Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1521

Conditional Hypothesis Generation for LLM-Based Text Analysis with Researcher-Specified Covariates

Paiheng Xu
University of Maryland, College Park

Jing Liu
University of Maryland, College Park

Wei Ai
University of Maryland, College Park

A core goal of computational social science is to discover interpretable differences in how language varies across outcomes of interest, such as political affiliation or instructional quality. Recent LLM-based hypothesis generation methods describe such differences in natural language, but select for globally discriminative patterns without accounting for covariates — identified through researchers' domain knowledge — that shape the data. When covariates are ignored, selected patterns can reflect confounds rather than differences of substantive interest. We introduce conditional hypothesis generation, a framework that incorporates researcher-specified covariates to steer hypothesis discovery toward differences that hold within relevant subgroups. Two challenges arise: the target subgroup may be underrepresented (stratum imbalance), and the direction of a difference may reverse across subgroups (sign reversal). We propose two econometrics-inspired methods: one introduces feature–covariate interactions to detect sign reversals, and the other applies within-stratum demeaning and inverse-frequency reweighting to equalize underrepresented strata. Synthetic experiments show each method outperforms global baselines in its targeted setting, and expert evaluation on two real-world datasets confirms that covariate-aware generation surfaces more useful hypotheses within relevant subgroups.

VERSION: July 2026

Suggested citation: Xu, Paiheng, Jing Liu, and Wei Ai. (2026). Conditional Hypothesis Generation for LLM-Based Text Analysis with Researcher-Specified Covariates. (EdWorkingPaper: 26-1521). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/knbg-hb51

Conditional Hypothesis Generation for LLM-Based Text Analysis with Researcher-Specified Covariates

Paiheng Xu, Jing Liu and Wei Ai

University of Maryland, College Park
{paiheng,jliu28,aiwei}@umd.edu

Abstract

A core goal of computational social science is to discover interpretable differences in how language varies across outcomes of interest, such as political affiliation or instructional quality. Recent LLM-based hypothesis generation methods describe such differences in natural language, but select for globally discriminative patterns without accounting for covariates that shape the data based on researchers' domain knowledge. When covariates are ignored, selected patterns can reflect confounds rather than differences of substantive interest. We introduce *conditional* hypothesis generation, a framework that incorporates researcher-specified covariates to steer hypothesis discovery toward differences that hold within relevant subgroups. Two challenges arise: the target subgroup may be underrepresented (*stratum imbalance*), and the direction of a difference may reverse across subgroups (*sign reversal*). We propose two econometrics-inspired methods: one introduces feature–covariate interactions to detect sign reversals, and the other applies within-stratum demeaning and inverse-frequency reweighting to equalize underrepresented strata. Synthetic experiments show each method outperforms global baselines in its targeted setting, and expert evaluation on two real-world datasets confirms that covariate-aware generation surfaces more useful hypotheses within relevant subgroups.

1 Introduction

A central goal of computational social science (CSS) is to understand how text relates to variables such as political affiliation, instructional quality, or social media engagement. Rather than predicting these outcomes, researchers seek interpretable hypotheses—natural-language descriptions of how textual patterns differ across outcome values—that can guide further investigation (Grimmer and Stewart, 2013; Card, 2019; Grimmer et al., 2022).

Recent LLM-based methods support this form of analysis by sampling labeled examples and prompting an LLM to propose natural-language hypotheses that characterize textual patterns associated with different outcome values (Zhong et al., 2022, 2023, 2024; Zhou et al., 2024; Movva et al., 2025). These methods typically select hypotheses by how well they discriminate between outcome groups.

However, global discrimination can be misleading. A globally discriminative feature may reflect a confound rather than a difference of substantive interest, a persistent concern in text-as-data research (Grimmer and Stewart, 2013; Gentzkow et al., 2019; Grimmer et al., 2022). For example, Taddy (2013) shows that national-park language appears as a party-predictive feature because public lands are unevenly distributed across states, though national park issues are not inherently partisan. The challenge is to steer discovery toward differences that hold within conditions researchers care about—and to let researchers specify those conditions.

We introduce *conditional* hypothesis generation, a framework for generating hypotheses that are discriminative within researcher-specified covariate strata. Covariates—such as policy area, time period, or classroom environment—encode the domain knowledge that researchers bring to text analysis: they define the conditions under which differences should be examined, without requiring the hypothesis itself to be known in advance.

Conditioning on covariates introduces two statistical challenges (Simpson, 1951; Gail and Simon, 1985). The target stratum may be underrepresented, so that its signal is dominated by larger strata (*stratum imbalance*); or the direction of a difference may reverse across strata (*sign reversal*), so that global aggregation cancels the conditional pattern.

We build on Movva et al. (2025), which maps documents to monosemantic Sparse Autoencoder (SAE) features and selects discriminative features via LASSO. Because the SAE features are fixed

before statistical selection, covariates can be incorporated directly into the selection step. Drawing on econometrics, we propose two complementary methods. *Interaction-lasso* augments the feature space with feature–covariate interactions, allowing a feature to be selected when it is discriminative within a single stratum, even if its global effect is zero. *Demeaned-reweighted-lasso* residualizes features and outcomes within covariate strata to isolate within-stratum variation, and applies inverse-frequency weighting so that underrepresented strata contribute comparably to feature selection.

In synthetic evaluations with known ground-truth hypotheses and covariate structure, demeaned-reweighted-lasso outperforms global baselines and approaches oracle performance across imbalance levels, while interaction-lasso is the only method that recovers differences under sign reversal.

We validate on two real-world datasets: CONGRESS, a longstanding testbed for political language (Gentzkow and Shapiro, 2010; Grimmer et al., 2021), and NCTE, a dataset of math classroom transcripts with rich annotations of instructional quality (Demszky and Hill, 2023; Hill et al., 2008; Pianta et al., 2012). Expert evaluations show that covariate-aware selection surfaces hypotheses that domain experts rate as more useful than those unique to the global baseline.

Our contributions are: (1) We formalize conditional hypothesis generation for text analysis with researcher-specified covariates. (2) We introduce two complementary covariate-aware methods, each targeting a distinct statistical challenge (i.e., stratum imbalance and sign reversal). (3) We design controlled synthetic evaluations covering these two challenges. (4) Expert evaluation on two real-world datasets shows that covariate-aware methods guide discovery toward hypotheses that domain experts find more useful.

2 Preliminaries

2.1 Task Formulation

We consider a dataset $\{(x_i, y_i)\}_{i\in[N]}$, where x_i is a text document and y_i is a target variable. In the simplest case, y_i indicates group membership of the document. The goal of *hypothesis generation* is to produce a set of natural-language statements $\mathcal{H}$ that characterize how y_i varies with text content, where $|\mathcal{H}|$ is a pre-specified number of hypotheses to generate. Each text is associated with a set of covariates $\mathcal{V}$, which may include metadata (e.g., time period, author demographics) or attributes inferable from text (e.g., topic). In our setting, a researcher may specify a subset $\mathbf{C}$ of these covariates based on domain knowledge to guide hypothesis generation within relevant covariate strata.

2.2 LLM-Based Hypothesis Generation

LLM-based hypothesis generation for comparing texts typically follows a sample-and-propose pattern: sample a small set of labeled examples, then ask an LLM to induce patterns and propose natural-language hypotheses from those examples. Existing methods differ in their sampling strategies. Zhong et al. (2022, 2023) train a classification model to randomly sample the most discriminative texts. Zhou et al. (2024) iteratively generate hypotheses from examples that the current candidate list mis-classifies. These methods then filter candidate hypotheses by their discriminative power.

Zhong et al. (2024); Movva et al. (2025) instead use statistical models to select the most predictive features over all N samples, then call an LLM to describe each selected feature from a sample of feature-related texts. Movva et al. (2025), the current state of the art, train a Sparse Autoencoder (SAE) on text embeddings to obtain a fixed set of interpretable features, then select those most predictive of y via LASSO, as detailed in §2.3. We build on this feature-selection view because it makes incorporating covariates $\mathbf{C}$ well-defined.

2.3 Hypothesis Generation with SAE

A SAE encourages each input to be reconstructed from only a small active subset of features (Huben et al., 2024). This training objective tends to produce monosemantic features, where each dimension captures a coherent, interpretable concept. Movva et al. (2025) exploit this property with three steps: **(1) SAE encoding**: a trained SAE maps each document embedding to a sparse vector of feature activations. Each dimension in this vector corresponds to one learned SAE feature. **(2) Feature selection**: a statistical model (e.g., LASSO) selects $|\mathcal{H}|$ features most predictive of y. **(3) LLM interpretation**: an LLM describes the texts that contrast most strongly activating input documents (positive examples) vs. zero-activating input documents (negative examples) for each selected feature as a natural-language hypothesis; Table 8 shows the prompt. Step 2 accepts any statistical model, making it straightforward to condition feature selection on covariates $\mathbf{C}$, as we show next.

3 Method

We define the goal of *conditional* hypothesis generation as discovering group differences that are discriminative within strata of $\mathbf{C}$, which brings two statistical challenges (Simpson, 1951; Blyth, 1972; Gail and Simon, 1985): (1) **stratum imbalance**: the relevant covariate stratum may be rare in the corpora, the targeted difference is dominated by irrelevant differences in other strata. (2) **sign reversal**: the direction of the group difference changes across strata, so the global difference can cancel or misrepresent the targeted differences. A classic example is Simpson's paradox (Simpson, 1951). The targeted hypothesis can only be reliably generated with within-stratum comparison. We adapt two well-studied econometric tools to SAE feature selection: interaction models and group fixed effects (implemented via within-stratum demeaning). Both subsume the global case: features that are discriminative across all strata remain selected alongside stratum-specific ones.

Setup. Let $Z \in \mathbb{R}^{N\times M}$ denote SAE activation matrix, where $z_{i,m}$ is the activation of feature m for document i and M is the number of SAE features. Let $\mathbf{C} = [C_1, \ldots, C_P] \in \mathbb{R}^{N\times P}$ denote the covariate matrix, where C_p is the pth covariate column and P is the number of covariates. For simplicity, we focus on binary corpus comparison ($y \in \{0,1\}^N$, encoding corpus A or B). The baseline LASSO fits L1-regularized logistic regression:

$$\hat{\beta} = \arg\min_{\beta} \ell(y, Z\beta) + \lambda\|\beta\|_1, \qquad (1)$$

and selects the top $|\mathcal{H}|$ features ranked by $|\hat{\beta}_m|$, capturing only globally discriminative features.

3.1 Interaction-lasso

A standard approach for modeling covariate-specific effects in regression is to include *interaction terms* (products of two variables) (Angrist and Pischke, 2009). When an SAE feature is interacted with a covariate, the feature's coefficient can vary with that covariate, capturing within-stratum differences. We apply this idea by augmenting the SAE activation matrix with the covariates $\mathbf{C}$ and feature–covariate interaction blocks $Z \odot C_p$ for each covariate $p \in [P]$, where $Z \odot C_p$ denotes row-wise multiplication of all M SAE features by covariate C_p. Let $\eta(\beta, \delta, \gamma) = Z\beta + \mathbf{C}\delta + \sum_{p=1}^{P}(Z \odot C_p)\gamma_p$. We fit LASSO on the full augmented space:

$$\hat{\beta}, \hat{\delta}, \hat{\gamma} = \arg\min_{\beta,\delta,\gamma} \ell(y, \eta(\beta, \delta, \gamma)) + \lambda\|(\beta, \delta, \gamma)\|_1. \qquad (2)$$

Here β are SAE feature main effects, δ are covariate main effects, and γ_p are interaction effects for covariate p. We rank features by $\max(|\hat{\beta}_m|, \max_{p\in[P]} |\hat{\gamma}_{p,m}|)$, and select the top $|\mathcal{H}|$ features. A feature qualifies if it is discriminative globally (large $|\hat{\beta}_m|$) or through any covariate-specific interaction (large $|\hat{\gamma}_{p,m}|$ for some p), directly implementing the stated goal. The covariate main effects δ are included as nuisance controls and are not ranked as potential hypotheses.

However, interaction-lasso faces two practical limitations. First, the feature space substantially expands by including $M \times P$ interaction features, which increases computational cost and can make feature selection less stable in high dimensions (Bien et al., 2013). Second, SAE features are inherently sparse, so each interaction term $Z \odot C_p$ is doubly sparse: nonzero only for samples with $C_{i,p} \neq 0$ and an active feature. When a covariate is rare, these terms become nearly all-zero, making the $\hat{\gamma}_{p,m}$ estimates noisy (Crump et al., 2009). We validate this empirically in §4.4.

3.2 Demeaned-Reweighted-Lasso

To overcome the practical limitations, we adapt another standard econometrics technique called demeaning, which subtracts group means from observations to remove between-group variation and isolate within-group differences (Angrist and Pischke, 2009). By the Frisch–Waugh–Lovell theorem (Lovell, 1963), this is analogous to including group fixed effects in the regression. Specifically, we group documents by their observed covariate values, residualize both SAE activations Z and y within each group, then run LASSO on the residuals. Unlike interaction-lasso, demeaning operates on the original M features without variable expansion, and estimates each feature's within-stratum difference from all N samples rather than sparse interaction terms, avoiding both limitations above.

Formally, let g_i denote the covariate stratum of document i, and let $\bar{z}_m^{(g_i)}$ and $\bar{y}^{(g_i)}$ be the mean feature activation and mean label within that stratum. We compute $\tilde{z}_{i,m} = z_{i,m} - \bar{z}_m^{(g_i)}$ and $\tilde{y}_i = y_i - \bar{y}^{(g_i)}$. The **demeaned-lasso** then selects features by running LASSO on $(\tilde{Z}, \tilde{y})$. The $\tilde{\beta}_m$ estimates capture the within-stratum difference associated with feature m, partialling out per-stratum mean shifts in

both the outcome and the feature activations.

Additionally, to address stratum imbalance, we propose **demeaned-reweighted-lasso**: set sample weights w_i proportional to the inverse frequency of stratum g_i and select features by the weighted lasso on $(\tilde{Z}, \tilde{y}, w)$. Reweighting equalizes stratum contributions to feature selection, allowing features from underrepresented target strata to be selected when their within-stratum differences are strong.

Demeaning-based methods assume **no qualitative interaction (NQI)**: the targeted difference has a consistent sign across strata (Gail and Simon, 1985). Under NQI, the covariates are treated as nuisance fixed effects rather than variables that change the direction of the text–outcome relationship. In the sign reversal scenario, NQI is violated and interaction-lasso should be used instead. We also include **demeaned-lasso** to isolate the contribution of the reweighting step.

4 Experiments on Synthetic Datasets

4.1 Synthetic Dataset Construction

We evaluate the methods by recovering *known ground-truth hypotheses* from synthetic datasets. We choose a dataset containing bill summaries from the 110-114th U.S. Congress (Hoyle et al., 2022). It has human-labeled high-level topics and granular subtopics, and metadata such as bill creation time, allowing us to repurpose it to construct synthetic corpora with *controlled covariate structure*. We instantiate the two data scenarios demonstrating the two challenges from §3: **stratum imbalance** and **sign reversal**. Figure 1 summarizes their corpus compositions.

4.1.1 Scenario 1: Stratum Imbalance

We investigate whether methods can recover a *targeted subset* of differences when many differences are simultaneously present. We split the topics into two themes: *Government & Economy*, $\mathcal{T}_{gov}$ and *Social policy*, $\mathcal{T}_{soc}$. For each seed run, we randomly select three topics from $\mathcal{T}_{gov}$ and three topics from $\mathcal{T}_{soc}$. For each topic T_i (e.g., "Health"), we randomly select one subtopic for A and one subtopic for B (e.g., "Mental" and "Drug Industry"). The example ground truth is in the form of "is about Health: Mental". The targeted difference is in the social policy theme, and we thus evaluate recovery of ground truth under $\mathcal{T}_{soc}$.

To simulate the case when relevant strata are rare, an imbalance ratio r controls the frequency of

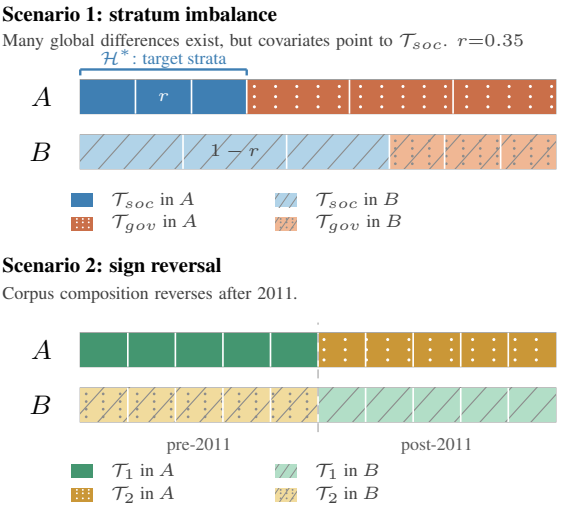

Figure 1: Corpus compositions for synthetic datasets.

$\mathcal{T}_{soc}$ topics in A: each $\mathcal{T}_{soc}$ topic contributes at rate r and each $\mathcal{T}_{gov}$ topic at rate $1-r$, with B receiving the mirror allocation. Lower r suppresses the $\mathcal{T}_{soc}$ signal in A more, making it harder to recover the targeted difference. We pass three binary topic covariates to covariate-aware methods, one indicator per $\mathcal{T}_{soc}$ topic (more data details in Appendix A). These covariates should steer the methods toward the relevant strata, but do not reveal which subtopic distinguishes A from B within each topic.

4.1.2 Scenario 2: Sign Reversal

We then investigate whether methods can recover targeted differences when a covariate masks the aggregate signal. The covariate is the bill creation time period (binary: pre/post 2011). For each seed run, we randomly select 10 subtopics, each from a distinct topic, and split them into two groups $\mathcal{T}_1$ and $\mathcal{T}_2$ of five each. In the pre-2011 period, $\mathcal{T}_1$ subtopics are assigned to A and $\mathcal{T}_2$ to B; in the post-2011 period, the assignment reverses. We pass one binary time-period covariate (pre-2011=1) to covariate-aware methods, and evaluate recovery of the five $\mathcal{T}_1$ subtopic differences. By construction, the difference reverses sign across periods, violating NQI assumption. We expect interaction-lasso to be the appropriate method, while demeaning-based methods are expected to fail.

4.2 Evaluation Metrics

We evaluate how well the methods recover the targeted ground-truth differences on the held-out set. For each scenario, we use an LLM to annotate generated hypotheses on held-out texts and match them to reference hypotheses using the Hungarian algo-

Method	r=0.50		r=0.35		r=0.20	
	Surf. Sim.	F1	Surf. Sim.	F1	Surf. Sim.	F1
oracle	.750±.12	.711±.08	.737±.11	.682±.06	.723±.15	.684±.05
random	.280±.18	.248±.10	.400±.10	.299±.09	.347±.10	.228±.08
LLM-direct						
llm-covariate	.330±.23	.410±.11	.137±.12	.178±.11	.057±.08	.080±.03
llm-global	.323±.16	.325±.08	.127±.11	.155±.06	.067±.09	.090±.03
SAE						
lasso	.557±.18	.394±.14	.383±.13	.235±.08	.393±.12	.230±.07
separation score	.487±.17	.340±.16	.397±.16	.221±.10	.333±.13	.228±.05
Proposed						
demeaned-reweighted-lasso	††**.700**±.11	††**.533**±.12	††**.730**±.14	††**.605**±.14	††**.730**±.11	††**.534**±.10
demeaned-lasso (ablation)	.530±.17	.406±.16	††.533±.15	††.429±.08	†.517±.13	††.378±.07
interaction-lasso	.407±.15	.280±.09	.363±.16	.226±.12	.357±.12	.211±.10

Table 1: Stratum Imbalance: $|\mathcal{H}|$=3. †/††: p<0.05/0.01 vs. lasso, exact paired sign-flip test.

rithm on reference-vs.-generated annotation correlations. See Appendix B for detailed process. We report two metrics computed over matched pairs. **Surface similarity**: for each matched reference/inferred hypothesis pair, we prompt `gpt-4.1` to assess whether the two hypotheses are the *same*, *related*, or *distinct*, with corresponding scores of 1.0, 0.5, and 0.0 (prompt in Table 10, examples in Table 9). To improve stability, we sample five outputs at temperature 0.7 and average them, then report the mean score across $|\mathcal{H}|$ pairs. **F1 similarity**: for each matched pair, we compute the F1-score between the reference and inferred hypothesis *text-level* annotations on the held-out set.

4.3 Baselines

Since no existing method handles covariates, we compare against SAE-based methods (Movva et al., 2025) as the primary baselines. For fair comparison, we use `gpt-4.1` for proposing hypotheses if not specified otherwise, and OpenAI's text-embedding-3-small for embeddings.

SAE-based. (1) **LASSO**: use Eq. 1 to select the top $|\mathcal{H}|$ features. (2) **separation score**: select features by the mean of y among the highest-activating texts vs. zero-activating texts.[1]

LLM-direct. We include direct prompting baselines to test whether covariate-aware discovery can be obtained simply by showing sampled texts to an LLM, without a corpus-level feature-selection step. **llm-global** is prompted with sampled corpus texts, while **llm-covariate** is also shown covariate-profile labels for each text. Because these baselines only observe the prompt sample, we use a generous sample budget in the main tables. Appendix C.3 gives the sensitivity analysis of the sample size.

Reference baselines. To calibrate the evaluation, we include: **Oracle** as an upper reference that gives the LLM 20 examples sampled from the exact positive and negative sides of each ground-truth contrast. **Random** samples 20 training texts into arbitrary positive and negative halves before the same interpretation step, measuring how much recovery can come from generic topical cues alone. We use 20 examples to match the SAE interpretation stage. More implementation details for these baseline methods are provided in Appendix C.

Proposed methods. We compare **interaction-lasso** (§3.1), **demeaned-reweighted-lasso** (§3.2), and **demeaned-lasso** as ablation with the baselines.

Statistical testing. Each proposed method is compared to **LASSO** on seed-level surface similarity and F1, with 10 seeds per setting. We test whether the mean paired difference differs from zero with an exact two-sided sign-flip test.

4.4 Results and Discussion

Demeaned-reweighted-lasso steers discovery toward specified covariates and recovers suppressed within-stratum signals. Table 1 evaluates conditional discovery as the target social-policy strata become increasingly underrepresented. Across all imbalance levels, demeaned-reweighted-lasso remains near the oracle in surface similarity and significantly improves F1 over LASSO, the strongest global SAE baseline. Even at r=0.50, it reaches .700 surface similarity, near the oracle's .750 and above LASSO's .557 (p=0.016), showing that covariates steer feature selection toward the target $\mathcal{T}_{soc}$ differences. The gap widens under

[1] https://github.com/rmovva/HypotheSAEs

Method	Surf. Sim.	F1
oracle	.716±.10	.788±.06
random	.418±.12	.381±.10
LLM-direct		
llm-global	.210±.10	.269±.05
llm-covariate	.224±.09	.325±.16
SAE		
lasso	.484±.06	.362±.09
sep. score	.460±.11	.379±.12
Proposed		
interaction-lasso	††**.600**±.10	††**.576**±.08
demeaned-reweighted-lasso	.448±.14	.335±.11
demeaned-lasso (ablation)	.454±.11	.398±.12

Table 2: Sign reversal: $|\mathcal{H}|$=5. ††: p<0.01 vs. lasso, exact paired sign-flip test.

stronger imbalance: LASSO falls to .383–.393 surface similarity, while demeaned-reweighted-lasso remains at .730. The demeaned-lasso ablation improves over LASSO under imbalance but remains below the reweighted variant, indicating that residualizing within strata helps but reweighting further helps rare target strata from being suppressed.

The covariate structure determines the appropriate selection model. Table 2 evaluates sign reversal, where the time-period covariate reverses the differences across strata and makes the aggregate signal misleading. As §3 notes, residualization is appropriate when the covariate induces fixed stratum-level shifts, but not when the difference itself changes sign. As expected, neither demeaned variant significantly improves over LASSO.

Interaction-lasso accounts for sign reversal. Interaction-lasso is the only covariate-aware method that significantly improves over LASSO in Table 2, increasing surface similarity from .484 to .600 (p=0.008) and F1 from .362 to .576 (p=0.002). This suggests modeling feature-by-time interactions allows a feature to be selected when it is discriminative within a period even if its global effect cancels.

Interaction-lasso faces computational and sparsity limitations. Despite this advantage in the sign-reversal setting, interaction-lasso does not improve over global SAE baselines in Table 1, despite being the more general model. This empirically validates our discussion in §3.1 that interaction-lasso faces feature-space expansion and sparsity limitations. These limitations may also help explain why interaction-lasso is best in Table 2 but still falls short of the oracle.

Statistical modeling of covariates, not prompt-level exposure, drives gains. Both tables show that LLM-direct methods fall behind SAE-based methods. The sensitivity analysis in Appendix C.3 further shows that llm-covariate does not consistently outperform llm-global across sample sizes or with a stronger reasoning model (`gpt-5.4`).

5 Validation on Real-World Datasets

The synthetic experiments above test recovery under known conditional structure. For real-world corpora, where no ground-truth hypotheses exist, we evaluate whether conditional generation surfaces hypotheses that domain experts find useful for comparing outcome groups within researcher-specified covariate strata. Because our goal is to generate hypotheses discriminative within strata of $\mathbf{C}$ while retaining globally discriminative findings, we focus on hypotheses unique to the covariate-aware method relative to the global baseline.

5.1 Real-World Datasets

We include two real-world datasets and instantiate $\mathbf{C}$ with one researcher-specified binary covariate for each, chosen from domain knowledge: (1) CONGRESS (Gentzkow and Shapiro, 2010): U.S. speeches from the 109th Congress (2005-07), a longstanding application in computational social science (Grimmer et al., 2021). The outcome variable y is party affiliation (Republican vs. Democrat). The covariate indicates whether a speech contains substantive public-policy discussion rather than procedural talk (e.g., scheduling and unanimous-consent requests). Congressional speeches mix these forms of activities, while procedural talk does not reflect the substantive policy disagreements that political scientists typically seek to measure (Gentzkow et al., 2019). Conditioning on this covariate allows us to guide hypothesis generation to reflect policy discussion beyond procedural activity. (2) NCTE (Demszky and Hill, 2023): the largest publicly available dataset of math classroom transcripts with rich expert annotations of instructional quality (Hill et al., 2008) and classroom environment (Pianta et al., 2012), collected by the National Center for Teacher Effectiveness (NCTE) between 2010-2013. The outcome variable y is low vs. high quality REMED, a measure of how well teachers remediate student math errors and difficulties in class. This teaching practice is central to math instruction as it provides students with opportunities for conceptual learning (Bray,

2011). We use behavioral management (e.g., organizing activities and redirecting students) as the covariate because management and instruction are intertwined: teachers often need to minimize time spent on behavioral management to focus on content instruction (Emmer and Stough, 2003). Conditioning on this covariate allows us to examine whether the model surfaces remediation-oriented teaching practices beyond behavioral management.

5.2 Study Setup

We expect the two covariates to steer hypothesis discovery toward within-stratum differences, analogous to the stratum-imbalance setting in §4.1.1. Prior work suggests that sign reversal is rare in practice (Higgins et al., 2024), so we compare the global LASSO baseline with demeaned-reweighted-lasso as the covariate-aware method. Our main metric is the *helpfulness* of each hypothesis, on a 1–5 scale, for understanding the group difference while accounting for the specified covariate. Each hypothesis is accompanied by its prevalence in the two outcome groups. After rating helpfulness, annotators reviewed the same hypotheses with global and covariate-stratified statistics and rated *conditional interpretive value* on a 1–5 scale: whether the breakdown by the specified covariate added interpretive value beyond the global finding. We set $|\mathcal{H}|=10$ to limit the cognitive burden on annotators.

We recruit research scholars for each domain. Specifically, we invited two computational social science scholars familiar with U.S. politics to annotate CONGRESS. And we invited two educational researchers who have used NCTE in prior research to annotate the NCTE dataset. More details on annotation instruction, expert recruitment, and covariate operationalization are in Appendix D.

5.3 Results and Discussion

To focus on complementary discoveries, we manually match semantically similar hypotheses across the two method outputs. Matched hypotheses capture patterns recovered by both methods and are therefore less diagnostic of the added value of covariate-aware selection. We report the full lists and matching decisions for both datasets in Appendix D.1. Differences among matched items may also reflect LLM interpretation noise, such as wording specificity. Table 3 lists the unique NCTE hypotheses by method as qualitative examples; aggregate expert ratings appear separately in Table 4.

Unique LASSO hypotheses
direct students for logistical tasks such as collecting items.
prompt student to write reflections about learning at the end.
use management language to enforce compliance.
Unique demeaned-reweighted-lasso hypotheses
Extended discussions of units of weight.
follow-up with specific students about needs.
organize to work in small groups and prompt students help and explain concepts to one another.

Table 3: NCTE hypotheses unique to each method after semantic matching. Hypotheses are shortened for readability. Full wording appears in Appendix D.1.

Covariate-aware selection surfaces useful within-stratum hypotheses. Table 3 illustrates this shift in NCTE: unique covariate-aware hypotheses move from classroom-management context toward remediation or other instructional activities. The global LASSO baseline surfaced two unique hypotheses about classroom management, the covariate-defined context we aim to account for. By contrast, the covariate-aware method surfaced more instruction-oriented hypotheses about units-of-weight discussion, individualized follow-up, and peer explanation. The small-group hypothesis is especially illustrative: rather than only identifying group organization, it emphasizes teachers prompting students to help and explain concepts to one another. The same pattern appears more modestly in CONGRESS: among the unique hypotheses, the global baseline retains a procedural hypothesis, while the covariate-aware method surfaces policy-related hypotheses about economic performance and border security. Together, these examples suggest that covariate-aware selection can surface unique hypotheses aligned with the researcher-specified strata. The averaged ratings in Table 4 further confirm that these hypotheses are more useful than their counterparts (3.10 vs. 2.50).

The value of the covariate depends on the dataset. Conditional interpretive value measures whether the $\mathbf{C}$-stratified statistics change how annotators interpret a generated hypothesis. In NCTE, behavioral management helps distinguish classroom organization from instructional activities, and unique covariate-aware hypotheses received higher conditional-value ratings than unique LASSO hypotheses (3.50 vs. 2.33). In CONGRESS, the policy/procedure covariate was less informative as a displayed breakdown: procedural talk remained

Unique hypotheses	Helpfulness	Cond.
Congress		
LASSO	2.75	2.50
demeaned-reweighted-lasso	3.25	1.50
NCTE		
LASSO	2.33	2.33
demeaned-reweighted-lasso	3.00	3.50
Overall		
LASSO	2.50	2.40
demeaned-reweighted-lasso	3.10	2.70

Table 4: Expert ratings for hypotheses unique to each method after semantic matching. Helpfulness and conditional interpretive value (Cond.) are 1–5 Likert means.

strong in both method outputs, and the two unique covariate-aware hypotheses were already clearly policy-related from their text. As one annotator noted, the policy/procedure breakdown often confirmed what was already expected rather than changing the interpretation. Accordingly, unique covariate-aware hypotheses in CONGRESS were rated as more helpful (3.25 vs. 2.75) but lower in conditional interpretive value (1.50 vs. 2.50).

6 Related Work

LLM-based hypothesis discovery. Complementing the task overview in §2.2, we situate LLM-based hypothesis generation among classical text-analysis tools and applications. Classical text analysis methods such as n-gram frequency comparisons (Gentzkow et al., 2019) and topic models, including LDA (Blei et al., 2003) and structural topic models (Roberts et al., 2014), remain popular tools for relating text to target variables (Grimmer et al., 2022), but typically produce word lists that require further interpretation. Recent LLM-based methods instead generate natural-language descriptions of corpus differences from sampled examples (Zhong et al., 2022, 2023), iterative error analysis (Zhou et al., 2024), literature-grounded candidates (Liu et al., 2025), or statistical modeling (Zhong et al., 2024; Movva et al., 2025; Agarwal et al., 2026). Related work has applied this idea to settings such as headline click rates (Batista and Ross, 2024), social media engagement (Xu et al., 2026), and causal measurement from text (Modarressi et al., 2025). OpenAI recently released a text measurement toolkit that also supports hypothesis discovery (Asirvatham et al., 2026), demonstrating its potential for real-world applications. Our work differs by incorporating domain knowledge through statistical models of covariate structure, steering discovery toward researcher-specified differences of interest.

Statistical methods for covariate adjustment. Our methods draw on standard statistical and econometric tools for making comparisons conditional on covariates. Fixed effects and within-group residualization are classical ways to remove nuisance variation (Angrist and Pischke, 2009), with the Frisch–Waugh–Lovell theorem providing the corresponding partialling-out justification (Lovell, 1963). Simpson's paradox (Simpson, 1951) shows why such adjustment matters: marginal associations can reverse relative to associations within each stratum (Blyth, 1972). When conditional effects change sign across strata, the problem becomes one of qualitative interaction (Gail and Simon, 1985). These results provide the statistical basis for using residualization, reweighting, and interaction terms as principled feature-selection mechanisms for conditional hypothesis generation.

7 Conclusion

We introduced conditional hypothesis generation for text analysis, a framework for steering natural-language hypothesis discovery toward differences that hold within researcher-specified covariate strata. This setting makes explicit two statistical challenges: *stratum imbalance*, where signals from underrepresented strata are suppressed, and *sign reversal*, where aggregation can cancel or misrepresent within-stratum differences. By incorporating covariates into the SAE feature selection, our methods let researchers encode domain knowledge about which strata should shape discovery.

We proposed two complementary covariate-aware methods. Interaction-lasso models feature–covariate interactions and is suited to sign reversal, while demeaned-reweighted-lasso removes stratum-level variation and reweights underrepresented strata to recover within-stratum signals under a consistent-direction assumption. Synthetic experiments show that these modeling choices matter: demeaned-reweighted-lasso recovers hypotheses suppressed by stratum imbalance, whereas interaction-lasso is needed when the direction of the difference reverses. Expert validation on two real-world datasets, CONGRESS and NCTE, further suggests that covariate-aware method can surface more useful hypotheses for researchers' stated comparisons than those unique to a global baseline.

Limitations

Dependence on researcher-specified covariates. Our framework intentionally makes covariate specification a design choice: researchers decide which contexts should shape hypothesis discovery based on domain knowledge and the comparison they want to make. This choice gives researchers control over the target of discovery, but it also means the methods are only as useful as the covariates supplied. They can reduce the influence of known covariate structure, but they cannot discover which covariates should matter, correct for omitted covariates, or make the resulting hypotheses causal claims. Poorly chosen or noisy covariates may steer discovery toward unhelpful strata, and overly fine-grained covariates may create sparse cells where within-stratum comparisons are unreliable.

Categorical strata. Both proposed methods are designed around observed covariate strata. In this paper we focus on binary or categorical covariates, such as policy/procedure status, behavior-management level, topic indicators, and time period. Continuous covariates would require discretization or a different residualization strategy, The quality of the conditional hypotheses therefore depends on the validity of the covariate operationalization, including any automated annotations used to define covariates.

Method-specific assumptions. The two methods address different statistical regimes. Demeaned-reweighted-lasso assumes no qualitative interaction: the relevant group difference should have a consistent direction across covariate strata. When the direction reverses, as in our sign-reversal experiment, residualization can remove nuisance variation but cannot recover the stratum-specific effects of interest. Interaction-lasso is better suited to such cases, but it expands the feature space and can be unstable when both SAE activations and covariate strata are sparse. In practice, researchers should inspect stratum-level statistics when possible and choose the selection model based on whether sign reversal is plausible.

Dependence on the SAE interpretation pipeline. Our implementation builds on the SAE-based hypothesis-generation pipeline of Movva et al. (2025). It therefore inherits the pipeline's requirements and failure modes (Peng et al., 2025): access to the full corpus for feature extraction, a trained SAE whose features are sufficiently interpretable for the domain, hyperparameter choices that affect the granularity of learned features, and an LLM interpretation step that can phrase selected features imperfectly. Our contribution modifies feature selection, not the underlying representation learning or natural-language verbalization stages.

Evaluation of discovery quality. The synthetic experiments provide controlled ground truth for stratum imbalance and sign reversal, but their reference hypotheses are constructed from bill-summary topics and may not cover the full range of linguistic phenomena or covariate choices encountered in applied text analysis. The real-world validation complements these experiments with expert judgments in two substantively different domains, CONGRESS and NCTE, while broader generalization will require additional datasets, covariates, and expert panels. Such validation is difficult because scientific discovery has no single objective metric: what counts as a useful hypothesis depends on domain expertise, the research question, and the implicit norms of the relevant research community rather than explicit deductive criteria alone (Polanyi et al., 2000). The ratings should therefore be interpreted as evidence that covariate-aware selection can surface useful hypotheses, not as a complete benchmark of performance across domains. Future evaluations should test more covariates, multi-valued and continuous settings, larger expert panels, and downstream research workflows where generated hypotheses are refined and validated by domain specialists.

References

E Scott Adler and John Wilkerson. 2018. Congressional bills project: 1995-2018.

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and 1 others. 2026. Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems*, 38:25181–25219.

Joshua D Angrist and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Hemanth Asirvatham, Elliott Mokski, and Andrei Shleifer. 2026. GPT as a measurement tool. Working Paper 34834, National Bureau of Economic Research.

Rafael M Batista and James Ross. 2024. Words that work: Using language to generate hypotheses. *Available at SSRN 4926398*.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. 2013. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Colin R Blyth. 1972. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366.

Wendy S Bray. 2011. A collective case study of the influence of teachers' beliefs and knowledge on error-handling practices during class discussion of mathematics. *Journal for Research in Mathematics education*, 42(1):2–38.

Dallas Card. 2019. *Accelerating Text-as-Data Research in Computational Social Science*. Ph.D. thesis, University of Washington.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Dorottya Demszky and Heather Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538.

Edmund T Emmer and Laura M Stough. 2003. Classroom management: A critical part of educational psychology, with implications for teacher education. In *Educational psychology*, pages 103–112. Routledge.

Mitchell Gail and Richard Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1):395–419.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. 2024. *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.5 edition. Cochrane.

Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4):430–511.

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.

Thomas J Kane and Douglas O Staiger. 2012. Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. research paper. met project. *Bill & Melinda Gates Foundation*.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2025. Literature meets data: A synergistic approach to hypothesis generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 245–281, Vienna, Austria. Association for Computational Linguistics.

Michael C Lovell. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Iman Modarressi, Jann Spiess, and Amar Venugopal. 2025. Causal inference on outcomes learned from text. *arXiv preprint arXiv:2503.00725*.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. In *International Conference on Machine Learning*, pages 44997–45023. PMLR.

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. 2025. Use sparse autoencoders to discover unknown concepts, not to act on known concepts. *arXiv preprint arXiv:2506.23845*.

Robert C Pianta, Bridget K Hamre, and Susan Mintz. 2012. *Classroom Assessment Scoring System: CLASS: Upper Elementary Manual*. Teachstone.

Michael Polanyi, John Ziman, and Steve Fuller. 2000. The republic of science: its political and economic theory minerva, i (1)(1962), 54-73. *Minerva*, 38(1):1–32.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Edward H Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Matt Taddy. 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4375–4389.

Paiheng Xu, Louiqa Raschid, and Vanessa Frias-Martinez. 2026. Does geo-co-location matter? a case study of public health conversations during covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 2518–2537.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Describing differences between text distributions with natural language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27099–27116. PMLR.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. 2024. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. 2023. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139.

A Synthetic Dataset Details

Topics are drawn from the BILLS dataset (Adler and Wilkerson, 2018; Hoyle et al., 2022), which provides 21 high-level topics and 114 low-level subtopics. For Scenario 1: conditional discovery in Section 4.1.1, we define two thematic pools: $\mathcal{T}_{gov}$ (Government & Economy) and $\mathcal{T}_{soc}$ (Social Policy). A topic is eligible if it has at least two subtopics each with ≥ 320 samples, ensuring sufficient data to support the imbalanced sampling under the largest imbalance ratio tested. The eligible topics retained for each theme are listed in Table 5. Each Scenario 1 corpus contains N=2,400 bill summaries, balanced across A and B (N_A=N_B=1,200). For each selected topic, 400 examples are allocated across the two corpora. For example, when r=0.20, corpus A contains $3\times400\times0.20$=240 examples from $\mathcal{T}_{soc}$ and $3\times400\times0.80$=960 examples from $\mathcal{T}_{gov}$, with corpus B receiving the mirror allocation. Each Scenario 2 corpus contains N=3,000 bill summaries (N_A=N_B=1,500), with equal-sized pre-2011 and post-2011 strata so that the within-period signals cancel at the aggregate level.

Theme	Eligible Topics
$\mathcal{T}_{gov}$ (Gov. & Econ.)	Government Operations, Defense, Domestic Commerce, Macroeconomics, Law and Crime, Energy, Transportation
$\mathcal{T}_{soc}$ (Social Policy)	Health, Education, Environment, Public Lands

Table 5: Eligible topic pools for conditional discovery. In each seed run, three topics are randomly sampled from each pool.

B Evaluation Metrics

We follow previous work (Zhong et al., 2024; Movva et al., 2025) in evaluating synthetic experiments by matching generated hypotheses to the known ground-truth differences. For each scenario, we repeat the experiment for 10 random seeds and report the mean and standard deviation of each metric. We split each synthetic dataset into training, validation, and test sets in a 60:20:20 ratio. The validation set is used for SAE hyperparameter selection, as described in Appendix C.5.

On the test set, reference annotations are determined directly by the synthetic construction. For each generated hypothesis, we use `Qwen3-30B-A3B-Instruct-2507` to produce bi-

nary annotations on the same test texts. Hypothesis annotation is an expensive operation. Movva et al. (2025) uses `GPT-4o-mini`, and we use this open-sourced model because we can host it locally to reduce costs, and it matches `GPT-4o-mini` in annotation performances. We internally evalutated the two models on a sample 20 hypotheses and 30 bill summaries per hypothesis, resulting in 600 text-concept pairs. The agreement is 95.3%, so we use `Qwen3-30B-A3B-Instruct-2507` for hypothesis annotation throughout the experiments. We hosted the model on two RTXA6000 GPUs with vLLM for fast inference.

Now we have one reference annotation vector (obtained from the dataset) and one generated annotation vector for each hypothesis. We compute the correlation between every reference/-generated pair, forming a $|\mathcal{H}| \times |\mathcal{H}|$ correlation matrix, and use the Hungarian algorithm to find the maximum-correlation one-to-one matching. The LLM used for judging surface similarity is `gpt-4.1-2025-04-14`.

C Method Implementation Details

C.1 SAE-Based Baselines

We compare against two global SAE feature-selection baselines. **SAE-lasso** uses Eq. 1 to select the top $|\mathcal{H}|$ features, using the implementation provided by Movva et al. (2025).[2] **SAE-separation score** ranks each feature by the mean of y among the highest-activating texts versus zero-activating texts, then selects the top $|\mathcal{H}|$ features. Both baselines use the same SAE activation matrix as the proposed methods.

C.2 Prompts

C.3 LLM-Direct Baselines

The LLM-direct baselines test whether covariate-aware discovery can be obtained simply by showing sampled texts to an LLM. **llm-global** generates $|\mathcal{H}|$ hypotheses from sampled corpus texts without covariate information. **llm-covariate** uses the same direct prompting setup but exposes covariate-profile labels. For fair comparison, we use `gpt-4.1-2025-04-14` for hypothesis proposal in the main tables and `Qwen3-30B-A3B-Instruct-2507` for hypothesis annotation.

The LLM-direct prompts differ from the SAE interpretation prompt in Table 8 in two ways. First, their positive and negative examples are sampled corpus texts, rather than examples chosen by activation of a selected SAE feature. Second, they ask the LLM to propose up to $|\mathcal{H}|$ corpus-level differences directly, rather than to name exactly one feature explaining one selected neuron. The **llm-covariate** prompt makes one additional change relative to **llm-global**: each text is prefixed with its covariate value(s), and the model is instructed to look for patterns that are strong either globally or within a covariate stratum.

We swept $n_{\text{prompt}} \in \{25, 50, 100\}$ sampled texts per corpus and proposal models `gpt-4.1-2025-04-14` and `gpt-5.4-2026-03-05`. Figures 2–3 visualize the seed-level distributions across the full prompt-budget sweep. The sweep shows that increasing the prompt budget does not consistently improve either LLM-direct baseline; covariate labels and the stronger proposal model also do not produce consistent gains. We therefore use $n_{\text{prompt}}=100$ with `gpt-4.1` as a generous fixed-budget setting rather than tuning the LLM-direct baseline per condition.

C.4 Reference Baselines

We include two reference baselines to calibrate the synthetic evaluation. **Oracle** is an upper reference: for each targeted ground-truth difference, 20 texts are randomly sampled from the exact positive and negative sides of the contrast. In Scenario 1, these are the paired A and B subtopics for the target social-policy topic. In Scenario 2, these are one $\mathcal{T}_1$ subtopic versus all $\mathcal{T}_2$ subtopics combined. These texts and their corpus labels (A or B) are directly provided to the LLM for interpretation using the prompt in Table 8. **Random** measures how much recovery can be obtained from general samples alone: 20 training texts are randomly sampled into positive and negative halves, then passed to the same prompt. This is not a pure lower bound, since sampled texts still contain topical cues that can receive partial credit. We set the sample size to 20 to be consistent with the SAE interpretation stage (Movva et al., 2025).

C.5 SAE Hyperparameter Tuning

We tune SAE hyperparameters on the validation split and use the selected SAE to extract feature activations for all SAE-based and proposed methods. Following the released HypotheSAEs imple-

[2] https://github.com/rmovva/HypotheSAEs

LASSO hypothesis	Δ	demeaned-reweighed-lasso hypothesis	Δ
Matched hypotheses			
the teacher repeatedly prompts students to explain their reasoning or clarify their answers at multiple steps in multi-step fraction or mixed number problems, often after an initial incorrect or incomplete response	+24.7	*the teacher prompts students to explain or justify their reasoning after a partially correct or incorrect answer, and facilitates peer discussion to resolve confusion*	+21.7
teacher prompts students to explicitly explain or justify their procedural steps or reasoning during problem-solving, guiding them through multi-step solutions by asking follow-up questions when confusion or errors occur	+21.9	*teacher repeatedly prompts students to show all intermediate steps in their math work (e.g. requiring students to write down each step, not just the final answer, asking them to explain or justify each step, correcting students who do steps in their head or skip showing borrowing/regrouping, explicitly tying process visibility to grading or real-life situations)*	+16.4
the teacher prompts students to explicitly explain or justify their mathematical thinking while solving arithmetic problems, with multiple follow-up questions driving students to clarify errors or reasoning (rather than simply confirming answers or having students display work)	+23.0	*teacher asks students to explain or justify their mathematical reasoning, often following up on errors or confusions, rather than just stating answers or procedures*	+21.0
the teacher asks students to explicitly explain or justify their reasoning or chosen strategy for solving a problem	+19.1	*the teacher prompts students to explain or compare their solution strategies to each other*	+11.2
teacher prompts students to explicitly identify and correct procedural mistakes in place value or decimal operations	+17.2	*the teacher frequently prompts students to explain their reasoning, describe their solution process, or justify place value choices during multi-digit arithmetic problems*	+15.3
the teacher assigns or reminds students about their math homework specifically during lesson closure or class transitions	+0.3	*teacher assigns or discusses specific math homework or practice problems for students to complete outside of class*	+10.2
teacher gives procedural classroom management instructions related to cleaning up materials, lining up, or transitioning between activities	-7.1	*teacher directing students to physically move between locations or groups for collaborative or group work*	-5.2
Unique hypotheses			
the teacher directs students to perform classroom management or logistical tasks, such as handing out papers, putting away materials, collecting items, or resetting groups	-9.7		
teacher prompts students to reflect in writing (e.g. journals or notebooks) on what they learned, their confusions, or their problem-solving process at the end of the lesson	-2.6		
frequent use of whole-class management language to enforce procedural compliance and orderly transitions, including directives for materials (e.g. whiteboards, slates), timing cues (e.g. minute warnings, counting down), and group routines (e.g. 'boards up', 'push your chair in'), delivered alongside academic activities	-6.2		
		teacher organizes students to work in small groups or pairs and prompts them to help and explain concepts to one another	+6.6
		teacher provides individualized follow-up or checks in with specific students about their understanding or needs during or after whole-class instruction	+20.6
		contains extended classroom discussions about units of weight (grams, kilograms, pounds, ounces, tons), including teacher-student exchanges involving conversions, comparisons, or real-life analogies related to weight measurement	-0.6

Table 6: NCTE full hypothesis list used in the expert evaluation, with matched and unique sections in one table. Color coding: green shades indicate ↑ mid/high REMED and red shades indicate ↑ low REMED; light/lighter shades alternate by row for readability. Δ = H/M − Low.

mentation [3], we treat the number of learned features M and the per-example sparsity K as the main SAE hyperparameters: M controls how many concepts the model can represent, while K controls how many concepts are active for each text. The implementation recommends vanilla Top-K choices of approximately $(M, K) = (64, 4)$ for $\sim$1K examples, $(256, 8)$ for $\sim$10K examples, and $(1024, 8)$ for $\sim$100K examples, with larger M yielding finer-grained concepts and K typically in the range 1–32. For the synthetic bill corpora ($N = 2400$) and the NCTE corpus ($\sim$10K examples), we therefore sweep a small grid around the rule-of-thumb setting: $M \in \{128, 256, 512\}$ and $K \in \{4, 8, 16\}$. For CONGRESS, we instead match the larger range used for the Congressional speech experiments in Movva et al. (2025), sweeping $M \in \{1024, 2048, 4096\}$ and $K \in \{8, 16, 32\}$. For each candidate pair, we train a vanilla Top-K SAE on the training embeddings with validation embeddings for early stopping, select predictive features, and generate candidate interpretations. On real-world datasets, the selected configuration is the one with the best validation predictive score (AUROC for binary classification, or R^2 for re-

[3] https://github.com/rmovva/HypotheSAEs

LASSO hypothesis	Δ	demeaned-reweighted-lasso hypothesis	Δ
Matched hypotheses			
requests unanimous consent for procedural actions in the Senate, such as granting floor privileges, scheduling committee meetings, agreeing to resolutions, making technical amendments, or permitting specific actions during Senate proceedings	+6.9	*requests unanimous consent for procedural actions within the Senate or House, such as granting floor privileges, modifying amendments, scheduling votes, printing materials in the record, or controlling debate time*	+7.4
requests unanimous consent to authorize a committee or subcommittee meeting or hearing in the Senate, specifying the committee/subcommittee, date, and topic or title of the meeting	+2.9	*announces or authorizes official Senate committee hearings or meetings, specifying committee names, dates, times, and locations*	+3.3
gives or discusses the schedule and timing of upcoming Senate votes, bill considerations, or legislative actions (including mentioning specific days, times, or order of business)	+7.5	*discusses the scheduling and process of votes and amendments on the Senate floor, including references to specific times, sessions, and expectations for voting activity*	+4.6
criticizes government or presidential leadership for mismanagement or policy failures, particularly in areas such as war, defense, budgets, or major federal programs	-20.0	*criticizes government or institutional incompetence or failure to act, particularly in relation to policies, disasters, or administration*	-25.1
discusses or criticizes tax benefits, tax cuts, or loopholes that disproportionately benefit the wealthiest Americans or top income percentiles (e.g. top 1%, billionaires, richest families), often contrasting with the needs of the middle or lower class	-3.3	*criticizes tax cuts or fiscal policies that disproportionately benefit wealthy individuals or families at the expense of the poor, the middle class, or public programs*	-8.5
discusses the United States national debt, including specific figures or references to the growing debt, debt ceiling, and its impact on citizens and the country	-2.8	*discusses the increase of the national debt or debt ceiling of the United States, with reference to specific amounts, time periods, or presidential administrations*	-2.9
criticizes proposed federal budget cuts, especially emphasizing negative impacts on socially vulnerable groups or essential services (such as veterans, students, low-income families, healthcare, or education)	-9.6	*discusses government spending, budget deficits, and the need for fiscal responsibility or limiting federal expenditures*	-4.3
criticizes or discusses actions, policies, or leadership of the Republican Party or House Republicans	-12.1	*criticizes House or congressional Republicans for their policies, actions, or leadership*	-11.7
Unique hypotheses			
discusses progress, specific events, or achievements related to democratization and government-building in Iraq, particularly referencing Iraqi elections, voting, establishment of new governmental bodies, and the training or accomplishments of Iraqi security forces	+0.1		
criticizes perceived procedural unfairness or lack of accountability within Congress or government institutions	-16.8		
		discusses United States border security, including measures such as border fences, patrols, or enforcement against illegal crossings	+1.6
		presents recent U.S. economic statistics or indicators (such as GDP growth, job creation, unemployment rate, or manufacturing output) to argue that the U.S. economy is strong or improving, often attributing this to Republican or presidential policies	+0.7

Table 7: Congress full hypothesis list used in the expert evaluation, with matched and unique sections in one table. Color coding: red shades indicate ↑ Republican and blue shades indicate ↑ Democrat; light/lighter shades alternate by row for readability. Δ = Rep. − Dem.

gression). On synthetic datasets, where reference hypotheses are available, the selected configuration is the one with the highest validation surface similarity to the ground-truth hypotheses.

D Real-World Validation Details

D.1 Matching Decisions and Full Hypothesis Lists

Tables 7 and 6 report the complete matched and unique lists used in the expert analysis. Colored cells indicate direction by global prevalence difference: Congress uses red for ↑ Republican and blue for ↑ Democrat; NCTE uses green for ↑ mid/high REMED and red for ↑ low REMED. All prevalence values are from the user-study annotation artifacts; Δ is the difference between the two group prevalences shown in each table.

D.2 Covariate operationalization

For CONGRESS, the covariate is a binary LLM annotation of whether a speech segment "contains substantive discussion of public policy, rather than only congressional procedure." We use the same CONGRESS split as Movva et al. (2025): approximately 114K training, 16K validation, and 12K held-out speeches. For NCTE, the covariate is derived from the CLASS Behavior Management score (CLBM) (Pianta et al., 2012): segments with CLBM<6 are marked as low-quality behavior management, while segments with CLBM≥ 6 are marked as high-quality behavior management. We expect to see less management language in the high-quallity segments. This threshold marks the low and mid ranges on the CLASS scale and captures roughly 30% of segments. This imbalanced distri-

SAE Interpretation Prompt

You are a machine learning researcher who has trained a neural network on a text dataset. You are trying to understand what text features cause a specific neuron in the neural network to fire.

You are given two sets of SAMPLES: POSITIVE SAMPLES that strongly activate the neuron, and NEGATIVE SAMPLES from the same distribution that do not activate the neuron. Your goal is to identify a feature that is present in the positive samples but absent in the negative samples.

`{goal}`

POSITIVE SAMPLES: `{positive_texts}`

NEGATIVE SAMPLES: `{negative_texts}`

Rules about the feature you identify:
The feature should be objective, focusing on concrete attributes rather than abstract concepts. The feature should be present in the positive samples and absent in the negative samples. Do not output a generic feature which also appears in negative samples. The feature should be as specific as possible, while still applying to all of the positive samples.

Do not output anything besides the feature. Please suggest exactly one feature, starting with - and surrounded by quotes.

Region	**`{goal}` field**
Synthetic bills	All of the texts are summaries of US legislative bills. Features should describe a specific aspect of the bill. For example: "is about healthcare"; "is about higher education."
CONGRESS	All of the texts are excerpts of speeches from the US Congress. Example features may include: "mentions that the economy is strong"; "discusses environmental issues or environmental policy."
NCTE	All texts are transcripts of ∼7.5-minute segments from K–12 mathematics lessons, rated on REMED (remediation of student errors and difficulties). Example features distinguishing high-vs-low quality remediation include: "the teacher follows up on student confusion with probing questions rather than re-explaining procedurally"; "the teacher revoices or builds on incorrect student contributions to guide conceptual understanding"; "students are given the opportunity to self-correct or revise their thinking."

Table 8: Prompt template used for the SAE interpretation step, with dataset-specific substitutions for the `{goal}` field.

Ground-truth hypothesis	**Generated hypothesis**	**Avg. score**
is about Health: Mental	is about mental health services or policies	1.0
is about Agriculture: Food Inspection & Safety	is about agriculture-related financial assistance	0.5
is about Technology: Broadcast	is about children's health or welfare	0.0

Table 9: Example matched ground-truth and generated hypotheses from the synthetic bills experiment, with averaged LLM surface-similarity judgment scores.

bution also allows us to test the stratum imbalance scenario in a real-world dataset, which is important for NCTE as high-quality teaching samples are rare in general (Kane and Staiger, 2012; Xu et al., 2024). The dataset size has around 10K samples in total with 5K rated as mid/high quality REMED and 5K rated as low quality REMED. We split the dataset into train/val/testsets in a 60:20:20 ratio. The scores were provided by trained annotators and come with NCTE dataset. These covariates were pre-specified before expert annotation.

D.3 Annotation instructions

Annotators reviewed method-blinded hypothesis sets from LASSO and demeaned-reweighted-lasso, shown as Set A and Set B. They were instructed to use their domain judgment; there were no right or wrong answers. First, they rated each hypothesis for helpfulness in understanding the outcome-group difference on a 1–5 scale (1: not helpful; 3: somewhat helpful; 5: very helpful). They then reviewed the same hypotheses with global and covariate-stratified statistics and rated whether the covariate breakdown added interpretive value beyond the global finding, also on a 1–5 scale (1: no value; 3: some value; 5: high value). Annotators could optionally provide free-text notes explaining

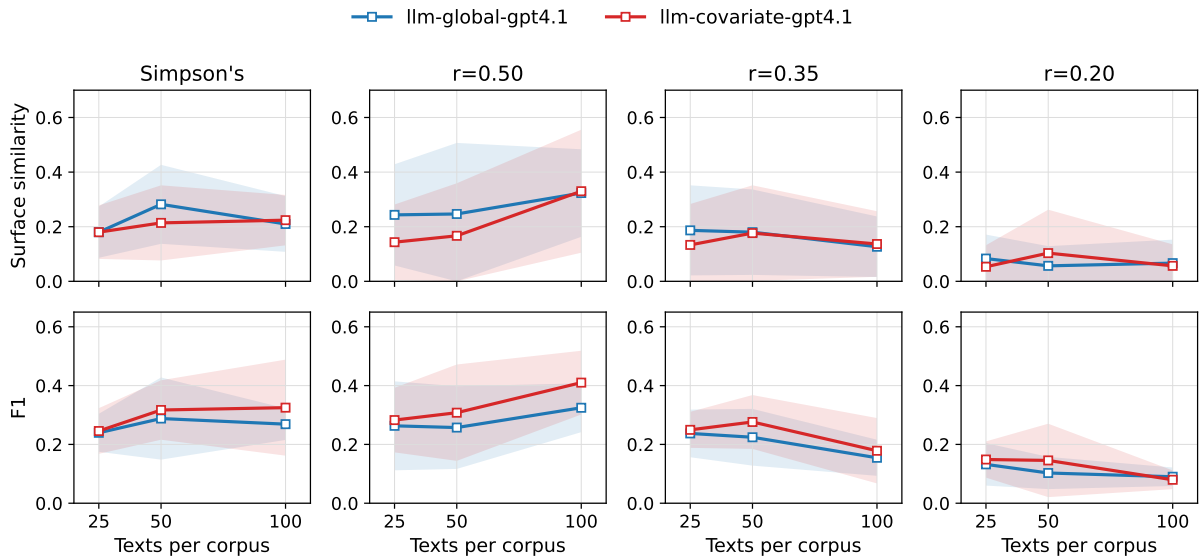

Figure 2: Prompt-budget sensitivity for LLM-direct baselines using `gpt-4.1`. Each panel varies the number of sampled texts per corpus used in the direct prompt. Lines show seed means; shaded bands show ± 1 standard deviation across seed runs. The trends are non-monotonic: larger prompt budgets and covariate labels do not consistently improve surface similarity or F1.

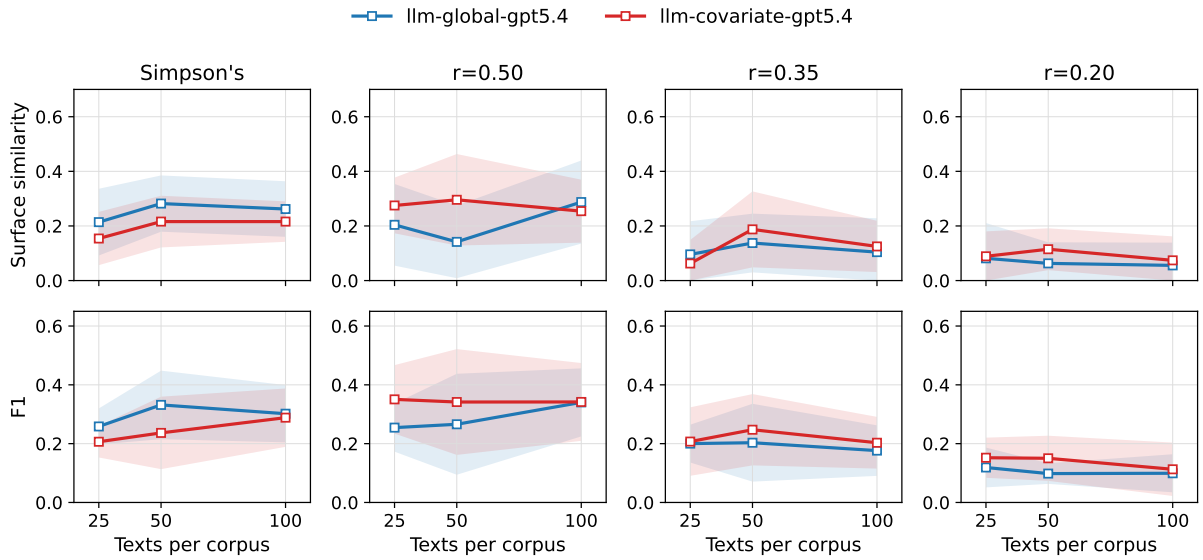

Figure 3: Prompt-budget sensitivity for LLM-direct baselines using `gpt-5.4`. Each panel varies the number of sampled texts per corpus used in the direct prompt. Lines show seed means; shaded bands show ± 1 standard deviation across seed runs. The qualitative pattern remains non-monotonic, and covariate-labeled prompting does not consistently dominate global prompting.

their ratings. For each dataset, we use manually curated one-to-one matched pairs and mark all remaining hypotheses as method-unique.

D.4 Recruitment and IRB.

Participation was voluntary. We invited research scholars with relevant domain expertise: computational social science researchers familiar with U.S. politics for CONGRESS, and education researchers familiar with NCTE for the classroom-transcript study. One NCTE annotator also had years of elementary teaching experience. The study was reviewed by an institutional review board and determined exempt under 45 CFR 46.104(d)(2)(ii).

Surface Similarity Prompt

Are text_a and text_b similar in meaning? Respond with yes, related, or no.

Here are a few examples.

Example 1:
text_a: has a topic of protecting the environment
text_b: has a topic of environmental protection and sustainability
output: yes

Example 2:
text_a: has a language of German
text_b: has a language of Deutsch
output: yes

Example 3:
text_a: has a topic of the relation between political figures
text_b: has a topic of international diplomacy
output: related

Example 4:
text_a: has a topic of the sports
text_b: has a topic of sports team recruiting new members
output: related

Example 5:
text_a: has a named language of Korean
text_b: uses archaic and poetic diction
output: no

Example 6:
text_a: has a named language of Korean
text_b: has a named language of Japanese
output: no

Example 7:
text_a: describes an important 20th century historical event
text_b: describes a 20th century European politician
output: no

Target:
text_a: {text_a}
text_b: {text_b}
output:

Table 10: Prompt used for judging the surface similarity between ground-truth and inferred hypotheses. Adapted from Zhong et al. (2024); Movva et al. (2025).