CONDITIONAL HYPOTHESIS GENERATION FOR LLM-BASED TEXT ANALYSIS WITH RESEARCHER-SPECIFIED COVARIATES

Anonymous authors
Paper under double-blind review

ABSTRACT

An important goal of text analysis is to discover interpretable hypotheses about how language varies with outcomes of interest. Recent LLM-based methods select globally discriminative patterns that do not necessarily answer a researcher's question. We introduce *conditional hypothesis generation*, the task of generating hypotheses that distinguish outcome groups within researcher-specified covariate strata. Covariates encode domain knowledge about which comparisons matter and can guide hypothesis discovery. We study two statistical challenges that arise in conditional discovery: relevant strata may be underrepresented within outcome groups (*stratum imbalance*), and differences may reverse direction across strata (*sign reversal*). Using sparse autoencoder features, we examine stratified regression, interaction models, and demeaning as covariate-aware methods. Controlled synthetic experiments show that these approaches can improve recovery of targeted differences overlooked by global methods, with effectiveness depending on the structure of conditional associations. On two real-world datasets, experts rate the covariate-aware method's unique hypotheses as more helpful on average than the global baseline's. We also illustrate how conditional comparisons can help researchers interpret generated hypotheses and formulate new ones.

1 INTRODUCTION

An important goal of text analysis is to generate interpretable hypotheses about how language varies with outcomes of interest, such as political affiliation, instructional quality, or social media engagement (Grimmer & 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, the most globally discriminative patterns do not necessarily answer a researcher's question. Meaningful comparisons may require focusing on particular contexts or accounting for covariates that shape the observed differences—a longstanding concern in text-as-data research (Grimmer & 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. The challenge is to let researchers specify these covariates and use them to guide which textual differences are discovered.

We introduce *conditional hypothesis generation*, the task of 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.

We study two statistical challenges in discovering conditional differences (Simpson, 1951; Gail & Simon, 1985). A relevant stratum may be underrepresented within an outcome group, so that its signal is dominated by more prevalent patterns (*stratum imbalance*); or a difference may reverse direction across strata (*sign reversal*), so that global aggregation misrepresents the conditional pattern.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We build on the approach of 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 study several statistical approaches—stratified regression, interaction models, and demeaning—that differ in how they represent feature–outcome associations within covariate strata and aggregate these associations across strata.

Controlled synthetic evaluations with known ground-truth hypotheses show that covariate-aware selection can improve recovery of targeted differences overlooked by global selection. Across the imbalance and sign-reversal settings, the approaches exhibit different recovery patterns, showing how modeling assumptions about shared versus context-specific associations shape discovery.

We further evaluate conditional generation on two real-world datasets: CONGRESS, a longstanding testbed for political language (Gentzkow & Shapiro, 2010; Grimmer et al., 2021), and NCTE, a dataset of math classroom transcripts with rich annotations of instructional quality and classroom environment (Demszky & Hill, 2023; Hill et al., 2008; Pianta et al., 2012). Expert evaluations show that hypotheses unique to the evaluated covariate-aware method receive higher average helpfulness ratings than those unique to the global baseline. We also examine how covariate breakdowns inform the interpretation of generated hypotheses. By showing how prevalence differences between outcome groups vary across strata, these breakdowns can help researchers refine their interpretation of generated hypotheses and formulate new ones.

Our contributions are: (1) We formalize conditional hypothesis generation for text analysis, using researcher-specified covariates to define the comparisons that guide discovery. (2) We study several statistical approaches to covariate-aware feature selection and evaluate their recovery of targeted differences under stratum imbalance and sign reversal. (3) On two real-world datasets, experts assess the usefulness of generated hypotheses and the interpretive value of covariate breakdowns.

2 PRELIMINARIES

2.1 LLM-BASED HYPOTHESIS GENERATION

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 the document's group membership. The goal of *hypothesis generation* is to produce a set of natural-language hypotheses $\mathcal{H}$ describing how textual patterns differ across outcome values, where the number of hypotheses, $|\mathcal{H}|$, is prespecified.

LLM-based methods can propose natural-language hypotheses directly from sampled labeled texts, then filter candidates by their discriminative power (Zhong et al., 2022; 2023; Zhou et al., 2024). Other approaches use statistical models to select predictive features over the full dataset, then ask an LLM to describe each selected feature from associated texts (Zhong et al., 2024; Movva et al., 2025). We build on Movva et al. (2025), who use a sparse autoencoder (SAE) trained on text embeddings to obtain a fixed set of interpretable features for statistical selection and subsequent LLM interpretation.

2.2 HYPOTHESIS GENERATION WITH SAE

An 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**: for each selected feature, an LLM generates a natural-language hypothesis describing the contrast between strongly activating documents (positive examples) and zero-activating documents (negative examples); Table 9 shows the prompt.

We incorporate researcher-specified covariates into the feature-selection step, allowing within-context associations to guide which features are interpreted while retaining the SAE encoding and LLM interpretation stages.

3 CONDITIONAL HYPOTHESIS GENERATION

3.1 PROBLEM FORMULATION

The textual attributes with the largest global prevalence differences do not necessarily answer a researcher's question. The question may call for examining differences within contexts of interest (e.g., partisan differences in speeches about education) or accounting for background variation (e.g., speakers' home states when studying national park discussion; §1). We let researchers specify these contexts as covariates C, drawn from metadata or measured text attributes, using domain knowledge to guide which differences are discovered.

Let $a_h(X) \in \{0, 1\}$ indicate whether text X exhibits the attribute described by hypothesis h. To control for C, we compare outcome groups within contexts $C{=}c$ represented in both groups, defining the conditional prevalence difference for a binary outcome $Y \in \{0, 1\}$ as

$$\Delta_h(c) = \mathbb{E}[a_h(X) \mid Y = 1, C = c] - \mathbb{E}[a_h(X) \mid Y = 0, C = c]. \tag{1}$$

Conditional hypothesis generation seeks textual attributes that distinguish outcome groups within contexts defined by researcher-specified covariates (C). These differences may persist across contexts, occur only in particular contexts, or reverse direction across contexts.

We study two statistical challenges in discovering conditional differences (Simpson, 1951; Blyth, 1972; Gail & Simon, 1985): (1) **stratum imbalance**: a relevant covariate stratum may be underrepresented within an outcome group, so that its signal is dominated by more prevalent patterns. (2) **sign reversal**: the direction of the group difference changes across strata, so the global difference can cancel or misrepresent the targeted differences.

3.2 COVARIATE-AWARE FEATURE SELECTION

We examine established statistical approaches to these challenges by applying them to SAE feature selection. Let $Z \in \mathbb{R}^{N \times M}$ contain the M feature activations for each document. Global LASSO selects features predictive of y from Z; we incorporate covariates into this selection step so that within-context associations influence which attributes are interpreted. Each approach below assigns a score q_m to feature m, and the top $|\mathcal{H}|$ features enter the same LLM interpretation stage (§2.2).

Stratified regression. Stratified regression estimates feature–outcome associations separately within each covariate stratum. Features are standardized once using pooled training data before fitting the stratum-specific models. Let Z_g and y_g denote the activations and outcomes in stratum g. **Stratified-max** fits a LASSO to (Z_g, y_g) in each supported stratum and scores features by $q_m = \max_g |\hat{\beta}_{g,m}|$. The maximum combines separately estimated associations into a feature ranking, without jointly modeling those associations across strata. Separate fits use only the observations in their respective strata, so coefficient estimates can be unstable in small strata; taking the maximum can promote such noisy estimates.

Interaction regression. A standard approach for modeling covariate-specific effects in regression is to include interaction terms, allowing a feature's association with the outcome to vary with context (Angrist & Pischke, 2009). We augment Z with the observed covariate matrix $\mathbf{C} = [C_1, \ldots, C_P]$ and the products $Z \odot C_p$, which multiply each feature activation by the corresponding document's covariate value. **Interaction-lasso** fits LASSO on this augmented matrix, with feature coefficients β, covariate coefficients δ, and interaction coefficients γ_p. It ranks features by $q_m = \max\{|\hat{\beta}_m|, \max_p |\hat{\gamma}_{p,m}|\}$. This requires fitting $M + P + MP$ predictors, increasing computational cost. Sparse activations in covariate groups can also destabilize interaction estimates.

Demeaning and reweighting. To address these limitations, we adapt another standard econometric technique, demeaning, which subtracts stratum means from observations to remove between-stratum variation and isolate within-stratum differences (Angrist & Pischke, 2009). For document i in stratum g_i, we subtract the stratum means from both activations and outcomes: $\tilde{z}_{i,m} = z_{i,m} - \bar{z}_m^{(g_i)}$ and $\tilde{y}_i = y_i - \bar{y}^{(g_i)}$. **Demeaned-lasso** then fits a LASSO to $(\tilde{Z}, \tilde{y})$ and ranks features by $q_m = |\hat{\beta}_m|$. **Demeaned-reweighted-lasso** additionally assigns observation weights $w_i \propto 1/n_{g_i}$, where n_{g_i} is

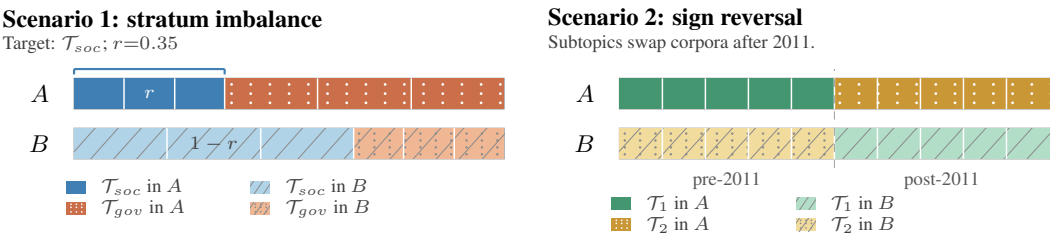

Figure 1: Corpus compositions for the synthetic datasets. In Scenario 1, r is the share of corpus A examples drawn from targeted social-policy topics.

the stratum size, so each stratum contributes equal total weight to the fit. Both regressions estimate shared coefficients from within-stratum variation across all strata; opposite-direction associations can therefore cancel during the fit.

Full method details and limitations are in Appendices C.1 and C.2.

4 Experiments on Synthetic Datasets

4.1 Synthetic Dataset Construction

We evaluate how well the methods recover *known ground-truth hypotheses* from synthetic datasets. We use a dataset containing bill summaries from the 110th–114th U.S. Congresses (Hoyle et al., 2022). It includes human-labeled high-level topics and granular subtopics, along with metadata such as bill creation time, allowing us to construct synthetic corpora with *controlled covariate structure*. We construct two data scenarios demonstrating the 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"). A ground-truth hypothesis takes the form "is about Health: Mental". The targeted differences are in the social policy theme, so we evaluate recovery of ground-truth hypotheses under $\mathcal{T}_{soc}$.

To simulate the case when relevant strata are rare, an imbalance ratio r controls the frequency of $\mathcal{T}_{soc}$ 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 (further dataset details are in Appendix A). These covariates should steer the methods toward the relevant strata, without 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. Our target is the set of subtopics that characterize A relative to B in the pre-2011 period. We pass one binary time-period covariate (pre-2011=1) to covariate-aware methods and evaluate recovery of the five $\mathcal{T}_1$ subtopics. By construction, the within-period differences have opposite signs and can cancel in the global comparison. This setting tests whether conditional selection preserves differences with opposing directions.

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

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
sep. score	.487±.17	.340±.16	.397±.16	.221±.10	.333±.13	.228±.05
Conditional						
stratified-max	†**.747**±.08	††**.610**±.08	††**.730**±.08	††.591±.12	††**.777**±.11	††**.573**±.09
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
conditional sep. score	†.677±.12	††.535±.11	.470±.10	†.363±.11	.347±.12	.244±.07

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 algorithm on correlations between reference and generated annotations. See Appendix B for the detailed procedure. We report two metrics computed over matched pairs. **Surface similarity**: for each matched reference/generated 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 12, examples in Table 10). 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 generated hypotheses' *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 a fair comparison, we use `gpt-4.1` to propose hypotheses unless otherwise specified, and OpenAI's `text-embedding-3-small` for embeddings.

SAE-based. (1) **LASSO**: use Eq. 2 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.

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.4 analyzes sensitivity to 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.5.

Conditional methods. We compare **stratified-max**, **interaction-lasso**, and **demeaned-reweighted-lasso** against the baselines and include **demeaned-lasso** as an ablation. We also include a **conditional separation score**, which computes the separation score within covariate strata

Table 2: Sign reversal: $|\mathcal{H}|$=5. †/††: p<0.05/0.01 vs. lasso, exact paired sign-flip test.

Group	Method	Surf. Sim.	F1
Reference	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
Conditional	interaction-lasso	††**.600**±.10	††**.576**±.08
	stratified-max	.548±.08	†.482±.10
	demeaned-reweighted-lasso	.448±.14	.335±.11
	demeaned-lasso (ablation)	.454±.11	.398±.12
	conditional sep. score	.442±.15	.380±.14

and ranks features by the largest absolute score across the global and within-stratum comparisons (Appendix C.2).

Statistical testing. Each conditional 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

Covariate-aware selection steers discovery toward specified contexts and recovers suppressed within-stratum signals. In Table 1, stratified-max and demeaned-reweighted-lasso significantly improve both surface similarity and F1 over global LASSO across all three ratios, including the balanced setting (r=0.50). These gains indicate that covariates help direct discovery toward the targeted within-stratum differences. As imbalance increases, global LASSO 's recovery declines, while both conditional approaches maintain near-oracle surface similarity. Demeaning helps recovery in the imbalanced settings, and reweighting further improves performance across all r. [1] Conditional separation score shows less consistent gains under imbalance.

The structure of conditional differences shapes which selection methods are effective. Table 2 evaluates sign reversal, where the time-period covariate reverses the differences across strata and makes the aggregate signal misleading. Interaction-lasso significantly improves both metrics over LASSO, while stratified-max significantly improves F1. Both approaches allow feature–outcome associations to vary across contexts, accommodating group differences that reverse direction. However, this property does not guarantee better recovery. Interaction-lasso does not improve over LASSO in Table 1. Its expanded feature space and sparse activations within contexts (§3.2) offer possible explanations for this weaker performance. The demeaned methods instead estimate shared coefficients, so opposite-direction associations can cancel. Neither demeaned variant significantly improves over LASSO, consistent with the limitation discussed in §3.2.

Corpus-level covariate modeling generally outperforms prompt-level exposure. Both tables show that LLM-direct methods fall behind SAE-based methods. The sensitivity analysis in Appendix C.4 further shows that llm-covariate does not consistently outperform llm-global across sample sizes or with a stronger reasoning model (`gpt-5.4`).

Intended use: guide hypothesis exploration. These results suggest that assumptions about the structure of conditional differences can guide method selection. Hypothesis generation serves as an exploratory tool (Zhong et al., 2023); similarly, we expect researchers to specify covariates using domain knowledge and use the generated hypotheses and their conditional breakdowns as inspiration for iteratively formulating and refining their own hypotheses. Researchers can compare outputs

[1] We also evaluate a squared-loss variant of global LASSO as a supplementary check of the regression specification. Details are in Appendix C.1.

across methods to explore how different modeling assumptions shape the hypotheses generated. We next examine the usefulness of these outputs in real-world datasets through expert assessment.

5 Expert Assessment 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 & 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 activity, 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 & 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 and 2013. The outcome variable y distinguishes low- from 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 & 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. We expect noisier conditional patterns in real-world corpora than in the controlled synthetic constructions, motivating a shared fit that combines within-stratum evidence across contexts. We therefore compare the global LASSO baseline with demeaned-reweighted-lasso, which estimates shared feature coefficients while giving each stratum equal total weight. 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 recruited scholars with expertise in each domain. Specifically, we invited two computational social science scholars familiar with U.S. politics to annotate CONGRESS and two educational researchers who had used NCTE in prior research to annotate the NCTE dataset. Further details on annotation instructions, expert recruitment, and covariate operationalization are in Appendix D.

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

Unique LASSO hypotheses	Unique covariate-aware hypotheses
direct students to perform classroom logistical tasks	*extended discussions of units of weight*
use management language to enforce compliance	*individual follow-up about student needs*
prompt students to write reflections about learning at the end of class	*organize small-group work and prompt peer help and explanations*

Table 4: Expert ratings for hypotheses unique to each method after semantic matching. Helpfulness (Help.) and conditional interpretive value (Cond.) are 1–5 Likert ratings. Cells show mean $\pm$ SD across expert–hypothesis ratings; DR-Lasso denotes demeaned-reweighted-lasso.

	CONGRESS		NCTE		Overall	
Method	Help.	Cond.	Help.	Cond.	Help.	Cond.
LASSO	2.75 ± 0.50	2.50 ± 1.29	2.33 ± 0.82	2.33 ± 1.51	2.50 ± 0.71	2.40 ± 1.35
DR-Lasso	3.25 ± 0.96	1.50 ± 0.58	3.00 ± 0.89	3.50 ± 1.64	3.10 ± 0.88	2.70 ± 1.64

5.3 Results and Discussion

To focus on complementary discoveries, we manually match semantically similar hypotheses across the two method outputs.[2] 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. As a rating check, 83% of within-annotator helpfulness comparisons for matched pairs differed by at most one Likert point (Appendix D.4). Table 3 lists the unique NCTE hypotheses by method as qualitative examples; aggregate expert ratings appear separately in Table 4.

Covariate-aware selection surfaces useful within-stratum hypotheses. Table 3 illustrates this shift in NCTE: unique covariate-aware hypotheses move from the 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 average ratings in Table 4 favor the covariate-aware hypotheses in this selected set (3.10 vs. 2.50).

The value of the covariate depends on the dataset. Conditional interpretive value measures whether the 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 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

[2]Experts rated both complete, method-blinded lists independently of matching. GPT-5.4 proposed one-to-one matches between anonymized lists; the authors reviewed and approved the matching.

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).

Conditional breakdowns enrich the interpretation of generated hypotheses. Table 11 reports global and stratified prevalence for all ten covariate-aware NCTE hypotheses. For example, prompting students to explain their reasoning after incorrect answers is 21.7 percentage points more prevalent in high-REMED than low-REMED segments overall. The difference is 22.9 percentage points in well-managed classrooms but only 7.3 points in less well-managed classrooms. This pattern motivates the hypothesis that *classroom-management demands may constrain opportunities for teachers to elicit student reasoning during remediation*. This hypothesis speaks to a recognized challenge in mathematics education: balancing classroom management with opportunities for mathematical discussion (Bray, 2011; Biza et al., 2015). Both experts rated the conditional interpretive value of this breakdown 5/5, underscoring the substantive importance of this insight.

6 Related Work

LLM-based hypothesis discovery. Complementing the task overview in §2.1, 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. Zhong et al. (2022; 2023) train a classifier to distinguish the two corpora and use its confidence to guide which examples are sampled for hypothesis generation. Zhou et al. (2024) iteratively generate new hypotheses from examples that existing hypotheses misclassify. Other approaches draw on 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 & 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 statistical models of covariate structure incorporate domain knowledge and steer 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 & 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 & 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 and Limitations

We introduce conditional hypothesis generation: researchers specify comparisons through covariates, and feature selection surfaces within-stratum textual differences. Controlled experiments show gains under stratum imbalance and sign reversal, with effectiveness depending on the structure of the associations. On CONGRESS and NCTE, experts rated hypotheses unique to the evaluated covariate-aware method as more helpful on average than those unique to the global baseline. Covariate breakdowns can help researchers interpret and refine these hypotheses.

These findings depend on meaningful, accurately measured covariates and sufficient support within strata. Future work can evaluate how sensitive the discovered hypotheses are to covariate choice and measurement quality. Our evaluation covers categorical covariates and expert judgments in two domains, limiting claims of generality. Imperfect SAE features or LLM descriptions may also affect outputs. Further limitations and selector assumptions appear in Appendix E.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology and experiment design and to assist with qualitative analysis: GPT-5.4 suggested semantic matches between generated hypotheses, which the authors reviewed and approved. The tools also helped debug and edit supporting code, locate references for statistical tools, and edit manuscript text for clarity. LLMs were also used within the experimental pipeline to interpret selected SAE features, annotate text–hypothesis pairs, and judge surface similarity. The models and procedures are described in the paper and appendix. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our real-world evaluation involved voluntary ratings by four domain experts. The study was reviewed by an institutional review board and determined exempt under 45 CFR 46.104(d)(2)(ii); recruitment and rating procedures are described in Appendix D.5. The congressional and classroom data are publicly available. We present these hypotheses as candidates for expert scrutiny, not as scientific findings.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the selection methods, synthetic constructions, baselines, and evaluation. The appendix documents topic pools and sampling (Appendix A), splits and evaluation metrics (Appendix B), implementation choices, model settings, and prompts (Appendix C), and covariate definitions, annotation instructions, and the full real-world hypothesis lists and matching decisions (Appendix D). Code will be made publicly available upon publication.

REFERENCES

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

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

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

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

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

Irene Biza, Elena Nardi, and Gareth Joel. Balancing classroom management with mathematical learning: Using practice-based task design in mathematics teacher education. *Mathematics Teacher Education and Development*, 17(2):182–198, 2015.

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

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

Wendy S Bray. 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, 2011.

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

Dorottya Demszky and Heather Hill. 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)*, pp. 528–538, 2023.

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

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

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

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

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

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

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

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

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. Are neural topic models broken? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5321–5344, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.390. URL https://aclanthology.org/2022.findings-emnlp.390/.

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

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. Literature meets data: A synergistic approach to hypothesis generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 245–281, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.12. URL https://aclanthology.org/2025.acl-long.12/.

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

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

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

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

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

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

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

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

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

Paiheng Xu, Louiqa Raschid, and Vanessa Frias-Martinez. 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, pp. 2518–2537, 2026.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27099–27116. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhong22a.html.

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

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

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

A Synthetic Dataset Details

Topics are drawn from the BILLS dataset (Adler & Wilkerson, 2018; Hoyle et al., 2022), which provides 21 high-level topics and 114 low-level subtopics. For Scenario 1 (§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 sampling at the greatest imbalance tested. The eligible topics retained for each theme are listed in Table 5. Each Scenario 1 dataset 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 3×400×0.20=240 examples from $\mathcal{T}_{soc}$ and 3×400×0.80=960 examples from $\mathcal{T}_{gov}$, with corpus B receiving the mirror allocation. Each Scenario 2 dataset 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.

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

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

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.6.

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 binary annotations on the same test texts. Hypothesis annotation is an expensive operation. Movva et al. (2025) use `GPT-4o-mini`, and we use this open-source model because we can host it locally to reduce costs, and it matches `GPT-4o-mini` in annotation performance. We internally evaluated the two models on a sample of 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 RTX A6000 GPUs with vLLM for fast inference.

This yields one annotation vector for each reference hypothesis (obtained from the dataset) and one for each generated hypothesis. We compute the correlation for 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`.

Surface-similarity judge repeatability. As a repeatability check, we randomly sampled 20 matched hypothesis pairs from the two synthetic experiments and repeated the `gpt-4.1` judgment five times per pair with the original prompt and temperature. All five judgments agreed for 17 of 20 pairs (85%). The mean within-pair sample variance was 0.0075, and the rerun averages differed from the originally reported scores by 0.055 on average. This check measures repeatability under the same judging setup.

C Method Implementation Details

C.1 Global Selection References

Each SAE-based method selects $|\mathcal{H}|$ features from the activation matrix $Z \in \mathbb{R}^{N \times M}$ of its validation-selected SAE for LLM interpretation; the hyperparameter grids and selection criteria are described in Appendix C.6. Following HypotheSAEs (Movva et al., 2025), the global **lasso** reference fits L1-regularized logistic regression:

$$(\hat{b}, \hat{\beta}) = \arg\min_{b,\beta} \ \ell_{\log}(y, b\mathbf{1} + Z\beta) + \lambda(\|\beta\|_1 + |b|), \tag{2}$$

where $\ell_{\log}$ is logistic loss. The **global-squared** reference uses squared loss with an unpenalized intercept. Both methods rank features by $|\hat{\beta}_m|$ after standardizing activations on the training set.

The global **separation score** evaluates each feature independently. For a set of texts I, let $T_m(I)$ contain the K highest-activating texts for feature m and let $O_m(I)$ contain its zero-activation texts.

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

LASSO hypothesis	Δ	demeaned-reweighted-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 prompts students to explain or compare their solution strategies to each other	+11.2	*the teacher asks students to explicitly explain or justify their reasoning or chosen strategy for solving a problem*	+19.1
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

Its score is

$$s_m(I) = \frac{1}{|T_m(I)|} \sum_{i \in T_m(I)} y_i - \frac{1}{|O_m(I)|} \sum_{i \in O_m(I)} y_i. \tag{3}$$

Global selection ranks features by $|s_m([N])|$.

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.

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

C.2 Conditional Selection Approaches

Let $\mathbf{C} \in \mathbb{R}^{N \times P}$ be the observed covariate matrix, g_i the joint covariate stratum of document i, and n_g the size of stratum g. Predictor columns are standardized using the training set before fitting. These scores guide feature selection; Eq. 1 defines prevalence differences.

Stratified-max. Stratified-max fits a squared-loss LASSO separately in strata with sufficient observations from both outcome groups:

$$(\hat{b}_g, \hat{\beta}_g) = \arg\min_{b_g, \beta_g} \frac{1}{2n_g} \|y_g - b_g \mathbf{1} - Z_g \beta_g\|_2^2 + \lambda \|\beta_g\|_1. \tag{4}$$

Using a shared penalty across strata, it ranks features by $q_m = \max_g |\hat{\beta}_{g,m}|$. Separate fits allow the association to change direction across contexts.

Interaction-lasso. Interaction-lasso augments the feature matrix with covariates and feature–covariate products. Writing $\eta(\beta, \delta, \gamma) = Z\beta + \mathbf{C}\delta + \sum_{p=1}^{P}(Z \odot C_p)\gamma_p$, it fits

$$(\hat{b}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) = \arg\min_{b,\beta,\delta,\gamma} \ell_{\log}(y, b\mathbf{1} + \eta(\beta, \delta, \gamma)) + \lambda\|(\beta, \delta, \gamma)\|_1. \tag{5}$$

The intercept is unpenalized, and covariate coefficients δ are fitted but not ranked as hypotheses. We rank features by $q_m = \max\{|\hat{\beta}_m|, \max_p |\hat{\gamma}_{p,m}|\}$, retaining either a main or covariate-specific association.

Demeaning and reweighting. Both variants subtract stratum means from activations and outcomes, yielding $\tilde{z}_{i,m} = z_{i,m} - \bar{z}_m^{(g_i)}$ and $\tilde{y}_i = y_i - \bar{y}^{(g_i)}$. They fit the squared-loss LASSO

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} w_i(\tilde{y}_i - \tilde{z}_i^\top \beta)^2 + \lambda\|\beta\|_1, \tag{6}$$

with $w_i = 1$ for **demeaned-lasso** and $w_i \propto 1/n_{g_i}$ for **demeaned-reweighted-lasso**, normalized to mean one. Both rank features by $|\hat{\beta}_m|$. For squared loss, within-stratum demeaning is equivalent to fitting unpenalized stratum intercepts (Lovell, 1963; Angrist & Pischke, 2009); inverse-size weighting then gives each stratum equal total weight.

Conditional separation score. Conditional separation score computes Eq. 3 globally and within strata with at least $K = 100$ texts, ranking each feature by the largest absolute score.

Table 8: Supplementary comparison of global regression specifications in the stratum-imbalance setting, $|\mathcal{H}|$=3. Values are means and standard deviations over 10 runs. Paired comparisons with lasso are not significant (two-sided exact sign-flip tests; all p>0.05).

Method	r=0.50		r=0.35		r=0.20	
	Surf. Sim.	F1	Surf. Sim.	F1	Surf. Sim.	F1
lasso	.557±.18	.394±.14	.383±.13	.235±.08	.393±.12	.230±.07
global-squared	.530±.13	.383±.14	.373±.16	.234±.07	.357±.12	.186±.09

C.3 PROMPTS

C.4 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 a 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 9 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 prompt budgets of $n_{\text{prompt}} \in \{25, 50, 100\}$ sampled texts per corpus using the 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_{prompt}=100 with `gpt-4.1` as a generous fixed-budget setting rather than tuning the LLM-direct baseline per condition.

Table 9: Prompt template used for the SAE interpretation step, with dataset-specific substitutions for the {goal} field.

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.

Dataset	**{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 10: Example matched ground-truth and generated hypotheses from the synthetic bills experiment, with averaged LLM surface-similarity judgment scores.

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 agricultural financial assistance	0.5
is about Technology: Broadcast	is about children's health or welfare	0.0

C.5 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

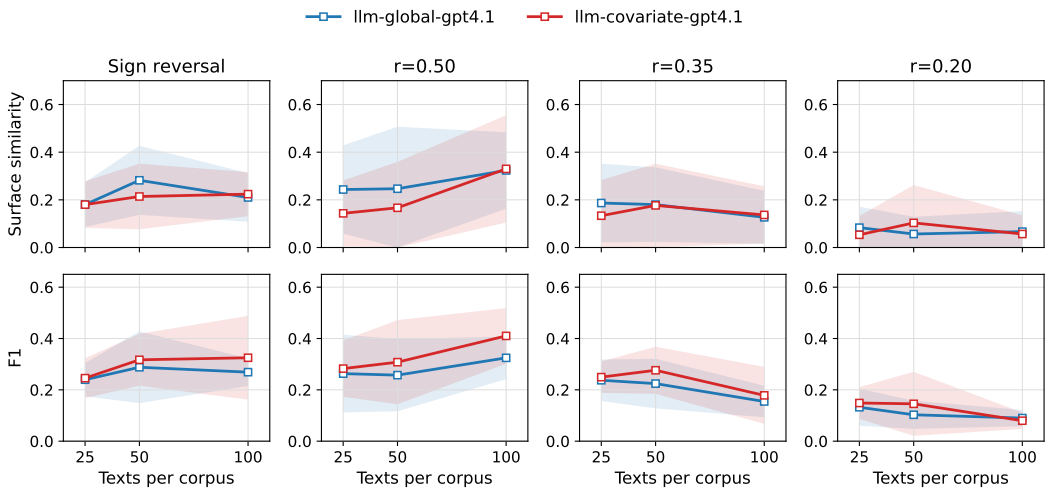

Figure 2: Prompt-budget sensitivity for LLM-direct baselines using `gpt-4.1`. Each panel varies the number of sampled texts per corpus used in the direct prompt. Lines show seed means; shaded bands show ± 1 standard deviation across seed runs. The trends are non-monotonic: larger prompt budgets and covariate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

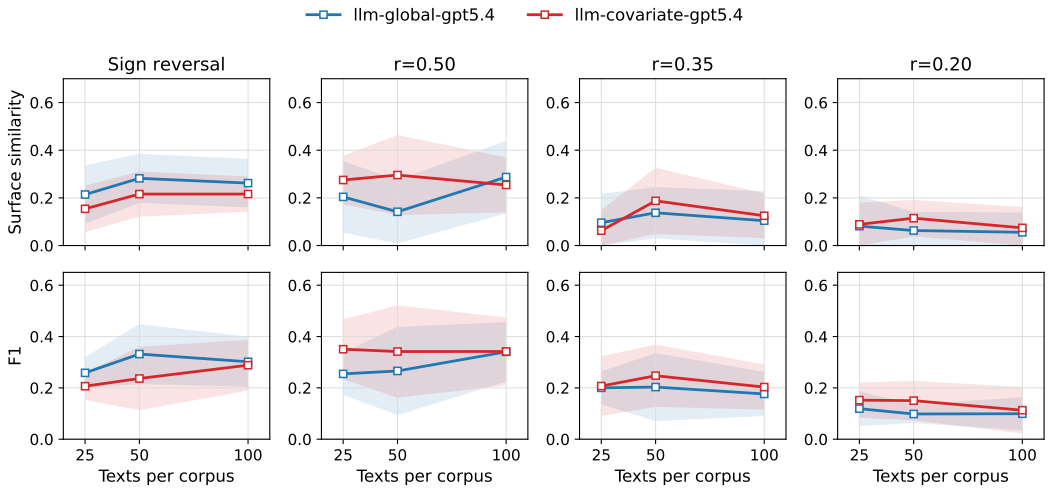

Figure 3: Prompt-budget sensitivity for LLM-direct baselines using `gpt-5.4`. Each panel varies the number of sampled texts per corpus used in the direct prompt. Lines show seed means; shaded bands show ± 1 standard deviation across seed runs. The qualitative pattern remains non-monotonic, and covariate-labeled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

combined. These texts and their corpus labels (A or B) are directly provided to the LLM for interpretation using the prompt in Table 9. **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.6 SAE HYPERPARAMETER TUNING

For each SAE-based method, we select the SAE hyperparameters from a dataset-specific grid using the validation split. Following the HypotheSAEs authors' guidance and released implementation [3], we vary the number of learned features M and the per-example sparsity K for datasets of different sizes: M controls how many concepts the model can represent, while K controls how many concepts are active for each text. For the synthetic bill corpora ($N = 2400$ in Scenario 1 and $N = 3000$ in Scenario 2), we sweep $M \in \{128, 256, 512\}$ and $K \in \{4, 8, 16\}$. For CONGRESS and NCTE, we sweep $M \in \{1024, 2048, 4096\}$ and $K \in \{8, 16, 32\}$, following the authors' guidance for the real-world experiments. 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 regression). 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 EXPERT ASSESSMENT 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.

Table 11 supplements the full lists with global and behavioral-management-stratified prevalences for all ten hypotheses generated by demeaned-reweighted-lasso on NCTE, together with the recorded expert ratings. The matched and unique groups follow Table 6.

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 having low-quality behavioral management, while segments with CLBM≥ 6 are marked as having high-quality behavioral management. We expect to see less management language in the high-quality segments. For the outcome variable, the dataset 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 training, validation, and test sets in a 60:20:20 ratio. The scores were provided by trained annotators and are included in the NCTE dataset. These covariates were 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 their ratings. For each dataset, we use manually curated one-to-one matched pairs and mark all remaining hypotheses as method-unique.

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

Table 11: Global and conditional breakdowns for all ten NCTE hypotheses generated by demeaned-reweighted-lasso. Each prevalence cell gives high versus low REMED percentages, with Δ in percentage points below. High pools the mid- and high-rated categories. Helpfulness (Help.) and conditional interpretive value (Cond.) are mean $\pm$ sample SD across two experts on a 1–5 scale. Hypotheses are shortened; full wording appears in Table 6.

Hypothesis	Overall	Behavioral-management quality		Expert ratings	
		Higher-rated	Lower-rated	Help.	Cond.
Matched hypotheses					
prompts explanation after incorrect or partially correct answers and facilitates peer discussion	70.3 vs. 48.6 (+21.7)	71.3 vs. 48.4 (+22.9)	58.3 vs. 51.0 (+7.3)	4.0 ± 0.00	5.0 ± 0.00
requires students to show and justify all intermediate steps in their math work	39.4 vs. 23.0 (+16.4)	40.6 vs. 23.0 (+17.6)	26.4 vs. 22.9 (+3.5)	3.0 ± 0.00	5.0 ± 0.00
asks students to justify mathematical reasoning, following up on errors or confusion	74.8 vs. 53.8 (+21.0)	75.6 vs. 53.5 (+22.1)	65.3 vs. 57.3 (+8.0)	4.0 ± 0.00	5.0 ± 0.00
asks students to explain or justify their reasoning or chosen problem-solving strategy	72.4 vs. 53.3 (+19.1)	73.5 vs. 53.5 (+20.0)	59.7 vs. 52.1 (+7.6)	3.0 ± 1.41	4.5 ± 0.71
elicits reasoning, solution processes, or place-value justifications during multi-digit arithmetic	50.2 vs. 34.9 (+15.3)	50.7 vs. 35.2 (+15.5)	44.4 vs. 32.3 (+12.1)	4.0 ± 0.00	3.0 ± 0.00
assigns or discusses math homework or practice outside class	57.0 vs. 46.8 (+10.2)	56.4 vs. 47.0 (+9.4)	63.9 vs. 44.8 (+19.1)	2.5 ± 0.71	4.0 ± 1.41
directs students to move between locations or groups for collaborative work	20.9 vs. 26.1 (−5.2)	21.0 vs. 26.2 (−5.2)	19.4 vs. 25.0 (−5.6)	3.0 ± 0.00	1.5 ± 0.71
Unique hypotheses					
organizes small-group work and prompts peer help and explanations	46.2 vs. 39.6 (+6.6)	46.2 vs. 39.6 (+6.6)	45.8 vs. 39.6 (+6.2)	2.5 ± 0.71	1.5 ± 0.71
provides individualized follow-up about students' understanding or needs	74.3 vs. 53.7 (+20.6)	74.8 vs. 53.3 (+21.5)	68.1 vs. 57.3 (+10.8)	3.0 ± 1.41	4.5 ± 0.71
conducts extended discussions of weight units, conversions, comparisons, or analogies	1.4 vs. 2.0 (−0.6)	1.6 vs. 1.8 (−0.2)	0.0 vs. 4.2 (−4.2)	3.5 ± 0.71	4.5 ± 0.71

Green/red shading indicates a positive/negative global difference, with alternating shades as in Table 6. Stratum labels follow the experimental covariate coding (Appendix D.2).

D.4 Rating Consistency

Given the small number of method-unique hypotheses, we report descriptive rating differences instead of an inter-annotator reliability coefficient. For each hypothesis, we compare the two experts' ratings within its domain; Table 13 summarizes these differences by matching status.

Table 13: Descriptive agreement between experts' 1–5 ratings for matched and method-unique hypotheses. Within 1 is the share of hypotheses for which the experts' ratings differ by at most one point; mean gap is the mean absolute difference.

Rating	Hypotheses	Within 1	Mean gap
Helpfulness	Matched	73.3%	1.00
Helpfulness	Unique	80.0%	0.80
Conditional value	Matched	83.3%	0.70
Conditional value	Unique	90.0%	1.10

We also compare each annotator's helpfulness ratings for the two hypotheses in each semantically matched pair. Across the two datasets, 83% of these comparisons differed by at most one Likert point; the mean absolute difference was 0.73. This similarity in ratings supports treating matched pairs as shared findings for the descriptive analysis.

D.5 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).

E 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. Our methods are designed around observed covariate strata. In this paper, we focus on binary or categorical covariates, such as policy/procedure status, behavioral-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 selection methods address different statistical regimes. Demeaned-reweighted-lasso combines within-stratum associations using shared coefficients. When the direction reverses, as in our sign-reversal experiment, opposing associations can cancel and make relevant features harder to recover. Interaction-lasso and stratified-max can accommodate such reversals, but have different limitations. Interaction-lasso expands the feature space and can be unstable when SAE activations are sparse, while stratified-max fits each stratum separately and can favor noisy estimates from small strata. In practice, researchers should inspect stratum-level statistics when possible and choose the selection model based on whether sign reversal is plausible and whether strata have sufficient support.

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 expert assessment complements these experiments with judgments in two substantively different domains, CONGRESS and NCTE, while broader generalization will require additional datasets, covariates, and expert panels. Such assessment 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.

Table 12: Prompt used for judging the surface similarity between ground-truth and generated hypotheses. Adapted from Zhong et al. (2024); Movva et al. (2025).

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: