Large Language Models for Statistical Inference: Context Augmentation with Applications to the Two-Sample Problem and Regression

Preliminary Draft:
Do not circulate or cite without permission

Marc Ratkovic
Professor of Social Data Science
Department of Political Science
Department of Data Science
University of Mannheim
Mannheim, DE
marc.ratkovic@uni-mannheim.de

Abstract

Text data pose fundamental challenges for statistical inference due to their high dimensionality and unstructured nature. Conventional methods based on embeddings or topic models often oversimplify linguistic complexity without formal inferential guarantees. We propose *context augmentation*, which treats LLM-generated contexts as auxiliary variables to build a structured mapping from text to inferential targets. A *clause function* scores interactions between each observed string and its contexts, and an *aggregation operator* summarizes these evaluations into robust statistics. Formally, under standard support, ignorability, and regularity conditions, we derive the limiting distribution of context-augmented estimators and extend the framework beyond two-sample tests to regression coefficients, quantiles, and ranks. In particular, we introduce a text-to-text regression—where both predictors and outcomes are strings mediated by latent contexts—that separates semantic from syntactic effects. When the target statistic is pivotal, repeated cross-fitting or bootstrap resampling achieves higher-order accuracy, reducing the required convergence rate from $n^{-1/4}$ to $n^{-1/8}$. We also supply finite-sample inequalities linking error rates to the number of contexts and cross-fits. Empirically, we apply context augmentation to a two-sample experiment with free-text responses and to our text-to-text regression. In replication, we recover string- and context-level treatment effects with interpretable diagnostics, thereby anchoring LLM outputs to classical estimands. Finally, we model structured dialogue to show how shifts in shared understanding yield more predictable syntax across turns.

1 Introduction

Large language models (LLMs), such as GPT (Brown et al., 2020; OpenAI, 2023) and BERT (Devlin et al., 2019), have significantly expanded the analytical capacity of text-based research across disciplines, including political science, law, and social sciences (Chen et al., 2024; Argyle et al., 2023). Despite their success in classification, prediction, and content generation, the integration of LLMs into formal statistical inference remains limited. Classical inferential methods—including hypothesis testing, regression modeling, and probability-based estimation—face challenges with textual data due to their high-dimensional dependencies and lack of intrinsic numeric structure (Blei et al., 2003; Gentzkow et al., 2019). Existing approaches typically rely on fixed embeddings or summary statistics that abstract away linguistic nuance, limiting their utility in rigorous statistical inference.

This paper introduces *context augmentation*, a method integrating LLM-generated contexts with semiparametric inference, drawing upon empirical process theory (van der Vaart, 1998a; Huber and Ronchetti, 2009) and classical data augmentation (Rubin, 1976; Tanner and Wong, 1987). Context augmentation leverages the distributional hypothesis from linguistics (Harris, 1954; Sahlgren, 2008), positing meaning as arising from contextual usage rather than fixed embeddings. By employing LLMs to dynamically generate contexts around observed texts, our approach constructs auxiliary linguistic structures that facilitate statistical estimation, maintaining textual richness without imposing restrictive parametric assumptions.

Our approach comprises three components. First, LLMs generate contextual expansions of observed text, conditional on experimental treatments or relevant covariates. Second, a *clause function* quantifies interactions between observed text and generated contexts, capturing probabilistic dependencies structurally. Third, an *aggregation operator* integrates context-specific evaluations into robust statistical summaries. This embedding within dynamically generated linguistic environments enables rigorous hypothesis testing, regression analyses, and dependence assessment. This formulation extends naturally to a regression setting with text-valued predictors and outcomes, enabling inference on the relationship between linguistic inputs and responses.

We establish conditions under which context-augmented estimators exhibit asymptotic normality (van der Vaart, 1998a). Crucially, we explicitly characterize and eliminate self-referential bias arising from context generation and evaluation using repeated cross-fitting (Chernozhukov et al., 2018). Additionally, by exploiting pivotal statistics—statistics whose asymptotic distributions do not depend on nuisance parameters—we achieve higher-order accuracy and provide new formal results

delineating computational-statistical tradeoffs. We also provide finite-sample inequalities that clarify how estimation error depends on the number of generated contexts and repeated cross-fits. This novel contribution ensures valid inference despite the inherent stochasticity of LLM-generated contexts.

Unlike embedding-based or topic-modeling methods, our framework directly incorporates linguistic variability into inference, mitigating dimensionality concerns and model dependence. To illustrate utility, we apply context augmentation to two canonical statistical tasks: a two-sample test from experimental data with a text outcome (Egami et al., 2022) and a regression model with textual predictors and outcomes. We show that context augmentation detects a stronger effect than the original topic model implementation, and offers a distinct set of insights and analyses. Our regression application operationalizes psycholinguistic theories of interactive alignment (Pickering and Garrod, 2004, 2013), showing that a shift in shared understanding in a dialog leads to the participants adopting a more predictable syntax.

2 Intuition Behind Context Augmentation

This section provides an intuitive overview of our approach before introducing the formal framework. In the two-sample setting, we compare two groups of text—say, Group A and Group B. For each observed text string, the LLM generates multiple contexts that enrich the raw text with additional semantic detail. For example, given the string "apple," the LLM might generate a context such as "The tart <str> tastes great in a pie." Such a context is clearly aligned with the food domain, resulting in a high probability for "apple," whereas unrelated strings like "Paris" or "scuba diving" would receive low probabilities. Using these generated contexts, we then construct a statistic for each string that quantifies whether it is more likely to appear under the contexts generated from members of one group versus the other. We avoid a self-referential bias by not comparing strings against contexts generated by that self-same string, implementing a repeated cross-fitting strategy, where we average over leave-n/2-out subsamples. In our implementation, the string-level statistic is a log-probabillity that compares the augmented probability of the string appearing in Group A contexts to that in Group B contexts. We illustrate using a difference-in-means t-test, and show that the method both properly recovers the null distribution when Group A and B are the same, but also differentiates between Groups when they are different. Just as bootstrapping a pivotal statistic leads to higher-order convergence, we show that repeated cross-fitting a pivotal statistic, like a t- or F-statistic allows for a similar gains.

We also apply our framework to the regression setting, where both the predictor s_x and the outcome s_y are text. Our goal is to measure whether observing s_x increases the probability of subsequently

observing s_y. To capture this effect, we generate latent contexts around the observed s_x; we then treat these contexts as mediating variables that link the syntactic and semantic features of s_x to s_y. To isolate the influence of s_x's content, we compare the LLM-assessed probability of s_y when paired with the informative s_x against that when s_x is replaced by a non-informative variant, $\tilde{s}_x$. We generate non-informative strings in three ways: in the masked variant, each token in s_x is replaced with a placeholder (preserving overall structure but removing specific lexical content); in the shuffled variant, the words are randomly permuted (disrupting syntactic order while retaining the words themselves); and in the jabberwocky variant, content words are replaced with invented nonsense words (stripping away semantic meaning while preserving an approximate syntactic scaffold). For example, consider the sentence "The director praised the outstanding performance." Its shuffled variant might be "performance the outstanding praised director the," while its jabberwocky version could be "The gormer flarked the slythering frimble." By holding the generated context fixed across these variants, we compare the LLM-assessed likelihoods of s_y under the informative and non-informative versions of s_x. A regression analysis of the outcome, regressing the log-probability of s_y with the informative context, on the log-probability of s_y on the non-informative predictors. This approach enables us to decompose and quantify the contributions of semantic content, syntax, and lexical details in driving the outcome, even when both the predictor and outcome are text.

3 Literature Review and Motivation

Existing text-based inference methods primarily utilize embedding transformations, such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and contextualized embeddings from models like BERT and GPT (Devlin et al., 2019; Brown et al., 2020). These approaches typically reduce text to fixed-length vector representations, enabling numerical analyses but often losing dynamic contextual information intrinsic to language. Similarity metrics such as cosine similarity or dot products lack probabilistic interpretation, limiting rigorous statistical inference.

Recent semiparametric estimation techniques address high-dimensional nuisance functions via regularization, orthogonal moment conditions, and machine learning (Chernozhukov, Fernández-Val, and Kowalski, Chernozhukov et al.; Belloni et al., 2014; Chernozhukov et al., 2022), building upon classical results on profiling, debiasing, and sample-splitting (Bickel, 1982; van der Vaart, 1998a; Robins and Rotznitzky, 1995; Murphy and van der Vaart, 2000). While these approaches robustly handle nuisance parameters, their direct application to textual data remains limited due to inherent linguistic variability. Our contribution departs from prior work by connecting text data to classical

estimands using LLM-generated contexts as auxiliary data, permitting inference on statistics like means, regression coefficients, and quantiles directly from text data.

In contrast, context augmentation integrates over LLM-generated contexts, effectively marginalizing nuisance dimensions. This aligns with integrated likelihood methods (Berger et al., 1999), traditionally restricted to parametric settings. Our framework generalizes integrated likelihood concepts to nonparametric, model-generated contexts, treating textual structure as a nuisance to be integrated out rather than explicitly profiled.

Efforts to integrate causal inference and text analysis (e.g., Egami et al., 2022; Roberts et al., 2020) commonly embed text into predefined numeric structures. Although these methods utilize split-sample strategies to avoid overfitting, their reliance on fixed embeddings constrains representational flexibility. Context augmentation moves beyond fixed embeddings, explicitly modeling linguistic variability through probabilistic integration over generated contexts, thus offering richer, uncertainty-aware modeling. Recent studies have used LLMs as black-box scoring mechanisms for prediction or plausibility ranking, but these approaches rarely yield interpretable estimands or theoretical guarantees.

Our approach is grounded in the distributional hypothesis (Harris, 1954; Sahlgren, 2008), which views meaning through context. While methods like latent Dirichlet allocation (Blei et al., 2003) and static embedding methods (e.g., Mikolov et al., 2013) provide valuable descriptive summaries of text corpora, they do not directly yield estimators with known frequentist properties or permit hypothesis testing without additional assumptions. Modern transformer-based architectures (BERT, GPT) advance contextual modeling but lack direct probabilistic interpretations necessary for rigorous inference.

By treating contexts as latent auxiliary variables—akin to classical data augmentation (Rubin, 1976; Tanner and Wong, 1987; Dempster et al., 1977)—our framework embeds texts within semantic environments, enabling explicit probabilistic inference. Our aggregation step, robust to outliers and heavy tails via ranks or quantiles, provides flexibility unmatched by embedding-based methods. Context augmentation thus bridges empirical process theory (van der Vaart, 1998a; Huber and Ronchetti, 2009; Newey, 1991; Chernozhukov et al., 2018) and NLP, addressing high-dimensional nuisance functions systematically (Wager and Athey, 2017; Athey et al., 2019).

4 Setup and Notation

We formalize the problem using notation tailored to text-valued data, distinguishing between observed strings, the contexts in which they are embedded, and the clause functions linking the two. Our

Variables	
$\mathbf{s}, s_i$	$\mathbf{s}$: d-vector of random strings; s_i: ith observed string
$\mathbf{x}_i$	Covariate vector (length p) for string s_i
n, n_c	n: number of strings; n_c: contexts generated per string
$\mathcal{E}$, $\mathcal{E}^{(k)}$	$\mathcal{E}$: set of conditioning events; $\mathcal{E}^{(k)}$: kth event
Latent Contexts	
$\mathbf{c}$, $\mathbf{c}_{j;i}$, $\mathbf{c}^{\mathcal{E}}$	$\mathbf{c}_{j;i}$: jth context for string s_i; $\mathbf{c}^{\mathcal{E}}$: contexts under event $\mathcal{E}$
Parameters and Models	
θ	Target estimand (scalar or vector)
$\widehat{\mathcal{M}}$, $\widehat{\mathcal{M}}^c$	$\widehat{\mathcal{M}}$: scoring model; $\widehat{\mathcal{M}}^c$: context-generation model
Operators & Functions	
$Cl(\mathbf{s}, \mathbf{c}, \mathcal{E}, \widehat{\mathcal{M}})$	Clause function: score of s in context $\mathbf{c}$ under $\mathcal{E}$
$Str(\mathbf{s}, \mathcal{E}, \widehat{\mathcal{M}})$	String-function: aggregates Cl over contexts to one statistic per string
$\mathcal{A}(\cdot)$, $T(\cdot)$	$\mathcal{A}$: aggregate over contexts; T: aggregate over strings to θ

Table 1: Key notation. Bold indicates vectors; scalar quantities in italics.

framework aggregates over contexts within each string and then across strings to the sample level, in a manner compatible with standard tools of statistical analysis. Table 1 provides a summary before the formal notation is introduced.

4.1 Data and Variables

We assume that the researcher observes n realizations of a d-dimensional vector-valued random variable $\mathbf{s}$, denoted $\{\mathbf{s}_i\}_{i=1}^n$, with $\mathbf{s}_i = (s_{i1}, s_{i2}, \ldots, s_{id})^\top$. This, along with a vector of p string-level covariates, $\mathbf{x}_i$, comprises the observed data.

We will measure how the string-probability changes across a set of conditioning events $\mathcal{E} = \{\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(q)}\}$. In the two sample problem, the event is sample membership; in the regression problem, it is presence of a predictor string. While we cannot evaluate the string-probability conditional on an event, we will use the large language model (LLM) $\widehat{\mathcal{M}}^c$ to generate *event-contexts* that capture the information in the event:

$$\mathbf{c}^{\mathcal{E}} = \mathbf{c} \mid \mathcal{E}, \widehat{\mathcal{M}}^c. \tag{1}$$

where $\mathbf{c}$ is a q-dimensional text-valued vector, with each element corresponding to an event. We will generate n_c contexts per string, with context j generated off string $\mathbf{s}_i$ as $\mathbf{c}_{j;i}$. These event-contexts will provide the auxiliary information connecting strings to events.

For the regression problem, we generate contexts off one set of strings, the predictor strings then evaluate them against another, the outcome strings. In the two-sample problem, the event–sample-membership–is operationalized by the strings in each event. A full-sample analysis, in which strings are evaluated against contexts generated by those same strings, induces a self-referential bias (analogous to

the "reflection problem" of Manski, 1993). We address this bias in two stages: cross-fitting eliminates the lead bias term and pivotality the second. First, though, we turn to constructing our estimand and estimate.

4.2 The Estimand

We construct our estimand, θ, in two steps. First, we aggregate over the context distribution to recover a string-level object, and then over the string distribution to obtain a population-level parameter. It will be useful at times to write this parameter as a functional of the joint distribution of strings and contexts by event, i.e. $\theta \doteq \theta(F_{\mathbf{s},\mathbf{c};\mathcal{E}})$.

Our basic object is the *clause function*, which evaluates each element of the d-dimensional string vector against each element of the q-dimensional context vector, and then maps each string-context pair to a real number.

$$Cl(\mathbf{s}, \mathbf{c}, \mathcal{E}, \widehat{\mathcal{M}}) : \mathcal{S}^d \times \mathcal{C}^q \mapsto \Re^{d \times q}$$

Evaluation will be done by a scoring model, $\widehat{\mathcal{M}}$, which need not coincide with the context-generation model, $\widehat{\mathcal{M}}^c$. In our examples, we generate and score with the same model, where the clause function returns the log string-probability. More generally, though, contexts can be generated off one LLM and evaluated off another, or contexts can be generated off of an LLM and then the resulting clauses may be scored by an auxiliary model that returns summary measures like sentiment or ideology.

In constructing θ, we first apply the clause function and aggregate over contexts with a user-specified aggregation operator $\mathcal{A}$,

$$Str(\mathbf{s}, \mathcal{E}, \widehat{\mathcal{M}}) = \mathcal{A} \circ Cl(\mathbf{s}, \mathbf{c}, \mathcal{E}, \widehat{\mathcal{M}}),$$

returning a string-level functional. Given the high-dimensional, and possibly erratic, nature of LLMs, this aggregation encompasses means but also robust alternatives, like medians, quantiles, ranks, trimmed means, or other robust location statistics. We then map this string-level object to our target parameter through the operator T, also user-specified,

$$\theta = T \circ Str(\mathbf{s}, \mathcal{E}, \widehat{\mathcal{M}})$$

The $T(\cdot)$ functional is also flexible, allowing for the same robust alternatives as $\mathcal{A}$ and smooth transformations thereof, such as logarithms. Formally, we will allow any transformation that admits a first-order Hadamard derivative, allowing for an asymptotic linearization.

4.3 The Estimate

Plug-in estimates are denoted with a hat. For observation $\mathbf{s}_i$ and its associated contexts $\{\mathbf{c}_{j;i}\}_{j=1}^{n_c}$, we construct the estimate

$$\widehat{Str}(\mathbf{s}_i, \mathcal{E}, \widehat{\mathcal{M}}) = \mathcal{A} \circ \{Cl(\mathbf{s}_i, \mathbf{c}_{j;i}, \mathcal{E}, \widehat{\mathcal{M}})\}_{j=1}^{n_c} \tag{2}$$

$$\doteq \widehat{Str}_i^{\mathcal{E}}. \tag{3}$$

where we adopt the compact notation in the second line when there is no ambiguity. The estimator for θ is then

$$\widehat{\theta} = T \circ \{\widehat{Str}_i^{\mathcal{E}}\}_{i=1}^{n}.$$

The variance in $\widehat{\theta}$ enters from two sources: the first within-string variance driven by contexts and the second cross-string variance over the sample.

5 Identification and Estimation

We next turn to four sets of results on the identification and estimation of our target parameter, θ. We first state a set of identification assumptions analogous to those from the literatures on missing data and causal inference (Rubin, 1976). Second, we derive the limiting distribution of our estimator in a way that characterizes bias induced through self-referential event-context generation. Third, we follow standard arguments in semiparametric theory and show that a subsampling strategy eliminates bias to first order (see, e.g. Chernozhukov et al., 2018; van der Vaart, 1998a). Finally, we provide two extensions. In the first, we extend results on higher-order efficiency by showing that cross-fitting a pivotal statistic can eliminate a second-order bias term—a new result in our setting (e.g., van der Vaart, 2014; Robins et al., 2008; Li et al., 2011; Pretorius and Swanepoel, 2018). In the second, we provide a bound balancing the rate requirement on the nuisance against the number of sampled contexts. We show how the rate can be met through a combination of computationally-expensive generated contexts and computationally inexpensive repeated cross-fits.

5.1 Identification

We connect strings to events using $\mathbf{c}^{\mathcal{E}}$ as auxiliary information. Under the following assumptions, differences in θ across separate sets of event-contexts imply differences in the underlying string-probabilities:

1. **Overlap:** For each event set $\mathcal{E}$,

$$\text{supp}(\mathbf{s} \mid \mathcal{E}) \cap \text{supp}(\mathbf{c}^{\mathcal{E}}) \neq \emptyset.$$

2. **Weak Ignorability:** For any two event sets $\mathcal{E}$ and $\mathcal{E}'$,

$$Cl(\mathbf{s}, \mathbf{c}^{\mathcal{E}}, \mathcal{E}, \widehat{\mathcal{M}}) = Cl(\mathbf{s}, \mathbf{c}^{\mathcal{E}}, \mathcal{E}', \widehat{\mathcal{M}}).$$

3. **Injectivity:** The mapping $\theta(\cdot) = T \circ \mathcal{A}$ is injective.

Lemma (Identification). *Under these assumptions,*

$$\theta(F_{\mathbf{s},\mathbf{c}^{\mathcal{E}}}; \mathcal{E}) \neq \theta(F_{\mathbf{s},\mathbf{c}^{\mathcal{E}'}}; \mathcal{E}') \quad \Rightarrow \quad F_{\mathbf{s}|\mathcal{E}} \neq F_{\mathbf{s}|\mathcal{E}'}.$$

Proof 1 See Appendix A.1.

Our identification assumptions adapt classic strategies utilizing latent variables and auxiliary information to LLM-generated contexts (Tanner and Wong, 1987; Rubin, 1976; Dempster et al., 1977). The Overlap condition is mild in this setting, since modern LLMs can generate virtually any string in response to a prompt—so the support of contexts will cover the support of observed strings. Weak Ignorability only requires that the clause score depend on the event's contexts, not on extraneous aspects of $\mathcal{E}$, which is weaker than the full conditional independence in Strong Ignorability (e.g. Rosenbaum and Rubin, 1984; Imbens and Rubin, 2015). We turn now to estimation and inference.

5.2 Estimation and Asymptotics

We present three main asymptotic results. First, we set up θ as the solution to an estimating equation and then generate a decomposition that isolates its sources of variance. Second, we discuss two standard ways to guarantee convergence and asymptotic unbiasedness: by a Donsker assumption or through a split-sample approach. Third, we show that, if the target statistic is pivotal, repeated cross-fitting generates a higher-order efficiency. Chernozhukov et al. (2018, Corollary 3.3 and subsequent text) analyze a fixed number of cross-fits; we show that one can let the number of repeated cross-fits grow with n and still control the error. We then balance this against the number of contexts per string n_c, needed to achieve a target estimation error rate. Rather than using plug-in corrections of higher order bias terms in U-statistics (e.g., van der Vaart, 2014; Robins et al., 2008; Li et al., 2011), we leverage pivotality to eliminate the leading term in an Edgeworth expansion—similar in spirit to Hall (1992). Our closest antecedent is Pretorius and Swanepoel (2018), who eliminate the lead bias term by using subsampled plug-in estimates in a Cornish-Fisher expansion. Our contribution extends this result to a general, semiparametric setting with application to text-valued data.

5.3 Setup and Influence Function Decomposition

We take as the target parameter, θ, defined above, and the nuisance the distribution of event-conditioned contexts, $\eta = F_{c^{\mathcal{E}}}$ and $\widehat{\eta} = \widehat{F}_{c^{\mathcal{E}}}$. The complete data consist of observed strings and generated contexts, $\{Z_i\}_{i=1}^n = \{\mathbf{s}_i, \{\mathbf{c}_{j;i}^{\mathcal{E}}\}_{j=1}^{n_c}\}_{i=1}^n$, with $\widehat{F}_{c^{\mathcal{E}};i}$ the estimated distribution of $\mathbf{c}_{j;i}^{\mathcal{E}}$. The following assumptions allow us to write θ as the solution to an estimating equation and then develop a von Mises expansion around it:

Assumption 1 *1. The string-level statistics $\widehat{Str}_i^{\mathcal{E}}$ are i.i.d. across observations, and the contexts used to construct each $\widehat{Str}_i^{\mathcal{E}}$ are i.i.d. conditional on the string $\mathbf{s}_i$.*

2. The parameter $\theta \in \Re^d$ solves

$$\Phi(\theta, \eta = \eta_0) = \mathbb{E}[\phi_{\theta,\eta=\eta_0}(Str_i^{\mathcal{E}})] = 0.$$

with Jacobian $\partial_\theta \Phi(\theta, \eta)$ invertible uniformly in a neighborhood of θ_0.

3. The map $\phi_{\theta,\eta}$ is Hadamard differentiable in η and pathwise (Gateaux) differentiable in θ, both tangentially to a subset of $L_2(P)$ that contains the closure of the image of $\mathcal{A}\dot{C}l$.

4. The model is differentiable in quadratic mean in θ, uniformly in a neighborhood of (θ_0, η_0).

5. The process $\widehat{\Phi}(\theta, \eta) = \frac{1}{n}\sum_{i=1}^n \phi_{\theta,\eta}(\widehat{Str}_{i;n_c}^{\mathcal{E}})$ is stochastically equicontinuous uniformly in a neighborhood of θ_0.

These mirror standard assumptions in Z-estimation see, van der Vaart (e.g. 1998a, Thm. 25.57) or Chernozhukov et al. (e.g. 2018, Sec 2.1), giving us a decomposition into three sources of error: sampling error, context-level error, and self-referential bias.

Result 1 Under Assumption 1

$$\begin{aligned}
\sqrt{n}(\widehat{\theta} - \theta_0) &= \sqrt{n}\Big\{ \underbrace{(\widehat{\theta}_{Str,\eta_0} - \theta_0)}_{\textit{Sampling variation}} + \underbrace{(\widehat{\theta}_{\widehat{Str},\eta_0} - \widehat{\theta}_{Str,\eta_0})}_{\textit{Context-level error}} + \underbrace{(\widehat{\theta}_{\widehat{Str},\widehat{\eta}} - \widehat{\theta}_{\widehat{Str},\eta_0})}_{\textit{Self-referential bias}} \Big\} \\
&= (\partial_\theta \Phi(\theta_0, \eta_0))^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big\{ \underbrace{\phi_{\theta_0,\eta_0}(Str_i^{\mathcal{E}})}_{\textit{Sampling variance}} \\
&\quad + \underbrace{\partial_{Str}\phi_{\theta_0,\eta_0}(Str_i^{\mathcal{E}}) \left[\widehat{Str}_{i;n_c}^{\mathcal{E}} - Str_i^{\mathcal{E}}\right]}_{\textit{Context-level error}} \\
&\quad + \underbrace{\partial_\eta \phi_{\theta_0,\eta_0}(\widehat{Str}_{i;n_c}^{\mathcal{E}}) \cdot \left\| \widehat{F}_{c^{\mathcal{E}};i} - F_{c^{\mathcal{E}};i} \right\|_\infty}_{\textit{Self-referential bias}} \Big\} + o_p(1)
\end{aligned} \tag{4}$$

Proof. *See Appendix A.2.*

This decomposition helps us distinguish between different directions in which the estimate may fluctuate around the population minimizer: sample-variation, context-variation, and estimation error on the event-context distribution. The first two influence functions can be handled through standard parametric arguments. The third one involves a nonparametric element, which may induce a persistent bias term.

The following assumption will guarantee a well-behaved limiting distribution:

Assumption 2 *1. Each influence function in the first order decomposition satisfies a Lyapunov condition.*

2. The string and error estimate converge pointwise as $\widehat{Str}_i^{\mathcal{E}} \xrightarrow{\text{p}} Str_i^{\mathcal{E}}$ *and, at each context* $\mathbf{c}$, $\widehat{F}_{c^{\mathcal{E}}}(\mathbf{c}) \xrightarrow{\text{p}} F_{c^{\mathcal{E}}}(\mathbf{c})$.

3. The error on the event context distribution vanishes uniformly over event-contexts and strings as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\eta \phi_{\theta_0,\eta_0}(\widehat{Str}_{i;n_c}^{\mathcal{E}}) \cdot \left\| \widehat{F}_{c^{\mathcal{E}};i} - F_{c^{\mathcal{E}};i} \right\|_\infty \xrightarrow{p} 0$$

Result 2 (Limit Theorem) Under Assumptions 1 and 2,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, \Omega),$$

where the asymptotic variance is given by

$$\Omega = (\partial_\theta \Phi(\theta_0, \eta_0))^{-1} \cdot \mathbb{E}\left[\phi_{\theta_0,\eta_0}(Str_i^{\mathcal{E}}) \phi_{\theta_0,\eta_0}(Str_i^{\mathcal{E}})^\top\right] \cdot (\partial_\theta \Phi(\theta_0, \eta_0))^{-1\top}.$$

Proof. *See Appendix A.3.*

Controlling the first two influence functions is a standard parametric problem. The third term requires a bit more care, as the nonparametric element may not achieve the parametric rate. Several strategies here include a classical no-bias condition for asymptotic normality (van der Vaart, 1998b, Condition 25.52), constraining the complexity of the function to be in a Donsker case.

In many common machine learning cases, this Donsker assumption is provably invalid. Recent years have found renewed interested in subsampling methods (Bickel, 1982; van der Vaart, 1998a; Politis et al., 1999; Chernozhukov et al., 2018), where disjoint sets of data are used to estimate the nuisance and conduct inference. Under this strategy, the rate requirement on the estimation error on the

nuisance term can be reduced from $o_p(n^{-1/2})$ to $o_p(n^{-1/4})$, such that the Assumptions 2 can hold under

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{F}_{c^{\mathcal{E}};i} - F_{c^{\mathcal{E}};i}\right\|_{\infty} = o_p(n^{-1/4}).$$

Key to achieving this rate is using different subsets of the data to estimate $\widehat{\theta} = \widehat{F}_{c^{\mathcal{E}}}$ and conduct inference on θ. In the regression problem, we do this naturally: contexts are generated off of predictor strings, while inference is done using a distinct set of outcome strings. In the two-sample problem, where strings are evaluated against contexts by strings in each group, we adjust for self-referential bias using through subsampling: we split the data in equal-sized subsamples, $\mathcal{I}^1, \mathcal{I}^2$. We use $\mathcal{I}^1$ for generating contexts, $\mathcal{I}^2$ for inference on θ, and then cross-fit, where we swap roles and average. We then repeatedly cross-fit, taking the mean or median over cross-fits in order to average over the peculiariaties of a given split, and then take a mean or median over repeated cross-fits to recover location and uncertainty estimates (Chernozhukov et al., 2018). The cross-fitting approach is motivated by a desire to reduce the rate-requirements on nonparametric nuisance terms. We turn next to a further refinement.

5.4 Higher-Order Accuracy and Computational Efficiency via Pivotal Statistics

The generation of contexts and evaluation of clauses is computationally intensive, while the downstream steps–regressions or averaging over contexts–are much less costly. To connect these tradeoffs with the sample size, we provide a result with two components. First, we show that when our estimator is pivotal, the nuisance-estimation rate can be halved, from $n^{-1/4}$ to $n^{-1/8}$. We then bound the lead bias, connecting the (expensive) number of generated contexts, n_c, and the inexpensive number of repeated cross-fits or bootstraps $M = R$ or B. This repeated cross-fitting or bootstrapping strategy can be used to reduce the number of contexts needed to guarantee the convergence rate necessary for valid inference.

Assumption 3 (Pivotal Limit) Beyond Assumptions 1–3, suppose:

1. ***(Pivot)*** *There exists a rate $r_n \to \infty$ and a known distribution G, such that*

$$r_n(\widehat{\theta}_n - \theta_0) \xrightarrow{d} G,$$

   and G involves no unknown parameters.

2. ***(Cramér's Condition on** G**)** *The characteristic function of G is non-zero in a neighborhood of the origin.*

Result 3 (Convergence Rates under Pivotality) Under Assumptions 1–3 and the Pivotal Limit condition, suppose we estimate θ by either R repeated cross-fits or B bootstrap replicates, drawing n_c contexts per string. Then:

1. ***Second-order convergence via pivot.*** *Pivotality reduces the nuisance convergence rate from $n^{-1/4}$ to $n^{-1/8}$.*

2. ***Uniform bias bound.*** *The lead bias term can be controlled uniformly as*

$$\text{Bias} = O_p\Big(\sqrt{\tfrac{\ln n_c \ \text{lnln}\, M}{n_c\, M}}\Big),$$

where $M = R$ (cross-fits) or $M = B$ (bootstrap).

Therefore, driving the bias below $n^{-\gamma}$ (with $\gamma = \frac{1}{4}$ in the standard semiparametric setting or $\gamma = \frac{1}{8}$ under pivotality), requires

$$(n_c\, M)^{1+\epsilon} \gg n^{2\gamma} \quad \text{for some small } \epsilon > 0.$$

Proof. *See Appendix A.4.*

Interpretation and Computational Guidance. These results supply a unified theoretical rationale for balancing the computational burden of costly context generation against virtually free inference repetitions. Once contexts are drawn and scored, increasing the number of cross-fits R or bootstrap replicates B incurs negligible overhead yet yields equivalent bias control. Under pivotality, the required convergence rate halves—from $n^{-1/4}$ to $n^{-1/8}$—so that a moderate n_c suffices when paired with a sufficiently large R or B.

Crucially, these conclusions hold for any pivot, including a normal limit but also the F or chi-square, making the approach broadly applicable in text settings where effect sizes are on a transformed scale. In practical terms, inexpensive inference repetitions can improve the accuracy and help achieve second-order accuracy under minimal computational expense.

6 The Two-Sample Problem via Context Augmentation

We apply context augmentation to test whether two groups of string-valued data differ in ways captured by an LLM. We denote as c^G clauses that could contain strings in group $G \in \{A, B\}$. The clause function is then

$$Cl(s, c, \mathcal{E} = G, \widehat{\mathcal{M}}) = \log \Pr(s|c^G),$$

the log probability of observing string s in context c^G, as scored by the model. We then construct the string function as a two-vector, with elements given by the mean log probability in each group,

$$Str(s, \mathcal{E} = \{A, B\}, \widehat{\mathcal{M}}) = \begin{bmatrix} \mathbb{E}_{\mathbf{c}^A}[\log \Pr(s|c^A)] \\ \mathbb{E}_{\mathbf{c}^B}[\log \Pr(s|c^B)] \end{bmatrix}.$$

Define

$$\theta^A = \mathbb{E}(\mathbb{E}_{\mathbf{c}^A}[\log \Pr(s|c^A)] - \mathbb{E}_{\mathbf{c}^B}[\log \Pr(s|c^B)]|s \in A)$$

and θ^B similarly for group B. Our parameter is then

$$\theta = \frac{\theta^A - \theta^B}{\sqrt{\text{Var}(\theta^A - \theta^B)}}$$

and we test the null hypothesis $\mathcal{H}_0 : \theta = 0$.

Estimation. We estimate the test statistic using a repeated cross-fitting procedure designed to eliminate self-referential bias. First, we randomly partition the data into two folds: one fold for generating contexts and the other for conducting inference. For each string in the inference fold, we evaluate its probability under the contexts in group A and group B generated from strings in the other fold. We then implement a t-statistic, the plug-in estimate for our statistic given above.

This process is repeated in both directions (i.e., swapping the roles of the two folds), and the resulting t-statistics are averaged and rescaled to correct for repeated sampling. To stabilize results and smooth out any randomness from the fold assignment, we repeat this procedure and average over 25 repeated cross-fits. For details, see Appendix B.

6.1 Empirical Demonstration

We evaluate our approach using synthetic data generated by GPT-4 across five categories (animals, body parts, cities, food, plants). For each category, we sample 100 multi-word strings and generate 10 contexts per string using FLAN-T5-XXL.

Contexts are generated using a procedure summarized in Table 2. Each stage builds on the previous, allowing us to build up contexts. The definition prompt enforces some similarity between the left and right contexts. Constructing, and storing, left- and right-contexts separately makes evaluation easier, making it simple to loop strings between the contexts on each side.

Results. We assess our method in two ways. First, for null calibration, strings from the semantic category are randomly split into two groups. The resulting p-values are plotted in a QQ-plot (Fig. 1), which adheres closely to the uniform(0,1) line, indicating valid Type I error control.

Prompt Stage	Description and Example Prompt
Definition	Prompt the model to define the target string, generating a short, standalone explanation. *Example:* `Provide a clear and concise definition of the word 'leopard'.`
Left Context ($\times n_c$)	Using the generated definition, prompt the model to generate sentence fragments that could logically precede the target string. *Example:* `Using the definition 'A large, spotted feline predator.', provide a sentence fragment that could logically come before the word 'leopard'.`
Right Context (per left)	For each left-hand context, prompt the model to generate a continuation that could follow the word, completing the sentence. *Example:* `Using the definition 'A large, spotted feline predator.', provide a sentence fragment that could logically come after the phrase 'After stalking its prey, the leopard'.`
Final Template	Assemble the final clause using a placeholder for substitution. `After stalking its prey, the «<STR»> pounced silently from the grass.`

Table 2: Context generation prompting pipeline used in simulation, structured to reflect the clause-building process in the implementation code.

Second, for power analysis, we compute t-statistics from pairwise comparisons across different categories. These are summarized in violin plots (Fig. 2). The results show strong and systematic differences, with most test statistics exceeding conventional thresholds ($|t| > 2$), demonstrating that the method is sensitive to meaningful semantic distinctions. Analytic variances closely matched bootstrap estimates, supporting the robustness of the inference.

7 Regression via Context Mediation

We consider a regression setting where both the predictor s_x and outcome s_y are text. Our goal is to measure whether observing s_x increases the probability of subsequently observing s_y. The effect of an input string s_x on an outcome string s_y is assumed to occur through latent contexts, which we generate using a large language model (LLM). These contexts, derived from the informative s_x, embed both the semantic and syntactic features of the text in a high-dimensional space. To assess how the characteristics of s_x influence s_y, we compare the likelihood of s_y when paired with the original (informative) s_x versus when s_x is replaced by a non-informative variant $\tilde{s}_x$. This transformation highlights the contribution of s_x's semantic or structural content. Because there is inherent ambiguity in removing information from text, we employ three baselines: masked, where each token in s_x is replaced by a placeholder, preserving token count and general syntactic structure; shuffled, where words in s_x are randomly permuted, thereby disrupting the original word order while retaining lexical

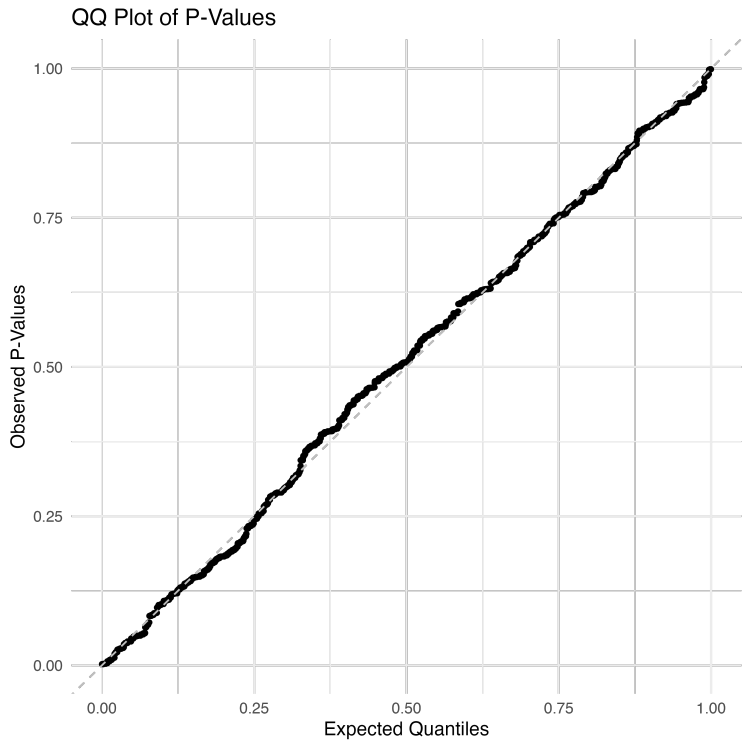

Figure 1: QQ-plot of p-values from within-category comparisons. The adherence to the diagonal confirms valid null calibration.

items; and jabberwocky, where content words are replaced with invented nonsense words, preserving an approximate syntactic scaffold.

Let $\{(s_{x,i}, s_{y,i})\}_{i=1}^{n}$ be n observations of predictor–outcome text pairs. For each pair, the LLM generates a set of latent contexts $\{c_{j;i}\}_{j=1}^{n_c}$ by conditioning on the informative s_x. Within the same set of contexts $\{c_{j;i}\}$, we compute the LLM-assessed probability of s_y when the predictor is replaced by each variant:

$$\Pr(s_y \mid c_{j;i}, s_x^{\text{inf}}), \quad \Pr(s_y \mid c_{j;i}, s_x^{\text{mask}}), \quad \Pr(s_y \mid c_{j;i}, s_x^{\text{shuffle}}), \quad \Pr(s_y \mid c_{j;i}, s_x^{\text{jabberwocky}}).$$

We define a clause function

$$Cl\big(s_y, s_x^{\nu}, c_{j;i}, \mathcal{M}\big) \;=\; \log \Pr(s_y \mid s_x^{\nu}, c_{j;i}, \mathcal{M}),$$

where $\nu \in \{\text{inf}, \text{mask}, \text{shuffle}, \text{jabberwocky}\}$ indexes the variant of s_x used. By holding the context fixed across variants, the differences in Cl between s_x^{inf} and each noninformative version isolate the impact of the textual features that have been modified. While not a formal causal mediation effect, as any manipulation is model-based, an analogue of the indirect effect can be expressed as the ratio

$$\frac{\Pr(s_y \mid s_x^{\text{inf}}, c_{j;i})}{\Pr\big(s_y \mid \tilde{s}_x, c_{j;i}\big)},$$

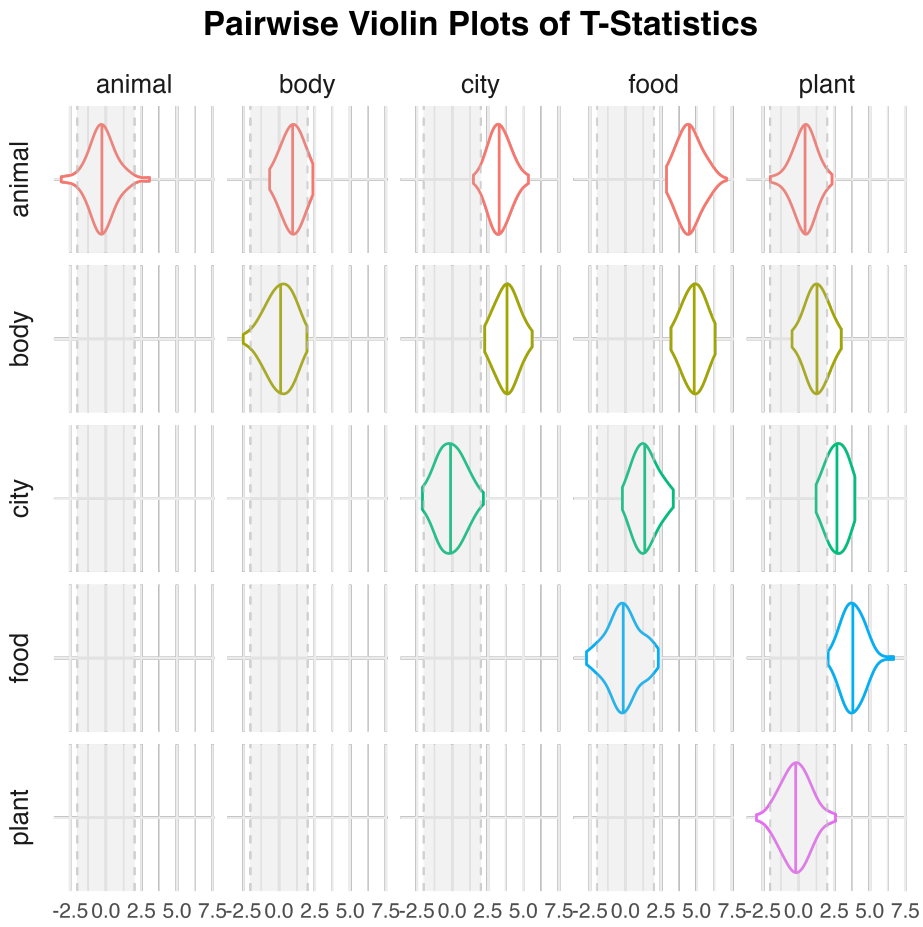

Figure 2: Violin plots of pairwise t-statistics for cross-category comparisons. Shaded regions represent non-significant t-statistics ($|t| < 2$). The distributions highlight clear differentiation among semantic categories.

where $\tilde{s}_x$ may represent any of the non-informative variants. Comparing these ratios for the different baselines quantifies the extent to which semantic content, word order, or lexical information contribute to the model's prediction of s_y. This setup follows the general structure of mediation analysis (Pearl, 2012; Imai et al., 2010), where the generated contexts serve as mediators linking the predictor to the outcome. By systematically altering the predictor while keeping the latent context fixed, we evaluate how semantic and syntactic modifications alter the probability of the outcome string.

To formally aggregate these effects, we implement a regression model of the form

$$\log \Pr(s_{y,i} \mid s_{x,i}^{\text{inf}}, c_{j;i}) = \beta_0 + \beta_1 \log \Pr(s_{y,i} \mid s_{x,i}^{\text{shuffle}}, c_{j;i}) + \beta_2 \log \Pr(s_{y,i} \mid s_{x,i}^{\text{jabberwocky}}, c_{j;i}) + \beta_3 \log \Pr(s_{y,i} \mid s_{x,i}^{\text{mask}}, c_{j;i}) + \mathbf{x}_i^\top \gamma + z_i^\top b + \varepsilon \tag{5}$$

where standard errors can then be clustered by string.

7.1 Empirical Demonstration

For a simple illustration, we generated a dataset consisting of 10 text-based predictor–outcome pairs, where predictors are movie-related phrases and outcomes either correspond to a semantically related movie review statement or a randomly assigned geographical fact (placebo condition). Table **??** presents the first three strings in each group.

From each predictor, we generate the three different baselines: the syntactic baseline, generated by scrambling word order while preserving vocabulary; the lexical baseline, which replaces each word or token with a non-informative mask token; and the semantic baseline, which uses a jabberwocky transformation to replace content words with nonsensical placeholders while retaining syntactic form. For example, given the predictor "The soundtrack was mesmerizing," the shuffled version might be "was mesmerizing the soundtrack," the masked version "<mask> <mask> <mask> <mask>," and the jabberwocky version "The ziggflorp was blarptastic." We then estimate the model in Equation 7, including random effects for each string.

The regression results, shown in Table 3, highlight the distinction between syntactic and semantic contributions to LLM-based text prediction. In both the informative and placebo settings, the shuffled predictor remains significant, confirming that syntactic coherence alone can contribute to predictability—if a sentence is grammatically well-formed, another well-formed sentence is more likely to follow, even in cases where the meaning is unrelated. However, a key distinction emerges with the jabberwocky transformation: in the informative case, where the outcome remains within the same semantic domain as the predictor, the jabberwocky effect is strongly positive and significant, suggesting that even when lexical content is removed, the syntactic structure of the predictor still carries meaningful information. In contrast, in the placebo setting, where the outcome is independent of the predictor's meaning, the jabberwocky transformation ceases to have any significant effect. This divergence indicates that the informative regression captures a genuine semantic relationship, whereas the placebo regression isolates only syntactic structure. The masked condition, which removes all lexical content while preserving the general length and shape of the sentence, remains significant in both cases, but with a much larger coefficient in the placebo setting, suggesting that when no real information is available, the model defaults to prior expectations rather than making meaningful inferences. Together, these findings confirm that while syntactic structure alone can generate statistical dependencies between texts, only in the informative regression does the model identify an effect that is truly semantic.

Table 3: Regression Table

	Informative	Placebo
	(1)	(2)
Syntactic Baseline (shuffled)	0.399**	0.278***
	(0.197)	(0.096)
Semantic Baseline (jabberwocky)	0.472***	−0.020
	(0.127)	(0.230)
Masked Baseline (mask)	0.339*	0.723***
	(0.196)	(0.194)
Constant	1.782*	−0.158
	(1.082)	(0.959)
Fixed Effects	Yes	Yes
Observations	100	100
R^2	0.867	0.860
F Statistic (df = 12; 87)	47.346***	44.402***

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the string level.

8 Replication: Estimating the Effect of a Treatment Given Text-Valued Outcomes

To evaluate the effectiveness of context augmentation for testing text-based treatment effects, we replicate Experiment 3 from Egami et al. (2022). In that study, respondents are randomly assigned to one of two prompts, each describing a 28-year-old man who illegally entered the U.S. with one of two strings included. The treatment condition includes the sentence "The man has two prior prison sentences (one for a violent crime) and has previously been deported," while the control condition includes "The man has no prior criminal history and has never been imprisoned." Participants respond to the question: *"Should this person go to jail?"* with free-text justifications. After removing observations with no open-ended response, we have $n = 1034$ with 518 treated and 516 control.

In the original study, the authors split the data into estimation and inference sets, then fit a Structural Topic Model (STM; Roberts et al., 2014) to the responses in the estimation set, using the treatment variable as a predictor for topic prevalence. Topics identified in the estimation set were then applied to the held-out inference set, with inference performed on 11 selected topics.

We apply our context augmentation method to the same data, preserving the train-test split. For each response, we construct prompts based on the assigned treatment or control vignette, appending the respondent's justification to the base prompt followed by a question-answer format:

```
You were asked, "Should this person go to jail?" You replied:
[response]. Why did you say this?
```

This composite prompt generates contexts evaluated through our context augmentation approach.

We compare context augmentation and STM on two dimensions: power and interpretability. Using Hotelling's T^2 statistic, adjusted for differences in degrees of freedom via the Wilson-Hilferty approximation, STM yields a test statistic of 38.8 on 10 degrees of freedom, whereas context augmentation yields 104.6 on 1 degree of freedom. Transformed to a common z-scale, context augmentation achieves 8.22, compared to STM's 3.95, indicating greater statistical precision.

This improved performance stems from context augmentation leveraging additional external linguistic knowledge from a pre-trained LLM, enabling fine-grained analysis at the context level. STM, in contrast, provides a thematic summary by aggregating word co-occurrences, offering insights into broader themes rather than specific context-level contributions.

We present leave-one-out context analyses (Tables table C.1–table C.2) in Appendix C. Contexts shown in Table table C.1, typically sympathetic or lenient, sharpen the observed treatment effect—omitting them attenuates effect size. Conversely, the punitive contexts in Table table C.2 attenuate the effect, as their omission increases the magnitude of the effect. These insights highlight precisely which narrative framings influence the overall effect.

Table C.4 further refines this analysis by providing string-level effect estimates. The top strings amplify the effect, depicting lenient or deportation-only responses (e.g., "They should simply be deported...this is simply a minor violation"; $t \approx 10.58$). Neutral strings near zero reflect ambivalent or compromise positions ("I don't even know the moral ethics of illegal entry anymore... exercise caution"; $t \approx 0.00$). The most negative strings depict the individual as a dangerous recidivist, strongly attenuating the effect ("This person has clearly demonstrated a pattern of violence...negative effect on national security"; $t \approx -12.33$).

Note that in estimating an effect for each observed string, we are utilizing string-level counterfactual estimates: we estimate the mean log-probability of each string under both its observed and counterfactual treatment condition, using contexts as auxiliary information. This allows for string-level predictions and analyses, and leaves a path open to future methods targeting causal estimands.

In this study, though, we restrict ourselves to descriptive and parametric parameters, in order to first get these correct. In this spirit, we turn next to a setting from psychology where outcomes and predictors are both textual, which is a new area of use in statistical inference.

9 Regression Analysis: Interactive Alignment and Integrative Repair

To illustrate the utility of context augmentation in modeling textual interactions, we apply our method to dialogue data from the DeliData corpus, a resource comprising multi-party deliberations designed explicitly for studying group decision-making dynamics (Karadzhov et al., 2023). Our analytical framework relies on the foundational psychological theories of Pickering and Garrod (Pickering and Garrod, 2004, 2013), who characterize dialogue as a joint activity involving tightly coupled processes of language production, comprehension, and prediction (ADD pickering and garrod 2007). Central to their theory is the concept of *interactive alignment*, wherein interlocutors' linguistic representations spontaneously synchronize across lexical, syntactic, and semantic dimensions. Such alignment enables efficient communication by minimizing cognitive demands, yet dialogues frequently encounter moments of misalignment, necessitating *integrative repair*. Such repair manifests through reliance on predictable linguistic structures, facilitating rapid restoration of common ground (Pickering and Garrod, 2004).

Empirical studies of deliberative dialogue across diverse contexts highlight the importance of alignment and repair mechanisms. Research on police encounters (Rho et al., 2023), gender dynamics in deliberative settings (Mendelberg et al., 2014), and semantic evolution in political discourse (Rodriguez et al., 2023) underscores how linguistic alignment shapes conversational trajectories. Yet, these studies often rely on human-coded interpretations of text, little or no attention paid to uncertainty estimation, or embeddings models that cannot capture syntactic attributes of text. In contrast, our context augmentation approach explicitly integrates linguistic variability into a formal inferential framework by leveraging large language model (LLM)-generated contexts, thus addressing methodological gaps noted by Le Mens et al. (2023).

Using our context augmentation regression, we find evidence of integrative repair within the DeliData, manifested as increased reliance on predictable syntactic structures when confronted with a shift either into our out of consensus within the pair. Specifically, we consider utterance transitions within dialogues, modeling each subsequent utterance's log-probability conditional on its predecessor. To disentangle syntactic, semantic, and lexical contributions, we use baseline transformations: a *syntactic baseline* (shuffled predictor), a *semantic baseline* (jabberwocky predictor), and a *lexical baseline* (masked predictor), isolating separate linguistic dimensions. These transformations isolate linguistic dimensions while preserving core textual properties.

The utterance, $s_{y;i}$ follows $s_{x;i}$ in the dialogue. We will also incorporate moderators that are measured between the $s_{x,i}$ statement and the one immedately previous. After fitting a baseline with no

moderator, we will consider three moderators: a time variable from 0 to 1 of how much of the dialogue is complete; a consensus indicator for whether both speakers agreed prior to the predictor utterance; and a shift indicator, for whether consensus was either established or lost immediately prior to the predictor utterance. ous We take this shfit indicator as a measure of integrative repair, as there is a shift in understanding in the previous period, and we expect alignment afterwards.

Using random effects for string ($u_{obs(i)}$), a dialogue-level random effect and time trend ($v_{dialogue(i)} + time \times v_{dialogue(i)}$) and an effect for each speaker $w_{speaker(i)}$, we estimate the following model.

$$
\begin{aligned}
&\log \Pr(s_{y,i} \mid s^{inf}_{x,i}, c_{j;i}) \\
&\quad = \beta_0 + \beta_1 \log \Pr(s_{y,i} \mid s^{shuffle}_{x,i}, c_{j;i}) + \beta_2 \log \Pr(s_{y,i} \mid s^{jabberwocky}_{x,i}, c_{j;i}) + \beta_3 \log \Pr(s_{y,i} \mid s^{mask}_{x,i}, c_{j;i}) \\
&\qquad + \gamma_1 consensus_i + \gamma_2 shift_i + \gamma_3 moderator_{ij} \\
&\qquad + \delta_1 (\log \Pr(s_{y,i} \mid s^{shuffle}_{x,i}, c_{j;i}) \times moderator_{ij}) + \delta_2 (\log \Pr(s_{y,i} \mid s^{jabberwocky}_{x,i}, c_{j;i}) \times moderator_i) \\
&\qquad + \delta_3 (\log \Pr(s_{y,i} \mid s^{mask}_{x,i}, c_{j;i}) \times shift_i) \\
&\qquad + u_{obs(i)} + v_{dialogue(i)} + time \times v_{dialogue(i)} + w_{speaker(i)} + \epsilon_i. \qquad (6)
\end{aligned}
$$

The results, summarized in Table 3, confirm several important theoretical predictions. First, all three baselines significantly predict future utterances, underscoring the combined contribution of syntactic, semantic, and lexical structures to linguistic alignment. Second, we observe no significant linear trend across dialogue exchanges nor a simple direct effect of consensus state. Most crucially, however, the interaction between consensus shifts and the syntactic baseline is strongly positive and statistically significant ($p < 0.01$). This robust interaction indicates that shifts between consensus and dissensus systematically amplify reliance on predictable syntactic forms. The interaction effects remain symmetric, demonstrating comparable magnitudes whether shifting toward agreement or disagreement, consistent with the integrative repair hypothesis of Pickering and Garrod (2004).

By contrast, semantic and lexical baselines reveal negligible interaction effects, indicating that integrative repair primarily involves syntactic adjustment rather than semantic or lexical modification. The asymmetry underscores that syntactic structure provides a crucial scaffolding mechanism facilitating rapid realignment during conversational disruptions. These findings are robust under stringent model specifications and clustering adjustments at the dialogue and speaker level.

Methodologically, our analysis offers significant innovations over existing approaches to text-based inference. By explicitly modeling the latent variability in LLM-generated contexts, context augmentation preserves statistical properties (asymptotic normality, unbiasedness) while accounting for

linguistic complexity. Unlike fixed embedding methods (Rodriguez et al., 2023) or purely descriptive semantic analyses (Le Mens et al., 2023), our approach provides a fully probabilistic framework accommodating textual uncertainty. This innovation represents a major advancement in bridging modern language modeling with classical empirical process theory.

Table 4: Regression Table

Dependent variable: Log Probability of Utterance given Previous Utterance					
Moderator:	None	Time	Consensus	Shift	Agree/Disagree
Syntactic Baseline	0.200***	0.200***	0.195***	0.189***	0.189***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Semantic Baseline	0.158***	0.158***	0.161***	0.160***	0.160***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Lexical Baseline	0.425***	0.425***	0.417***	0.423***	0.423***
	(0.015)	(0.015)	(0.017)	(0.018)	(0.018)
Moderator		0.108	0.514	0.390	
		(0.550)	(0.393)	(0.364)	
Syntactic Baseline x Moderator		−0.003	0.021	0.045***	
		(0.019)	(0.014)	(0.013)	
Semantic Baseline x Moderator		0.0003	−0.016	−0.008	
		(0.021)	(0.015)	(0.014)	
Lexical Baseline x Moderator		0.001	0.040	0.006	
		(0.052)	(0.037)	(0.034)	
Agree					0.838*
					(0.485)
Disagree					−0.028
					(0.471)
Syntactic Baseline x Agree					0.052***
					(0.017)
Semantic Baseline x Agree					−0.014
					(0.019)
Lexical Baseline x Agree					0.043
					(0.046)
Syntactic Baseline x Disagree					0.037**
					(0.017)
Semantic Baseline x Disagree					−0.003
					(0.018)
Lexical Baseline x Disagree					−0.028
					(0.045)
Constant	−1.391***	−1.387***	−1.504***	−1.498***	−1.498***
	(0.162)	(0.164)	(0.183)	(0.189)	(0.189)
Observations	7,390	7,390	7,390	7,390	7,390
Akaike Inf. Crit.	7,989.413	7,994.170	7,992.423	7,984.979	7,990.185
Bayesian Inf. Crit.	8,058.492	8,090.880	8,089.133	8,081.690	8,114.527
Note x					*p<0.1; **p<0.05; ***p<0.01
					Random effects included for dialogue, speaker, and utterance.

10 Discussion

Our aim is to connect LLMs to simple yet essential statistical models in a way that facilitates valid statistical inference. With this in hand, a variety of different directions open. The key challenge we

address is that text-valued data lack a natural probability structure, making standard inferential techniques difficult to apply. Context augmentation provides a principled approach by leveraging LLM-generated contexts as auxiliary information, which allows us to recover structured dependencies from unstructured text. This framework permits estimation of quantities such as likelihood ratios, regression coefficients, and other inferential targets using well-established statistical techniques.

A key technical challenge in implementing context augmentation lies in the engineering complexity of working with large language models. Unlike classical parametric estimation, where the computational burden is often minimal, generating context-augmented probabilities involves multiple layers of modeling choices. The choice of prompts, temperature settings, number of contexts per observation, and even minor variations in sampling strategies can influence the results. Importantly, our theoretical framework remains robust across these variations—identification and estimation hold under different prompt formulations and generation settings, provided that the required support and ignorability assumptions remain valid. However, the computational cost of generating context samples remains a nontrivial consideration. Unlike conventional parametric estimators, which can be computed in seconds, generating sufficiently rich sets of contexts requires substantially more computational effort. In practice, this creates a tradeoff between statistical precision and computational efficiency, particularly in large-scale applications.

The two-sample and regression settings highlight deeper connections between context augmentation and causal inference. In the two-sample problem, we effectively estimate a counterfactual likelihood for each string—how likely it would have been under the other group's contexts. This mirrors standard counterfactual-based approaches in causal inference, where potential outcomes under different treatments are estimated. Likewise, the regression framework exhibits an explicit mediation-like structure: the LLM-generated contexts serve as mediators that link predictor and outcome text. The core assumptions—support inclusion and weak ignorability—are foundational in causal inference, reinforcing the deep conceptual link between context augmentation and causal modeling. While our focus has been on inference rather than causal identification, these structural similarities suggest that context augmentation could be extended to formal causal estimands in future work.

Future research can move in several directions. One immediate avenue is to extend this framework to nonparametric settings, where context augmentation could be used in kernel methods, density estimation, or clustering applications. Another is to apply context augmentation to causal inference more explicitly, using LLMs to generate counterfactual text sequences or define text-based instrumental variables. Further methodological refinements could involve Bayesian extensions, where priors over

context distributions are incorporated directly into the estimation procedure. Finally, a deeper integration with robust statistical methods, such as quantile-based estimators, could further enhance the resilience of this approach to outliers and model misspecification.

11 Conclusion

This paper develops a framework for integrating LLMs with statistical inference via context augmentation, a method that introduces model-generated contexts as auxiliary structures to enable valid estimation. By leveraging generated contexts, we recover structured probabilistic relationships between text observations while maintaining compatibility with standard inferential techniques. This approach allows for hypothesis testing, regression analysis, and likelihood-based estimation directly on text data, bridging contemporary deep learning with classical statistical methodology.

We establish theoretical conditions under which context augmentation leads to valid inference, showing that under support inclusion and ignorability conditions, estimators exhibit asymptotic normality with variance contributions from both text sampling and context generation. Empirical applications to two-sample testing and text regression demonstrate the method's utility in extracting structured dependencies from unstructured text, with clear distinctions between syntactic and semantic effects.

More broadly, this work contributes to the growing intersection of statistical inference, natural language processing, and causal analysis. By treating generated contexts as structured latent variables, context augmentation offers a principled alternative to fixed embedding-based approaches, allowing for more flexible, model-based inference on text-valued data. As language models continue to improve, we anticipate that context augmentation will provide a foundation for increasingly sophisticated applications in both social science and computational statistics.

References

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis 31*(3), 337–351.

Athey, S., J. Tibshirani, and S. Wager (2019, April). Generalized random forests. *Annals of Statistics 47*(2), 1148–1178.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Berger, J. O., B. Liseo, and R. L. Wolpert (1999, February). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science 14*(1), 1–28.

Bickel, P. J. (1982). On adaptive estimation. *Annals of Statisitcs 10*(3), 647–671.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Brown, T., B. Mann, N. Ryder, et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

Chen, S., Y. Li, S. Lu, H. Van, H. J. W. L. Aerts, G. K. Savova, and D. S. Bitterman (2024). Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association : JAMIA 31*(4), 940–948.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–1535.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski. Quantile regression with censoring and endogeneity. *Journal of Econometrics 178*(1), 293–318.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 4171–4186.

Egami, N., C. J. Fong, J. Grimmer, and M. E. Roberts (2022). How to make causal inferences using texts. *Science Advances 8*(6), eabg2652.

Gentzkow, M., B. Kelly, and M. Taddy (2019, September). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Huber, P. J. and E. M. Ronchetti (2009). Robust statistics. *Wiley Series in Probability and Statistics*.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological methods 15*(4), 309.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Karadzhov, G., T. Stafford, and A. Vlachos (2023). Delidata: A dataset for deliberation in multi-party problem solving.

Le Mens, G., B. Kovács, M. T. Hannan, and G. Pros (2023). Uncovering the semantics of concepts using gpt-4. *Proceedings of the National Academy of Sciences 120*(49), e2309350120.

Li, L., E. T. Tchetgen, A. van der Vaart, and J. M. Robins (2011, Jul). Higher order inference on a treatment effect under low regularity conditions. *Statistics & probability letters 81*(7), 821–828.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Mendelberg, T., C. F. Karpowitz, and J. B. Oliphant (2014). Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics 12*(1), 18–44.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association 95*(450), 449–465. Received 01 Apr 1998, Published online: 17 Feb 2012.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*(4), 1161–1167.

OpenAI (2023, March). GPT-4 Technical Report.

Pearl, J. (2012). The causal mediation formula: a guide to the assessment of pathways and mechanisms. *Prevention Science 13*(4), 426–436.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *ACL*, pp. 1532–1543.

Pickering, M. J. and S. Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*(2), 169–225.

Pickering, M. J. and S. Garrod (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences 36*(4), 329–347.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Series in Statistics. New York: Springer.

Pretorius, C. and J. W. H. Swanepoel (2018). On the asymptotic theory of new bootstrap confidence bounds. *The Annals of Statistics 46*(1), 438–456.

Rho, E. H., M. Harrington, Y. Zhong, and J. L. Eberhardt (2023). Escalated police stops of black men are linguistically and psychologically distinct in their earliest moments. *Proceedings of the National Academy of Sciences*.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020, October). Adjusting for confounding with text matching. *American Journal of Political Science 64*(4), 887–903.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014, October). Structural topic models for open-ended survey responses. *American Journal of Political Science 58*(4), 1064–1082.

Robins, J., L. Li, E. Tchetgen, and A. van der Vaart (2008). "higher order influence functions and minimax estimation of nonlinear functionals". In *Probability and Statistics: Essays in Honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics.

Robins, J. M. and A. Rotznitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the Americal Statistical Association 90*(429), 122–129.

Rodriguez, P., A. Spirling, and B. Stewart (2023, November). *American Political Science Review 117*(4), 1255–1274.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*(387), 516–524.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Sahlgren, M. (2008, 01). The distributional hypothesis. *Italian Journal of Linguistics 20*.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

van der Vaart, A. (1998a). *Asymptotic Statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

van der Vaart, A. (2014). Higher order tangent spaces and influence functions. *Statistical Science 29*(4), 679–686.

van der Vaart, A. W. (1998b). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

A Proofs

A.1 Identification

.

Proof 2 Overlap allows the auxiliary event-context distribution to inform the distribution of strings under each event. Weak Ignorability ensures that the event-contexts isolate the effect of the event itself when evaluated with the clause function. The result follows directly from injectivity.

A.2 Influence Function Decomposition

.

Assumption 1 allows us an estimating equation that is a sample average over i.i.d. strings and their associated contexts. Assumption 2 guarantees that the estimating problem is well-posed. Assumptions 3 and 4 provide the regularity conditions required for the functional delta method (see van der Vaart, 1998a, Thm. 20.8). Assumption 5 ensures a uniform approximation in a neighborhood of the true value. Together, these conditions yield the reported von Mises expansion.

A.3 Limit Theorem

.

The Lyapunov condition ensures asymptotic normality of the leading term of the von Mises expansion from the previous result. The uniform continuous mapping theorem and Slutsky's method justify the limiting distribution. The result is directly comparable to the central limit theorem for regular Z-estimators with nuisance functions (see van der Vaart, 1998a, Thm. 25.54 and 25.57).

A.4 Pivotality and Rate Result

Define the pivot

$$T_n \;=\; r_n(\widehat{\theta}_n - \theta_0).$$

Under Cramér's condition it admits the first-order Edgeworth expansion

$$\Pr\{T_n \le t\} = G(t) \;+\; \frac{1}{r_n}\, A(t)\, g(t) \;+\; o\big(r_n^{-1}\big),$$

where G is a known limit law with density g and polynomial bias A. Hall (1992) shows that the bootstrap—or, by the same argument, repeated cross-fitting—yields an identical expansion

$$\overset{*}{\Pr}\{T_n^* \le t\} = G(t) \;+\; \frac{1}{r_n}\, A(t)\, g(t) \;+\; o_p\big(r_n^{-1}\big).$$

Since G involves no unknowns, the $O(r_n^{-1})$ term cancels exactly, delivering the higher-order gain.

Next, in our von Mises decomposition the only non-parametric remainder is the self-referential bias, which factors into

$$\underbrace{\|\widehat{F} - F\|_\infty}_{\text{context error}} \quad \text{and} \quad \underbrace{\frac{1}{M}\sum_{m=1}^{M} R^{(m)}}_{\substack{\text{self-referential} \\ \text{remainder}}},$$

where each $R^{(m)}$ is i.i.d. mean-zero, finite-variance conditional on the fixed contexts. By Dvoretzky–Kiefer–Wolfowitz,

$$\|\widehat{F} - F\|_\infty = O_p\Big(\sqrt{\tfrac{\ln n_c}{n_c}}\Big),$$

and by the Kolmogorov LIL,

$$\frac{1}{M}\sum_{m=1}^{M} R^{(m)} = O_p\Big(\sqrt{\tfrac{\ln\ln M}{M}}\Big).$$

Hadamard differentiability plus our Lyapunov and equicontinuity conditions keep everything in $L_2(P)$, so that re-using the same contexts across folds maintains a finite variance. Because the two rates multiply in the remainder, the total bias is

$$O_p\Big(\sqrt{\tfrac{\ln n_c}{n_c}}\Big) \;\times\; O_p\Big(\sqrt{\tfrac{\ln\ln M}{M}}\Big) \;=\; O_p\Big(\sqrt{\tfrac{\ln n_c}{n_c}\,\tfrac{\ln\ln M}{M}}\Big).$$

Equating this to $o(n^{-\gamma})$ and absorbing the log factors into an $\epsilon > 0$ gives

$$(n_c\, M)^{1+\epsilon} \;\gg\; n^{2\gamma},$$

as reported in the text.

B Estimation

To avoid self-referential bias, we implement a repeated cross-fitting procedure. Let $\mathcal{I}^1$ and $\mathcal{I}^2$ denote a random partition of the sample into two folds. For each string $s_i \in \mathcal{I}^1$, we evaluate it on out-of-fold contexts generated by strings in $\mathcal{I}^2$, and vice versa. Specifically, define the string-level statistic as

$$\widehat{Str}_i^{\mathcal{E};\mathcal{I}^2} = \begin{bmatrix} \frac{1}{|C^A|}\sum_{c\in C^A} \log \Pr(s_i \mid c) \\ \frac{1}{|C^B|}\sum_{c\in C^B} \log \Pr(s_i \mid c) \end{bmatrix}, \quad \text{for } s_i \in \mathcal{I}^1,$$

where C^G denotes the set of contexts generated from group G's strings in fold $\mathcal{I}^2$. The difference-in-means estimator for group A is then

$$\widehat{\theta}^{A;\mathcal{I}^2} = \frac{1}{n_{A1}} \sum_{i\in\mathcal{I}^1\cap A} \left(\widehat{Str}_{i,A}^{\mathcal{E};\mathcal{I}^2} - \widehat{Str}_{i,B}^{\mathcal{E};\mathcal{I}^2}\right),$$

and similarly for group B using $\mathcal{I}^2 \cap B$.

We then reverse the roles of the folds and average across directions:

$$\widehat{\theta}^{\text{avg}} = \frac{1}{2}\left(\widehat{\theta}^{A;\mathcal{I}^2} - \widehat{\theta}^{B;\mathcal{I}^2} + \widehat{\theta}^{A;\mathcal{I}^1} - \widehat{\theta}^{B;\mathcal{I}^1}\right).$$

To account for sampling variability, we compute a t-statistic in each direction and average:

$$\widehat{t} = \sqrt{2} \cdot \frac{1}{2} \left(t_{\mathcal{I}^1 \to \mathcal{I}^2} + t_{\mathcal{I}^2 \to \mathcal{I}^1} \right),$$

where $t_{\mathcal{I}^1 \to \mathcal{I}^2}$ is the t-statistic computed by evaluating strings in $\mathcal{I}^1$ on contexts from $\mathcal{I}^2$, and vice versa.

This estimator is repeated across multiple random splits, and final inference is based on the distribution of $\widehat{t}$ across cross-fits.

C Context and String Two-Sample Analysis

Effect	Context
0.0158	[ii] Usually illegal immigrants are refugees and we should make an effort to help and not condemn him. We are in fact a land of immagrints. We «<STR»> [ii] Usually illegal immigrants are refugees and we should make an effort to help and not condemn him. We are in fact a land of immagrints. We
0.0158	I do not know the law in this regard. «<STR»> i.
0.0152	They should simply be deported. They have never committed another crime, and their actions were not violent. This is simply a minor violation «<STR»> They should simply be deported. They have never committed another crime, and their actions were not violent. This is simply a minor violation
0.0152	i am not in favor of putting a citizen of a country illegally into prison for the crime of illegally entering for food and work. i think we have a «<STR»> i am not in favor of putting a citizen of a country illegally into prison for the crime of illegally entering for food and work. i think we have a
0.0144	I think he should be sent back to whatever country he came from, Why should the taxpayers have to support him for whatever length of sentence he would be given? «<STR»> i.
0.0139	This person is a known criminal who has been convicted of violent crimes. In addition, they are in the country illegally. They should definitely go to prison as it is likely they will «<STR»> They need to be sent to prison for their crimes so it will stop them from breaking more laws
0.0137	No he should not go to prison because he doesn't have a prior record and his only crime is entering the United States illegally. He should be allowed to stay in the «<STR»> He should be allowed to stay in the United States and try to make a life for himself as long as he follows the law
0.0137	I think he shouldn't be sent to prison because A, he has no previous record and B, and more importantly, he committed no serious (or violent) offense here. «<STR»> I think he shouldn't be sent to prison because A, he has no previous record and B, and more importantly, he committed no serious (or violent) offense here
0.0133	It depends on the nature of the violent action from before. Not enough information to decide. «<STR»> Not enough information to decide.
0.0128	i think that once in the US, people should only be put in jail if they commit an offense that would send a US citizen to jail. the reason the man came to the «<STR»> I think that once in the US, people should only be put in jail if they commit an offense that would send a US citizen to jail. the reason the man came to the

Table C.1: **Top Contexts that Sharpen the Effect**

Effect	Context
-0.0251	I believe the offender should go to prison at minimum and should also be deported. The offender has broken the law and illegally entered the United States. He also has a violent «<STR»> I believe the offender should go to prison at minimum and should also be deported. The offender has broken the law and illegally entered the United States. He also has a violent
-0.0186	This person should serve prison time because they have history of being deported and serving previous prison sentences. This person it also a violent criminal and they should not be given any sort of lee «<STR»> This person should serve prison time because they have history of being deported and serving previous prison sentences. This person it also a violent criminal and they should not be given any sort of lee
-0.0185	This person should serve prison time because they have history of being deported and serving previous prison sentences. This person it also a violent criminal and they should not be given any sort of lee «<STR»> This person should serve prison time because they have history of being deported and serving previous prison sentences. This person it also a violent criminal and they should not be given any sort of lee
-0.0171	The person should serve a prison sentence to pay for his crime. He clearly does not respect the laws of our country so he should receive and serve a harsher sentence before deport «<STR»> The person should serve a prison sentence to pay for his crime. He clearly does not respect the laws of our country so he should receive and serve a harsher sentence before deport
-0.0170	I think this man should be deported. America should not incur the cost of food/clothes/shelter for a criminal. Depending on the severity of the crime «<STR»> I think this man should be deported. America should not incur the cost of food/clothes/shelter for a criminal. Depending on the severity of the crime
-0.0167	He should be jailed because obviously he cannot be trusted to follow the law since he has been previously convicted of a violent crime and entered the country illegally. No one «<STR»> He should be jailed because obviously he cannot be trusted to follow the law since he has been previously convicted of a violent crime and entered the country illegally. No one

Table C.2: **Top Contexts that Attenuate the Effect**

Effect	String
10.5754	They should simply be deported. They have never committed another crime, and their actions were not violent. This is simply a minor violation
9.5168	He should not go to prison, since he did not commit a crime which harmed another person or harmed property. He doesn't seem to be a danger to others or himself. The U.S. government should deport him back to his home country.
9.0263	I think the man should be sent back to his country since he came here illegally. Since the man hasn't been incarcerated before I don't think he should be sent to prison.
8.9130	He has been guilty of a violent crime. He has been to prison before.
8.4869	He tried to enter the country with no criminal record. Just because he is not a legal entrant does not mean he should get prison time.
8.2690	This person was a repeat offender so we do not want him in our country. As long as he did not commit a crime in our country we have nothing to convict him for and he should be deported.
8.1170	His only violation in this case is that he was crossing the border illegally. I don't believe crossing the border illegally is a case where prison should be the punishment, regardless of prior criminal record. And so it follows that prison isn't the proper punishment here.
8.0016	With no record, I think this person should be sent back to the country he came. I do not think he should be penalised with any jail time.
-12.3304	This person has clearly demonstrated a pattern of violence and criminal activity. Allowing him into the United States would have a negative effect on national security.
-11.3786	Because of his prior history of offenses, it shows that he should be placed in prison. Because he has a tendency to be a violent criminal, it would be better if he were behind bars.
-9.7921	He should be sent to prison because he was convicted of a crime, and obviously deportation didn't keep him out of the US. Also, he is a violent criminal and should be incarcerated, not simply deported.
-9.7467	I believe he should be sent to prison. Mostly due to his violent past and the U.S government should crack down on illegal immigrants committing crimes in our county. They should imprison him because if they deport him his past shows he will come back and commit more violent crimes.
-9.3869	He has show a disregard for the laws of our nation and without serving a prison sentence will very likely just come right back to the country upon his inevitable re-deportation. He should serve a minimum sentence of 1 year and upon completion should immediately be deported to his home country (preferably at his own expense if possible). Without consequences for illegally entering the country, people will just come back again.
-9.1911	He should be sent to prison because he is obviously not learning his lesson. He has already served time in prison and keeps coming back in this country.
-8.9385	This offender should be sent back to his home country. The U.S. government should not be responsible for using tax payer money to support this man in prison. If he is caught again, he should be executed.
-8.9186	He's a prior offender who should have learned his lesson by now. Having been convicted of a violent crime he's clearly demonstrated that he's a threat to society. After he serves his sentence he should be deported back to his home country.

Table C.3: String Level Effect Estimates (Positive = Largest Treatment Effect)

group	Effect Size	string
Top 7	10.575	They should simply be deported. They have never committed another crime, and their actions were not violent. This is simply a minor violation
Top 7	9.517	He should not go to prison, since he did not commit a crime which harmed another person or harmed property. He doesn't seem to be a danger to others or himself. The U.S. government should deport him back to his home country.
Top 7	9.026	I think the man should be sent back to his country since he came here illegally. Since the man hasn't been incarcerated before I don't think he should be sent to prison.
Top 7	8.913	He has been guilty of a violent crime. He has been to prison before.
Top 7	8.487	He tried to enter the country with no criminal record. Just because he is not a legal entrant does not mean he should get prison time.
Top 7	8.269	This person was a repeat offender so we do not want him in our country. As long as he did not commit a crime in our country we have nothing to convict him for and he should be deported.
Top 7	8.117	His only violation in this case is that he was crossing the border illegally. I don't believe crossing the border illegally is a case where prison should be the punishment, regardless of prior criminal record. And so it follows that prison isn't the proper punishment here.
Neutral 7	0.004	I don't even know the moral ethics of illegal entry anymore. The government should let him in but if the sentences he had prior to this were very severe then they should exercise caution.
Neutral 7	-0.006	I don't believe that someone trying to get into the country looking for opportunity should be punished especially if they have no criminal history. I believe they should be fined and be given information about a path to legal citizenship.
Neutral 7	0.007	get ride of them, send them back to their own country and let them puttheir own people in harms way
Neutral 7	-0.007	He should be made to return to his own country because he came in illegally. It's not our responsibility to pay for him in prison
Neutral 7	-0.011	He should be sent back to his country of origin. Since it is his first time, that is all the action that needs to be taken.
Neutral 7	0.012	I think that he should have to pay a fine. The immigration system in the US is now setting up immigrants to fail. They make it very difficult for people to come here legally to work.
Neutral 7	0.013	He should be sent back to his native country. He should not be imprisioned on the tax payers dime in the country he illegally entered.
Bottom 7	-12.330	This person has clearly demonstrated a pattern of violence and criminal activity. Allowing him into the United States would have a negative effect on national security.
Bottom 7	-11.379	Because of his prior history of offenses, it shows that he should be placed in prison. Because he has a tendency to be a violent criminal, it would be better if he were behind bars.
Bottom 7	-9.792	He should be sent to prison because he was convicted of a crime, and obviously deportation didn't keep him out of the US. Also, he is a violent criminal and should be incarcerated, not simply deported.
Bottom 7	-9.747	I believe he should be sent to prison. Mostly due to his violent past and the U.S government should crack down on illegal immigrants committing crimes in our county. They should imprison him because if they deport him his past shows he will come back and commit more violent crimes.
Bottom 7	-9.387	He has show a disregard for the laws of our nation and without serving a prison sentence will very likely just come right back to the country upon his inevitable re-deportation. He should serve a minimum sentence of 1 year and upon completion should immediately be deported to his home country (preferably at his own expense if possible). Without consequences for illegally entering the country, people will just come back again.
Bottom 7	-9.191	He should be sent to prison because he is obviously not learning his lesson. He has already served time in prison and keeps coming back in this country.
Bottom 7	-8.938	This offender should be sent back to his home country. The U.S. government should not be responsible for using tax payer money to support this man in prison. If he is caught again, he should be executed.

Table C.4: String-level effects: Top, Neutral, and Bottom 7