcal{G}_J], \qquad q_{Ji}(a) := \mathbb{E}[|\zeta_i(a)|^3 \mid \mathcal{G}_J].$$

By Lemma 3 and the definition of $\Lambda_J(a)$,

$$\sum_{i=1}^{n} v_{Ji}(a) = v_{\varepsilon,J}(a), \qquad \sum_{i=1}^{n} q_{Ji}(a) = \Lambda_J(a).$$

When $n = 0$, all sums and products below are interpreted using the usual empty-sum and empty-product conventions.

For each active annotator, define

$$\phi_{Ji}(t) := \mathbb{E}\Big[e^{it\zeta_i(a)} \mid \mathcal{G}_J\Big], \qquad g_{Ji}(t) := \exp\Big\{-\frac{t^2}{2} v_{Ji}(a)\Big\}.$$

The second-order Taylor remainder for the complex exponential gives

$$\left|e^{ix} - \left(1 + ix - \frac{x^2}{2}\right)\right| \le \frac{|x|^3}{6}, \qquad x \in \mathbb{R};$$

see Billingsley (1995, equation (26.4)). Using $\mathbb{E}[\zeta_i(a) \mid \mathcal{G}_J] = 0$ and the definition of $v_{Ji}(a)$, we obtain

$$\begin{aligned}
&\left|\phi_{Ji}(t) - \left\{1 - \frac{t^2}{2} v_{Ji}(a)\right\}\right| \\
&\quad = \left|\mathbb{E}\left[e^{it\zeta_i(a)} - 1 - it\zeta_i(a) + \frac{t^2}{2}\zeta_i(a)^2 \,\middle|\, \mathcal{G}_J\right]\right| \\
&\quad \le \mathbb{E}\left[\left|e^{it\zeta_i(a)} - 1 - it\zeta_i(a) + \frac{t^2}{2}\zeta_i(a)^2\right| \,\middle|\, \mathcal{G}_J\right] \\
&\quad \le \frac{|t|^3}{6} q_{Ji}(a),
\end{aligned}$$

where the last inequality applies the Taylor bound with $x = t\zeta_i(a)$.

Moreover, for $x \ge 0$,

$$0 \le e^{-x} - (1 - x) = \int_0^x (1 - e^{-s})\, ds \le \int_0^x s\, ds = \frac{x^2}{2}.$$

Applying this with $x = t^2 v_{Ji}(a)/2$ yields

$$\left|g_{Ji}(t) - \left\{1 - \frac{t^2}{2} v_{Ji}(a)\right\}\right| \le \frac{t^4}{8} v_{Ji}(a)^2.$$

The same factorwise exponential remainder is used in the characteristic-function proof of the Berry–Esseen theorem; see Durrett (2019, proof of Theorem 3.4.17, p. 160). Combining the last two bounds gives

$$|\phi_{Ji}(t) - g_{Ji}(t)| \le \frac{|t|^3}{6} q_{Ji}(a) + \frac{t^4}{8} v_{Ji}(a)^2. \tag{C.3}$$

Conditional independence gives

$$\mathbb{E}\Big[e^{it\sum_{i=1}^n \zeta_i(a)} \,\Big|\, \mathcal{G}_J\Big] = \prod_{i=1}^{n} \phi_{Ji}(t), \qquad \exp\left\{-\frac{t^2}{2}\sum_{i=1}^{n} v_{Ji}(a)\right\} = \prod_{i=1}^{n} g_{Ji}(t).$$

Replacing the factors one at a time gives the exact identity

$$\prod_{i=1}^{n} \phi_{Ji}(t) - \prod_{i=1}^{n} g_{Ji}(t) = \sum_{k=1}^{n} \left\{ \prod_{i<k} \phi_{Ji}(t) \right\} \{\phi_{Jk}(t) - g_{Jk}(t)\} \left\{ \prod_{i>k} g_{Ji}(t) \right\}.$$

Because $|\phi_{Ji}(t)| \le 1$ and $0 < g_{Ji}(t) \le 1$, taking absolute values yields

$$\left| \prod_{i=1}^{n} \phi_{Ji}(t) - \prod_{i=1}^{n} g_{Ji}(t) \right| \le \sum_{i=1}^{n} |\phi_{Ji}(t) - g_{Ji}(t)|.$$

This is the product comparison in Billingsley (1995, equation (27.5)). Compare also the proof of Theorem 27.2, especially equations (27.11)–(27.15). Consequently,

$$\begin{aligned} &\left| \mathbb{E}\left[ e^{it\sum_{i=1}^{n} \zeta_i(a)} \,\middle|\, \mathcal{G}_J \right] - \exp\left\{ -\frac{t^2}{2} \sum_{i=1}^{n} v_{Ji}(a) \right\} \right| \\ &\qquad \le \frac{|t|^3}{6} \sum_{i=1}^{n} q_{Ji}(a) + \frac{t^4}{8} \sum_{i=1}^{n} v_{Ji}(a)^2. \end{aligned} \tag{C.4}$$

The identity and inequality remain valid when $n = 0$ under the stated empty-product convention.

Finally, conditional Lyapunov's moment inequality gives

$$v_{Ji}(a)^{3/2} = \left\{ \mathbb{E}[\zeta_i(a)^2 \mid \mathcal{G}_J] \right\}^{3/2} \le \mathbb{E}[|\zeta_i(a)|^3 \mid \mathcal{G}_J] = q_{Ji}(a).$$

Together with monotonicity of finite-dimensional $\ell^p$ norms, this implies

$$\sum_{i=1}^{n} v_{Ji}(a)^2 \le \left\{ \sum_{i=1}^{n} v_{Ji}(a)^{3/2} \right\}^{4/3} \le \left\{ \sum_{i=1}^{n} q_{Ji}(a) \right\}^{4/3} = \Lambda_J(a)^{4/3}.$$

Substitution into (C.4), together with the exact identities for the two sums, proves (C.1). If $\Lambda_J(a) \xrightarrow{p} 0$, the right-hand side of (C.1) is $o_p(1)$. This proves (C.2). □

### C.1.3 Combining the two terms

We now combine the sampling limit with the conditional measurement approximation. Variance stabilization makes the limiting conditional characteristic function nonrandom, allowing the characteristic function of the sum to factor asymptotically.

*Proof of Theorem 1.* Write

$$S_J := \sqrt{\bar{B}}\, T_J^s(a), \qquad M_J := \sqrt{\bar{B}}\, T_J^m(a), \qquad \varphi_J(t) := \mathbb{E}[e^{itM_J} \mid \mathcal{G}_J].$$

Let $R_Y := \sup_{y,y' \in \mathcal{Y}} |y - y'|$. Balanced spending and the annotator-degree bound give

$$|\zeta_i(a)| \le \frac{K\bar{a}R_Y}{c_\lambda \sqrt{\bar{B}}}.$$

Since the number of active annotators is at most $B$, Lemma 5 implies

$$\Lambda_J(a) \le B \left( \frac{K \bar{a} R_Y}{c_\lambda \sqrt{\bar{B}}} \right)^3 = O_p(\bar{B}^{-1/2}) = o_p(1).$$

Because $S_J$ is $\mathcal{G}_J$-measurable, the tower property gives

$$\mathbb{E}[e^{it(S_J + M_J)}] = \mathbb{E}[e^{itS_J} \varphi_J(t)]. \tag{C.5}$$

Let

$$c_t := \exp\left\{ -\frac{t^2}{2} v_\varepsilon(a) \right\}, \qquad D_J(t) := \varphi_J(t) - c_t.$$

Lemma 6 therefore gives

$$\varphi_J(t) - \exp\left\{ -\frac{t^2}{2} v_{\varepsilon,J}(a) \right\} \xrightarrow{p} 0,$$

while Condition **(R4)** and continuity of the exponential give

$$\exp\left\{ -\frac{t^2}{2} v_{\varepsilon,J}(a) \right\} - c_t \xrightarrow{p} 0.$$

Hence $D_J(t) \xrightarrow{p} 0$. Moreover, $|D_J(t)| \le 2$. Thus, for every $\eta > 0$,

$$\mathbb{E}|D_J(t)| \le \eta + 2\mathbb{P}\{|D_J(t)| > \eta\}.$$

Taking the limit superior and then letting $\eta \downarrow 0$ shows that $\mathbb{E}|D_J(t)| \to 0$. Therefore

$$\left| \mathbb{E}[e^{itS_J} D_J(t)] \right| \le \mathbb{E}|D_J(t)| \longrightarrow 0.$$

Substitution into (C.5) yields

$$\mathbb{E}[e^{it(S_J + M_J)}] = \exp\left\{ -\frac{t^2}{2} v_\varepsilon(a) \right\} \mathbb{E}[e^{itS_J}] + o(1). \tag{C.6}$$

By Lemma 1,

$$\mathbb{E}[e^{itS_J}] \longrightarrow \exp\left\{ -\frac{t^2}{2} v_r(a) \right\}.$$

Combining this with (C.6) and applying Lévy's continuity theorem (Billingsley, 1995, Theorem 26.3) gives

$$\sqrt{\bar{B}} \{ \widehat{\Phi}_J(a) - \Phi_J(a) \} = S_J + M_J \Rightarrow \mathcal{N}(0, v_r(a) + v_\varepsilon(a)).$$

Finally, if $1 \le m_j \le \bar{m}$, Lemma 5 gives $\sqrt{B/\bar{B}} \xrightarrow{p} 1$. Hence

$$\sqrt{B} \{ \widehat{\Phi}_J(a) - \Phi_J(a) \} = \sqrt{B/\bar{B}} \, \sqrt{\bar{B}} \{ \widehat{\Phi}_J(a) - \Phi_J(a) \}$$

has the same limit by Slutsky's theorem. □

## C.2 Why allocation regularity is needed

Bounded degrees and regular document arrays do not by themselves stabilize the graph-specific measurement variance. The following fixed-budget example uses a sample-level imbalance to select between two bounded-load allocation rules. That imbalance also determines the first-order sampling error.

**Proposition 3** (Failure under sample-adaptive allocation)**.** *There exists a fixed-corpus sequence under fixed-size simple random sampling with*

$$B = \bar{B} = N_J, \qquad m_j = a_j = \lambda_j = 1, \qquad n = N_J/2, \qquad K = 2,$$

*satisfying the bounded-response model, Assumptions **(M1)**–**(M3)** and **(D3)**, Assumption **(R1)**, the leverage bound* (4.5)*, the condition $\Lambda_J(a) \xrightarrow{p} 0$, the corresponding document-sampling CLT and variance limit, and the two document-array limits used for variance stabilization, but not Condition **(R4)**. Its normalized estimator has a non-Gaussian limit. Neither exact integrated-variance standardization nor oracle hybrid studentization using the graph-specific measurement variance has a standard-normal limit.*

*Fixed-size SRS replaces Assumption **(D1)**, and the sample-adaptive allocation intentionally falls outside Assumption **(D2)**.*

*Proof.* Let $J$ and $N_J$ be even and suppose

$$N_J \to \infty, \qquad J - N_J \to \infty, \qquad f_J := N_J/J \to f \in (0,1).$$

Divide the fixed corpus into two equally sized classes,

$$\chi_j = \begin{cases} 1, & j \le J/2, \\ -1, & j > J/2. \end{cases}$$

Set $\pi_j = f_J$, $m_j = a_j = 1$, and $g_j = c_0$. Thus $\bar{B} = Jf_J = N_J$ and $\lambda_j := J\pi_j m_j/\bar{B} = 1$.

Let $F_i, G_{i,1}, G_{i,-1}$, and $U_{ij}$ be mutually independent Rademacher variables, with the annotator blocks i.i.d. across $i$, and set

$$Y^*_{ij} = c_0 + \delta\chi_j + \tau_0 F_i + \tau_1 G_{i,\chi_j} + \eta U_{ij}, \qquad \delta, \tau_1 > 0.$$

Choose the constants so that the maintained outcome bound holds. This is the maintained response model with

$$\theta_i = (F_i, G_{i,1}, G_{i,-1}), \qquad \varepsilon_{ij} = U_{ij}.$$

Consequently,

$$\mu_j = c_0 + \delta\chi_j, \qquad r_j = \delta\chi_j,$$

and

$$\nu_j = \eta^2 + \tau_0^2 + \tau_1^2, \qquad \tau_{jk} = \tau_0^2 + \tau_1^2 \mathbf{1}\{\chi_j = \chi_k\}. \tag{C.7}$$

Draw exactly $N_J$ documents by SRS without replacement and define

$$N_{+,J} := \sum_{j \in \mathcal{S}_J} \mathbf{1}\{\chi_j = 1\}, \quad N_{-,J} := \sum_{j \in \mathcal{S}_J} \mathbf{1}\{\chi_j = -1\},$$
$$D_J := N_{+,J} - N_{-,J}.$$

Pair the selected documents and assign the $N_J/2$ pairs to distinct annotators. If $D_J \geq 0$, draw uniformly from the matchings with as many within-class pairs as parity permits. If $D_J < 0$, draw a uniform perfect matching of all selected documents. Both the SRS draw and all matching randomization are independent of the annotator-response blocks, so Assumption **(D3)** holds. Both branches have document degree one and annotator degree two.

Write

$$S_J := \sqrt{N_J}\, T_J^s(a), \qquad M_J := \sqrt{N_J}\, T_J^m(a).$$

Because the corpus class counts are equal and the number of selected positive-class documents has a hypergeometric distribution, we have

$$S_J = \delta \frac{D_J}{\sqrt{N_J}}, \qquad \mathrm{Var}(D_J) = \frac{N_J(J - N_J)}{J - 1}.$$

The finite-population central limit theorem (Li and Ding, 2017, Theorem 1) therefore gives

$$S_J \Rightarrow S, \qquad S \sim \mathcal{N}(0, v_r), \qquad v_r := \delta^2(1 - f). \tag{C.8}$$

Let $\rho_J$ denote the fraction of annotators whose paired documents have the same class, and set $V_J := v_{\varepsilon,J}(a)$. Lemma 3 and (C.7) give

$$V_J = N_J \,\mathrm{Var}\{T_J^m(a) \mid \mathcal{G}_J\} = \eta^2 + 2\tau_0^2 + \tau_1^2(1 + \rho_J). \tag{C.9}$$

In the within-class branch, at most one pair is cross-class, so

$$0 \leq 1 - \rho_J \leq \frac{2}{N_J}.$$

In the pooled branch,

$$\mathbb{E}[\rho_J \mid \mathcal{S}_J] = \frac{N_{+,J}(N_{+,J} - 1) + N_{-,J}(N_{-,J} - 1)}{N_J(N_J - 1)} = \frac{1}{2} + \frac{D_J^2 - N_J}{2N_J(N_J - 1)} \xrightarrow{p} \frac{1}{2}.$$

For $N_J \geq 4$, concentration in the pooled branch follows by writing $\rho_J = 1 - 2C_J/N_J$, where $C_J$ is the number of cross-class pairs. For each selected + document $j$, let $H_j$ indicate that it is matched to a $-$ document, so that $C_J = \sum_j H_j$. Conditional on $\mathcal{S}_J$, the one- and two-indicator probabilities are $N_{-,J}/(N_J - 1)$ and $N_{-,J}(N_{-,J} - 1)/\{(N_J - 1)(N_J - 3)\}$, respectively. It follows that

$$\mathrm{Var}(\rho_J \mid \mathcal{S}_J) = \frac{8N_{+,J}(N_{+,J} - 1)N_{-,J}(N_{-,J} - 1)}{N_J^2(N_J - 1)^2(N_J - 3)} = O(N_J^{-1}).$$

It follows that

$$V_J - \{v_H \mathbf{1}\{D_J \geq 0\} + v_P \mathbf{1}\{D_J < 0\}\} \xrightarrow{p} 0, \qquad \begin{cases} v_H = \eta^2 + 2\tau_0^2 + 2\tau_1^2, \\ v_P = \eta^2 + 2\tau_0^2 + \frac{3}{2}\tau_1^2. \end{cases} \tag{C.10}$$

Because $v_H \neq v_P$ and $\mathbb{P}(D_J \geq 0), \mathbb{P}(D_J < 0) \to 1/2$, $V_J$ does not converge in probability to a constant, so Condition **(R4)** fails.

Each of the $N_J/2$ annotator contributions to $M_J$ is bounded by $C/\sqrt{N_J}$. Hence, uniformly in $\mathcal{G}_J$,

$$\Lambda_J(a) = \sum_i \mathbb{E}[|\zeta_i(a)|^3 \mid \mathcal{G}_J] = O(N_J^{-1/2}),$$

so $\Lambda_J(a) \xrightarrow{p} 0$. By Lemma 6,

$$\mathbb{E}(e^{itM_J} \mid \mathcal{G}_J) = e^{-t^2 V_J/2} + o_p(1).$$

The error is bounded and hence is also $o(1)$ in $L^1$. Therefore, for every $u, t, w \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}[e^{iuS_J + itM_J + iwV_J}] &= \mathbb{E}\big[e^{iuS_J + iwV_J}\mathbb{E}(e^{itM_J} \mid \mathcal{G}_J)\big] \\ &= \mathbb{E}[e^{iuS_J + iwV_J - t^2V_J/2}] + o(1) \\ &\longrightarrow \mathbb{E}[e^{iuS + iwv(S) - t^2 v(S)/2}], \end{aligned}$$

where

$$v(S) := v_H \mathbf{1}\{S \geq 0\} + v_P \mathbf{1}\{S < 0\}.$$

The final convergence follows from (C.8)–(C.10) and the extended continuous mapping theorem, since $\operatorname{sign}(S_J) = \operatorname{sign}(D_J)$ and $\mathbb{P}(S = 0) = 0$. The limiting characteristic function identifies

$$(S_J, M_J, V_J) \Rightarrow \big(S, \sqrt{v(S)}Z, v(S)\big), \qquad Z \sim \mathcal{N}(0,1), \quad Z \perp S.$$

Consequently,

$$\sqrt{N_J}\{\widehat{\Phi}_J(a) - \Phi_J(a)\} = S_J + M_J \Rightarrow X := S + \sqrt{v(S)}Z. \tag{C.11}$$

Although $\mathbb{E}[X] = 0$, symmetry of $S$ gives

$$\mathbb{E}[X^3] = 3\mathbb{E}[Sv(S)] = 3(v_H - v_P)\sqrt{\frac{v_r}{2\pi}} \neq 0,$$

so $X$ is not Gaussian.

The exact scaled SRS sampling variance is

$$v_{s,J}^{\mathrm{SRS}} := N_J \operatorname{Var}\{T_J^s(a)\} = \delta^2 \frac{J - N_J}{J - 1} \longrightarrow v_r.$$

Conditional centering gives exact orthogonality of the sampling and measurement components.

Hence their exact integrated variance is

$$v_{\mathrm{int},J} := N_J \operatorname{Var}\{\widehat{\Phi}_J(a) - \Phi_J(a)\} = v_{s,J}^{\mathrm{SRS}} + \mathbb{E}[V_J].$$

Since $V_J$ is uniformly bounded and each branch has limiting probability $1/2$,

$$v_{\mathrm{int},J} \longrightarrow v_r + \frac{v_H + v_P}{2}.$$

It follows that

$$\frac{\sqrt{N_J}\{\widehat{\Phi}_J(a) - \Phi_J(a)\}}{\sqrt{v_{\mathrm{int},J}}} \Rightarrow \frac{X}{\sqrt{v_r + (v_H + v_P)/2}},$$

which remains non-Gaussian. The preceding joint convergence also yields the oracle hybrid limit

$$\frac{\sqrt{N_J}\{\widehat{\Phi}_J(a) - \Phi_J(a)\}}{\sqrt{v_{s,J}^{\mathrm{SRS}} + V_J}} \Rightarrow \frac{S + \sqrt{v(S)}Z}{\sqrt{v_r + v(S)}}.$$

The mean of the limiting variable is

$$\sqrt{\frac{v_r}{2\pi}} \left\{ \frac{1}{\sqrt{v_r + v_H}} - \frac{1}{\sqrt{v_r + v_P}} \right\} \neq 0,$$

so this limit is not standard normal either.

Finally, the document arrays are regular:

$$\frac{1}{J}\sum_{j=1}^{J} \frac{a_j^2 \nu_j}{\lambda_j} = \eta^2 + \tau_0^2 + \tau_1^2, \qquad \frac{1}{J^2}\sum_{j \neq k} a_j a_k \tau_{jk} \longrightarrow \tau_0^2 + \frac{1}{2}\tau_1^2,$$

while

$$\lambda_j = 1, \qquad \delta_J^{\mathrm{lev}}(a) = \frac{\sqrt{N_J}}{J f_J} = \frac{1}{\sqrt{N_J}} \longrightarrow 0.$$

Thus the failure is caused by the sample-adaptive global allocation switch, not by irregular document moments, excessive degrees, or failure of the conditional annotator Gaussian approximation. □

### C.3 Primitive conditions for variance stabilization

This subsection proves Theorem 2. Let

$$\mathcal{E}_J^{\mathrm{feas}} := \{B = 0\} \cup \left\{ B > 0,\ \mathfrak{A}_J(\boldsymbol{d}_J, \boldsymbol{k}_J^{\mathrm{cf}}) \neq \varnothing \right\}$$

be the event on which the capacity-filled graph is feasible (with the empty graph used when $B = 0$), and define

$$\mathcal{H}_J := \mathcal{E}_J^{\mathrm{feas}} \cap \left\{ \frac{\bar{B}}{2} \leq B \leq 2\bar{B} \right\}. \tag{C.12}$$

We first rigorously verify that, whenever $B \geq K\bar{m}$, the requested annotation depths and prescribed annotator workloads can be satisfied without assigning the same document to an anno-

tator twice. For $B > 0$, let $m_{(1)} \geq \cdots \geq m_{(|\mathcal{S}_J|)}$ be the selected document degrees in decreasing order, and let $k_1, \ldots, k_{n_0}$ be the prescribed capacity-filled annotator degrees. Since both degree sequences sum to $B$, the Gale–Ryser criterion states that a simple bipartite graph with these degrees exists if and only if

$$\sum_{\ell=1}^{t} m_{(\ell)} \leq \sum_{i=1}^{n_0} \min(t, k_i), \qquad t = 1, \ldots, |\mathcal{S}_J|.$$

The left-hand side is the evaluation demand of the $t$ most heavily annotated documents; the right-hand side is the total capacity annotators can contribute to those documents without repeated evaluations. For $t \geq K$, the right-hand side equals $B$, so the inequality holds automatically. For $t < K$, capacity filling provides $\lfloor B/K \rfloor$ annotators with workload $K$. If $B \geq K\bar{m}$, then

$$\sum_{\ell=1}^{t} m_{(\ell)} \leq t\bar{m} \leq t\lfloor B/K \rfloor \leq \sum_{i=1}^{n_0} \min(t, k_i).$$

Thus the capacity-filled allocation is feasible whenever $B \geq K\bar{m}$.

Since $K\bar{m}$ is fixed and $\bar{B} \to \infty$, we have $\bar{B}/2 \geq K\bar{m}$ for all sufficiently large $J$. The preceding feasibility argument therefore gives $\mathcal{H}_J = \{\bar{B}/2 \leq B \leq 2\bar{B}\}$ for such $J$. By Lemma 5,

$$\mathbb{P}(\mathcal{H}_J^c) = \mathbb{P}\big(B \notin [\bar{B}/2, 2\bar{B}]\big) \longrightarrow 0. \tag{C.13}$$

The event $\mathcal{H}_J$ is determined by $\mathcal{S}_J$. On this event, $B \asymp \bar{B} \to \infty$, and, conditional on $\mathcal{S}_J$, Assumption **(D2)** makes $A_J$ uniform over the nonempty class of simple bipartite graphs with the prescribed degrees. We apply this uniform law below only on $\mathcal{H}_J$.

We introduce some notation. For nonnegative integers $x$ and $r$, write

$$(x)_r := x(x-1)\cdots(x-r+1), \qquad (x)_0 := 1.$$

The next lemma approximates the probability that a particular pattern of annotator–document assignments occurs under our allocation scheme. This is directly relevant to $N_{jk}$: a given annotator evaluating both documents $j$ and $k$ forms a two-edge pattern, and $N_{jk}$ counts the annotators for whom this pattern occurs. For patterns involving a bounded number of assignments, we show that their probabilities under our allocation scheme can be approximated by those obtained by uniformly matching document slots to annotator slots without imposing simplicity.

**Lemma 7** (Local patterns in a uniform fixed-degree graph)**.** *Consider a sequence of nonempty classes of simple bipartite graphs with left degrees $\{m_j\}$, right degrees $\{k_i\}$, and common degree sum*

$$B = \sum_j m_j = \sum_i k_i \longrightarrow \infty.$$

*Suppose that the maximum degree on both sides is uniformly bounded. For each class, let $H$ be a specified simple bipartite pattern on its labelled vertex classes, with $|E(H)|$ uniformly bounded,*

*and let $x_j(H)$ and $y_i(H)$ denote its degrees on the two sides. If $\mathcal{A}$ is uniform over the corresponding graph class, then*

$$\mathbb{P}(H \subseteq \mathcal{A}) = \frac{\prod_j (m_j)_{x_j(H)} \prod_i (k_i)_{y_i(H)}}{(B)_{|E(H)|}} \{1 + O(B^{-1})\}.$$

*The remainder is uniform over the degree sequences and patterns satisfying these bounds.*

*Proof.* If a degree of $H$ exceeds the corresponding prescribed degree, the claim is immediate. Otherwise, let $D$ be a common bound on the prescribed degrees. Each vertex class then contains at least $B/D$ positive-degree vertices, so both classes diverge as required by McKay's asymptotic convention. Apply McKay (2010, Theorem 2.2) with

$$s_j = m_j, \quad t_i = k_i, \quad x_j = x_j(H), \quad y_i = y_i(H), \quad E = B, \quad X = |E(H)|.$$

Writing $s_{\max}, t_{\max}, x_{\max}, y_{\max}$ for the corresponding maximum degrees, McKay's condition is $\Delta' X = o(E)$, and his relative error is $O(\Delta' X/E)$, where

$$\Delta' = (s_{\max} + t_{\max})(s_{\max} + t_{\max} + x_{\max} + y_{\max}).$$

Under the stated bounds, $\Delta' X = O(1) = o(B)$. Since $E = B$, McKay's formula gives the stated $O(B^{-1})$ remainder, uniformly under these bounds. □

We first show that the annotator reuse part of the variance concentrates around its conditional mean given the document sample. Thus, for a given sample, variation in the measurement variance across annotator allocations becomes asymptotically negligible.

The bounded-response-scale assumption allows us, after enlarging the bound if necessary, to fix $C_Y < \infty$, uniformly in the array indices, such that

$$|Y_{ij}^* - \mu_j| \le C_Y, \qquad \nu_j \le C_Y, \qquad |\tau_{jk}| \le C_Y.$$

For $j \neq k$, define

$$\beta_{jk}^{\mathrm{ov}} := \frac{\bar{B}}{J^2} \frac{a_j a_k \tau_{jk}}{\pi_j \pi_k m_j m_k} = \frac{a_j a_k \tau_{jk}}{\lambda_j \lambda_k \bar{B}}. \tag{C.14}$$

**Lemma 8** (Pairwise mean and concentration of the weighted overlap)**.** *Suppose the conditions of Theorem 2 hold. On $\mathcal{H}_J$, uniformly over distinct selected documents $j \neq k$,*

$$\mathbb{E}[N_{jk} \,|\, \mathcal{S}_J] = (K-1)\frac{m_j m_k}{B} + O\Big(\frac{m_j m_k}{B^2}\Big). \tag{C.15}$$

*Define the centered allocation remainder*

$$R_J^{\mathrm{alloc}}(a) := \sum_{j \neq k} I_j I_k \beta_{jk}^{\mathrm{ov}} \left\{ N_{jk} - \mathbb{E}[N_{jk} \,|\, \mathcal{S}_J] \right\}, \tag{C.16}$$

*where the sum is over ordered document pairs. Then*

$$R_J^{\mathrm{alloc}}(a) \xrightarrow{p} 0.$$

*Proof.* We first prove the mean statement. Since $B \geq \bar{B}/2 \to \infty$ uniformly over realizations in $\mathcal{H}_J$, the McKay expansion applies to every such realization for all sufficiently large $J$. Fix one such realization. For distinct selected documents $j \neq k$ and an active annotator $i$, the event $\{i \in M_j \cap M_k\}$ is the containment event for the two-edge pattern $\{(j,i),(k,i)\}$. Lemma 7 therefore gives

$$\mathbb{P}(i \in M_j \cap M_k \,|\, \mathcal{S}_J) = \frac{m_j m_k \, k_i(k_i-1)}{B(B-1)}\{1 + O(B^{-1})\},$$

uniformly over $i, j, k$ and the stated realizations. If $k_i < 2$, both sides are zero.

On $\mathcal{H}_J$, the annotator degrees are fixed conditional on $\mathcal{S}_J$. Capacity filling assigns load $K$ to all but the last annotator, whose load is $r_B \in \{1, \ldots, K\}$. Hence

$$\sum_{i=1}^{n} k_i(k_i-1) = (K-1)B - r_B(K-r_B) = (K-1)B + O(1),$$

where the remainder is uniformly bounded because $K$ is fixed.

Since $N_{jk} = \sum_{i=1}^{n} \mathbf{1}\{i \in M_j \cap M_k\}$, summing the uniform error bounds gives

$$\begin{aligned}
\mathbb{E}[N_{jk} \,|\, \mathcal{S}_J] &= \frac{m_j m_k}{B(B-1)} \sum_{i=1}^{n} k_i(k_i-1) + O\left( \frac{m_j m_k}{B^2(B-1)} \sum_{i=1}^{n} k_i(k_i-1) \right) \\
&= (K-1)\frac{m_j m_k}{B-1} + O\Big(\frac{m_j m_k}{B^2}\Big) \\
&= (K-1)\frac{m_j m_k}{B} + O\Big(\frac{m_j m_k}{B^2}\Big).
\end{aligned}$$

The second equality uses the capacity-filling identity above and $\sum_i k_i(k_i-1) \leq (K-1)B$. The last equality uses $(B-1)^{-1} = B^{-1} + O(B^{-2})$. All remainder bounds are uniform over distinct selected documents $j \neq k$ and sampling realizations in $\mathcal{H}_J$, by Lemma 7, the fixed bound $K$, and $B \geq \bar{B}/2 \to \infty$. This proves (C.15).

We next establish concentration around the conditional mean. We first rewrite the annotator-reuse component of the variance from a sum over document pairs to a sum of annotator contributions. We then show that these contributions are small and weakly dependent. Concentration around the conditional mean then follows from an LLN-type argument.

Let

$$M_J^{\mathrm{ov}} := \max_{j \neq k} |\beta_{jk}^{\mathrm{ov}}|.$$

The bounds $\lambda_j \geq c_\lambda$ and $|a_j| \leq \bar{a}$ give

$$M_J^{\mathrm{ov}} \leq \frac{\bar{a}^2 C_Y}{c_\lambda^2 \bar{B}} = O(\bar{B}^{-1}). \tag{C.17}$$

Here and below, $C_\star < \infty$ may change from line to line but depends only on the fixed bounds.

All sums over $j \neq k$ below are over ordered pairs. For each active annotator $i$, define

$$U_i := \sum_{\substack{j,k \in \mathcal{J}_i \\ j \neq k}} \beta_{jk}^{\mathrm{ov}}.$$

Then $|U_i| \leq K(K-1)M_J^{\mathrm{ov}}$. This establishes that each annotator contribution is small.

Counting each annotator shared by an ordered document pair once gives

$$R_J^{\mathrm{alloc}}(a) = \sum_{i=1}^{n} \{U_i - \mathbb{E}[U_i \mid \mathcal{S}_J]\}. \tag{C.18}$$

We next establish weak dependence between the document sets assigned to two annotators, which we refer to as annotator neighborhoods. On $\mathcal{H}_J$, condition on $\mathcal{S}_J$. All patterns below have at most $2K$ edges and degrees bounded in terms of $(\bar{m}, K)$, so the remainder in Lemma 7 is uniform over all annotators, neighborhoods, and selection realizations considered below. For $A \subseteq \mathcal{S}_J$ with $|A| = k_i$, containment of the full star from annotator $i$ to $A$ is exactly the event $\{\mathcal{J}_i = A\}$, because the star exhausts annotator $i$'s prescribed degree. Hence

$$p_i(A) := \mathbb{P}(\mathcal{J}_i = A \mid \mathcal{S}_J) = \frac{k_i! \prod_{j \in A} m_j}{(B)_{k_i}} \{1 + O(B^{-1})\}. \tag{C.19}$$

To bound the conditional covariance between annotator contributions, we compare the joint distribution of their assigned document sets with a distribution under which the two sets are independent, while preserving each annotator's assignment probabilities given the document sample. Since each contribution is a bounded function of its assigned document set, a bound on the total variation distance between these distributions gives a bound on the conditional covariance.

For distinct active annotators $i \neq \ell$, let $P_{i\ell}$ be the conditional joint law of $(\mathcal{J}_i, \mathcal{J}_\ell)$, and let $Q_{i\ell}$ be the product of its exact conditional marginals, so that $Q_{i\ell}(A, C) = p_i(A)p_\ell(C)$. If $|A| = k_i$, $|C| = k_\ell$, and $A \cap C = \varnothing$, the union of the two full stars exhausts both annotator degrees. Lemma 7 therefore gives

$$P_{i\ell}(A, C) = \frac{k_i! k_\ell! \prod_{j \in A \cup C} m_j}{(B)_{k_i + k_\ell}} \{1 + O(B^{-1})\}.$$

If $Q_{i\ell}(A, C) = 0$, then $P_{i\ell}(A, C) = 0$ as well, because the joint event is contained in each marginal event. Otherwise, dividing by the two marginal formulas in (C.19), and using disjointness to cancel the document factors, gives

$$\begin{aligned}
\frac{P_{i\ell}(A, C)}{Q_{i\ell}(A, C)} &= \frac{(B)_{k_i}(B)_{k_\ell}}{(B)_{k_i + k_\ell}} \frac{1 + O(B^{-1})}{\{1 + O(B^{-1})\}\{1 + O(B^{-1})\}} \\
&= \prod_{r=0}^{k_\ell - 1} \left(1 + \frac{k_i}{B - k_i - r}\right) \{1 + O(B^{-1})\} \\
&= 1 + O(B^{-1}).
\end{aligned}$$

The last equality uses $k_i, k_\ell \le K$, with $K$ fixed.

Consequently,

$$|P_{i\ell}(A,C) - Q_{i\ell}(A,C)| \le \frac{C_\star}{B} Q_{i\ell}(A,C), \qquad A \cap C = \varnothing. \tag{C.20}$$

Let $D := \{(A,C) : A \cap C = \varnothing\}$. For each $j \in \mathcal{S}_J$, the one- and two-edge cases of Lemma 7 give

$$\mathbb{P}(j \in \mathcal{J}_i \mid \mathcal{S}_J) = \frac{k_i m_j}{B}\{1 + O(B^{-1})\},$$
$$\mathbb{P}(j \in \mathcal{J}_i \cap \mathcal{J}_\ell \mid \mathcal{S}_J) = \frac{k_i k_\ell (m_j)_2}{(B)_2}\{1 + O(B^{-1})\}.$$

Under $Q_{i\ell}$, the two assigned document sets are independent. A union bound and $k_i, k_\ell \le K$ therefore give

$$\begin{aligned} Q_{i\ell}(D^c) &\le \sum_{j \in \mathcal{S}_J} \mathbb{P}(j \in \mathcal{J}_i \mid \mathcal{S}_J)\mathbb{P}(j \in \mathcal{J}_\ell \mid \mathcal{S}_J) \\ &\le \frac{C_\star}{B^2} \sum_{j \in \mathcal{S}_J} m_j^2. \end{aligned}$$

Under $P_{i\ell}$, the corresponding union bound gives

$$\begin{aligned} P_{i\ell}(D^c) &\le \sum_{j \in \mathcal{S}_J} \mathbb{P}(j \in \mathcal{J}_i \cap \mathcal{J}_\ell \mid \mathcal{S}_J) \\ &\le \frac{C_\star}{B^2} \sum_{j \in \mathcal{S}_J} (m_j)_2. \end{aligned}$$

Since $m_j \le \bar{m}$ and $\sum_{j \in \mathcal{S}_J} m_j = B$, it follows that

$$Q_{i\ell}(D^c) + P_{i\ell}(D^c) \le \frac{C_\star}{B^2} \sum_{j \in \mathcal{S}_J} \{m_j^2 + (m_j)_2\} \le \frac{C_\star(2\bar{m} - 1)}{B} = O(B^{-1}).$$

Combining this bound with (C.20) yields

$$\begin{aligned} 2\|P_{i\ell} - Q_{i\ell}\|_{\mathrm{TV}} &= \sum_{(A,C) \in D} |P_{i\ell}(A,C) - Q_{i\ell}(A,C)| \\ &\quad + \sum_{(A,C) \in D^c} |P_{i\ell}(A,C) - Q_{i\ell}(A,C)| \\ &\le \frac{C_\star}{B} Q_{i\ell}(D) + P_{i\ell}(D^c) + Q_{i\ell}(D^c) = O(B^{-1}). \end{aligned} \tag{C.21}$$

The bound is uniform over $i \ne \ell$ and the selection realizations in $\mathcal{H}_J$.

Because $Q_{i\ell}$ is the product of the exact marginals of $P_{i\ell}$,

$$\mathrm{Cov}(U_i, U_\ell \mid \mathcal{S}_J) = \mathbb{E}_{P_{i\ell}}[U_i U_\ell] - \mathbb{E}_{Q_{i\ell}}[U_i U_\ell].$$

The standard inequality

$$|\mathbb{E}_P h - \mathbb{E}_Q h| \le 2\|h\|_\infty \|P - Q\|_{\mathrm{TV}},$$

together with $|U_i| \le K(K-1)M_J^{\mathrm{ov}}$, therefore gives

$$|\operatorname{Cov}(U_i, U_\ell \mid \mathcal{S}_J)| \le \frac{C_\star (M_J^{\mathrm{ov}})^2}{B}, \qquad \operatorname{Var}(U_i \mid \mathcal{S}_J) \le C_\star (M_J^{\mathrm{ov}})^2.$$

The sums in (C.18) run only over active annotators, so $n \le \sum_i k_i = B$. It follows that, on $\mathcal{H}_J$,

$$\begin{aligned}\operatorname{Var}(R_J^{\mathrm{alloc}}(a) \mid \mathcal{S}_J) &\le C_\star \left\{ n(M_J^{\mathrm{ov}})^2 + \frac{n^2 (M_J^{\mathrm{ov}})^2}{B} \right\} \\ &\le C_\star B (M_J^{\mathrm{ov}})^2 \le \frac{C_\star}{\bar{B}}, \end{aligned} \tag{C.22}$$

where the last inequality uses (C.17) and $B \le 2\bar{B}$.

Finally, $R_J^{\mathrm{alloc}}(a)$ has conditional mean zero given $\mathcal{S}_J$. Since $\mathcal{H}_J \in \sigma(\mathcal{S}_J)$, conditional Chebyshev's inequality gives, for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(|R_J^{\mathrm{alloc}}(a)| > \epsilon\Big) &\le \mathbb{P}(\mathcal{H}_J^c) + \mathbb{E}\Big[\mathbf{1}_{\mathcal{H}_J} \mathbb{P}\Big(|R_J^{\mathrm{alloc}}(a)| > \epsilon \,\Big|\, \mathcal{S}_J\Big)\Big] \\ &\le \mathbb{P}(\mathcal{H}_J^c) + \frac{C_\star}{\epsilon^2 \bar{B}} \longrightarrow 0. \end{aligned}$$

□

*Proof of Theorem 2.* Recall from (C.13) that $\mathbb{P}(\mathcal{H}_J^c) \to 0$. Throughout the proof, $C_0 < \infty$ denotes a constant depending only on the fixed bounds and may change from line to line.

By Lemma 3 and the definitions of $\lambda_j$ and $\beta_{jk}^{\mathrm{ov}}$, we can decompose the variance into a diagonal term $D_J(a)$ and an annotator-reuse term $C_J(a)$

$$v_{\varepsilon,J}(a) = D_J(a) + C_J(a),$$

where

$$D_J(a) := \frac{1}{J} \sum_{j=1}^{J} \frac{I_j}{\pi_j} \frac{a_j^2 \nu_j}{\lambda_j}, \qquad C_J(a) := \sum_{j \ne k} \beta_{jk}^{\mathrm{ov}} I_j I_k N_{jk}.$$

The diagonal term depends only on the document sample and is unaffected by annotator allocation. We establish its concentration around its mean using the usual Horvitz–Thompson variance formula under Poisson sampling. By Assumption **(D1)**, the selection indicators are independent, so

$$\mathbb{E}[D_J(a)] = \frac{1}{J} \sum_{j=1}^{J} \frac{a_j^2 \nu_j}{\lambda_j} \longrightarrow v_{\mathrm{diag}}(a; \pi, m),$$

and

$$\operatorname{Var}(D_J(a)) = \frac{1}{J^2} \sum_{j=1}^{J} \frac{1 - \pi_j}{\pi_j} \left( \frac{a_j^2 \nu_j}{\lambda_j} \right)^2 \le \frac{C_0}{\bar{B}},$$

where we used $1/\pi_j = J m_j / (\lambda_j \bar{B}) \le J \bar{m} / (c_\lambda \bar{B})$. Consequently,

$$D_J(a) = v_{\mathrm{diag}}(a; \pi, m) + o_p(1).$$

For the Annotator reuse, or the off-diagonal term, add and subtract the conditional mean of $N_{jk}$ given $\mathcal{S}_J$ to isolate fluctuation from allocation:

$$C_J(a) = R_J^{\text{alloc}}(a) + \sum_{j \neq k} \beta_{jk}^{\text{ov}} I_j I_k \mathbb{E}[N_{jk} \mid \mathcal{S}_J].$$

Lemma 8 gives $R_J^{\text{alloc}}(a) = o_p(1)$. On $\mathcal{H}_J$, for distinct selected documents $j, k \in \mathcal{S}_J$, define

$$\delta_{jk} := \mathbb{E}[N_{jk} \mid \mathcal{S}_J] - (K-1)\frac{m_j m_k}{B}.$$

By (C.15),

$$|\delta_{jk}| \leq C_0 \frac{m_j m_k}{B^2},$$

uniformly over these document pairs and sampling realizations in $\mathcal{H}_J$.

Write $b_{jk} := a_j a_k \tau_{jk}$ and define

$$U_J(a) := \frac{1}{J^2} \sum_{j \neq k} \frac{I_j I_k}{\pi_j \pi_k} b_{jk}.$$

The array $\{b_{jk}\}$ is symmetric and uniformly bounded. Define the remainder from approximating the conditional overlaps by

$$R_J^{\text{mean}}(a) := \begin{cases} \displaystyle\sum_{\substack{j,k \in \mathcal{S}_J \\ j \neq k}} \beta_{jk}^{\text{ov}} \delta_{jk}, & \text{on } \mathcal{H}_J, \\ 0, & \text{on } \mathcal{H}_J^c. \end{cases}$$

Substituting the conditional-overlap approximation into the decomposition of $C_J(a)$ gives, on $\mathcal{H}_J$,

$$C_J(a) = R_J^{\text{alloc}}(a) + (K-1)\frac{\bar{B}}{B} U_J(a) + R_J^{\text{mean}}(a).$$

To bound the remainder, recall $M_J^{\text{ov}} = \max_{j \neq k} |\beta_{jk}^{\text{ov}}|$. Since $\sum_{j \in \mathcal{S}_J} m_j = B$, on $\mathcal{H}_J$,

$$\begin{aligned} |R_J^{\text{mean}}(a)| &\leq \frac{C_0 M_J^{\text{ov}}}{B^2} \sum_{\substack{j,k \in \mathcal{S}_J \\ j \neq k}} m_j m_k \\ &\leq C_0 M_J^{\text{ov}} \leq \frac{C_0}{\bar{B}}, \end{aligned}$$

where the last inequality follows from (C.17). Because $R_J^{\text{mean}}(a) = 0$ outside $\mathcal{H}_J$, it follows that

$$R_J^{\text{mean}}(a) = O_p(\bar{B}^{-1}) = o_p(1).$$

It remains to control the document-selection fluctuation in $U_J(a)$. Define

$$T_J(a) := \mathbb{E}[U_J(a)] = \frac{1}{J^2} \sum_{j \neq k} b_{jk}, \qquad \xi_j := \frac{I_j}{\pi_j} - 1, \qquad \bar{b}_j := \frac{1}{J} \sum_{k \neq j} b_{jk}.$$

Here $\xi_j$ is the centered selection weight of document $j$, and $\bar{b}_j$ summarizes its pairwise contributions across the corpus. Expanding $(1+\xi_j)(1+\xi_k)-1 = \xi_j+\xi_k+\xi_j\xi_k$, and using symmetry and ordered-pair summation, gives

$$U_J(a) - T_J(a) = \frac{2}{J}\sum_{j=1}^{J} \bar{b}_j \xi_j + \frac{1}{J^2}\sum_{j\neq k} b_{jk}\xi_j\xi_k. \tag{C.23}$$

This is the Hoeffding decomposition with respect to the independent selection indicators. As for a second-order U-statistic, the first term collects the single-document projections: $\mathbb{E}[U_J(a) \mid I_j] - T_J(a) = (2/J)\bar{b}_j\xi_j$. The second term captures the remaining pairwise interactions. It is degenerate: its conditional mean given any individual $I_j$ is zero, because every product retains an independent centered factor.

The $\xi_j$ are independent and centered, with $\mathbb{E}[\xi_j^2] = (1-\pi_j)/\pi_j \leq C_0 J/\bar{B}$. Since the $\bar{b}_j$ are uniformly bounded, the linear term satisfies

$$\mathbb{E}\left[\left(\frac{2}{J}\sum_{j=1}^{J}\bar{b}_j\xi_j\right)^2\right] = \frac{4}{J^2}\sum_{j=1}^{J}\bar{b}_j^2\mathbb{E}[\xi_j^2] \leq \frac{C_0}{\bar{B}}.$$

For the quadratic term, independence and centering eliminate every cross-product containing an index that appears only once. In particular, two pairs sharing just one document contribute zero; only identical or reversed pairs remain. Symmetry therefore gives

$$\mathbb{E}\left[\left(\frac{1}{J^2}\sum_{j\neq k} b_{jk}\xi_j\xi_k\right)^2\right] = \frac{2}{J^4}\sum_{j\neq k} b_{jk}^2\mathbb{E}[\xi_j^2]\mathbb{E}[\xi_k^2] \leq \frac{C_0}{\bar{B}^2}.$$

Thus the linear and quadratic terms are $O_p(\bar{B}^{-1/2})$ and $O_p(\bar{B}^{-1})$, respectively. Both vanish, so $U_J(a) - T_J(a) = o_p(1)$. The maintained array limit $T_J(a) \to \tau(a)$ then yields

$$U_J(a) = \tau(a) + o_p(1) = O_p(1).$$

Finally, Lemma 5 gives $B/\bar{B} \xrightarrow{p} 1$. On $\mathcal{H}_J$, we have $\bar{B}/B \leq 2$ and $|\bar{B}/B - 1| \leq 2|B/\bar{B} - 1|$. The preceding decomposition of $C_J(a)$ therefore implies, on this event,

$$\begin{aligned}|C_J(a) - (K-1)\tau(a)| \leq{}& |R_J^{\text{alloc}}(a)| + |R_J^{\text{mean}}(a)| \\ &+ 2(K-1)\left\{|U_J(a)-\tau(a)| + |\tau(a)|\left|\frac{B}{\bar{B}} - 1\right|\right\}.\end{aligned}$$

Every term on the right converges to zero in probability. Since $\mathbb{P}(\mathcal{H}_J^c) \to 0$, it follows that, for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}(|C_J(a) - (K-1)\tau(a)| > \epsilon) \leq{}& \mathbb{P}(\mathcal{H}_J^c) \\ &+ \mathbb{P}(\mathcal{H}_J \cap \{|C_J(a) - (K-1)\tau(a)| > \epsilon\}) \longrightarrow 0.\end{aligned}$$

Together with the diagonal convergence, this proves

$$v_{\varepsilon,J}(a) \xrightarrow{p} v_{\mathrm{diag}}(a;\pi,m) + (K-1)\tau(a).$$

Thus Condition **(R4)** holds with $v_\varepsilon(a) = v_{\mathrm{diag}}(a;\pi,m)+(K-1)\tau(a)$, completing the proof. □

# D Variance estimation and feasible inference

This section proves the unbiasedness and asymptotic consistency of the variance estimators introduced in the main text.

## D.1 Unbiasedness

For notational convenience, set $\widehat{r}_j^2 = 0$ when $I_j = 0$.

**Lemma 9** (Unbiasedness of the variance estimators)**.** *Suppose Assumptions **(M1)**–**(M3)** and **(D3)** hold, the allocation is simple, and $m_j \geq 2$ whenever $\pi_j > 0$. Then, for every selected document $j$,*

$$\mathbb{E}[\widehat{\sigma}_j^2 \mid \mathcal{G}_J] = \nu_j, \qquad \mathbb{E}[\widehat{r}_j^2 \mid \mathcal{G}_J] = r_j^2,$$

*and, for every pair of distinct selected documents,*

$$\mathbb{E}[\widehat{\Gamma}_{jk} \mid \mathcal{G}_J] = N_{jk}\tau_{jk}.$$

*Consequently,*

$$\mathbb{E}[\widehat{v}_{\varepsilon,J}(a) \mid \mathcal{G}_J] = v_{\varepsilon,J}(a). \tag{D.1}$$

*If Assumption **(D1)** also holds, then*

$$\mathbb{E}[\widehat{v}_{r,J}(a)] = v_{r,J}(a), \tag{D.2}$$

*and hence*

$$\mathbb{E}[\widehat{v}_J(a)] = v_{r,J}(a) + \bar{v}_{\varepsilon,J}(a) = \bar{B}\ \mathrm{Var}\Big(\widehat{\Phi}_J(a) - \Phi_J(a)\Big). \tag{D.3}$$

*Proof.* Condition on $\mathcal{G}_J$, and write $U_{ij} := Y_{ij}^* - \mu_j$. By (B.4) and the maintained response model, response vectors belonging to distinct annotators are conditionally independent, and, for $j \neq k$, the relevant conditional moments satisfy

$$\mathbb{E}[U_{ij} \mid \mathcal{G}_J] = 0, \qquad \mathbb{E}[U_{ij}^2 \mid \mathcal{G}_J] = \nu_j, \qquad \mathbb{E}[U_{ij}U_{ik} \mid \mathcal{G}_J] = \tau_{jk}.$$

For each selected document $j$, graph simplicity ensures that its $m_j \geq 2$ responses come from distinct annotators. Conditional on $\mathcal{G}_J$, these responses are independent with common mean $\mu_j$ and variance $\nu_j$. Hence the usual sample-variance identity gives

$$\mathbb{E}[\widehat{\sigma}_j^2 \mid \mathcal{G}_J] = \nu_j.$$

Moreover,

$$\mathbb{E}[(\bar{Y}_j - g_j)^2 \mid \mathcal{G}_J] = \mathrm{Var}(\bar{Y}_j \mid \mathcal{G}_J) + (\mu_j - g_j)^2 = \frac{\nu_j}{m_j} + r_j^2.$$

Since $\widehat{r}_j^2 = (\bar{Y}_j - g_j)^2 - \widehat{\sigma}_j^2/m_j$, these two identities give $\mathbb{E}[\widehat{r}_j^2 \mid \mathcal{G}_J] = r_j^2$.

For the overlap identity, fix distinct selected documents $j, k$, and put $N := N_{jk}$. The claim is immediate if $N = 0$. If $N \geq 1$, fix $i \in M_j \cap M_k$, and define

$$\bar{U}_{j,-i} := \frac{1}{m_j - 1} \sum_{\ell \in M_j \setminus \{i\}} U_{\ell j}, \qquad \bar{U}_{k,-i} := \frac{1}{m_k - 1} \sum_{q \in M_k \setminus \{i\}} U_{qk}.$$

Since $Y_{ij}^* - \bar{Y}_{j,-i} = U_{ij} - \bar{U}_{j,-i}$ (and analogously for document $k$), the two mixed terms have conditional expectation zero: annotator $i$'s responses are independent of the centered responses entering either leave-out average. The two leave-out averages can nevertheless be correlated. In their product, only terms involving the same annotator can contribute to the conditional expectation. There are exactly $N - 1$ such terms, one for each remaining shared annotator, each with conditional expectation $\tau_{jk}$. Therefore

$$\begin{aligned}
&\mathbb{E}\big[(Y_{ij}^* - \bar{Y}_{j,-i})(Y_{ik}^* - \bar{Y}_{k,-i}) \,\big|\, \mathcal{G}_J\big] \\
&\quad = \mathbb{E}[U_{ij}U_{ik} \mid \mathcal{G}_J] + \mathbb{E}[\bar{U}_{j,-i}\bar{U}_{k,-i} \mid \mathcal{G}_J] \\
&\qquad - \underbrace{\mathbb{E}[U_{ij}\bar{U}_{k,-i} \mid \mathcal{G}_J]}_{=0} - \underbrace{\mathbb{E}[\bar{U}_{j,-i}U_{ik} \mid \mathcal{G}_J]}_{=0} \\
&\quad = \tau_{jk} + \frac{N-1}{(m_j - 1)(m_k - 1)}\tau_{jk} \\
&\quad = \left\{1 + \frac{N-1}{(m_j - 1)(m_k - 1)}\right\}\tau_{jk}.
\end{aligned}$$

Summing over the $N$ shared annotators and applying the definition of $\kappa_{jk}^{\mathrm{loo}}$ gives

$$\mathbb{E}[\widehat{\Gamma}_{jk} \mid \mathcal{G}_J] = \kappa_{jk}^{\mathrm{loo}} N \left\{1 + \frac{N-1}{(m_j - 1)(m_k - 1)}\right\}\tau_{jk} = N\tau_{jk}.$$

Conditional on $\mathcal{G}_J$, all coefficients and overlap counts in $\widehat{v}_{\varepsilon,J}(a)$ are fixed. Applying the local identities term by term in (4.16) therefore proves (D.1).

Now suppose Assumption **(D1)** also holds. Since $I_j$ is $\mathcal{G}_J$-measurable, the local identity and the convention for unselected documents give

$$\mathbb{E}[I_j\widehat{r}_j^2 \mid \mathcal{G}_J] = I_j r_j^2, \qquad \mathbb{E}[I_j\widehat{r}_j^2] = \pi_j r_j^2.$$

Substitution into (4.15) proves (D.2). Taking expectations in (D.1), adding the two components, and applying Lemma 4 proves (D.3). □

## D.2 Concentration

Recall that

$$\lambda_j := \frac{J\pi_j m_j}{\bar{B}}$$

compares document $j$'s expected label expenditure with $\bar{B}/J$, the expected expenditure under an equal division of the expected annotation budget across documents.

**(R5) Feasible-inference regularity.** The expected budget satisfies $\bar{B} \to \infty$, and, for fixed finite constants $\bar{m}$, $\bar{a}$, and $\bar{r}$,

$$2 \le m_j \le \bar{m}, \qquad |a_j| \le \bar{a}, \qquad |r_j| \le \bar{r}$$

uniformly over $j$ and $J$. There is a deterministic sequence $c_J \in (0, 1]$ such that

$$\min_{j \le J} \lambda_j \ge c_J, \qquad c_J^3 \bar{B} \longrightarrow \infty.$$

The condition permits the minimum relative expenditure to approach zero. Under the balanced-spending condition in Theorem 2, $c_J$ can be fixed, so the rate reduces to $\bar{B} \to \infty$. Bounded outcomes and Assumption **(R1)** imply the stated residual bound.

**Proposition 4** (Concentration of the variance estimators)**.** *Suppose the bounded-response-scale assumption and Assumptions **(M1)**–**(M3)**, **(D1)**, and **(D3)** hold. Suppose the allocation is simple and assigns at most $K < \infty$ documents to each annotator. Under **(R5)**,*

$$\widehat{v}_{r,J}(a) - v_{r,J}(a) \xrightarrow{p} 0, \qquad \widehat{v}_{\varepsilon,J}(a) - v_{\varepsilon,J}(a) \xrightarrow{p} 0.$$

*Proof.* Throughout the proof, $C_0 < \infty$ denotes a constant depending only on the fixed bounds and may change from line to line. Let $R_Y < \infty$ be the width of the maintained outcome interval. Then $|Y_{ij}^* - g_j| \le R_Y + \bar{r}$.

*Sampling component.* Introduce the infeasible estimator where we suppose we can observe $r_j$

$$\widehat{v}_{r,J}^{\mathrm{or}}(a) := \frac{\bar{B}}{J^2} \sum_{j=1}^{J} \frac{I_j(1-\pi_j)}{\pi_j^2} a_j^2 r_j^2.$$

We first establish that the infeasible estimator is consistent.

Independent Bernoulli selection gives $\mathbb{E}[\widehat{v}_{r,J}^{\mathrm{or}}(a)] = v_{r,J}(a)$ and

$$\mathrm{Var}(\widehat{v}_{r,J}^{\mathrm{or}}(a)) = \frac{\bar{B}^2}{J^4} \sum_{j=1}^{J} \frac{(1-\pi_j)^3}{\pi_j^3} a_j^4 r_j^4 \le \frac{C_0}{c_J^3 \bar{B}} \longrightarrow 0, \tag{D.4}$$

where we used $\pi_j = \lambda_j \bar{B}/(J m_j)$.

Now, we replace $r_j^2$ by $\widehat{r}_j^2$ and establish that our estimator converges to the infeasible estimator

in probability. Write

$$R_J^{\mathrm{res}}(a) := \widehat{v}_{r,J}(a) - \widehat{v}_{r,J}^{\mathrm{or}}(a) = \sum_{j=1}^{J} \omega_j^{\mathrm{res}} I_j(\widehat{r}_j^2 - r_j^2), \qquad \omega_j^{\mathrm{res}} := \frac{\bar{B}}{J^2} \frac{1-\pi_j}{\pi_j^2} a_j^2.$$

Lemma 9 gives

$$\mathbb{E}[R_J^{\mathrm{res}}(a) \mid \mathcal{G}_J] = 0. \tag{D.5}$$

For the replacement bound, expand the corrected residual as

$$\widehat{r}_j^2 = \frac{1}{m_j(m_j-1)} \sum_{\substack{i,\ell \in M_j \\ i \neq \ell}} (Y_{ij}^* - g_j)(Y_{\ell j}^* - g_j). \tag{D.6}$$

For an active annotator $i$, let $\mathbf{Y}_i := (Y_{ij}^* : j \in \mathcal{J}_i)$. Conditional on $\mathcal{G}_J$, these annotator response vectors are independent by (B.4).

Let $R_J^{\mathrm{res},(i)}(a)$ denote the value of $R_J^{\mathrm{res}}(a)$ obtained by replacing annotator $i$'s responses with those of a new annotator drawn independently from the reference population, evaluating the same assigned documents with fresh response shocks. The graph and all other annotators' responses are held fixed. Under the conditional law, this replaces $\mathbf{Y}_i$ by an independent copy. We use this construction to apply the Efron–Stein inequality (Boucheron et al., 2013, Theorem 3.1) below.

Because $|Y_{ij}^* - g_j| \leq R_Y + \bar{r}$, replacing one label on document $j$ changes only $2(m_j - 1)$ products in (D.6), each by at most $2(R_Y + \bar{r})^2$. Hence

$$|R_J^{\mathrm{res}}(a) - R_J^{\mathrm{res},(i)}(a)| \leq C_0 \sum_{j \in \mathcal{J}_i} \frac{\omega_j^{\mathrm{res}}}{m_j}.$$

Applying the Efron–Stein inequality (Boucheron et al., 2013, Theorem 3.1) conditionally on $\mathcal{G}_J$, and then using the preceding replacement bound, gives

$$\begin{aligned}
\mathrm{Var}(R_J^{\mathrm{res}}(a) \mid \mathcal{G}_J) &\leq \frac{1}{2} \sum_i \mathbb{E}\Big[\{R_J^{\mathrm{res}}(a) - R_J^{\mathrm{res},(i)}(a)\}^2 \,\Big|\, \mathcal{G}_J\Big] \\
&\leq C_0 \sum_i \left( \sum_{j \in \mathcal{J}_i} \frac{\omega_j^{\mathrm{res}}}{m_j} \right)^2 \\
&\leq C_0 \sum_i \sum_{j \in \mathcal{J}_i} \frac{(\omega_j^{\mathrm{res}})^2}{m_j^2} \\
&= C_0 \sum_{j \in \mathcal{S}_J} \frac{(\omega_j^{\mathrm{res}})^2}{m_j}.
\end{aligned} \tag{D.7}$$

The third inequality uses Cauchy–Schwarz and $|\mathcal{J}_i| \leq K$, with the fixed factor $K$ absorbed into $C_0$. The last equality counts each selected document in exactly $m_j$ annotator neighborhoods.

Taking expectations in (D.7) and using the conditional centering in (D.5) gives

$$\begin{aligned}
\mathbb{E}[R_J^{\mathrm{res}}(a)^2] &= \mathbb{E}[\mathrm{Var}(R_J^{\mathrm{res}}(a) \mid \mathcal{G}_J)] \\
&\le C_0\, \mathbb{E}\left[\sum_{j=1}^{J} I_j \frac{(\omega_j^{\mathrm{res}})^2}{m_j}\right] \\
&= C_0 \sum_{j=1}^{J} \pi_j \frac{(\omega_j^{\mathrm{res}})^2}{m_j} \\
&= C_0 \frac{\bar{B}^2}{J^4} \sum_{j=1}^{J} \frac{(1-\pi_j)^2 a_j^4}{\pi_j^3 m_j} \\
&= \frac{C_0}{J\bar{B}} \sum_{j=1}^{J} \frac{(1-\pi_j)^2 a_j^4 m_j^2}{\lambda_j^3} \\
&\le \frac{C_0}{c_J^3 \bar{B}} \longrightarrow 0. \qquad \text{(D.8)}
\end{aligned}$$

Here we used $\mathbb{E}[I_j] = \pi_j$, the definition of $\omega_j^{\mathrm{res}}$, and $\pi_j = \lambda_j \bar{B}/(Jm_j)$. The final inequality uses $\lambda_j \ge c_J$, the boundedness of $a_j$ and $m_j$, and $(1-\pi_j)^2 \le 1$, with fixed constants absorbed into $C_0$. The limit follows from $c_J^3 \bar{B} \to \infty$.

Thus $R_J^{\mathrm{res}}(a) = o_p(1)$. Combining this with (D.4) gives

$$\widehat{v}_{r,J}(a) - v_{r,J}(a) = R_J^{\mathrm{res}}(a) + \left\{\widehat{v}_{r,J}^{\mathrm{or}}(a) - v_{r,J}(a)\right\} = o_p(1),$$

which proves the sampling claim.

*Measurement component.* Let $\Delta_{\varepsilon,J}(a) := \widehat{v}_{\varepsilon,J}(a) - v_{\varepsilon,J}(a)$. Lemma 9 gives

$$\mathbb{E}[\Delta_{\varepsilon,J}(a) \mid \mathcal{G}_J] = 0. \qquad \text{(D.9)}$$

We first show that replacing one annotator's responses changes each $\widehat{\sigma}_j^2$ or $\widehat{\Gamma}_{jk}$ by at most a constant independent of $J$.

Responses lie in an interval of width $R_Y$, and every leave-out mean lies in the same interval. The sample variance can be written in terms of squared differences between responses:

$$\widehat{\sigma}_j^2 = \frac{1}{2m_j(m_j-1)} \sum_{\substack{\ell,\ell' \in M_j \\ \ell \ne \ell'}} (Y_{\ell j}^* - Y_{\ell' j}^*)^2.$$

Therefore,

$$0 \le \widehat{\sigma}_j^2 \le \frac{1}{2m_j(m_j-1)} \sum_{\substack{\ell,\ell' \in M_j \\ \ell \neq \ell'}} R_Y^2 = \frac{R_Y^2}{2},$$

$$\begin{aligned} |\widehat{\Gamma}_{jk}| &\le \kappa_{jk}^{\mathrm{loo}} \sum_{i \in M_j \cap M_k} |Y_{ij}^* - \bar{Y}_{j,-i}| \, |Y_{ik}^* - \bar{Y}_{k,-i}| \\ &\le \kappa_{jk}^{\mathrm{loo}} N_{jk} R_Y^2 \le N_{jk} R_Y^2 \le \bar{m} R_Y^2. \end{aligned} \tag{D.10}$$

The first line uses the $m_j(m_j-1)$ ordered pairs of distinct annotators. The last line uses $0 \le \kappa_{jk}^{\mathrm{loo}} \le 1$ and $N_{jk} \le \min(m_j, m_k) \le \bar{m}$. These bounds hold before and after replacement, so the changes in the two estimates are bounded by $R_Y^2/2$ and $2\bar{m}R_Y^2$, respectively.

We next show that replacing one annotator's responses affects only a bounded number of these estimates. For each selected document $j$, let

$$\mathcal{N}_j := \{k \neq j : N_{jk} > 0\}$$

denote the set of documents sharing at least one annotator with $j$. Each of the $m_j$ annotators assigned to $j$ labels at most $K-1$ other documents, so

$$|\mathcal{N}_j| \le m_j(K-1) \le \bar{m}(K-1). \tag{D.11}$$

Replacing annotator $i$'s responses can change $\widehat{\sigma}_j^2$ only for the at most $K$ documents in $\mathcal{J}_i$. To count the covariance estimates $\widehat{\Gamma}_{jk}$ that can change, define

$$\mathcal{A}_i := \{(j,k) : j,k \in \mathcal{S}_J,\ j \neq k,\ N_{jk} > 0,\ i \in M_j \cup M_k\}.$$

Since the pairs are ordered,

$$|\mathcal{A}_i| \le 2 \sum_{j \in \mathcal{J}_i} |\mathcal{N}_j| \le 2K\bar{m}(K-1).$$

With the same block replacement, let $\widehat{v}_{\varepsilon,J}^{(i)}(a)$ denote the estimator recomputed after replacing $\mathbf{Y}_i$ by an independent copy from its conditional law given $\mathcal{G}_J$, with the graph fixed, and write

$$\omega_{j,J}(a) := \frac{\bar{B}|a_j|}{J\pi_j m_j} = \frac{|a_j|}{\lambda_j}.$$

The coefficients satisfy

$$\frac{\bar{B}a_j^2}{J^2\pi_j^2 m_j} = \frac{m_j}{\bar{B}}\,\omega_{j,J}(a)^2, \qquad \frac{\bar{B}|a_j a_k|}{J^2\pi_j\pi_k m_j m_k} = \frac{1}{\bar{B}}\,\omega_{j,J}(a)\omega_{k,J}(a).$$

Thus an affected diagonal summand changes by at most $C_0\omega_{j,J}(a)^2/\bar{B}$, while an affected off-diagonal summand $(j,k)$ changes by at most $C_0\omega_{j,J}(a)\omega_{k,J}(a)/\bar{B}$. Summing the absolute

changes over the affected terms and applying Cauchy–Schwarz, together with (D.11), gives

$$\begin{aligned}
&|\widehat{v}_{\varepsilon,J}(a) - \widehat{v}^{(i)}_{\varepsilon,J}(a)|^2 \\
&\quad\le \frac{C_0}{\bar{B}^2}\left\{\sum_{j\in\mathcal{J}_i}\omega_{j,J}(a)^2 + \sum_{(j,k)\in\mathcal{A}_i}\omega_{j,J}(a)\omega_{k,J}(a)\right\}^2 \\
&\quad\le \frac{C_0\big(|\mathcal{J}_i| + |\mathcal{A}_i|\big)}{\bar{B}^2}\left\{\sum_{j\in\mathcal{J}_i}\omega_{j,J}(a)^4 + \sum_{(j,k)\in\mathcal{A}_i}\omega_{j,J}(a)^2\omega_{k,J}(a)^2\right\} \\
&\quad\le \frac{C_0}{\bar{B}^2}\left\{\sum_{j\in\mathcal{J}_i}\omega_{j,J}(a)^4 + \sum_{(j,k)\in\mathcal{A}_i}\omega_{j,J}(a)^2\omega_{k,J}(a)^2\right\}. \qquad \text{(D.12)}
\end{aligned}$$

The second inequality applies Cauchy–Schwarz to the $|\mathcal{J}_i| + |\mathcal{A}_i|$ terms. The final inequality uses

$$|\mathcal{J}_i| + |\mathcal{A}_i| \le K + 2K\bar{m}(K-1),$$

with this fixed bound absorbed into $C_0$.

The first term in (D.12) satisfies

$$\sum_i \sum_{j\in\mathcal{J}_i}\omega_{j,J}(a)^4 = \sum_{j\in\mathcal{S}_J} m_j\omega_{j,J}(a)^4 \le \bar{m}\sum_{j\in\mathcal{S}_J}\omega_{j,J}(a)^4.$$

Each ordered overlapping pair $(j,k)$ belongs to at most $|M_j \cup M_k| \le 2\bar{m}$ of the sets $\mathcal{A}_i$. Moreover, symmetry of the overlap relation and $2\omega_{j,J}(a)^2\omega_{k,J}(a)^2 \le \omega_{j,J}(a)^4 + \omega_{k,J}(a)^4$ give

$$\sum_{\substack{j\neq k\\ N_{jk}>0}}\omega_{j,J}(a)^2\omega_{k,J}(a)^2 \le \sum_{j\in\mathcal{S}_J}|\mathcal{N}_j|\omega_{j,J}(a)^4 \le \bar{m}(K-1)\sum_{j\in\mathcal{S}_J}\omega_{j,J}(a)^4.$$

Conditional Efron–Stein applied to (D.12) therefore yields

$$\operatorname{Var}(\widehat{v}_{\varepsilon,J}(a) \mid \mathcal{G}_J) \le \frac{C_0}{\bar{B}^2}\sum_{j\in\mathcal{S}_J}\omega_{j,J}(a)^4. \qquad \text{(D.13)}$$

Since $v_{\varepsilon,J}(a)$ is $\mathcal{G}_J$-measurable, (D.9) and (D.13) give

$$\begin{aligned}
\mathbb{E}[\Delta_{\varepsilon,J}(a)^2] &= \mathbb{E}[\operatorname{Var}(\widehat{v}_{\varepsilon,J}(a) \mid \mathcal{G}_J)] \\
&\le \frac{C_0}{\bar{B}^2}\sum_{j=1}^{J}\pi_j\frac{|a_j|^4}{\lambda_j^4} = \frac{C_0}{J\bar{B}}\sum_{j=1}^{J}\frac{|a_j|^4}{m_j\lambda_j^3} \le \frac{C_0}{c_J^3\bar{B}} \longrightarrow 0.
\end{aligned}$$

The first inequality uses $\mathbb{E}[\sum_{j\in\mathcal{S}_J}\omega_{j,J}(a)^4] = \sum_{j=1}^{J}\pi_j\omega_{j,J}(a)^4$. Thus $\Delta_{\varepsilon,J}(a) = o_p(1)$, proving the measurement claim and the proposition. □

## D.3 Joint inference for weighted corpus averages

In this section, we establish joint inference results for a fixed-dimensional vector of weighted corpus averages and its smooth transformations.

### D.3.1 Moment vector and linear combinations

Let $a^{(1)}, \ldots, a^{(q)}$ be target-weight arrays, each satisfying the maintained boundedness and pre-specification conditions on $a$, where $q < \infty$ is fixed. Define

$$\mathbf{\Phi}_J := \big(\Phi_J(a^{(1)}), \ldots, \Phi_J(a^{(q)})\big)^\top, \qquad \Phi_J(a^{(r)}) = \frac{1}{J}\sum_{j=1}^{J} a_j^{(r)} \mu_j.$$

For each coordinate $r$, let $g_j^{(r)}$ satisfy the maintained conditions on frozen predictions. Using the same document sample and annotator allocation, apply (4.1) coordinatewise:

$$\widehat{\Phi}_{Jr} := \frac{1}{J}\sum_{j=1}^{J} a_j^{(r)} g_j^{(r)} + \frac{1}{J}\sum_{j \in \mathcal{S}_J} \frac{a_j^{(r)}}{\pi_j}\big(\bar{Y}_j - g_j^{(r)}\big).$$

Collect these estimates in $\widehat{\mathbf{\Phi}}_J := (\widehat{\Phi}_{J1}, \ldots, \widehat{\Phi}_{Jq})^\top$.

For any $t \in \mathbb{R}^q$ fixed independently of $J$, let

$$a_j(t) := \sum_{r=1}^{q} t_r a_j^{(r)}.$$

Linearity gives

$$\begin{aligned} t^\top \widehat{\mathbf{\Phi}}_J = \frac{1}{J}\sum_{j=1}^{J}\sum_{r=1}^{q} t_r a_j^{(r)} g_j^{(r)} \\ + \frac{1}{J}\sum_{j \in \mathcal{S}_J} \frac{1}{\pi_j}\left\{ a_j(t)\bar{Y}_j - \sum_{r=1}^{q} t_r a_j^{(r)} g_j^{(r)} \right\}. \end{aligned} \tag{D.14}$$

This is the scalar estimator (4.1) applied to responses $a_j(t)Y_{ij}^*$, frozen predictions $\sum_{r=1}^{q} t_r a_j^{(r)} g_j^{(r)}$, and unit outer weights. Its target is

$$\frac{1}{J}\sum_{j=1}^{J} a_j(t)\mu_j = t^\top \mathbf{\Phi}_J.$$

For each fixed $t$, these transformed responses and predictions remain uniformly bounded and preserve the maintained response and design conditions. The preceding scalar arguments therefore apply under the corresponding sampling- and measurement-variance limit conditions.

### D.3.2 Covariance estimation

The scalar variance estimator also supplies the covariances between coordinates. The identity

$$\mathrm{Cov}(X, Y) = \frac{1}{2}\{\mathrm{Var}(X + Y) - \mathrm{Var}(X) - \mathrm{Var}(Y)\}$$

shows that it is enough to estimate the variance of each coordinate and each coordinate sum.

For each fixed $t \in \mathbb{R}^q$, let $\widehat{V}_J(t)$ denote the raw scalar variance estimator in (4.17), evaluated with responses $a_j(t)Y_{ij}^*$, frozen predictions $\sum_{r=1}^q t_r a_j^{(r)} g_j^{(r)}$, and unit outer weights. Let $e_r$ denote the $r$th standard basis vector in $\mathbb{R}^q$. Define

$$\begin{aligned}
\widehat{\Sigma}_{J,rr} &:= \widehat{V}_J(e_r), \\
\widehat{\Sigma}_{J,rs} &:= \frac{1}{2}\left\{\widehat{V}_J(e_r + e_s) - \widehat{V}_J(e_r) - \widehat{V}_J(e_s)\right\}, \qquad r \neq s.
\end{aligned} \tag{D.15}$$

All evaluations use the same observed annotations. Consequently, the covariance estimates account for the dependence induced by shared documents and annotator reuse through the existing scalar variance routine.

### D.3.3 Joint limit and covariance consistency

The preceding scalar results yield joint normality and consistency of the covariance estimator.

**Corollary 2** (Joint inference for weighted means)**.** *Maintain the conditions of Theorem 2, with the limits in* (4.10) *required for each fixed linear combination constructed above. Suppose also that the sampling-variance limit* (4.6) *holds for each such combination and that $m_j \geq 2$ for every document.*

*For each fixed $t \in \mathbb{R}^q$, let $V(t)$ denote the sum of the limiting sampling and measurement variances of the corresponding scalar estimator. Define*

$$\Sigma_{rr} := V(e_r), \qquad \Sigma_{rs} := \frac{1}{2}\{V(e_r + e_s) - V(e_r) - V(e_s)\}, \quad r \neq s. \tag{D.16}$$

*Then $\Sigma$ is nonnegative semidefinite, and*

$$\sqrt{\bar{B}}\,(\widehat{\mathbf{\Phi}}_J - \mathbf{\Phi}_J) \Rightarrow \mathcal{N}_q(0, \Sigma), \qquad \widehat{\Sigma}_J \xrightarrow{p} \Sigma. \tag{D.17}$$

*Proof.* For each fixed $t$, the linear combination in (D.14) has the scalar prediction-assisted form, with bounded transformed responses and frozen predictions. Theorem 2 establishes its measurement-variance limit. Corollary 1 therefore gives the scalar Gaussian limit whenever $V(t) > 0$. If $V(t) = 0$, both component variance limits vanish; Chebyshev's inequality for the sampling term and conditional Chebyshev's inequality for the measurement term give convergence of the scaled error to zero in probability.

Both variance components are quadratic in $t$. Their limits therefore satisfy $V(t) = t^\top \Sigma t$, with $\Sigma$ defined in (D.16). Since $V(t) \geq 0$ for every $t$, $\Sigma$ is positive semidefinite. The Cramér–Wold device gives the joint Gaussian limit.

The scalar variance-consistency result gives

$$\widehat{V}_J(e_r) \xrightarrow{p} V(e_r), \qquad \widehat{V}_J(e_r + e_s) \xrightarrow{p} V(e_r + e_s).$$

Applying the covariance identity in (D.15) gives consistency of every entry of $\widehat{\Sigma}_J$. Because $q$ is fixed, this establishes $\widehat{\Sigma}_J \xrightarrow{p} \Sigma$. □

## D.4 Smooth transformations and implicit parameters

Throughout this subsection, suppose the conclusions of Corollary 2 hold.

**Proposition 5** (Smooth transformations)**.** *Suppose $\boldsymbol{\Phi}_J \to \boldsymbol{\Phi}_0$, and let $H : \mathcal{O} \subseteq \mathbb{R}^q \to \mathbb{R}^p$ be continuously differentiable on an open neighborhood of $\boldsymbol{\Phi}_0$. Define*

$$\boldsymbol{\beta}_J := H(\boldsymbol{\Phi}_J), \qquad \widehat{\boldsymbol{\beta}}_J := H(\widehat{\boldsymbol{\Phi}}_J),$$

*and write*

$$G := DH(\boldsymbol{\Phi}_0), \qquad \widehat{G}_J := DH(\widehat{\boldsymbol{\Phi}}_J).$$

*Then*

$$\sqrt{\bar{B}}\,(\widehat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J) \Rightarrow \mathcal{N}_p(0, \Omega), \qquad \Omega := G\Sigma G^\top, \tag{D.18}$$

*and*

$$\widehat{\Omega}_J := \widehat{G}_J \widehat{\Sigma}_J \widehat{G}_J^\top \xrightarrow{p} \Omega. \tag{D.19}$$

*Proof.* The joint limit gives $\widehat{\boldsymbol{\Phi}}_J - \boldsymbol{\Phi}_J = O_p(\bar{B}^{-1/2})$. Continuous differentiability therefore yields

$$\sqrt{\bar{B}}\,(\widehat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J) = G\sqrt{\bar{B}}\,(\widehat{\boldsymbol{\Phi}}_J - \boldsymbol{\Phi}_J) + o_p(1).$$

The Gaussian limit follows by Slutsky's theorem. Continuity of $DH$ gives $\widehat{G}_J \xrightarrow{p} G$; together with $\widehat{\Sigma}_J \xrightarrow{p} \Sigma$, this proves covariance consistency. □

**Implicitly defined parameters.** The same argument applies to a parameter defined by

$$F_J(\boldsymbol{\beta}_J, \boldsymbol{\Phi}_J) = 0,$$

where $F_J : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ is known under the fixed-corpus convention. Let $\widehat{\boldsymbol{\beta}}_J$ be a consistent local solution of $F_J(\boldsymbol{b}, \widehat{\boldsymbol{\Phi}}_J) = 0$. Suppose $F_J$ is continuously differentiable on neighborhoods of a common positive radius around $(\boldsymbol{\beta}_J, \boldsymbol{\Phi}_J)$, with first derivatives locally equicontinuous uniformly in $J$. A sufficient condition is that $F_J$ is twice continuously differentiable on these neighborhoods, with second derivatives uniformly bounded over the neighborhoods and across $J$. Assume their values at the target converge, with a nonsingular limit for $D_\beta F_J$.

The standard first-order expansion for estimating equations (Newey and McFadden, 1994, Section 3.1) then gives (D.18) with

$$G = -\lim_{J\to\infty} \{D_\beta F_J(\boldsymbol{\beta}_J, \boldsymbol{\Phi}_J)\}^{-1} D_\Phi F_J(\boldsymbol{\beta}_J, \boldsymbol{\Phi}_J).$$

A consistent covariance estimator is

$$\begin{aligned}\widehat{G}_J &:= -\left\{D_\beta F_J(\widehat{\boldsymbol{\beta}}_J, \widehat{\boldsymbol{\Phi}}_J)\right\}^{-1} D_\Phi F_J(\widehat{\boldsymbol{\beta}}_J, \widehat{\boldsymbol{\Phi}}_J), \\ \widehat{\Omega}_J &:= \widehat{G}_J \widehat{\Sigma}_J \widehat{G}_J^\top \xrightarrow{p} G\Sigma G^\top.\end{aligned} \tag{D.20}$$

For either construction, a coordinate with positive limiting variance has estimated standard error

$$\widehat{\mathrm{se}}(\widehat{\beta}_{Jr}) = \sqrt{\widehat{\Omega}_{J,rr}/\bar{B}}.$$

Wald inference for fixed linear restrictions follows whenever the limiting covariance of the tested restrictions is positive definite.

# E Graph implementation

## E.1 An edge-pair implementation of uniform allocation

When the rejection sampler described in the main text is computationally costly, we approximate the uniform graph allocation using the standard edge-switch chain (Carstens and Kleer, 2018, Section 6). We apply existing mixing results to this implementation and show that vanishing graph-sampling error preserves the baseline design's asymptotic conclusions.

**Implementation by degree-preserving swaps.** Conditional on a realized document sample $\mathcal{S}_J$ with $B = \sum_{j \in \mathcal{S}_J} m_j > 0$ for which the prescribed document depths and capacity-filled annotator loads are feasible, write

$$\mathfrak{A}_J^{\mathrm{cf}} := \mathfrak{A}_J(\boldsymbol{d}_J, \boldsymbol{k}_J^{\mathrm{cf}})$$

for the nonempty set of simple bipartite graphs with these labelled document and annotator vertices and these degree sequences, and let $\mathsf{U}_J^A := \mathrm{Unif}(\mathfrak{A}_J^{\mathrm{cf}})$. A feasible starting graph may, for example, be found by a deterministic bipartite-flow construction.

If $\mathfrak{A}_J^{\mathrm{cf}}$ is a singleton, return its unique graph and set $P_J^e = I$. Otherwise, starting from any feasible graph, use the standard edge-switch chain (Carstens and Kleer, 2018, Section 6) with the following attempted transition:

1. With probability one half, leave the graph unchanged.
2. Otherwise, choose an unordered pair of distinct current edges uniformly from the $\binom{B}{2}$ possibilities. If the edges are vertex-disjoint and replacing them by the two cross-edges preserves simplicity, perform that swap; otherwise leave the graph unchanged.

Denote this edge-pair kernel by $P_J^e$. Choose the starting graph and the number of attempted transitions from the sample and margins before running the chain, and use randomization independent of annotator response blocks. Every transition preserves the prescribed document depths, annotator loads, and simplicity. Thus the implementation retains these features of the theoretical design at every run length; mixing controls how closely the final graph follows the

required uniform law.

We now verify that the edge-pair procedure has the same guarantees from the standard swap-chain theory. Its kernel is symmetric, hence a uniformly distributed graph remains uniformly distributed after another attempted transition. Any feasible graph can be reached from any other through a sequence of valid swaps (Erdős et al., 2018). The probability of staying put rules out a fixed periodic alternation between groups of graphs. These properties imply that the graph distribution converges to $\mathsf{U}_J^A$ from every feasible starting graph, by the standard convergence theorem for finite Markov chains (Levin and Peres, 2017, Theorem 4.9).

The chain studied by Erdős et al. (2018) proposes vertex pairs, whereas the implementation above proposes edge pairs. The following comparison allows us to use their mixing theorem for this proposal rule. Suppose $\mathfrak{A}_J^{\mathrm{cf}}$ is non-singleton (hence $B \geq 2$, $|\mathcal{S}_J| \geq 2$, and $n \geq 2$), and let $P_J^v$ denote their lazy vertex-pair swap chain. For distinct graphs $G, H \in \mathfrak{A}_J^{\mathrm{cf}}$ that differ by one valid swap,

$$P_J^e(G,H) = \varrho_J P_J^v(G,H), \qquad \varrho_J := \frac{\binom{|\mathcal{S}_J|}{2}\binom{n}{2}}{\binom{B}{2}}; \tag{E.1}$$

If $G \neq H$ and $H$ cannot be obtained from $G$ by one valid swap, both transition probabilities are zero. Thus both procedures allow exactly the same changes of graph.

For each current graph $G$, the probability of staying put is one minus the sum of the probabilities of moving to other graphs. The preceding comparison therefore also determines the holding probabilities:

$$P_J^e(G,G) = 1 - \varrho_J\big[1 - P_J^v(G,G)\big].$$

Combining the probabilities of moving and staying gives

$$P_J^e = I + \varrho_J(P_J^v - I).$$

For either chain, define its spectral gap by

$$\operatorname{gap}(P) := 1 - \lambda_2(P),$$

where $\lambda_2(P)$ is the second-largest eigenvalue. For these lazy chains, the gap controls how quickly departures from the uniform distribution decay. The matrix identity above transforms each eigenvalue $\lambda$ of $P_J^v$ into $1 + \varrho_J(\lambda - 1)$ for $P_J^e$. Since $\varrho_J > 0$, their ordering is preserved, giving

$$\operatorname{gap}(P_J^e) = \varrho_J \operatorname{gap}(P_J^v). \tag{E.2}$$

To transfer the polynomial mixing guarantee, it is enough to bound the change in the inverse spectral gap by a polynomial in $B$. Since the non-singleton case has at least two document vertices and two annotator vertices,

$$\varrho_J^{-1} = \frac{\binom{B}{2}}{\binom{|\mathcal{S}_J|}{2}\binom{n}{2}} \leq \binom{B}{2}.$$

The gap comparison therefore gives

$$\frac{1}{\mathrm{gap}(P_J^e)} \le \binom{B}{2} \frac{1}{\mathrm{gap}(P_J^v)}.$$

Thus any polynomial bound on the inverse spectral gap of the vertex-pair chain also holds, with at most an additional quadratic factor, for the edge-pair chain.

The same degree bounds also ensure that the published sufficient condition for rapid mixing eventually holds. Put

$$\Delta_J := \max\left\{\max_{j\in\mathcal{S}_J} m_j, \max_{i\le n} k_i\right\}.$$

If $K \ge 2$, then on $\mathcal{H}_J$, for all sufficiently large rows, capacity filling gives $\Delta_J \ge K \ge 2$. Bounded depths and loads, together with the growing label count on this event, give $\Delta_J \le \sqrt{B/2}$. Thus Erdős et al. (2018, Theorem 2) applies to $P_J^v$, and the gap comparison transfers its polynomial mixing guarantee to $P_J^e$. In particular, for $\eta \in (0,1)$, the mixing time

$$\tau_J^e(\eta) := \min\left\{t : \sup_{G\in\mathfrak{A}_J^{\mathrm{cf}}} \left\|\delta_G (P_J^e)^t - \mathsf{U}_J^A\right\|_{\mathrm{TV}} \le \eta\right\}$$

is bounded by a polynomial in $B$ and $\log(\eta^{-1})$ on the regular event for all sufficiently large rows. It should be noted that this is an asymptotic implementability guarantee. It does not supply a universal finite-sample burn-in constant or a stopping diagnostic based on the observed chain.

**Consequences for the statistical theory.** The mixing result concerns the assignment graph. The next proposition translates that approximation into a guarantee for the complete design-and-response experiment. It separates conclusions that require only a simple exogenous graph, which remain exact, from the asymptotic conclusions established under uniform allocation, which are preserved when the mixing error vanishes.

**Proposition 6** (Finite-run graph implementation)**.** *Suppose $K \ge 2$, $\Pr(\mathcal{H}_J^c) \to 0$, and let $0 < \varepsilon_J^{\mathrm{mix}} \downarrow 0$. On the regular allocation event $\mathcal{H}_J$, choose a feasible starting graph and an integer $t_J$ measurably from the sample and margins, with $t_J \ge \tau_J^e(\varepsilon_J^{\mathrm{mix}})$, and run $P_J^e$ for $t_J$ attempted transitions. Off $\mathcal{H}_J$, use any simple exogenous allocation with annotator loads bounded by $K$, retaining the standing fallback when the capacity-filled margins are infeasible. If $\widetilde{\mathbb{P}}_J$ and $\mathbb{P}_J$ denote the laws of the complete design-and-response experiment (the sample, margins, final assignment graph, and responses) under, respectively, this implementation and exact uniform graph allocation, then*

$$\left\|\widetilde{\mathbb{P}}_J - \mathbb{P}_J\right\|_{\mathrm{TV}} \le \varepsilon_J^{\mathrm{mix}} + \Pr(\mathcal{H}_J^c) \longrightarrow 0. \tag{E.3}$$

*Consequently every baseline-design probability limit, weak limit, studentization result, and coverage conclusion proved under exact uniform allocation also holds under the finite-run implementation.*

*Proof.* On $\mathcal{H}_J$, given the sampled documents and the prescribed document and annotator degrees, we run the swap procedure for $t_J \ge \tau_J^e(\varepsilon_J^{\mathrm{mix}})$ attempted transitions. By the previous

result, the resulting graph distribution is therefore within $\varepsilon_J^{\mathrm{mix}}$ of the uniform distribution $\mathsf{U}_J^A$ in total variation. On $\mathcal{H}_J^c$ the discrepancy is at most one. The feasible implementation procedure above and the ideal uniform draw use the same distribution of sampled documents and annotation depths, on which we conditioned. Weighting the error bounds by the corresponding probabilities then gives the right-hand side of (E.3).

Given the realized graph and preceding design variables, responses are generated in the same way under the finite-run implementation and ideal uniform allocation. The same error bound therefore remains valid after including the responses. In particular, the probabilities of any event defined by the resulting data or statistics—such as a confidence interval covering the target—differ between the two procedures by at most $\varepsilon_J^{\mathrm{mix}} + \Pr(\mathcal{H}_J^c)$. Since this bound tends to zero, the asserted asymptotic conclusions continue to hold. □

# F First-order allocation for weighted means

This section gives the complete allocation characterization for the scalar weighted-mean target under the baseline Poisson design in Section 5. We maintain the design restrictions stated there.

Documents with zero target weight are removed before assigning the budget, with the remaining weights rescaled to preserve the target.[4] Henceforth, $J$ and $a_j$ denote the resulting corpus size and weights, so $a_j \neq 0$.

Motivated by the variance expansion in (5.1), we define the first-order oracle as a minimizer of

$$L_J(\pi, m; a) := \sum_{j=1}^{J} a_j^2 \left\{ \frac{1-\pi_j}{\pi_j} r_j^2 + \frac{\nu_j}{\pi_j m_j} \right\}, \tag{F.1}$$

where $r_j = \mu_j - g_j$.

Let $s_j := \pi_j m_j$ denote expected label expenditure on document $j$. We consider the continuous relaxation with $2 \leq m_{\min} \leq \bar{m} < \infty$ and

$$0 < \pi_j \leq 1, \qquad m_{\min} \leq m_j \leq \bar{m}, \qquad s_j \geq c_\lambda \frac{\bar{B}}{J}, \qquad \sum_{j=1}^{J} s_j = \bar{B}.$$

Write

$$\underline{s}_J := c_\lambda \frac{\bar{B}}{J}, \qquad \overline{s}_J := \bar{m}. \tag{F.2}$$

Assume $0 < \bar{B} \leq J\bar{m}$ and $0 < c_\lambda \leq 1$, so this continuous policy class is nonempty.

[4] Let $J_0$ be the original corpus size, $\mathcal{J}_+ := \{j : a_j \neq 0\}$, and $J_+ := |\mathcal{J}_+|$. For $j \in \mathcal{J}_+$, set $a_j^+ := (J_+/J_0)a_j$. Then

$$\frac{1}{J_+} \sum_{j \in \mathcal{J}_+} a_j^+ \mu_j = \frac{1}{J_0} \sum_{j=1}^{J_0} a_j \mu_j.$$

If all weights are zero, the target is known and no allocation is needed.

To state the result, define the coverage-first implementation

$$\pi_j^\star(s) = \min\left\{1, \frac{s}{m_{\min}}\right\}, \qquad m_j^\star(s) = \frac{s}{\pi_j^\star(s)}, \tag{F.3}$$

and its associated loss

$$\phi_j(s) := \frac{\nu_j}{s} + r_j^2\left(\frac{m_{\min}}{s} - 1\right)_+ = \begin{cases} \dfrac{\nu_j + m_{\min} r_j^2}{s} - r_j^2, & s \le m_{\min}, \\ \dfrac{\nu_j}{s}, & s \ge m_{\min}. \end{cases} \tag{F.4}$$

**Proposition 7** (Continuous first-order allocation)**.** *An optimal policy can be implemented using* (F.3)*, with expenditures solving the convex program*

$$\min_{\{s_j\}} \sum_{j=1}^J a_j^2 \phi_j(s_j) \quad \textit{subject to} \quad \sum_{j=1}^J s_j = \bar{B}, \qquad \underline{s}_J \le s_j \le \overline{s}_J. \tag{F.5}$$

*There is a common multiplier $\xi \ge 0$ such that each $s_j^\star$ minimizes $a_j^2\phi_j(s) + \xi s$ over $[\underline{s}_J, \overline{s}_J]$, with $\sum_j s_j^\star = \bar{B}$.*

*Define*

$$b_{j,-} := a_j^2(\nu_j + m_{\min} r_j^2), \qquad b_{j,+} := a_j^2 \nu_j.$$

*For $\xi > 0$, let*

$$s_j^{\mathrm{unc}}(\xi) := \begin{cases} \sqrt{b_{j,-}/\xi}, & \xi > b_{j,-}/m_{\min}^2, \\ m_{\min}, & b_{j,+}/m_{\min}^2 \le \xi \le b_{j,-}/m_{\min}^2, \\ \sqrt{b_{j,+}/\xi}, & 0 < \xi < b_{j,+}/m_{\min}^2. \end{cases} \tag{F.6}$$

*Impose the spending limits by raising values below $\underline{s}_J$ to that lower bound and reducing values above $\overline{s}_J$ to that upper bound:*

$$s_j^\star = \min\left\{\overline{s}_J, \max\left\{\underline{s}_J, s_j^{\mathrm{unc}}(\xi)\right\}\right\}, \qquad \sum_{j=1}^J s_j^\star = \bar{B}. \tag{F.7}$$

*If $\xi = 0$, choose expenditures that attain each document's minimum loss over its allowed interval while meeting the total budget:*

$$s_j^\star \in \arg\min_{\underline{s}_J \le s \le \overline{s}_J} a_j^2 \phi_j(s), \qquad \sum_{j=1}^J s_j^\star = \bar{B}.$$

*Proof.* At fixed expenditure $s = \pi_j m_j$, the derivative of document $j$'s contribution with respect to coverage is

$$\frac{\partial}{\partial \pi_j} a_j^2 \left\{\frac{\nu_j}{s} + r_j^2\left(\frac{1}{\pi_j} - 1\right)\right\} = -\frac{a_j^2 r_j^2}{\pi_j^2} \le 0.$$

The depth restrictions give the feasible coverage interval

$$\frac{s}{\bar{m}} \le \pi_j \le \min\left\{1, \frac{s}{m_{\min}}\right\}.$$

Thus the largest feasible coverage minimizes the loss, giving (F.3). This choice is unique when $r_j \neq 0$; when $r_j = 0$, all feasible coverage–depth combinations at the same expenditure have the same loss. Substitution gives (F.4) and reduces the problem to (F.5).

For $\xi \ge 0$, define

$$h_j(s;\xi) := a_j^2 \phi_j(s) + \xi s.$$

Away from $m_{\min}$,

$$\frac{\partial h_j(s;\xi)}{\partial s} = \begin{cases} \xi - b_{j,-}/s^2, & s < m_{\min}, \\ \xi - b_{j,+}/s^2, & s > m_{\min}. \end{cases}$$

The derivative is nondecreasing on each branch and increases by

$$\frac{b_{j,-} - b_{j,+}}{m_{\min}^2} = \frac{a_j^2 r_j^2}{m_{\min}} \ge 0$$

at their common endpoint. Since $h_j$ is continuous there, it is convex. Taking $\xi = 0$ also shows that the weighted loss $a_j^2 \phi_j$ is nonincreasing.

Since the objective is convex and finite on $(0,\infty)^J$ and all constraints are affine, the subgradient KKT conditions are necessary and sufficient for global optimality (Rockafellar, 1970, Corollaries 28.2.2 and 28.3.1). We verify these conditions by minimizing $h_j(s;\xi)$ over each document's spending interval and choosing the common multiplier to satisfy the budget.

Suppose $\xi > 0$ and $b_{j,-} > 0$. Setting the derivative equal to zero on each smooth branch gives the square-root cases in (F.6). The minimum occurs at $m_{\min}$ when the left derivative is nonpositive and the right derivative is nonnegative:

$$\xi - \frac{b_{j,-}}{m_{\min}^2} \le 0 \le \xi - \frac{b_{j,+}}{m_{\min}^2},$$

which gives the middle case. Convexity then implies that an unrestricted minimum outside the allowed spending interval is replaced by the nearest endpoint, as in (F.7). If $b_{j,-} = 0$, then $h_j(s;\xi) = \xi s$, whose minimum is at $\underline{s}_J$, also as given by that formula. Write $s_j(\xi)$ for this bounded spending choice.

We next show that a multiplier can be chosen to meet the budget. Let $d_j$ be the smallest expenditure attaining the minimum of $a_j^2 \phi_j$ over its allowed interval. Continuity and monotonicity imply that the full set of minimizing expenditures is $[d_j, \overline{s}_J]$. The explicit spending rule is continuous and nonincreasing in $\xi > 0$, with

$$\lim_{\xi \downarrow 0} s_j(\xi) = d_j.$$

Moreover, $s_j(\xi) = \underline{s}_J$ for every $j$ whenever

$$\xi > \max_{k \le J} \frac{b_{k,-}}{\underline{s}_J^2}.$$

If $\bar{B} < \sum_j d_j$, continuity and $J\underline{s}_J = c_\lambda \bar{B} \le \bar{B}$ therefore give a positive multiplier for which $\sum_j s_j(\xi) = \bar{B}$. The case $\bar{B} = J\underline{s}_J$ is included because the lower bounds are attained at a finite multiplier. If instead $\bar{B} \ge \sum_j d_j$, take $\xi = 0$ and choose $s_j^\star \in [d_j, \overline{s}_J]$ whose sum is $\bar{B}$; such a choice exists because $\bar{B} \le J\overline{s}_J$.

Finally, for any feasible allocation $s$, the document-level minimizing property and the common budget give

$$\sum_{j=1}^{J} a_j^2 \phi_j(s_j^\star) + \xi \bar{B} \le \sum_{j=1}^{J} a_j^2 \phi_j(s_j) + \xi \bar{B}.$$

Cancelling $\xi \bar{B}$ proves global optimality. Implementing these expenditures through (F.3) completes the proof. □

The allocation problem has the bounded separable convex form studied by van der Klauw et al. (2017), whose Lemma 1 gives the corresponding optimality conditions using one-sided derivatives. For the present objective, (F.6) and (F.7) give each document's expenditure as a function of the common multiplier. In the case $\xi > 0$, the budget equation $\sum_j s_j(\xi) = \bar{B}$ can be solved by bisection because total expenditure is continuous and nonincreasing in $\xi$. For a discussion of multiplier-search methods, see Patriksson (2008, Section 3.1).

### F.1 A default without document-specific difficulty estimates

The first-order oracle depends on the unknown document-specific quantities $r_j^2$ and $\nu_j$. To describe a simple allocation when these quantities cannot be estimated from a pilot or comparable annotated data, let $\mathcal{I}_J$ denote the information available when the design is chosen, and let $\mathbb{E}_0(\cdot \mid \mathcal{I}_J)$ denote the researcher's working pre-annotation expectation. This expectation is used only to choose the design; it does not alter the fixed-corpus inferential model. The conditional optimization below is pointwise in $\mathcal{I}_J$; in particular, candidate expenditures are $\mathcal{I}_J$-measurable.

On the minimum-depth portion of the allocation problem, the relevant combined difficulty is

$$d_j^{\mathrm{diff}} := \nu_j + m_{\min} r_j^2.$$

Suppose the working expectation does not distinguish documents by this quantity:

$$\mathbb{E}_0(d_j^{\mathrm{diff}} \mid \mathcal{I}_J) = d_J^0 \in (0, \infty), \qquad j = 1, \dots, J. \tag{F.8}$$

Define

$$\|a\|_{1,J} := \sum_{k=1}^{J} |a_k|, \qquad w_j := \frac{J|a_j|}{\|a\|_{1,J}}, \qquad s_j^0 := \frac{\bar{B}}{J} w_j = \bar{B} \frac{|a_j|}{\|a\|_{1,J}}.$$

**Corollary 3** (Default under common expected difficulty)**.** *Suppose* (F.8) *holds and*

$$c_\lambda \le w_j, \qquad j = 1, \dots, J, \qquad \frac{\bar{B}}{J m_{\min}} \max_{1 \le j \le J} w_j \le 1. \tag{F.9}$$

*Then* $s^0 = (s_1^0, \dots, s_J^0)$ *uniquely minimizes the working expected first-order criterion*

$$\sum_{j=1}^{J} a_j^2 \mathbb{E}_0\{\phi_j(s_j) \mid \mathcal{I}_J\} \tag{F.10}$$

*over the feasible set in* (F.5)*. A coverage-first solution is*

$$m_j^0 = m_{\min}, \qquad \pi_j^0 = \frac{s_j^0}{m_{\min}}, \qquad j = 1, \dots, J. \tag{F.11}$$

*Proof.* For every $s > 0$,

$$\phi_j(s) \ge \frac{d_j^{\text{diff}}}{s} - r_j^2,$$

with equality whenever $s \le m_{\min}$. Taking the working expectation conditional on $\mathcal{I}_J$ and using (F.8) gives

$$\mathbb{E}_0\{\phi_j(s) \mid \mathcal{I}_J\} \ge \frac{d_J^0}{s} - \mathbb{E}_0(r_j^2 \mid \mathcal{I}_J).$$

For any feasible expenditures, Cauchy–Schwarz implies

$$\sum_{j=1}^{J} \frac{a_j^2}{s_j} \ge \frac{\left(\sum_{j=1}^{J} |a_j|\right)^2}{\sum_{j=1}^{J} s_j} = \frac{\|a\|_{1,J}^2}{\bar{B}},$$

with equality only when $s_j = \bar{B}|a_j|/\|a\|_{1,J} = s_j^0$ for every $j$. The first part of (F.9) gives $\underline{s}_J \le s_j^0$, while the second gives $s_j^0 \le m_{\min} \le \bar{m} = \overline{s}_J$. Thus $s^0$ is feasible and attains equality in both preceding inequalities. It is therefore the unique minimizer of (F.10). Equation (F.3) gives (F.11). □