What Moved? Interpretable Decompositions of Effects on Text with Sparse Autoencoders

Elisa Maria Wirsching*

November 3, 2026

Abstract

Social scientists increasingly estimate effects on unstructured data, such as open-ended responses, speeches or social media posts. Researchers face a tradeoff between capturing the multidimensional change in texts and describing it in interpretable terms. We propose SAE regression to combine both. It estimates the shift in the average document embedding, learns concepts with a sparse autoencoder (SAE), projects the shift onto the SAE's directions, and labels them with a large language model. It reports the movement along each direction and its share of the effect. In simulations, the estimates are unbiased, well calibrated and recover true concepts. Applied to a survey experiment on immigration and to congressional tweets after mass shootings, SAE regression captures the reasoning and rhetoric behind effects on text and outperforms existing methods, including the structural topic model. SAE regression thus makes multidimensional effects on text both measurable and interpretable.

Word count: 9,860

*Assistant Professor, London School of Economics and Political Science.
Correspondence: e.m.wirsching@lse.ac.uk
I thank Tak-Huen Chau, Carl Müller-Crepon, Florian Sichart, Arthur Spirling, and audience members of the LSE Methodology Seminar for helpful comments and suggestions. I used Anthropic's Claude Opus 5.5 to help write and check code (including the `saereg` package), find and verify references, and draft and edit text. I directed and reviewed all of this work, and all ideas and interpretations are my own.

1 Introduction

Social scientists have built an impressive toolkit for analyzing unstructured data. In text-as-data applications, one of the most common strategies is to classify texts into categories fixed in advance—e.g., whether a household's open-ended account of rising inflation blames an energy crisis (Andre et al., 2026) or whether a campaign ad is negative (Fowler et al., 2021)—and then to use these codes as variables in downstream analyses. Because large language models have made such classification cheap, a growing literature advocates and uses them to automate it (Gilardi et al., 2023; Linegar et al., 2023; Ziems et al., 2024). Classification is attractive because its output is easy to interpret and to use, but this interpretability comes at a cost. Reducing each document to a few binary codes discards much of the richness of text, and prediction errors bias downstream estimates even when the annotations are highly accurate (Egami et al., 2026). Classification also requires researchers to fix the categories in advance, so changes in the text they did not anticipate remain unmeasured, and it is unclear how much of the overall change their categories capture.

These limitations matter because scholars often use text *inductively*, for example, to learn how a treatment or group shapes what people think or say and the reasoning behind it (Grimmer et al., 2022; Egami et al., 2022). Suppose a survey experiment finds that respondents are more willing to support a welfare recipient when a vignette says the recipient lost their job than when it says they stopped looking for work. To learn why support changed, the researcher asks respondents to explain their answer. Those who read about the job loss may, for example, attribute the situation to the economy, judge the recipient as more deserving, or write about their own experiences of unemployment. Ideally, the researcher would discover the various reasons and quantify their relative importance. In short, we often want to capture a multidimensional change, but still describe it in interpretable terms.

Existing tools force a tradeoff between these two goals. Classification is interpretable but captures only the dimensions the researcher fixed in advance. Embedding regression captures

the full change in meaning by estimating how far the treatment shifts the average document embedding (Rodriguez et al., 2023). Yet, the estimated change is a vector of thousands of dimensions, and the nearest-neighbor words or contexts used to interpret it are often open to several readings. Topic models, such as structural topic models (Roberts et al., 2014), sit in between. They learn interpretable topics inductively, but by representing texts as clusters of words, they can merge different arguments that share words and split one argument made in different words. Moreover, none of these approaches says how much of the overall change in the texts the identified dimensions account for.

We propose SAE regression, which captures the multidimensional change in texts and decomposes it in interpretable dimensions. As Figure 1 illustrates, it first estimates how far the treatment moves the average text in a high-dimensional embedding space (step 1), building on embedding regression (Rodriguez et al., 2023; Wirsching et al., 2025). It then learns the key directions in the corpus with a sparse autoencoder (SAE, step 2). An SAE is a sparse neural network that rebuilds each text's embedding from a handful of features, so that each feature—a direction in the embedding space—tends to capture one recurring concept, such as an argument or a theme (Cunningham et al., 2024; Bricken et al., 2023). Finally, the method projects the effect onto these directions (step 3) and labels them with an LLM or a human reader (step 4). For each direction, the method estimates how far the treatment moves the texts along it and what share of the overall effect it explains.

This paper makes several contributions. First, we offer an estimator that decomposes any difference between groups of texts into interpretable directions. Second, we provide inference for the whole effect and its parts, with a global test of whether the texts changed at all and intervals for each direction. Third, we draw on simulation exercises against a ground truth and a comparison of LLM and human labels to validate the method. We show that the estimates are unbiased, well calibrated, and sufficiently powered to detect small effects, and mechanisms that differ in content are attributed to separate directions. From these results, we derive guidance on sample size, the size of the SAE and the reading of labels. Fourth, two

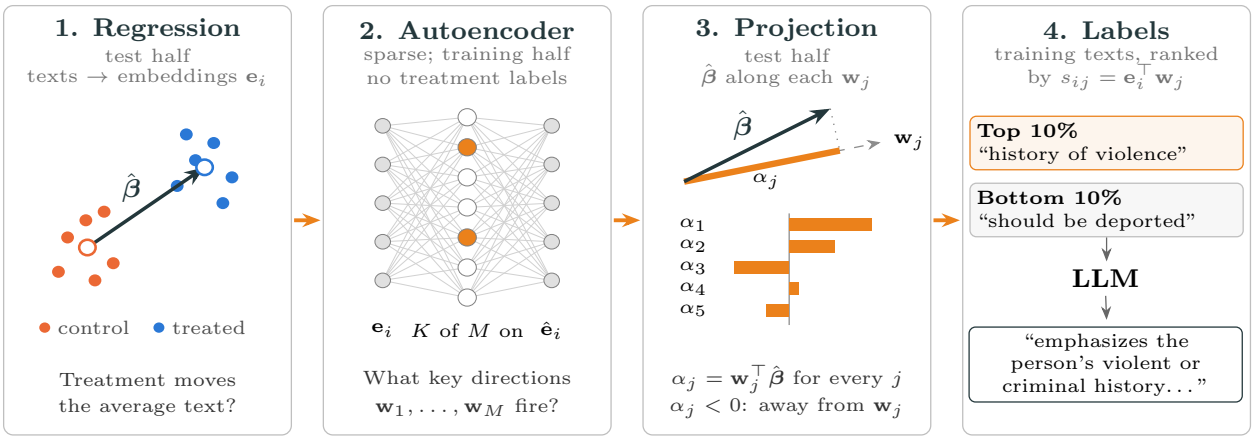


Figure 1: Overview of SAE regression. The texts are split into a training and a test half. (1) In the test half, the embedding regression estimates the effect $\hat{\boldsymbol{\beta}}$ on the average text. (2) A sparse autoencoder learns directions $\mathbf{w}_1, \ldots, \mathbf{w}_M$ from the training half, without treatment labels. (3) The effect is projected onto each direction. (4) An LLM labels each direction from training texts ranked by their score on it; panel 4 shows feature F2 of the application in Section 4.

applications illustrate what the decomposition adds. In a survey experiment on immigration and punishment (Egami et al., 2022), we find that SAE regression captures the underlying reasoning respondents give, while topic models primarily name respondents' positions. In a difference-in-differences design, we decompose how members of Congress turn to tragedy talk on social media after a mass shooting in their district or state, primarily through condolences and prayers for victims and, to a lesser extent, calls for legislative action. Finally, we provide an R package (`saereg`) to implement the method.

2 Related Literature

SAE regression builds on work using learned measures for causal inference with text as outcome (Egami et al., 2022; Feder et al., 2022). In this work, a learned function g maps each text to a few dimensions, such as topic proportions, and the effects on these dimensions are the result. SAE regression also uses such a codebook, where each SAE direction defines one dimension of g as a linear function of the embedding, but it adds a common reference: the effect on the whole embedding, estimated by applying embedding regression (Rodriguez et al., 2023; Wirsching et al., 2025) to whole texts instead of the contexts of a focal word.

Because g is linear, the effect along each direction is a projection of this overall effect, so we can estimate how much of it the directions jointly explain. Topic models—whether built on word counts (Roberts et al., 2014) or on embeddings (Grootendorst, 2022)—offer no such total effect.

A second literature uses machine learning to generate hypotheses from unstructured data (Ludwig and Mullainathan, 2024). Several methods learn concepts from texts with an SAE or an LLM, keep those that best predict treatment or outcome groups, and often test them on held-out texts (Movva et al., 2025; Modarressi et al., 2025; Xu et al., 2026; Wang et al., 2026). Because the concepts are selected for their predictive power in separating the groups, the estimand benefits salient, obvious differences. As a result, it can favor rare, clean features over those that account for most of the difference. For example, a phrase that 5% of treated respondents but no control respondent uses separates the groups almost perfectly, but it moves the average text very little. Finally, most of these methods use stochastic LLMs or human coding for estimation beyond training, which makes the estimates hard to reproduce and can add systematic measurement error (Egami et al., 2026). SAE regression learns its directions without the treatment and estimates how much of the effect each one explains. Additionally, the labels only name the directions without entering the estimates. The closest work in this literature is Carlson (2026), who passes texts through a language model with a pretrained SAE and treats thousands of its concepts as binary outcomes, testing the effect on each with a generalized familywise error rate. SAE regression instead trains a small SAE on the study's own texts, so that its features describe the corpus, decomposes one overall effect linearly, and reports how complete that decomposition is.

Third, SAE regression borrows tools from research on the interpretability of machine learning models (Doshi-Velez and Kim, 2017). SAEs are often deployed to decompose the internal representations of language models into sparse, monosemantic features that are easier to interpret (Makhzani and Frey, 2014; Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2024), and LLMs are used to describe these features at scale

(Bills et al., 2023; Paulo et al., 2024). More recent work trains SAEs on text embeddings to analyze corpora (O'Neill et al., 2024; Jiang et al., 2025). Our strategy to project an effect onto the SAE's directions relates to the idea that concepts can be represented as linear directions in the representation spaces of language models (Park et al., 2024). We contribute to this literature by moving these tools from the interpretation of a model to the interpretation of an effect, which requires an estimand, statistical inference and validation.

3 Combining Embedding Regression with Sparse Autoencoders

3.1 The Framework

To state what the method estimates, we use the potential outcomes framework with text as an outcome. Consider units $i = 1, \ldots, n$ and a binary treatment T_i. Following Egami et al. (2022), we assume that unit i would write the text $Y_i(1)$ under treatment and $Y_i(0)$ under control, and that a unit's text depends only on its own treatment, i.e., the observed text is $Y_i = Y_i(T_i)$. Since a text cannot be subtracted or averaged as it is, we define an effect on text through a codebook g, a fixed function that maps a text to a number or a vector, as

$$\tau_g = E\big[g(Y_i(1)) - g(Y_i(0))\big]. \tag{1}$$

Every quantity in this paper is such an effect for a particular g: the embedding of the text (Section 3.2), and its score along an SAE direction or its coordinates in the basis of SAE directions (Section 3.4).

While τ_g is a causal quantity, SAE regression applies to *any* contrast that a linear regression of the embeddings defines—descriptive or causal—and the projections take on the interpretation of that regression as linear functions of the embeddings. For example, a comparison of two parties' speeches in parliament describes along which dimensions and by how much the parties' speeches differ. The contrast becomes a causal estimand under the usual

identifying assumptions of the design. In a randomized experiment with text as the outcome, for instance, random assignment secures ignorability and positivity, leaving SUTVA as the key assumption to make (Egami et al., 2022). In a difference-in-differences design, the assumption is parallel trends in the embeddings, which then hold along every direction the SAE learns. For readability, we describe the method in terms of a binary treatment and use "treated" for texts with $T_i = 1$ throughout.

While the embedding model is fixed ex ante, the SAE and its labels are learned from the study's own texts. If the same texts were then used to estimate effects, the directions would partly reflect the very differences they are meant to measure (Egami et al., 2022). We therefore split the documents at random, learn the SAE and the labels on one half without access to the treatment, and estimate the overall effect and its projections on the other.[1] In this section, we walk through the key steps in the method to introduce the various estimands (see Table 1 for a summary).

3.2 Embedding Regression

A pretrained embedding model $e(\cdot)$ maps any text to a vector of unit length in $\mathbb{R}^D$, and $\mathbf{e}_i = e(Y_i)$ is the embedding of unit i's text. Following Rodriguez et al. (2023), we regress these embeddings on the variable of interest,

$$\mathbf{e}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}\, T_i + \boldsymbol{\varepsilon}_i, \tag{2}$$

which amounts to one linear regression per embedding dimension. Covariates and fixed effects can enter the design in the usual way. The coefficient $\boldsymbol{\beta}$ is the first estimand of interest—the effect on the average text in the embedding space. With a randomized binary treatment it is the average effect on the codebook $g(Y) = e(Y)$, $\boldsymbol{\beta} = E\big[e(Y_i(1)) - e(Y_i(0))\big]$, and its estimate is the difference between the mean embeddings of the two arms. The effect

[1] We write $\mathcal{S}_{\text{train}}$ and $\mathcal{S}_{\text{test}}$ for the units in the two halves.

records every change in the texts that the embedding registers, while its squared length $\|\boldsymbol{\beta}\|^2$ is a scalar that measures how far the treatment moves the texts overall. Because each of the D coordinates of $\hat{\boldsymbol{\beta}}$ is estimated with error, its squared length overstates $\|\boldsymbol{\beta}\|^2$ by the summed sampling variance, which we subtract following Green et al. (2024),

$$\|\hat{\boldsymbol{\beta}}\|^2_{\text{cor}} = \sum_{\ell=1}^{D} \left(\hat{\beta}_\ell^2 - \widehat{V}\left[\hat{\beta}_\ell\right]\right). \tag{3}$$

For inference, we test whether the treatment moved the texts at all, $H_0\colon \boldsymbol{\beta} = \mathbf{0}$, using $\|\hat{\boldsymbol{\beta}}\|^2$ as the test statistic. Under random assignment, permuting the treatment labels gives an exact test of Fisher's sharp null that the treatment changes no unit's embedding; with covariates, one can permute the residuals of the model without the treatment (Freedman and Lane, 1983), and without random assignment (e.g., in a DiD setting), we use a cluster bootstrap. Standard errors and confidence intervals for $\|\hat{\boldsymbol{\beta}}\|^2_{\text{cor}}$, and for the projections below, come from a leave-one-out jackknife over documents or over clusters when units are clustered. These intervals are conservative when the effect is small, but they come closer to nominal coverage than bootstrap intervals, which fail for squared means (Dodd and Korn, 2007; Green et al., 2024).

3.3 Interpretable Features using Sparse Autoencoders

The coefficient $\hat{\boldsymbol{\beta}}$ from Equation (2) is a high-dimensional estimate that is hard to interpret on its own. Existing applications read it through nearest neighbors, the words or contexts closest to the fitted embedding in each group, such as $\hat{\boldsymbol{\beta}}_0$ for $T_i = 0$ and $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}$ for $T_i = 1$ (Rodriguez et al., 2023; Wirsching et al., 2025). Comparing these lists describes how meaning differs across groups, but it remains ad hoc with substantial room for interpretation and does not decompose $\hat{\boldsymbol{\beta}}$ into separate concepts or quantify their importance. Here, we use sparse autoencoders (SAEs) instead. SAEs have been found to discover unknown concepts and provide interpretable features in high-dimensional data (Gao et al., 2024; Peng et al., 2025;

Movva et al., 2025; Bouyamourn and Spirling, 2026), which fits our inductive approach to uncovering reasoning in text.

An SAE consists of an encoder, which maps each embedding to a small set of active features, and a decoder, which rebuilds the embedding from them (Makhzani and Frey, 2014; Gao et al., 2024). The encoder maps the embedding $\mathbf{e}_i$ to a hidden layer of M features shared across the corpus and keeps only the K largest,

$$\mathbf{z}_i = \mathrm{ReLU}\left(\mathrm{TopK}\left(W_{\mathrm{enc}}(\mathbf{e}_i - \mathbf{b}_{\mathrm{pre}}) + \mathbf{b}_{\mathrm{enc}}\right)\right) \in \mathbb{R}^M_{\geq 0}, \tag{4}$$

where TopK sets all but the K largest entries to zero. That is, each text switches on at most K of the M features, and $z_{ij} \geq 0$ measures how strongly feature j is present in text i. The decoder then maps the active features in the hidden layer back to the embedding space,

$$\hat{\mathbf{e}}_i = \mathbf{b}_{\mathrm{pre}} + W\mathbf{z}_i = \mathbf{b}_{\mathrm{pre}} + \sum_{j=1}^{M} z_{ij}\,\mathbf{w}_j, \qquad \|\mathbf{w}_j\| = 1, \tag{5}$$

so that each reconstructed text is a sum of at most K decoder directions—the columns of W—weighted by how strongly the text activates them. The encoder, the decoder and the bias $\mathbf{b}_{\mathrm{pre}}$ are learned together by minimizing the reconstruction error relative to the variance of the embeddings (i.e., the fraction of variance unexplained),

$$\mathcal{L} = \frac{\sum_{i \in \mathcal{S}_{\mathrm{train}}} \left\|\mathbf{e}_i - \hat{\mathbf{e}}_i\right\|^2}{\sum_{i \in \mathcal{S}_{\mathrm{train}}} \left\|\mathbf{e}_i - \bar{\mathbf{e}}\right\|^2}, \tag{6}$$

with a small auxiliary term that revives features that stop firing (Gao et al., 2024; O'Neill et al., 2024). We train the SAE on the training half only, using the implementation of Movva et al. (2025). Unlike SAEs trained on the internal activations of a language model (Carlson, 2026), which have thousands of features—many of them lexical and generic—ours is trained on the study's own text embeddings. Hence, M is small, the features fit the corpus, and each feature describes texts, such as an argument or a theme.

The SAE is related to principal component analysis (PCA) but better suited to our purpose. With a linear encoder and decoder and no sparsity constraint, minimizing the reconstruction error indeed recovers the subspace of the first M principal components (Baldi and Hornik, 1989; Peng et al., 2025). However, PCA captures the largest variation across all texts and, as a result, mixes several concepts into one dimension. SAE, in contrast, gives a concept its own direction even when it occurs only in a subset of texts and explains little of the overall variance, because each text uses only K of the M directions. SAE features therefore capture directly what "fires" in a text, are more interpretable and narrow than principal components (Cunningham et al., 2024), and each can be described by a single label (as described in Section 3.5).

3.4 Projection

Conveniently, the SAE's decoder directions—the columns of W—live in the same space as the embeddings and the effect $\boldsymbol{\beta}$. We can therefore project the overall effect onto them and describe it in terms of the features, asking along which directions the average text moves, to what extent, and how much of the overall effect these movements account for.

3.4.1 Simple Projection

We start from a simple projection, taking one direction $\mathbf{w}_j$ at a time. We can think of the effect $\boldsymbol{\beta}$ as an arrow in the embedding space and of a decoder direction $\mathbf{w}_j$ as a second arrow of unit length. The projection $\alpha_j = \mathbf{w}_j^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\| \cos\theta_j$, where θ_j is the angle between the two arrows, is the signed length of the shadow that the effect casts on the direction.

Alternatively, we can write this projection as a regression of the score of text i along the direction, $s_{ij} = \mathbf{e}_i^\top \mathbf{w}_j$, on the variable of interest as in Equation (2),[2]

$$s_{ij} = \alpha_{0j} + \alpha_j T_i + u_{ij} \quad \Rightarrow \quad \hat{\alpha}_j = \mathbf{w}_j^\top \hat{\boldsymbol{\beta}}. \tag{7}$$

[2]See Lemma A.1 for the proof.

The coefficient α_j is the effect along $\mathbf{w}_j$, that is, how far the average text moves along the direction, in the same units for every feature, since every direction has unit length. A positive α_j means that treated texts move towards feature j, and a negative α_j that they move away from it. This regression setting again allows for additional controls or fixed effects, and because α_j is the effect on a fixed codebook, $g(Y) = e(Y)^\top \mathbf{w}_j$, it has the same interpretation as $\boldsymbol{\beta}$ under the design's assumptions.

How much of the overall effect $\boldsymbol{\beta}$ does each feature explain? The share of the effect that lies along $\mathbf{w}_j$ can be expressed as

$$r_j = \frac{\alpha_j^2}{\|\boldsymbol{\beta}\|^2} = \cos^2 \theta_j, \tag{8}$$

which equals the R^2 of regressing $\boldsymbol{\beta}$ on $\mathbf{w}_j$ without an intercept and lies between 0 and 1.[3] Yet the decoder directions are not orthogonal, and two directions that point in similar ways pick up similar movement. As a result, the r_j of different features overlap and can sum to more than 1. The multivariate projection removes this overlap.

3.4.2 Multivariate Projection

Rather than casting the effect onto each direction separately, we can project it onto a combination of directions, essentially regressing $\boldsymbol{\beta}$ on all columns of W at once, $\boldsymbol{\beta} = W\boldsymbol{\delta} + \boldsymbol{\nu}$, where $\boldsymbol{\nu}$ is the part of the effect that no combination of the directions describes. The joint coefficient δ_j is the effect along $\mathbf{w}_j$ with the other directions held fixed. Per document, this is a regression of each text's coordinates in the basis of decoder directions on the variable of interest,

$$\tilde{s}_{ij} = \delta_{0j} + \delta_j T_i + v_{ij}, \qquad \tilde{\mathbf{s}}_i = G^{-1}\mathbf{s}_i \quad \Rightarrow \quad \hat{\boldsymbol{\delta}} = G^{-1}\hat{\boldsymbol{\alpha}}, \tag{9}$$

[3] As for the norm, $\hat{\alpha}_j^2$ overstates α_j^2 by the sampling variance of $\hat{\alpha}_j$, so we estimate r_j by $\big(\hat{\alpha}_j^2 - \widehat{V}[\hat{\alpha}_j]\big)/\|\hat{\boldsymbol{\beta}}\|^2_{\text{cor}}$.

where $\mathbf{s}_i = W^\top \mathbf{e}_i$ stacks the scores s_{ij} of text i from the simple projection, and $G = W^\top W$ is the Gram matrix of the directions, whose off-diagonal entries g_{jk} are the cosines between them. The coordinates $\tilde{\mathbf{s}}_i$ are the scores with the overlap between the directions removed. As for α_j, $\boldsymbol{\delta}$ is an effect on a fixed codebook.

The two projections differ only through the overlap of the directions and coincide in the case of orthogonal directions ($G = I$). The simple projection α_j measures how far the texts move towards $\mathbf{w}_j$ in a bivariate regression sense, while δ_j says how much of the effect runs along $\mathbf{w}_j$ once the other features are held fixed. Thus, unlike α_j, δ_j depends on which other directions the SAE learns. Proposition 1 states what both projections inherit from the embedding regression; the proofs of the propositions are in Appendix A.

Proposition 1 (Identification along the directions)**.** Let the decoder W be learned on the training half, so that it is fixed given $\mathcal{S}_{\text{train}}$, let its columns be linearly independent, and let $\mathbf{b} = E[\hat{\boldsymbol{\beta}} \mid \mathcal{S}_{\text{train}}] - \boldsymbol{\beta}$ be the bias of the embedding regression in a given design.

(a) Unbiasedness. The projections inherit this bias along their directions,

$$E\big[\hat{\alpha}_j \mid \mathcal{S}_{\text{train}}\big] - \alpha_j = \mathbf{w}_j^\top \mathbf{b}, \qquad E\big[\hat{\boldsymbol{\delta}} \mid \mathcal{S}_{\text{train}}\big] - \boldsymbol{\delta} = G^{-1} W^\top \mathbf{b}.$$

If the design's assumptions make $\hat{\boldsymbol{\beta}}$ unbiased, $\hat{\alpha}_j$ and $\hat{\boldsymbol{\delta}}$ are therefore unbiased for the effects on the codebooks $e(Y)^\top \mathbf{w}_j$ and $G^{-1} W^\top e(Y)$. More generally, $\hat{\alpha}_j$ is unbiased whenever $\mathbf{b}$ is orthogonal to $\mathbf{w}_j$, and $\hat{\boldsymbol{\delta}}$ whenever $\mathbf{b}$ is orthogonal to every column of W.

(b) Consistency and inference. If $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ as the test half grows, so are $\hat{\alpha}_j$ and $\hat{\boldsymbol{\delta}}$. If $\hat{\boldsymbol{\beta}}$ is asymptotically normal, so are they, and the jackknife of Section 3.2, applied to the same deletions, gives their standard errors.

We flag a feature when the global test of Section 3.2 rejects and the feature's 95% interval excludes zero.[4]

[4]With many features, the intervals can be adjusted for multiple testing, for instance with a max-T critical value taken from the resampling distribution of the largest studentized $|\hat{\alpha}_j|$ (Westfall and Young, 1993).

By accounting for the cosines between the columns of W, the multivariate projection also splits the effect into shares that add up to 1. The joint share

$$R^2 = \frac{\|W\boldsymbol{\delta}\|^2}{\|\boldsymbol{\beta}\|^2} = \frac{\boldsymbol{\alpha}^\top G^{-1}\boldsymbol{\alpha}}{\|\boldsymbol{\beta}\|^2} \tag{10}$$

captures the share of the effect that the directions describe together, and lies between 0 and 1. It divides among the features as

$$q_j = \frac{\delta_j\,\alpha_j}{\|\boldsymbol{\beta}\|^2} = \frac{\delta_j^2 + \sum_{k\neq j} g_{jk}\,\delta_j\delta_k}{\|\boldsymbol{\beta}\|^2}, \qquad \sum_{j=1}^{M} q_j = R^2, \tag{11}$$

because $\boldsymbol{\delta}^\top\boldsymbol{\alpha} = \boldsymbol{\alpha}^\top G^{-1}\boldsymbol{\alpha}$. Each q_j captures the feature's own part, $\delta_j^2/\|\boldsymbol{\beta}\|^2$, as well as half of every overlap term $2g_{jk}\delta_j\delta_k/\|\boldsymbol{\beta}\|^2$ it shares with another feature, so that every part of R^2 is assigned exactly once. It equals r_j when the directions are orthogonal, and it can be negative when the overlap with other features reverses the sign of α_j relative to δ_j.[5] Proposition 2 collects these properties.

Proposition 2 (Decomposition of the effect)**.** Let the columns of W be linearly independent and $\boldsymbol{\beta} \neq \mathbf{0}$. Then (a) $R^2 = \|W\boldsymbol{\delta}\|^2/\|\boldsymbol{\beta}\|^2$ lies in $[0,1]$, and $(1-R^2)\|\boldsymbol{\beta}\|^2$ is the squared length of the part of $\boldsymbol{\beta}$ orthogonal to all directions; (b) $\sum_{j=1}^{M} q_j = R^2$, and $q_j = r_j$ for every j when the directions are orthogonal; (c) if $\boldsymbol{\beta} \neq \mathbf{0}$ and $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, the estimated shares r_j, q_j and R^2 are consistent.

Earlier applications of SAEs in social science mainly work with the sparse activations z_{ij} rather than the decoder directions (Movva et al., 2025; Bouyamourn and Spirling, 2026; Carl-

[5] We correct R^2 and q_j for sampling noise as we do r_j, subtracting $\operatorname{tr}\left(G^{-1}\widehat{V}[\hat{\boldsymbol{\alpha}}]\right)$ and $\left(G^{-1}\widehat{V}[\hat{\boldsymbol{\alpha}}]\right)_{jj}$ from the numerators and dividing by $\|\hat{\boldsymbol{\beta}}\|^2_{\mathrm{cor}}$. Shapley values give an alternative split of R^2 that is never negative but averages over all 2^M subsets of features.

son, 2026).[6] The natural counterpart for our estimates is the effect on the mean activation, γ_j, from the same regression with the activations z_{ij} of Equation (4) as the outcome,

$$z_{ij} = \gamma_{0j} + \gamma_j T_i + \eta_{ij}. \tag{12}$$

We nonetheless treat the projections as the core estimands, because the encoder's nonlinearity can make γ_j misstate how the texts move along $\mathbf{w}_j$. First, an activation cannot fall below zero, so γ_j misses movement away from $\mathbf{w}_j$ among texts that do not use feature j. Binarizing the activations, as in Movva et al. (2025) or Carlson (2026), also discards how strongly a text expresses the feature and adds classification error. Second, under TopK, a feature's activation can fall because another feature took its slot, even when the texts do not move along its direction. Third, the activations are nonlinear in the embeddings, so they do not decompose the effect into the shares r_j and q_j. In practice, when the SAE reconstructs $\hat{\boldsymbol{\beta}}$ well, $\hat{\delta}_j$ and $\hat{\gamma}_j$ are close for every feature (Appendix Figure A4), and researchers can use the activations while bearing these limits in mind.[7]

3.5 Labeling

While the projections help decompose β into key features, they do not make the directions interpretable for humans yet. Following related work (Bricken et al., 2023; Paulo et al., 2024), we label each direction $\mathbf{w}_j$ from the training texts ranked by their score along it, $s_{ij} = \mathbf{e}_i^\top \mathbf{w}_j$, i.e., the same score the projections are built on. The labeler sees seven randomly drawn texts from each of three bins of the score, the bottom 10%, the middle 45–55% and the top 10%,

[6] For example, Movva et al. (2025) select the features whose activations predict a variable attached to each text, such as a speech's party. The LLM label of each selected feature becomes a hypothesis, for instance that Republican speeches discuss illegal immigration, which they test by having an LLM code whether each held-out text expresses the concept and regressing the variable on these binary codes. Bouyamourn and Spirling (2026) regress quantities from a model of LLM annotations, such as each document's predicted probability, on the documents' pooled activations.

[7] Since the decoder writes every reconstructed text as a sum of the directions (Equation (5)), $\hat{\boldsymbol{\beta}} = W\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\rho}}$, where $\hat{\boldsymbol{\rho}}$ is the effect on the part of the embeddings that the SAE does not reconstruct. Hence $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\gamma}} + G^{-1}W^\top \hat{\boldsymbol{\rho}}$, and the two differ by the movement that the encoder drops (see Appendix A).

Quantity	Eq.	Question it answers	How to read it
$\boldsymbol{\beta}$	(2)	How does the variable of interest change the average text?	A vector in the embedding space, not read directly.
$\|\boldsymbol{\beta}\|^2$	(3)	Do the texts move at all, and how far?	Tested against H_0 by the global test.
$\alpha_j = \mathbf{w}_j^\top \boldsymbol{\beta}$	(7)	How far do the texts move along feature j's direction?	Positive towards $\mathbf{w}_j$, negative away from it. Includes movement along overlapping directions.
$r_j = \alpha_j^2 / \|\boldsymbol{\beta}\|^2$	(8)	What share of the effect lies along $\mathbf{w}_j$?	Between 0 and 1. Overlapping features share the same movement, so the r_j do not add up.
$\boldsymbol{\delta} = G^{-1} W^\top \boldsymbol{\beta}$	(9)	How far do the texts move along $\mathbf{w}_j$ with the other directions held fixed?	Signs as for α_j. The feature's own movement; depends on which other directions the SAE learns.
$R^2 = \boldsymbol{\alpha}^\top G^{-1} \boldsymbol{\alpha} / \|\boldsymbol{\beta}\|^2$	(10)	How much of the effect do the features describe together?	Between 0 and 1.
$q_j = \delta_j \alpha_j / \|\boldsymbol{\beta}\|^2$	(11)	How much of R^2 goes to feature j?	Adds up to R^2 over all features; can be negative.

Table 1: The estimands. The squared quantities are corrected for sampling noise (Sections 3.2 and 3.4). A feature is flagged when the global test rejects and its 95% interval excludes zero.

ordered from the lowest, and writes one short label for what separates texts in the top bin from the texts in the lower bins (see Figure 2 for an example). Because activations cannot be negative, low scores come from texts that use features pointing away from $\mathbf{w}_j$. A label therefore names the top end of a direction, and a negative α_j or δ_j means that the texts move away from what the label describes (not necessarily in the "opposite" direction).

We use an LLM as the labeler, GPT-4.1 at temperature zero (the prompt is in Appendix B). It's important to note that none of its output enters the estimands. The projections, their intervals and the shares depend only on the embeddings and the decoder directions, while the label only assigns a name to a direction. Hence, the labeling could equally be done by a human reading the same texts. Section 5.6 checks how well the LLM labels pick out the top-scoring texts and compares them with labels written by humans.

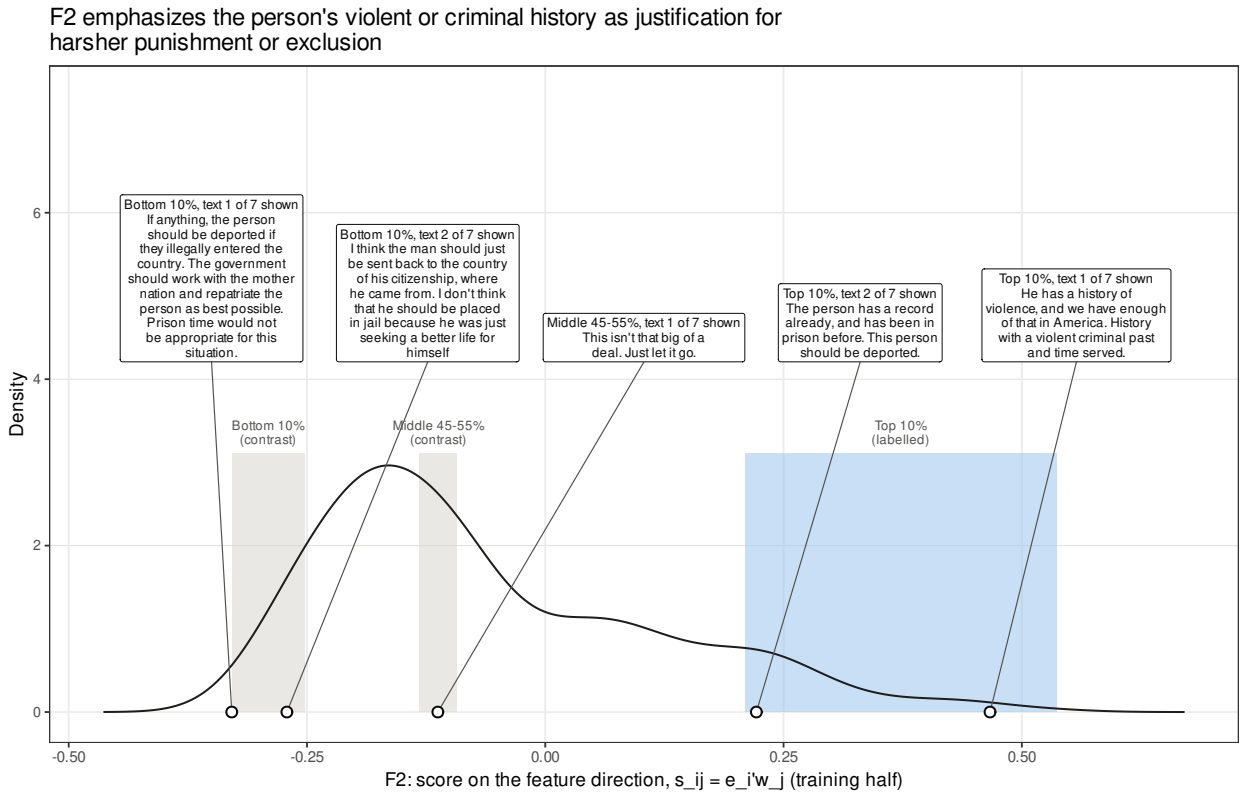


Figure 2: What the labeler sees, for feature F2 of the Egami et al. application. The curve is the distribution of the training texts' scores s_{ij}; shaded are the three bins the labeler samples from (bottom 10%, middle 45–55%, top 10%). The boxes are texts from the prompt itself, and the title is the resulting label.

4 Applications

4.1 Experimental Data

Our first application compares SAE regression against the workhorse model to analyze effects on text: the structural topic model (STM). We revisit Experiment 3 of Egami et al. (2022), a survey experiment on immigration and punishment in which 1,094 respondents on Amazon Mechanical Turk read about a 28-year-old man convicted of entering the United States illegally. In the treatment condition, respondents learned that he had previously served two prison sentences—one of them for a violent crime—and had been deported. In the control condition, he had never been imprisoned before. Respondents then answered whether the offender should be sent to prison and explained what the US government should do and

why. We use these open-ended answers as our texts. Egami et al. (2022) fit an STM with 11 topics on a random half of the answers and estimate the treatment's effect on the topic proportions in the other half. The authors find that respondents who read about a criminal history write more about maximal punishment and deportation, while those in the control condition write more about letting the person stay or about the circumstances of the crime (left side of Figure 3).

We reproduce the authors' data split exactly, with 521 in the training half and 513 in the test half, and a median length of 32 words.[8] We embed the answers with OpenAI's text-embedding-3-large ($D = 3{,}072$), train the SAE on the training half with $M = 7$ features and $K = 2$ active features per answer, and label each feature from the training answers (Section 3.5). We set $K = 2$ because the answers are short, and the decomposition is similar with $M = 11$ (Appendix Figure A1). As the benchmark, we apply the authors' own STM.

The STM and the SAE decompositions differ in various ways. First, as a bag-of-words approach, the STM models which words an answer uses, so one argument made in different words can spread over several topics, and different arguments that share words—such as prison, deportation and punishment—can merge into one. The SAE instead depends on embeddings, which represent the meaning of the entire answer. Second, topic proportions sum to one in every answer by design, and a rise in one topic forces falls elsewhere. In contrast, the SAE's features are not exclusive, and their effects are not tied together.[9] Finally, while the STM's effects are changes in topic proportions that cannot decompose a total effect, the SAE decomposes the single effect $\hat{\boldsymbol{\beta}}$ and q_j and R^2 state how much of it each feature and all features together describe.

Empirically, both representations detect the effect, with permutation $p < 0.001$ for the

[8] The treatment raises the share of empty answers by 4.5 percentage points (95% CI 1.8 to 7.1), and 54 of the 60 respondents without an answer also skipped the closed-ended question. As in Egami et al. (2022), the estimates therefore concern respondents who wrote an answer.

[9] The SAE's sparsity constraint limits how many features an answer activates, but this only affects which directions the SAE learns, not the effects along their directions.

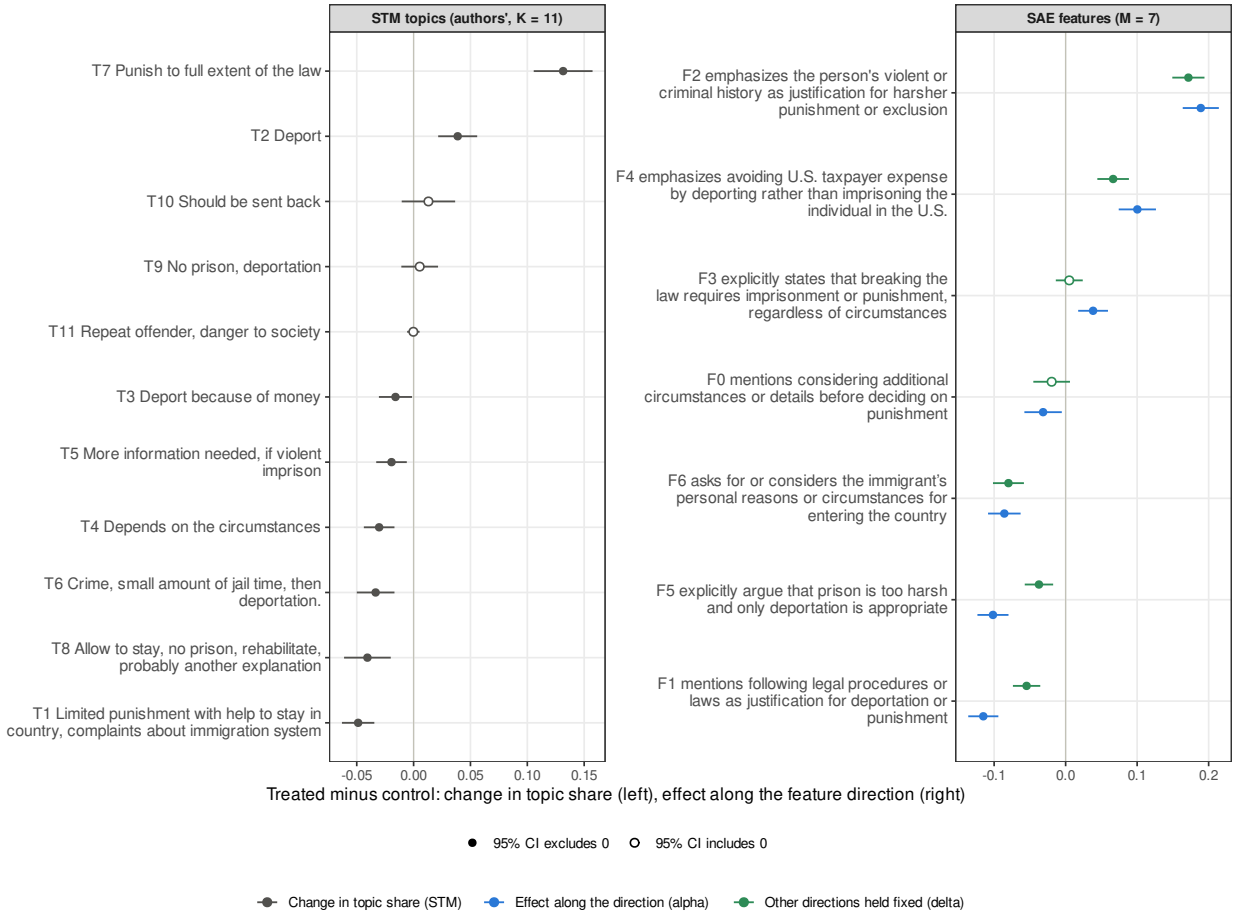


Figure 3: Treatment effects on the answers in Egami et al. Experiment 3 (test half). Left: change in the share of each STM topic (the authors' model with 11 topics). Right: for each SAE direction ($M = 7$), the effect along the direction $\hat{\alpha}_j$ and the joint coefficient $\hat{\delta}_j$ (other directions held fixed); 95% jackknife intervals.

embeddings ($\|\hat{\boldsymbol{\beta}}\|^2_{\text{cor}} = 0.061$, 95% CI 0.049 to 0.072) and for the topic proportions,[10] and they tell the same broad story (Figure 3). In both cases, respondents move towards harsher punishment and away from leniency. For example, the topics on letting the person stay (T8), on limited punishment (T1) and on the circumstances of the crime (T4) lose 0.03 to 0.05 of their shares. Similarly, the answers move away from the SAE directions that consider the immigrant's personal reasons for entering the country ($\hat{\delta}_{F6} = -0.08$), invoke legal procedures ($\hat{\delta}_{F1} = -0.06$) and argue that prison is too harsh and deportation enough ($\hat{\delta}_{F5} = -0.04$).

[10] Egami et al. (2022) report topic-by-topic intervals. For a joint test, we apply the permutation test of Section 3.2 to the 11 topic proportions, with the sum of squared differences in mean proportions as the statistic.

However, the SAE's estimates are arguably more informative and easier to interpret. The topic with the STM's largest effect names a *position* on maximal punishment ("Punish to full extent of the law"), whose share rises by 0.13 with the treatment (95% CI 0.11 to 0.16). The SAE's strongest feature, in contrast, names the *reason* respondents give related to the man's criminal records: treated individuals move towards the direction that "emphasizes the person's violent or criminal history as justification for harsher punishment or exclusion" ($\hat{\delta}_{F2} = 0.17$, 95% CI 0.15 to 0.19), which alone carries about half of the effect ($q_{F2} = 0.53$). The STM's topic that matches the treatment's content, "Repeat offender, danger to society", is rare (4% of the average answer) and does not move.

Additionally, the STM and SAE disagree about one key concept. While the SAE finds that answers move towards deporting the man to spare taxpayers the cost of prison ($\hat{\delta}_{F4} = 0.07$, 95% CI 0.04 to 0.09), the STM's topic "Deport because of money" *falls* slightly (-0.02). Figure 4 shows why by plotting each test answer by its share of the topic (T3) and its score on the feature (F4). Interestingly, the correlation between the two measures is low (0.14). T3's most probable words are *just, send, back, countri, jail, come, prison, let, harm*, with *money* only being its tenth most probable word. Hence, an answer can score high on T3 without any argument about cost, such as "Give him a warning and send him back." What is more, the STM cannot capture synonyms for *money*, while the SAE F4 picks up the cost argument in whatever words it is made. For example, an answer reasoning that "he should not be imprisioned (sic) on the tax payers *dime* in the country he illegally entered" scores high on the SAE direction, but low on the STM topic. The topic thus misses much of what F4 captures, and its slight fall is largely mechanical. Since topic shares sum to one, the rise of "Punish to full extent of the law" from 0.09 to 0.22 of the average answer crowds out the other topics, including T3.

Finally, we leverage the fact that the survey also included a closed-ended response on whether the offender should be sent to prison, which respondents answered right before the open-ended text responses. Particularly, we check how well each representation (STM vs.

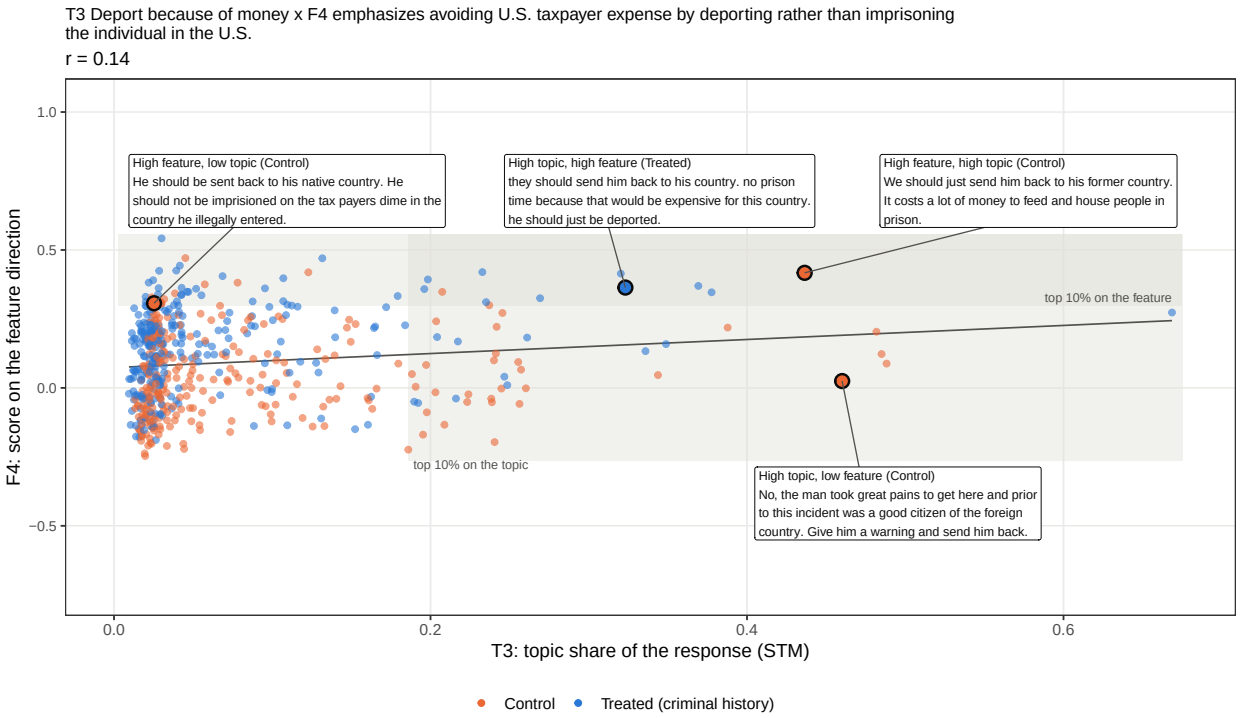


Figure 4: One topic against one feature, Egami et al. Experiment 3 (test half). Each point is an answer: its share of topic T3 ("Deport because of money") and its score on feature F4 ("emphasizes avoiding U.S. taxpayer expense…"). T3's ten highest-probability words are just, send, back, countri, jail, come, prison, let, harm, money. Shaded bands mark the top 10% on the topic and on the feature; boxes show example answers.

SAE) predicts the respondent's yes/no answer by regressing it on its dimensions and the treatment in the training half, and compute the AUC of the resulting text index for the 472 test answers (Figure 5). The seven SAE features separate yes from no almost perfectly (AUC 0.97, 95% CI 0.95 to 0.98), against 0.79 (0.74 to 0.83) for the 11 STM topics and 0.75 for the treatment alone, and they do so within each arm (0.96 among control and 0.95 among treated answers). The advantage holds for the single dimension with the largest effect,[11] where the SAE's feature on criminal history reaches 0.92 (0.89 to 0.94) and the STM's topic "Punish to full extent of the law" 0.81 (0.77 to 0.85).

One might criticize that this is an unfair comparison, since embedding models have many more dimensions to help with predictive performance. However, BERTopic (Grootendorst,

[11] We choose this dimension by its effect in the training half, so that, like the SAE and the labels, it is fixed before the test half is used. F2 and "Punish to full extent of the law" also have the largest effects in the test half.

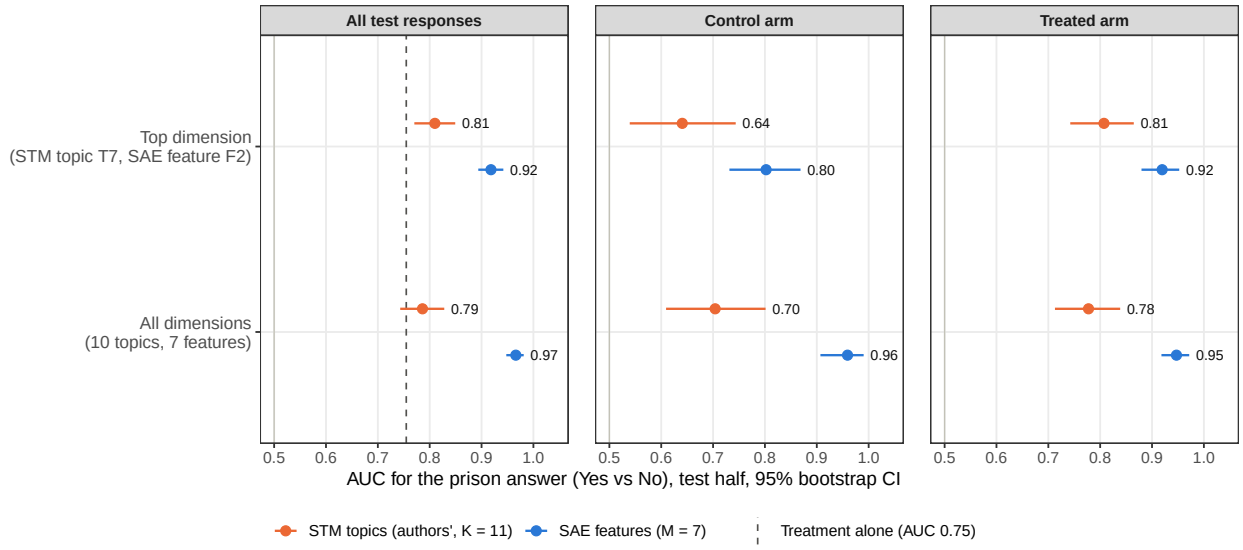


Figure 5: How well each representation predicts the respondent's own closed-ended answer ("Should this offender be sent to prison?"), Egami et al. Experiment 3. AUC on the test half, with weights fitted on the training half; 95% bootstrap intervals.

2022) uses the same embeddings as the SAE (Appendix Figures A2 and A3). Its topics together nearly match the SAE's features (AUC 0.94 against 0.97), but its single most affected topic predicts the prison answer poorly (0.65 against 0.92), because assigning each answer to a single cluster mixes arguments. Similarly, principal components describe the effect about as fully as the SAE (0.90 of $\|\hat{\boldsymbol{\beta}}\|^2$ for seven components), but each has two ends, and the component with the largest effect contrasts answers that propose both prison and deportation with answers that support legal pathways or leniency, two groups that each combine several arguments (Appendix Table A2).

4.2 Observational Data

SAE regression can also be applied to estimate and decompose effects in observational text data, including ATT estimates of a DiD design. To illustrate this, we study how members of Congress change what they tweet after a mass shooting in their district or state, a question earlier work has studied for the public's tweets (Zhang et al., 2019; Demszky et al., 2019). The shootings come from Mother Jones' database of public mass shootings with at least

three people killed (Follman et al., 2026). We compile 51 shootings between July 2017 and June 2023.[12] A shooting treats the House member of the district where it happened and the state's two senators, which gives 155 treated member-events. The tweets come from the office and campaign accounts of all members collected by congresstweets (Litel, 2026) (dropping pure retweets), and we embed the 1.4 million tweets posted within 14 days of a shooting with OpenAI's text-embedding-3-large.

We use a stacked difference-in-differences design (Cengiz et al., 2019; Wing et al., 2024), where each event forms a stack whose window runs from 14 days before its first shooting to 14 days after its last, with day 0 the day of the first shooting. The controls of a stack are the members in office on day 0 who are not treated by any shooting whose window overlaps the stack's, leaving out House members from the affected state, since they are arguably partly exposed. An observation is a member-day on which the member tweeted, and its outcome is the mean embedding of that day's tweets, $\mathbf{e}_{ist}$ for member i, stack s and day t. We estimate

$$\mathbf{e}_{ist} = \boldsymbol{\mu}_{is} + \boldsymbol{\lambda}_{ts} + \boldsymbol{\beta}\, D_{is}\, \mathrm{Post}_{st} + \boldsymbol{\varepsilon}_{ist}, \tag{13}$$

with member × stack and day × stack fixed effects, where D_{is} marks the treated members of stack s and Post_{st} the days from day 0 on. Under parallel trends in the embeddings,[13] $\boldsymbol{\beta}$ is the average effect of a shooting on the treated members' tweets,[14] and its projections decompose this effect as in Section 3.4. The SAEs are trained on the tweets of a random half of the members who are never treated (314 members), and these members are left out of the estimation, so that the SAE does not depend on the treated members' tweets. All

[12] Shootings in the same state at most 28 days apart are merged into one shooting.

[13] The global test needs parallel trends in every dimension of the embedding, a projection only along its direction, and the joint coefficients only within the span of the decoder directions. Note that the linear projections inherit parallel trends from the embeddings, but the nonlinear activations don't (Roth and Sant'Anna, 2023).

[14] As in any stacked regression without corrective weights, this average weights each stack in proportion to its size and to the variance of treatment within it (Wing et al., 2024). A collapsed 2 × 2 design, which compares each treated member's change from before to after the shooting with the controls' change and gives every treated member-event the same weight, finds the same features with somewhat larger effects (bias-corrected squared norm 5.5×10^{-3} against 4.2×10^{-3}; $\hat{\delta}_{M3} = 0.063$ against 0.054).

features of the SAEs are in Appendix Table A3. We use a cluster-bootstrap over members for the global test, and the intervals come from a leave-one-member-out jackknife.

Shootings change what the treated members tweet (Figure 6). The bias-corrected squared norm of the effect on all tweets is 4.2×10^{-3} (95% CI 2.1 to 6.2×10^{-3}; bootstrap $p = 0.001$), and one feature of the main SAE carries about half of it. Looking at the projections, members' tweets move towards M3, whose top tweets "address large-scale tragedies or societal problems, often calling for action or expressing collective mourning" ($\hat{\delta}_{M3} = 0.054$, 95% CI 0.042 to 0.067; $q_{M3} = 0.53$). The other features move little ($|\hat{\delta}_j| \leq 0.013$), mostly away from routine communication such as economic and tax policy (M17) and legislative announcements (M6). Members also post more such tweets. The number of tweets on which M3 fires rises by 0.45 per member-day (95% CI 0.28 to 0.63) from 0.62 before the shooting, while the total number of tweets does not change.

The main SAE shows *that* and *to what extent* members turn to tragedy talk, but not *how* they talk about that concept. To decompose it further, we restrict the outcome to tragedy tweets—i.e., those on which M3 is activated by the SAE—and decompose the effect on them with a separate SAE trained on such tweets ($M = 12$, $K = 3$). The estimate compares each member's tragedy talk with their own before the shooting and with the controls' on the same day. The effect within the tragedy dimension is clear and substantively interesting (6.8×10^{-3}, 95% CI 2.2 to 11.3×10^{-3}; $p = 0.001$). Treated members' tragedy tweets primarily move towards T7, "expresses condolences, sympathy, or prayers for victims or their families after a tragedy or loss" ($\hat{\delta}_{T7} = 0.054$, 95% CI 0.038 to 0.069; $q_{T7} = 0.42$), and towards condolences and gratitude for victims and first responders ($\hat{\delta}_{T5} = 0.017$; $q_{T5} = 0.09$). The movement towards gun violence and calls for action ($\hat{\delta}_{T1} = 0.022$; $q_{T1} = 0.06$) is smaller, and its interval barely touches zero (0.000 to 0.043). They move away from tweets about local incidents of national crises such as fentanyl or the border ($\hat{\delta}_{T11} = -0.030$; $q_{T11} = 0.08$).

As one would expect, the two parties react differently. Democrats move towards M3 more than twice as far as Republicans (0.068 against 0.030; test of the difference in the all-

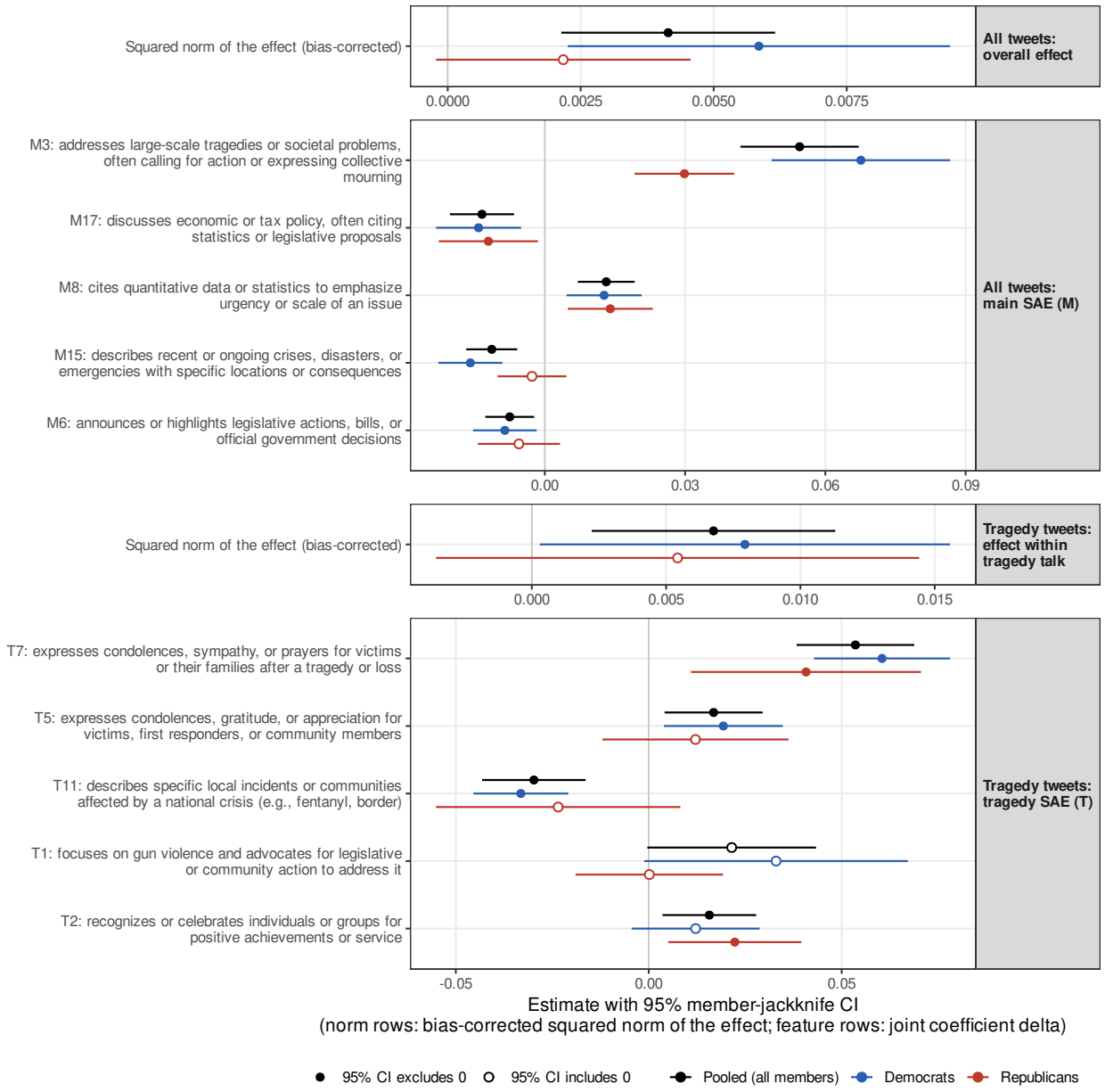


Figure 6: Effect of a mass shooting in a member's district or state on the member's tweets. Stacked difference-in-differences (51 shootings, 2017–2023; 155 treated member-events) with member × stack and day × stack fixed effects. Main SAE ($M = 20$, $K = 4$): the five features with the largest bias-corrected shares q_j in the pooled estimate. Tragedy tweets are those on which the main SAE's feature M3 fires, and their SAE ($M = 12$, $K = 3$) is trained on them. Both SAEs are trained on members never treated and left out of the estimate; all features are in Appendix Table A3. Party estimates come from one regression with treatment × post × party against the common day effects, so each includes its party's national reaction; within-party estimates are in Appendix Figure A5.

tweets effect $p = 0.032$). Within the tragedy direction, both move towards condolences and prayers ($\hat{\delta}_{T7} = 0.061$ and 0.041), but only Democrats move towards calls for action related

to gun violence ($\hat{\delta}_{T1} = 0.033$, 95% CI -0.001 to 0.067; Republicans 0.000). These party estimates compare treated members with all controls and so include each party's national reaction to the shooting. Compared only with controls from their own party (Appendix Figure A5), treated Democrats still move towards condolences and prayers ($\hat{\delta}_{T7} = 0.059$, 95% CI 0.042 to 0.075) but not clearly towards gun violence ($\hat{\delta}_{T1} = 0.026$, -0.005 to 0.057). Treated Republicans move towards gun violence and calls for action ($\hat{\delta}_{T1} = 0.034$, 0.016 to 0.051) relative to their co-partisans, although the global test for their tragedy tweets is not significant ($p = 0.064$). This suggests that Democrats' turn to legislative action on gun violence largely reflects their party's national reaction, whereas Republicans exposed to a shooting turn to it more than their party colleagues.

The event study (Figure 7) shows no pre-trend (lead bin $p = 0.28$ for all tweets and 0.51 for tragedy tweets) and a reaction concentrated in the first three days (squared norm 21.5×10^{-3} on days 0 to 2, against 3.3×10^{-3} on days 3 to 7). Additionally, Congress members also adjust the timing of specific arguments emphasized in their tweets: condolences appear instantaneously ($\hat{\delta}_{T7} = 0.121$ on days 0 to 2), while calls for action seem more prevalent a few days later ($\hat{\delta}_{T1} = 0.006$ on days 0 to 2 and 0.033, 95% CI 0.004 to 0.061, on days 3 to 7).

We report several robustness checks in the appendix. On placebo dates a year earlier, the design finds no effect (global $p = 0.31$ for all tweets and 0.55 for tragedy tweets; Figure A8). An SAE trained only on tweets from before the shootings, and SAEs with $M = 10$ and $M = 40$, each have a feature close to M3 that carries the effect (Figure A7). Comparing treated members only with controls from their own party, which removes each party's national reaction to the shooting, gives nearly the same pooled estimate ($\hat{\delta}_{M3} = 0.054$; $\hat{\delta}_{T7} = 0.049$; Figure A5). And the House member of the affected district reacts more than twice as strongly as the state's senators (Figure A6).

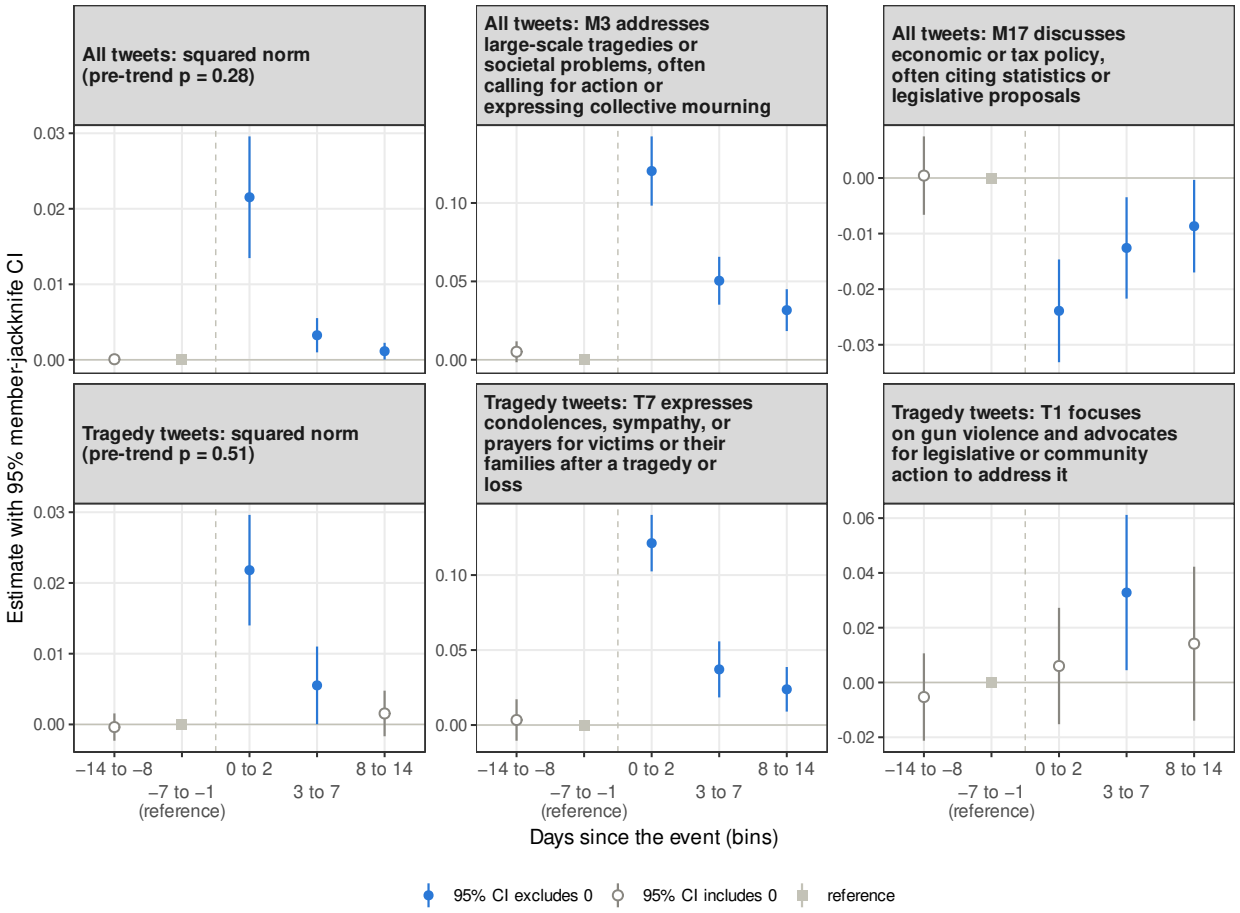


Figure 7: The effect by days since the shooting, against the week before (days −7 to −1). Left: the bias-corrected squared norm; the pre-trend p-value is a member-bootstrap test of the lead bin (days −14 to −8). Other panels: the joint coefficients of the tragedy feature M3 and of M17 (economic and tax policy) for all tweets, and of T7 (condolences and prayers) and T1 (gun violence and calls for action) for tragedy tweets. Design as in Figure 6; 95% member-jackknife intervals.

5 Validation and Guidance for Researchers

5.1 Ground Truth: Restaurant Reviews with Known Mechanisms

Having shown how SAE regression applies to experimental and observational data, we now use simulations to assess how well the decomposition recovers true effects. This requires texts whose potential outcomes are both observed, which the CEBaB data provide (Abraham et al., 2022). Crowdworkers edited 2,299 restaurant reviews to change how each evaluates one aspect of the restaurant (food, service, ambiance or noise), leaving the rest of the text as is (12,791 edits). Because we observe every original review and each of its edits, we can simulate experiments in which a treatment works through known mechanisms, and compute the true effect and its projections exactly.

The simulated treatment makes reviews more negative through four mechanisms. It turns praise of the food into a complaint, or adds a complaint about service, ambiance or noise to a review that did not mention the aspect. Let $Y_i^{(0)}$ be the original text of review i and $Y_i^{(c)}$ its edit for mechanism c, which exists for $c \in \mathcal{A}_i$, with $n_i = |\mathcal{A}_i|$ and $\boldsymbol{\tau}_{ic} = e(Y_i^{(c)}) - e(Y_i^{(0)})$. A control review appears as written. A treated review is edited with probability π, the dose, and receives one of its mechanisms at random, so $Y_i(1) = Y_i^{(C_i)}$ with $P(C_i = c) = \pi/n_i$. Averaged over the N reviews from which the experiments are drawn, the true effect is

$$\boldsymbol{\beta}^* = \frac{1}{N}\sum_{i=1}^{N} E\big[e(Y_i(1)) - e(Y_i(0))\big] = \sum_c \boldsymbol{\beta}_c^*, \qquad \boldsymbol{\beta}_c^* = \frac{\pi}{N}\sum_{i:\, c\in\mathcal{A}_i} \frac{\boldsymbol{\tau}_{ic}}{n_i}, \tag{14}$$

where $\boldsymbol{\beta}_c^*$ is the part of the effect that runs through mechanism c. For a given SAE, the true projections are $\boldsymbol{\alpha}^* = W^\top \boldsymbol{\beta}^*$ and $\boldsymbol{\delta}^* = G^{-1}\boldsymbol{\alpha}^*$. The main feature of mechanism c is the feature with the largest share q_{cj} of $\boldsymbol{\beta}_c^*$.

We split the reviews into a training half and a test half of $N = 1{,}150$. Each experiment draws $2n$ distinct test reviews, assigns n to each arm and estimates $\hat{\boldsymbol{\beta}}$ as the difference in mean embeddings, so that $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}^*$. The SAEs are trained on texts drawn in the same way

from the training half. Unless stated otherwise, each condition uses 10 SAEs with $M = 10$ and $K = 4$, 200 experiments and 500 permutations for the global test.

Notably, this simulation is demanding. Even at a dose of one, only 81% of treated texts are edited at all—with 23% showing the food edit, 12% the service edit, 20% the ambiance edit and 25% the noise edit[15]—and the treatment explains only 0.64% of the variance in the embeddings, less than a quarter of the 2.8% explained by the treatment in Egami et al. (2022).

5.2 The Decomposition Is Unbiased, Well Calibrated and Meaningful

We first demonstrate that SAE regression is unbiased and calibrated in finite samples. Figure 8A compares the average of each estimated $\hat{\alpha}_j$ and $\hat{\delta}_j$ over 200 experiments with their true value, for every feature of every SAE in every experiment condition with a nonzero effect. The points cleanly line up on the diagonal, with an average error below 10^{-4}, and the 95% intervals cover the true value in 95% of experiments on average (88% in the worst case). When $\boldsymbol{\beta}^* = \mathbf{0}$, the global test rejects in 3 to 5% of experiments, and no mechanism's main feature is flagged in more than 2.2% of them.

Figure 8B shows how detection depends on a feature's true movement (measured as the standardized movement in the feature's score: $d = \alpha_j^*/\mathrm{SD}(s_{ij})$ for $\hat{\alpha}_j$ and $d = \delta_j^*/\mathrm{SD}(\tilde{s}_{ij})$ for $\hat{\delta}_j$). With 500 or 550 test texts per arm, $\hat{\delta}_j$ flags a feature that moves by 0.12 standard deviations in about half of experiments and one that moves by 0.25 almost always. Flagged estimates have the wrong sign mainly for features that barely move (13 to 28% of those that move less than 0.05 standard deviations, and less than 1% of those that move more than 0.1).

How informative are the directions and their labels in decomposing the induced treatment effect and mechanisms? To assess this, Figure 9 holds one SAE fixed and varies the dose of

[15] A review with several edits receives only one, so mechanism c reaches a treated review with probability πs_c, where $s_c = \frac{1}{N}\sum_{i:\, c\in\mathcal{A}_i} 1/n_i$. Also, the edits differ in how far they move a review: food accounts for 37% of $\|\boldsymbol{\beta}^*\|^2$, noise for 35%, ambiance for 20% and service for 9% ($\boldsymbol{\beta}_c^{*\top}\boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|^2$).

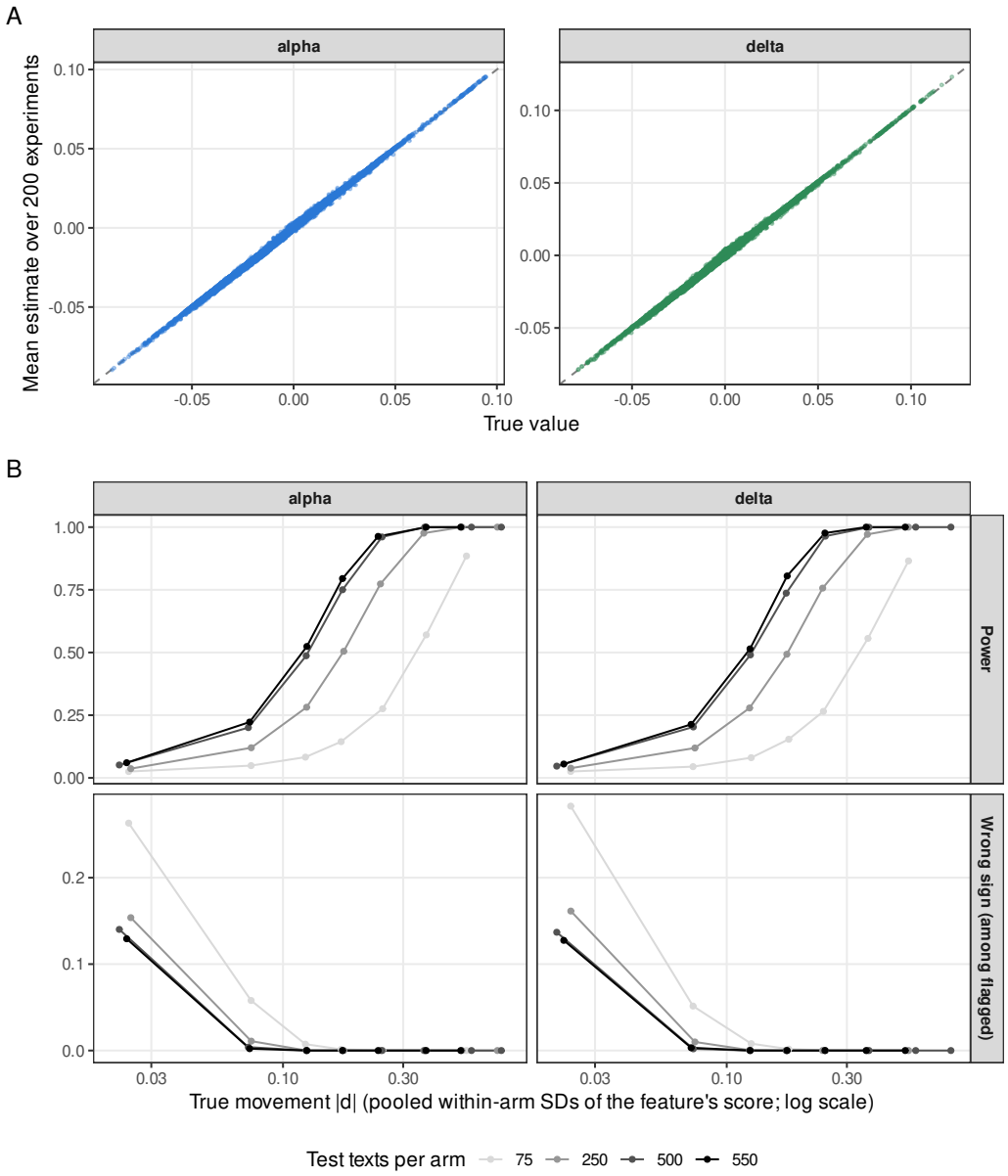


Figure 8: Unbiasedness and calibration, CEBaB ground truth. (A) Mean estimate over 200 experiments against the true value, for every feature, SAE and condition with a nonzero effect; the dashed line is the diagonal. (B) Power and wrong signs against the true movement $|d|$, the true effect on the feature's score (s_{ij} or $\tilde{s}_{ij}$) in pooled within-arm standard deviations, by test texts per arm. A feature is flagged when the global test rejects and its 95% interval excludes zero.

a treatment with two mechanisms, turning praise of the food into a complaint and adding a complaint about noise. Each panel is a feature, ordered by the true share r_j^* of the effect along its direction. The estimates are again unbiased, following their true values at every dose. Additionally, the top features capture the manipulated mechanisms: the feature carrying

the largest share comments on noise, and the treatment moves the embeddings towards its direction. The change in food appears as a change in tone: two features labeled as an intent to return or recommend the restaurant carry the next largest shares, and the treatment moves the embeddings away from them, while it moves them towards a feature labeled as strong negative judgments. Reassuringly, aspects that the manipulation left untouched, such as complaints about prices, are estimated to move little or not at all (bottom row of the figure).

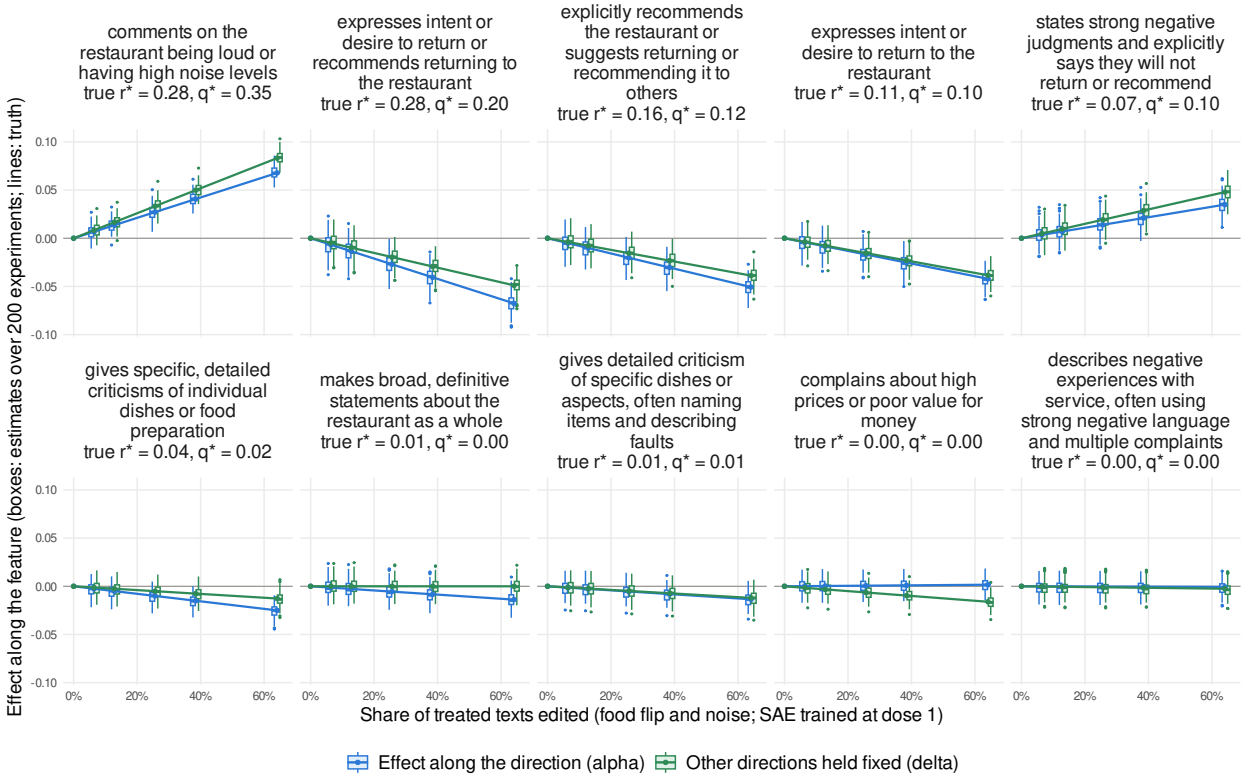


Figure 9: One SAE across doses, CEBaB ground truth. Food praise turned into complaints and added complaints about noise; the SAE ($M = 10$, $K = 4$) is trained at a dose of one. Panels are features, titled with their LLM label and true shares r_j^* and q_j^*. Boxes show the estimates over 200 experiments per dose (1,000 test texts); lines show the true values.

5.3 Mechanisms That Differ in Content Separate

Figure 10 asks whether a change in one mechanism is attributed to that mechanism alone. Complaints about noise are held at a dose of 0.4, while we vary the dose of a second mech-

anism (either ambiance or food) from 0 to 0.8, holding the SAE fixed throughout. Reassuringly, when the second mechanism turns praise of the food into complaints, the estimate for noise's main feature stays flat, whereas the estimate for food's main feature grows sixfold. This separation is harder for complaints about ambiance and noise that change the text in similar ways: as the ambiance dose rises, the estimate for noise's main feature grows by 19%, although nothing about noise changed. Yet, the leak is small (about half of a single experiment's margin of error) and, importantly, this is a limit of the DGP rather than a bias concern for the estimator. In these SAEs, part of the ambiance edits lies along noise's feature itself (their mean edits have a cosine of 0.40, $\cos(\boldsymbol{\beta}^*_{\text{ambiance}}, \boldsymbol{\beta}^*_{\text{noise}})$), and $\hat{\delta}_j$ cannot separate two mechanisms that inherently share a direction.[16]

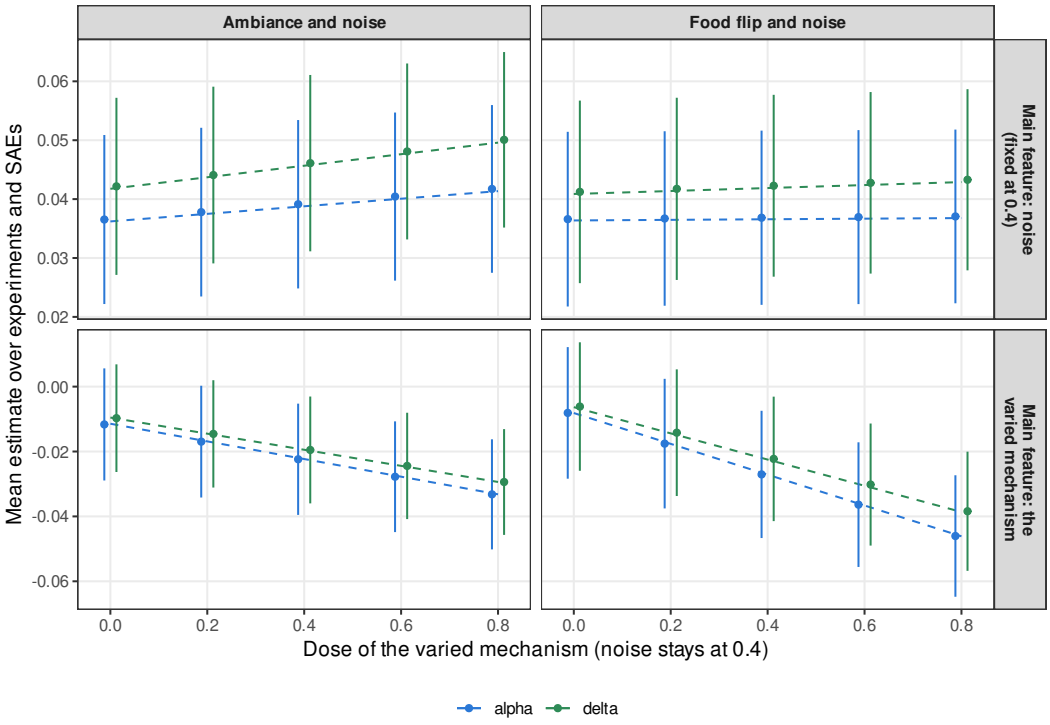


Figure 10: Selectivity, CEBaB ground truth. Noise complaints are held at a dose of 0.4 while the dose of food or ambiance complaints rises; one SAE throughout ($M = 10$, $K = 4$, trained at a dose of one). Top: noise's main feature; bottom: the varied mechanism's. Points are mean estimates, bars a single experiment's 95% interval, and dashed lines the true values.

[16] With SAEs trained on each condition's own texts instead of a fixed SAE (Appendix Figure A9), the results hold for food, but ambiance's main feature is flagged in up to 0.11 fewer experiments at low doses.

5.4 Informativeness by Treatment Strength and Sample Size

What treatment strength is necessary for the full decomposition to be informative? Figure 11 varies the dose of the four-mechanism treatment for samples of 300, 1,000 and 2,200 texts. Detection is the bottleneck. Yet, with enough texts SAE regression already detects small effects. With 2,200 texts, the global test rejects in 87% of experiments when 24% of treated texts are edited—a treatment that explains only 0.06% of the variance in the embeddings. With 300 texts, it reaches 90% when 65% of treated texts are edited. When half of the treated texts are edited, the main feature is flagged in 93% and 99% of experiments with 1,000 and 2,200 texts, and the significant features carry 63% and 77% of the true effect. Expressibility—the share of the true effect the SAE's directions span—hardly depends on the dose (0.65 to 0.85) and marginally rises with the size of the training half.[17]

5.5 Choice of M and K

Scholars may wonder what the optimal choice for M and K is for the SAE. Figure 12 varies M from 5 to 60 and K from 1 to 8, with the SAE trained on 250 or 1,000 texts and the effect estimated on 1,000 others. The decomposition is robust to M and K over a moderate range. On the one hand, a larger M spans more of the effect and gives more mechanisms a feature of their own, but beyond $M = 20$ it adds little: its significant features carry 0.80 of the effect at every M from 20 to 60 (with 1,000 training texts, half of treated texts edited and $K = 4$).[18] On the other hand, a higher M and K is costly. The share of flags on features that barely move—those that came with wrong signs in Figure 8—rises from 3% at $M = 10$ to 16% at $M = 60$ with half of treated texts edited. More flexibility also risks overfitting features to the training text. With 250 training texts, for example, the features become less

[17] This partly reflects the texts the SAE is trained on. Complaints also occur in untreated reviews, so an SAE that saw no edited text still spans most of the effect (expressibility of 0.80 and 0.75 for two pairs of mechanisms, against 0.90 and 0.88 at full strength; Appendix Figure A10). A treatment that introduces new content would make the SAE more dependent on treated texts.

[18] Features are flagged with pointwise 95% intervals. A correction for testing M features, such as max-T (Westfall and Young, 1993), becomes stricter as M grows and further raises the cost of a larger M.

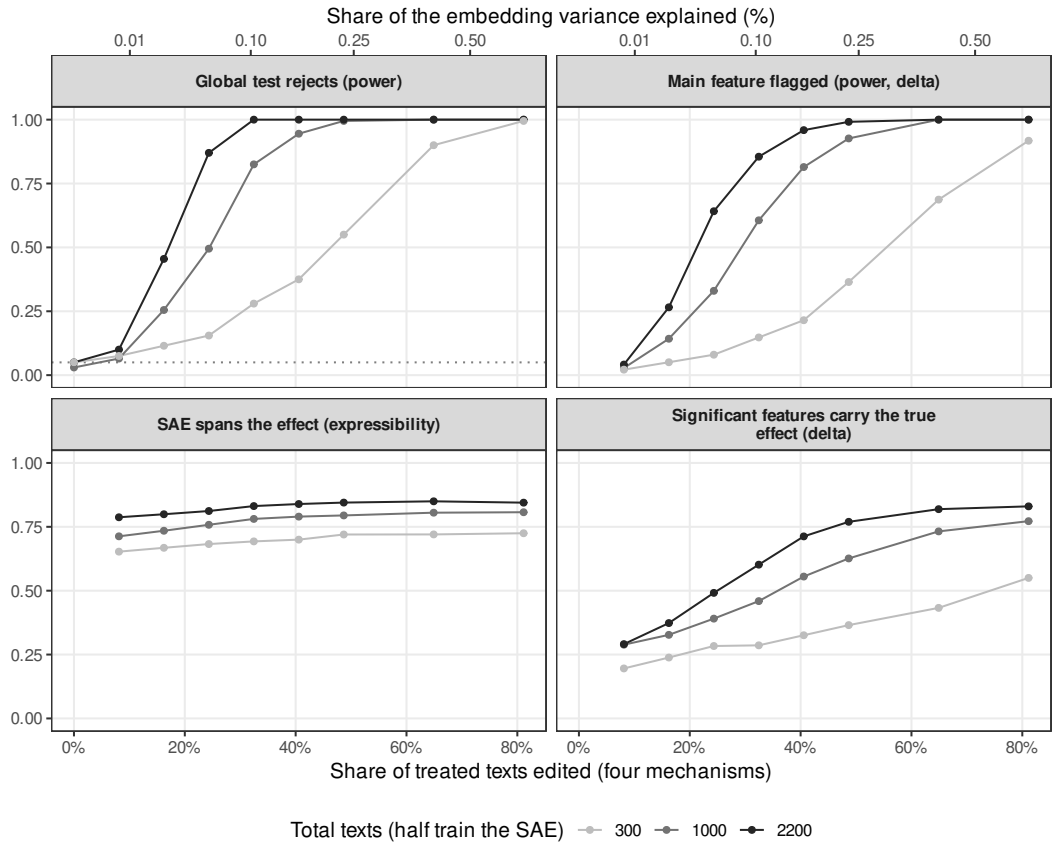


Figure 11: Strength of the treatment, CEBaB ground truth with four mechanisms ($M = 10$, $K = 4$; half of the texts train the SAE at the experiment's dose). Top: rejection rate of the global test (dotted line at 0.05) and power for the main feature. Bottom: expressibility and, when the global test rejects, the share of the true effect carried by the features with significant $\hat{\delta}_j$. Lower axis: share of treated texts edited; upper axis: share of the variance explained.

reproducible across SAEs as M grows.[19] Researchers should therefore fix M in advance at a moderate value relative to the number of training texts (10 to 20 for 1,000 texts here, 10 to 15 for 250), with K between 2 and 4, and report neighboring values as robustness. Researchers who need to tell similar concepts apart may also report a larger M.

[19]Cross-fitting does not help to choose M or K. The stability of features across SAEs favors small, very sparse SAEs, and the reconstruction error on the test texts is lowest past the point where the significant features stop carrying more of the effect, since the treatment explains well under 1% of the texts' variance (Appendix Figure A11). Additionally, cross-fitting would lead to more data splits, which cost additional power.

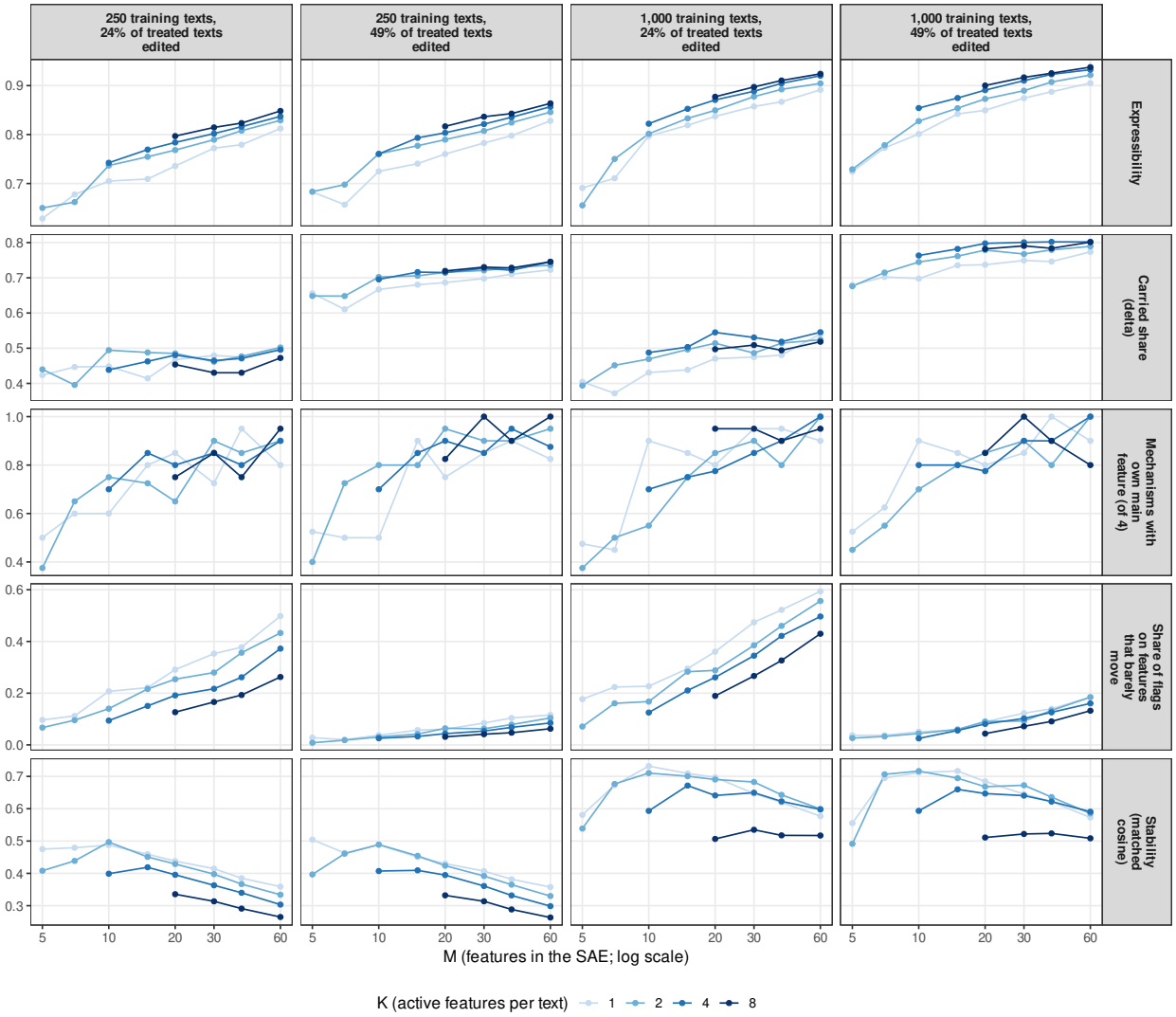


Figure 12: Choice of M and K, CEBaB ground truth with four mechanisms (effect estimated on 1,000 texts). Columns: the SAE trained on 250 or 1,000 texts, with a quarter or a half of treated texts edited. Rows: expressibility; the share of the true effect carried by the features with significant $\hat{\delta}_j$; the mechanisms with a main feature of their own (share of four); the share of flags on features that barely move (true movement below 0.05 standard deviations); and stability, the mean matched decoder cosine between SAEs trained on different samples.

5.6 The Validity of the Labels

The labels of the SAE decoder do not enter the estimates, but they determine how researchers read them. We validate the LLM labeling with a new corpus, namely the 47,448 distinct written questions that members of the UK House of Commons tabled in 2023 (UK Parliament, 2026). An SAE with $M = 20$ and $K = 4$ is trained on half of the questions, and

the contrast between Labour and Conservative members is estimated on the other half as $\boldsymbol{\beta}$.

First, we measure each label's fidelity, i.e., whether it picks out its feature's texts among texts the labeler did not see. A second LLM call reads 100 held-out training questions per feature, half from the top bin of its score and half from the lower bins, together with its label, and decides whether the label applies (Movva et al., 2025; Paulo et al., 2024). The LLM labels reach a mean precision of 0.81 and a mean recall of 0.67 (F-score of 0.71; Appendix Table A1). Notably, fidelity is lowest where the labeler gave several features similar, generic labels about requests for quantitative data (F-score of 0.64, against 0.77 for the other ten features).

Second, we labeled all 20 features blind, using the labeler's exact prompts (Table 2). For 11 of the 20 features, our label's embedding has a higher cosine similarity with the same feature's LLM label than with any other LLM label. Compared with the LLM label of a randomly chosen other feature, the same feature's LLM label is the more similar one in 80% of cases. In terms of the F-score, the human labels fit slightly better (F of 0.75), mainly because they are more precise (0.92 against 0.81), and they name the topics that the generic LLM labels miss, such as budget figures, healthcare and housing. Labels are therefore best read as candidate names for a direction, which researchers should check against the top-scoring texts, especially for the features that carry most of the effect.

Table 2: Human against LLM labels, UK House of Commons written questions (47,448 distinct questions tabled in 2023, at most 255 characters each; $M = 20$, $K = 4$). Score: share of the other 19 LLM labels further from the human label than its own LLM label (embedding cosine); 0.5 is chance. q_j: the feature's share of the Labour–Conservative contrast on the test half. F: F-score of each label set's fidelity on the same 100 questions per feature.

	Human label	LLM label	Score	q_j	F, LLM	F, human
F10	asks about specific steps a department or minister is taking on an issue	asks what steps are being taken to address or improve a specific issue	1.00	0.17	0.94	0.92
F12	refers to and follows up on a specific report or action with a certain date	explicitly references previous official documents, reports, or answers in the question text	1.00	0.11	0.99	0.98
F8	asks for specific policy actions, such as amendments to existing policies or assessments	requests a policy change, action, or assessment from a government official or body	1.00	0.05	0.91	0.92
F1	asks about a specific public budget figure (funding, debt, spending)	requests specific monetary amounts or financial figures	1.00	0.05	0.36	0.88
F9	asks for a specific number (numerical figure or size) on an issue	requests specific quantitative data or statistics (e.g., numbers, proportions, sizes) about a subject	1.00	0.04	0.79	0.76
F17	asks about an assessment made by the minister/department on a specific policy or question	asks for an assessment or evaluation of adequacy, impact, or implications of a policy or event	1.00	0.03	0.98	0.96
F3	mentions the health, wellbeing and protection of vulnerable groups	asks about assessments or plans regarding health, social care, or LGBT-related policies and services	1.00	0.02	0.22	0.20
F0	mentions a foreign policy issue with a particular country and its relations to the UK	references foreign countries, international relations, or UK actions involving other nations	1.00	0.01	0.96	0.70
F4	asks about government timelines, commitments and publications	asks about future plans, policies, or publication of reports or responses	1.00	0.00	0.86	0.84
F7	asks about a healthcare issue	asks for specific quantitative data or timelines about healthcare services or medical issues	1.00	0.00	0.48	0.85
F19	asks about specific consultations and cooperation the minister/department has had with other actors (yes/no questions)	mentions discussions or consultations with specific people or organizations	1.00	0.00	0.82	0.84
F16	lists two aspects in the question, numerated by a) and b) or i) and ii)	asks for specific quantitative data, such as counts, proportions, or measurements	0.89	0.06	0.77	0.91
F2	asks about educational policies and childcare, including issues in primary and secondary schools	requests data or statistics broken down by categories such as region, school type, or condition	0.74	0.01	0.43	0.80
F13	asks about individual-level service delivery by the state and accountability (especially around asylum)	requests quantitative data or statistics about specific outcomes, populations, or timeframes	0.68	0.14	0.67	0.36
F18	asks about issues related to employment, staff management and work programs	requests specific quantitative data or statistics about government programs or expenditures	0.63	0.10	0.81	0.80
F14	discusses domestic social issues and justice and how institutions treat people	requests quantitative data or statistics about specific issues or departmental activities	0.63	-0.01	0.80	0.38
F5	asks about the maintenance and building of public infrastructure	mentions specific locations, facilities, or infrastructure in the context of operational or logistical details	0.63	-0.01	0.70	0.52
F6	asks about environmental policies, particularly energy use, net zero policies and climate events	requests recent or future actions, estimates, or statistics rather than past assessments or consultations	0.37	0.00	0.49	0.61
F11	mentions law enforcement or the military	requests specific factual information about past events, actions, or statistics	0.32	0.06	0.69	0.83
F15	asks about an issue related to public estates, physical assets, and housing	requests specific quantitative data such as costs, numbers, proportions, or sizes	0.16	0.01	0.48	0.94
Mean			0.80		0.71	0.75

6 Conclusion

Effects on unstructured data such as text rarely run along a single dimension. SAE regression estimates the overall effect on the texts and decomposes it into interpretable directions, each with its own estimate, interval and share of the effect. It allows researchers to discover dimensions they did not anticipate and estimate their relevance. The labels name the directions without entering the estimates, so the results do not depend on the LLM, and human readers can check or replace its labels. The method applies to any regression design that identifies an effect on the average embedding, such as survey experiments or difference-in-differences. In our applications, the decomposition captured the reasoning in respondents' answers and the rhetoric of members of Congress after mass shootings. Having proved and validated the method for text data, the method could likely be applied and extended to other high-dimensional, unstructured data, such as images, videos or maps.

Two caveats qualify these results. First, the SAE is only as good as the texts it learns from. It assigns its directions to what varies most in the training texts, so a treatment that introduces rare content may get no direction of its own, and its effect then spreads over features whose labels describe something else. In our simulations this mattered little, because complaints also occur in untreated reviews, but a treatment that introduces new content would make the SAE more dependent on treated texts. Detection is unaffected, since the overall effect and the global test never use the SAE, and the projections remain unbiased for the SAE at hand. However, interpretation becomes weaker, and a low R^2 warns that the labels can only describe part of the effect. Notably, the same limit applies to any measure learned from the study's texts, including topic models.

Second, the SAE builds each text by adding directions with nonnegative weights, so a direction captures one emphasis without an opposite pole. A text scores low on $\mathbf{w}_j$ when the other features it uses point away from $\mathbf{w}_j$, and a low score does not mean that the text expresses the opposite of feature j. A negative projection thus means that treated texts

move away from the feature, but not that they express the "opposite" of a feature.

With a causal research design, SAE regression estimates causal effects along learned directions. A natural next step is to specify when these estimates capture causal mechanisms (Slough and Tyson, 2024), which can mean two things here: it can be a process that links the treatment to the texts themselves, such as an argument treated respondents begin to make, or a mediator that links the treatment to a separate outcome, such as the respondent's punishment judgment. Our simulations, in which known edits define the mechanisms, suggest that the leading directions can capture mechanisms in the first sense. Establishing when they do so more generally, and when their effects speak to mediation (Imai et al., 2011; Blackwell et al., 2024; Fu and Slough, 2026), is a promising avenue for future work.

References

Abraham, Eldar David, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. "CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior". In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. arXiv:2205.14140.

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart. 2026. "Narratives about the macroeconomy". *Review of Economic Studies*: rdag014.

Baldi, Pierre and Kurt Hornik. 1989. "Neural networks and principal component analysis: Learning from examples without local minima". *Neural Networks* 2 (1): 53–58.

Bills, Steven, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. "Language models can explain neurons in language models". OpenAI. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Blackwell, Matthew, Ruofan Ma, and Aleksei Opacic. 2024. "Assumption smuggling in intermediate outcome tests of causal mechanisms". arXiv:2407.07072.

Bouyamourn, Adam and Arthur Spirling. 2026. "Interpretable aggregation of correlated LLM annotators". Working paper.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. "Towards monosemanticity: Decomposing language models with dictionary learning". Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features.

Carlson, Jacob. 2026. "Interpretable discovery from unstructured data: A high-dimensional approach". arXiv:2511.01680.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs". *Quarterly Journal of Economics* 134 (3): 1405–1454.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. "Sparse autoencoders find highly interpretable features in language models". In *International Conference on Learning Representations (ICLR)*. arXiv:2309.08600.

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2970–3005.

Dodd, Lori E. and Edward L. Korn. 2007. "The bootstrap variance of the square of a sample mean". *The American Statistician* 61 (2): 127–131.

Doshi-Velez, Finale and Been Kim. 2017. "Towards a rigorous science of interpretable machine learning". arXiv:1702.08608.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. "How to make causal inferences using texts". *Science Advances* 8 : eabg2652.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2026. "Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses". Conditionally accepted at the *American Journal of Political Science*. https://naokiegami.com/paper/dsl_ss.pdf.

Feder, Amir, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond". *Transactions of the Association for Computational Linguistics* 10 : 1138–1158.

Follman, Mark, Gavin Aronsen, and Deanna Pan. 2026. "US mass shootings, 1982–2026: Data from Mother Jones' investigation". Mother Jones. Dataset, last updated 2 August 2026. https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/, accessed 11 October 2026.

Fowler, Erika Franklin, Michael M. Franz, Gregory J. Martin, Zachary Peskowitz, and Travis N. Ridout. 2021. "Political advertising online and offline". *American Political Science Review* 115 (1): 130–149.

Freedman, David and David Lane. 1983. "A nonstochastic interpretation of reported significance levels". *Journal of Business & Economic Statistics* 1 (4): 292–298.

Fu, Jiawei and Tara Slough. 2026. "Heterogeneous treatment effects and causal mechanisms". *American Political Science Review*. Published online.

Gao, Leo, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. "Scaling and evaluating sparse autoencoders". arXiv:2406.04093.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks". *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Green, Breanna, William Hobbs, Sofia Avila, Pedro Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2024. "Measuring distances in high dimensional spaces: Why average group vector comparisons exhibit bias, and what to do about it". Working paper, June 26, 2024.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". arXiv:2203.05794.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies". *American Political Science Review* 105 (4): 765–789.

Jiang, Nick, Xiaoqing Sun, Lisa Dunlap, Lewis Smith, and Neel Nanda. 2025. "Interpretable embeddings with sparse autoencoders: A data analysis toolkit". arXiv:2512.10092.

Linegar, Mitchell, Rafal Kocielnik, and R. Michael Alvarez. 2023. "Large language models and political science". *Frontiers in Political Science* 5 : 1257092.

Litel, Alex. 2026. "congresstweets: Daily archive of tweets by members of Congress". GitHub repository. Data from 21 June 2017 to 11 July 2023, commit f5c837de. https://github.com/alexlitel/congresstweets, accessed 11 October 2026.

Ludwig, Jens and Sendhil Mullainathan. 2024. "Machine learning as a tool for hypothesis generation". *The Quarterly Journal of Economics* 139 (2): 751–827.

Makhzani, Alireza and Brendan Frey. 2014. "k-sparse autoencoders". In *International Conference on Learning Representations (ICLR)*. arXiv:1312.5663.

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

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. "Sparse autoencoders for hypothesis generation". In *Proceedings of the 42nd International Conference on Machine Learning*, Volume 267 of *Proceedings of Machine Learning Research*. arXiv:2502.04382.

O'Neill, Charles, Christine Ye, Kartheek Iyer, and John F. Wu. 2024. "Disentangling dense embeddings with sparse autoencoders". arXiv:2408.00657.

Park, Kiho, Yo Joong Choe, and Victor Veitch. 2024. "The linear representation hypothesis and the geometry of large language models". In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Paulo, Gonçalo, Alex Mallen, Caden Juang, and Nora Belrose. 2024. "Automatically interpreting millions of features in large language models". arXiv:2410.13928.

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

Roberts, Margaret E., 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.

Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart. 2023. "Embedding regression: Models for context-specific description and inference". *American Political Science Review* 117 (4): 1255–1274.

Roth, Jonathan and Pedro H. C. Sant'Anna. 2023. "When is parallel trends sensitive to functional form?". *Econometrica* 91 (2): 737–747.

Slough, Tara and Scott A. Tyson. 2024. *External Validity and Evidence Accumulation*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge: Cambridge University Press.

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2024. "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet". Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity.

UK Parliament. 2026. "Written questions, answers and statements API". Written questions tabled in the House of Commons in 2023. https://questions-statements-api.parliament.uk, accessed 6 October 2026. Open Parliament Licence v3.0.

@unitedstates Project. 2026a. "congress-legislators: Members of the United States Congress, 1789–present". GitHub repository. Commit d5af3d2d. https://github.com/unitedstates/congress-legislators, accessed 11 October 2026.

@unitedstates Project. 2026b. "districts: Congressional district boundaries". GitHub repository. North Carolina's 2020 districts, commit b84d0cbd. https://github.com/unitedstates/districts, accessed 11 October 2026.

U.S. Census Bureau. 2026. "Cartographic boundary files: Congressional districts of the 115th, 116th and 118th Congresses". Files cb_2017_us_cd115_500k, cb_2018_us_cd116_500k and cb_2022_us_cd118_500k. https://www2.census.gov/geo/tiger/, accessed 11 October 2026.

Wang, Jenny S., Aliya Saperstein, and Emma Pierson. 2026. "In your own words: Computationally identifying interpretable themes from free-text survey data". arXiv:2603.26930.

Westfall, Peter H. and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley.

Wing, Coady, Seth M. Freedman, and Alex Hollingsworth. 2024. "Stacked difference-in-differences". NBER Working Paper 32054, National Bureau of Economic Research.

Wirsching, Elisa M., Pedro L. Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2025. "Multilanguage word embeddings for social scientists: Estimation, inference, and validation resources for 157 languages". *Political Analysis* 33 (2): 156–163.

Xu, Paiheng, Jing Liu, and Wei Ai. 2026. "Conditional hypothesis generation for LLM-based text analysis with researcher-specified covariates". arXiv:2606.03029.

Zhang, Yini, Dhavan Shah, Jordan Foley, Aman Abhishek, Josephine Lukito, Jiyoun Suk, Sang Jung Kim, Zhongkai Sun, Jon Pevehouse, and Christine Garlough. 2019. "Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014". *Journal of Computer-Mediated Communication* 24 (4): 182–202.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can large language models transform computational social science?". *Computational Linguistics* 50 (1): 237–291.

Appendix: Supporting Information for *What Moved? Interpretable Decompositions of Effects on Text with Sparse Autoencoders*

Table of Contents

A Proofs **A1**

A.1 Projections as regressions A1

A.2 Proof of Proposition 1 . . . . . A2

A.3 Proof of Proposition 2, and remarks on the shares A2

A.4 The effect on the activations z_{ij} A4

B Labeling and Fidelity Prompts **A6**

C Additional Results on Applications **A9**

D Additional Results on Validations **A18**

E Data Sources **A20**

A Proofs

Throughout, W is the $D \times M$ matrix whose unit-length columns are the decoder directions $\mathbf{w}_1, \ldots, \mathbf{w}_M$, $G = W^\top W$ is their Gram matrix with entries $g_{jk} = \mathbf{w}_j^\top \mathbf{w}_k$, and $\boldsymbol{\alpha} = W^\top \boldsymbol{\beta}$ and $\boldsymbol{\delta} = G^{-1}\boldsymbol{\alpha}$ are the projections of Section 3.4. G is invertible because the columns of W are linearly independent.

A.1 Projections as regressions

Lemma 1. Let X be the design of the embedding regression on the test half, containing the variable of interest and any covariates or fixed effects, and let $\hat{\boldsymbol{\beta}}$ be the coefficient of the variable of interest when the embeddings are regressed on X. For any fixed $D \times m$ matrix A, the coefficient of the variable of interest when the m outcomes $A^\top \mathbf{e}_i$ are regressed on X is $A^\top \hat{\boldsymbol{\beta}}$.

Proof. Stack the embeddings in the $n \times D$ matrix E. The coefficients of the embedding regression are $\hat{B} = (X^\top X)^{-1} X^\top E$, and $\hat{\boldsymbol{\beta}}^\top$ is the row of $\hat{B}$ that belongs to the variable of interest. The regression of the outcomes EA on the same design has coefficients

$(X^\top X)^{-1}X^\top EA = \hat{B}A$, whose row for the variable of interest is $\hat{\boldsymbol{\beta}}^\top A$. Fixed effects are columns of X, so the argument covers them, including the stacked design of Equation (13). □

With $A = \mathbf{w}_j$, the outcome is the score s_{ij} and the coefficient is $\hat{\alpha}_j = \mathbf{w}_j^\top \hat{\boldsymbol{\beta}}$ (Equation (7)). With $A = WG^{-1}$, the outcomes are the coordinates $\tilde{\mathbf{s}}_i = G^{-1}W^\top \mathbf{e}_i$ and the coefficients are $G^{-1}W^\top \hat{\boldsymbol{\beta}} = G^{-1}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\delta}}$ (Equation (9)). The same $\hat{\boldsymbol{\delta}}$ is the coefficient vector of the regression of $\hat{\boldsymbol{\beta}}$ on the columns of W without an intercept, since minimizing $\|\hat{\boldsymbol{\beta}} - W\boldsymbol{\delta}\|^2$ gives $\boldsymbol{\delta} = (W^\top W)^{-1}W^\top \hat{\boldsymbol{\beta}} = G^{-1}\hat{\boldsymbol{\alpha}}$.

A.2 Proof of Proposition 1

Given $\mathcal{S}_{\text{train}}$, the directions are fixed, and $\hat{\alpha}_j = \mathbf{w}_j^\top \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}} = G^{-1}W^\top \hat{\boldsymbol{\beta}}$ are fixed linear maps of $\hat{\boldsymbol{\beta}}$.

(a) Taking expectations conditional on the training half,

$$E\big[\hat{\alpha}_j \mid \mathcal{S}_{\text{train}}\big] = \mathbf{w}_j^\top\big(\boldsymbol{\beta} + \mathbf{b}\big) = \alpha_j + \mathbf{w}_j^\top \mathbf{b},$$
$$E\big[\hat{\boldsymbol{\delta}} \mid \mathcal{S}_{\text{train}}\big] = G^{-1}W^\top\big(\boldsymbol{\beta} + \mathbf{b}\big) = \boldsymbol{\delta} + G^{-1}W^\top \mathbf{b}.$$

With $\mathbf{b} = \mathbf{0}$ and a randomized binary treatment, linearity of the expectation gives $\alpha_j = \mathbf{w}_j^\top E[e(Y_i(1)) - e(Y_i(0))] = E[e(Y_i(1))^\top \mathbf{w}_j - e(Y_i(0))^\top \mathbf{w}_j]$, the effect on the codebook $e(Y)^\top \mathbf{w}_j$, and in the same way $\boldsymbol{\delta}$ is the effect on $G^{-1}W^\top e(Y)$. Finally, $\mathbf{w}_j^\top \mathbf{b} = 0$ exactly when $\mathbf{b}$ is orthogonal to $\mathbf{w}_j$, and, since G is invertible, $G^{-1}W^\top \mathbf{b} = \mathbf{0}$ exactly when $W^\top \mathbf{b} = \mathbf{0}$.

(b) Consistency and asymptotic normality follow from the continuous mapping theorem, because $\mathbf{w}_j^\top$ and $G^{-1}W^\top$ are fixed linear maps given the training half. For the same reason, each leave-one-out estimate is $\hat{\alpha}_{j,(-i)} = \mathbf{w}_j^\top \hat{\boldsymbol{\beta}}_{(-i)}$, so the jackknife of $\hat{\alpha}_j$ is the jackknife of a fixed linear combination of $\hat{\boldsymbol{\beta}}$, and likewise for $\hat{\boldsymbol{\delta}}$. □

A.3 Proof of Proposition 2, and remarks on the shares

Let $P_W = W(W^\top W)^{-1}W^\top = WG^{-1}W^\top$. It is symmetric, because G is, and idempotent, because $P_W^2 = WG^{-1}(W^\top W)G^{-1}W^\top = WG^{-1}W^\top = P_W$. So P_W is the orthogonal projection matrix onto the span of the decoder directions: it leaves every vector in the span unchanged ($P_W W = W$), and it maps every vector orthogonal to the span to zero.

Part (a). Since $\boldsymbol{\delta} = G^{-1}W^\top \boldsymbol{\beta}$, the fitted combination of directions is $W\boldsymbol{\delta} = P_W\boldsymbol{\beta}$, and $\boldsymbol{\nu} = \boldsymbol{\beta} - W\boldsymbol{\delta} = (I - P_W)\boldsymbol{\beta}$. Because P_W is symmetric and idempotent, its squared length is

$$\|W\boldsymbol{\delta}\|^2 = \boldsymbol{\beta}^\top P_W^\top P_W \boldsymbol{\beta} = \boldsymbol{\beta}^\top P_W \boldsymbol{\beta} = (W^\top \boldsymbol{\beta})^\top G^{-1}(W^\top \boldsymbol{\beta}) = \boldsymbol{\alpha}^\top G^{-1}\boldsymbol{\alpha},$$

which gives the second form of R^2 in Equation (10). Write $\boldsymbol{\beta} = P_W\boldsymbol{\beta} + (I - P_W)\boldsymbol{\beta}$ and expand the square:

$$\|\boldsymbol{\beta}\|^2 = \|P_W\boldsymbol{\beta}\|^2 + \|(I - P_W)\boldsymbol{\beta}\|^2 + 2\,(P_W\boldsymbol{\beta})^\top (I - P_W)\boldsymbol{\beta}.$$

The cross term vanishes because P_W is symmetric and idempotent,

$$(P_W\boldsymbol{\beta})^\top(I-P_W)\boldsymbol{\beta} = \boldsymbol{\beta}^\top P_W^\top(I-P_W)\boldsymbol{\beta} = \boldsymbol{\beta}^\top(P_W - P_W^2)\boldsymbol{\beta} = 0,$$

so the two parts are orthogonal and, by Pythagoras,

$$\|\boldsymbol{\beta}\|^2 = \|P_W\boldsymbol{\beta}\|^2 + \|(I-P_W)\boldsymbol{\beta}\|^2 = \|W\boldsymbol{\delta}\|^2 + \|\boldsymbol{\nu}\|^2.$$

Moreover, $\boldsymbol{\nu}$ is orthogonal to every direction of W: since P_W is symmetric and $P_W W = W$,

$$W^\top\boldsymbol{\nu} = W^\top(I-P_W)\boldsymbol{\beta} = W^\top\boldsymbol{\beta} - (P_W W)^\top\boldsymbol{\beta} = W^\top\boldsymbol{\beta} - W^\top\boldsymbol{\beta} = \mathbf{0},$$

so $\mathbf{w}_j^\top\boldsymbol{\nu} = 0$ for all j. Both terms in the Pythagorean identity are non-negative, so $R^2 = \|W\boldsymbol{\delta}\|^2/\|\boldsymbol{\beta}\|^2 \in [0,1]$. It is the uncentered R^2 of the regression of $\boldsymbol{\beta}$ on the columns of W without an intercept, with the D coordinates as observations. $(1-R^2)\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\nu}\|^2$ is the squared length of the part of $\boldsymbol{\beta}$ orthogonal to all directions, the part that no combination of the directions describes. $R^2 = 1$ exactly when $\boldsymbol{\nu} = \mathbf{0}$, that is, when $\boldsymbol{\beta}$ lies in the span of the directions.

Part (b). Since $G\boldsymbol{\delta} = \boldsymbol{\alpha}$,

$$\|W\boldsymbol{\delta}\|^2 = \boldsymbol{\delta}^\top G\boldsymbol{\delta} = \boldsymbol{\delta}^\top\boldsymbol{\alpha} = \sum_{j=1}^{M}\delta_j\alpha_j,$$

so the shares $q_j = \delta_j\alpha_j/\|\boldsymbol{\beta}\|^2$ sum to R^2. Written out, the squared length of the described part consists of an own part for each feature and an overlap term for each pair of directions,

$$\boldsymbol{\delta}^\top G\boldsymbol{\delta} = \sum_j \delta_j^2 + \sum_{j\neq k} g_{jk}\,\delta_j\delta_k,$$

where each pair appears twice in the second sum, so its overlap term is $2g_{jk}\delta_j\delta_k$. Writing $\alpha_j = \sum_k g_{jk}\delta_k$ with $g_{jj} = 1$ gives $q_j\|\boldsymbol{\beta}\|^2 = \delta_j\alpha_j = \delta_j^2 + \sum_{k\neq j} g_{jk}\delta_j\delta_k$, the form in Equation (11). Since G is symmetric, the two features of a pair receive the same $g_{jk}\delta_j\delta_k$, half of the pair's overlap term, so every part of $\|W\boldsymbol{\delta}\|^2$ is assigned exactly once. If $G = I$, there are no overlap terms, $\boldsymbol{\delta} = \boldsymbol{\alpha}$, and $q_j = \alpha_j^2/\|\boldsymbol{\beta}\|^2 = r_j$.

Part (c). The shares are continuous functions of $\hat{\boldsymbol{\beta}}$ wherever $\boldsymbol{\beta} \neq \mathbf{0}$, so the continuous mapping theorem applies. The bias corrections subtract estimated sampling variances, which vanish as the test half grows, so the corrected shares have the same limits. □

Where the overlap enters. Since $\boldsymbol{\alpha} = G\boldsymbol{\delta}$, the summed simple shares and the joint share differ by

$$\Big(\sum_j r_j - R^2\Big)\|\boldsymbol{\beta}\|^2 = \boldsymbol{\delta}^\top G^2\boldsymbol{\delta} - \boldsymbol{\delta}^\top G\boldsymbol{\delta} = \boldsymbol{\delta}^\top G(G-I)\boldsymbol{\delta},$$

which is zero for orthogonal directions. With two directions at cosine g_{12}, it equals $g_{12}^2(\delta_1^2 + \delta_2^2) + 2g_{12}\delta_1\delta_2$: the sum overstates the joint share whenever the overlap term $2g_{12}\delta_1\delta_2$ is non-negative, and can understate it only when that term is negative and large.

Negative shares. An overlap term is negative when two directions overlap but the effect moves towards one and away from the other. If feature j's half of such terms outweighs its own part, $q_j < 0$, which happens exactly when δ_j and α_j have opposite signs, and another feature's share can then exceed one. For example, assume $\boldsymbol{\beta} = (0.4, 0.8)$ with $\|\boldsymbol{\beta}\|^2 = 0.8$, $\mathbf{w}_1 = (1, 0)$ and $\mathbf{w}_2 = (0.6, 0.8)$. Then, $\boldsymbol{\alpha} = (0.40, 0.88)$ and $\boldsymbol{\delta} = (-0.2, 1)$. The own parts are 0.05 and 1.25, the overlap term is $2 \cdot 0.6 \cdot (-0.2) \cdot 1/0.8 = -0.30$, and $q = (-0.10, 1.10)$, which sums to $R^2 = 1$. The simple projection shows the texts moving towards $\mathbf{w}_1$ ($\alpha_1 = 0.40$) only because $\mathbf{w}_1$ overlaps with $\mathbf{w}_2$, along which they move strongly; with $\mathbf{w}_2$ held fixed they move slightly away from $\mathbf{w}_1$ ($\delta_1 = -0.2$). The simple shares $r = (0.20, 0.97)$ sum to 1.17 and count the overlap twice. Shapley values offer an alternative split of R^2 in which no share is negative.

The simple shares r_j. With a single direction, $W = \mathbf{w}_j$, $G = 1$, and $P_j = \mathbf{w}_j\mathbf{w}_j^\top$, so part (a) gives $r_j = \alpha_j^2/\|\boldsymbol{\beta}\|^2 \in [0, 1]$, the R^2 of the regression of $\boldsymbol{\beta}$ on $\mathbf{w}_j$ without an intercept. Since $\alpha_j = \|\boldsymbol{\beta}\| \cos\theta_j$, also $r_j = \cos^2\theta_j$. Summed over features, $\sum_j r_j = \boldsymbol{\alpha}^\top\boldsymbol{\alpha}/\|\boldsymbol{\beta}\|^2$, which equals $R^2 = \boldsymbol{\alpha}^\top G^{-1}\boldsymbol{\alpha}/\|\boldsymbol{\beta}\|^2$ when $G = I$ but not otherwise. The sum is bounded by the largest eigenvalue of G, $\sum_j r_j \le \lambda_{\max}(G)$, which exceeds one for overlapping directions.

A.4 The effect on the activations z_{ij}

This subsection relates the projections to the effect on the SAE's activations, $\hat{\gamma}_j$ (Equation (12)).

The SAE model as regression. Write $\boldsymbol{\xi}_i = \mathbf{e}_i - \hat{\mathbf{e}}_i$ for the part of text i's embedding that the SAE does not reconstruct. By Equation (5), every embedding is then exactly

$$\mathbf{e}_i = \mathbf{b}_{\text{pre}} + \sum_{k=1}^{M} z_{ik}\,\mathbf{w}_k + \boldsymbol{\xi}_i,$$

and its score along direction j splits into four parts,

$$s_{ij} = \mathbf{w}_j^\top\mathbf{e}_i = \underbrace{\mathbf{w}_j^\top\mathbf{b}_{\text{pre}}}_{\text{same for all texts}} + \underbrace{z_{ij}}_{\text{own activation}} + \underbrace{\sum_{k\neq j} g_{jk}\,z_{ik}}_{\text{overlapping features}} + \underbrace{\mathbf{w}_j^\top\boldsymbol{\xi}_i}_{\text{not reconstructed}},$$

because $\mathbf{w}_j^\top\mathbf{w}_j = 1$ and $\mathbf{w}_j^\top\mathbf{w}_k = g_{jk}$. Since $z_{ij} \ge 0$, a text's own activation can only raise its score. A text scores low along $\mathbf{w}_j$ only through other features whose directions point away from $\mathbf{w}_j$ ($g_{jk} < 0$) or through the part the SAE does not reconstruct, which is why the labels name the top end of a direction (Section 3.5).

The effects as a function of activation and direction. The coefficient of the variable of interest is a weighted sum of the outcomes, $\hat{\boldsymbol{\beta}} = \sum_i \lambda_i \mathbf{e}_i$, where $\boldsymbol{\lambda}^\top$ is the row of $(X^\top X)^{-1} X^\top$ that belongs to the variable of interest, and the same weights give the coefficient for any other outcome; since X contains an intercept or fixed effects that sum to one, $\sum_i \lambda_i = 0$. Applying them to the decomposition of each embedding,

$$\hat{\boldsymbol{\beta}} = \underbrace{\Big(\sum_i \lambda_i\Big)}_{0} \mathbf{b}_{\text{pre}} + \sum_{k=1}^{M} \underbrace{\Big(\sum_i \lambda_i z_{ik}\Big)}_{\hat{\gamma}_k} \mathbf{w}_k + \underbrace{\sum_i \lambda_i \boldsymbol{\xi}_i}_{\hat{\boldsymbol{\rho}}} = W\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\rho}}.$$

The effect on the embeddings is each decoder direction weighted by the effect on its feature's activation, plus the effect on what the SAE does not reconstruct. The identity holds exactly in every sample, and by taking expectations also for the estimands. Projecting it onto the directions gives the simple and the multivariate projection,

$$\hat{\alpha}_j = \mathbf{w}_j^\top \hat{\boldsymbol{\beta}} = \hat{\gamma}_j + \sum_{k \neq j} g_{jk}\, \hat{\gamma}_k + \mathbf{w}_j^\top \hat{\boldsymbol{\rho}}, \qquad \hat{\boldsymbol{\alpha}} = G\hat{\boldsymbol{\gamma}} + W^\top \hat{\boldsymbol{\rho}},$$

$$\hat{\delta}_j = (G^{-1}\hat{\boldsymbol{\alpha}})_j = \hat{\gamma}_j + \sum_{k=1}^{M} \big(G^{-1}\big)_{jk}\, \mathbf{w}_k^\top \hat{\boldsymbol{\rho}}, \qquad \hat{\boldsymbol{\delta}} = G^{-1}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\gamma}} + G^{-1}W^\top \hat{\boldsymbol{\rho}}.$$

The simple projection adds to the effect on feature j's own activation the effects on the overlapping features, weighted by their cosines with $\mathbf{w}_j$, and the effect on the reconstruction error along $\mathbf{w}_j$. In the joint coefficient, G^{-1} cancels the G that brought in the overlapping features, so $\hat{\delta}_j$ keeps only the feature's own activation and the effect on the reconstruction error, attributed to the directions jointly, since $\big(G^{-1}W^\top \hat{\boldsymbol{\rho}}\big)_j$ is the coefficient on $\mathbf{w}_j$ when $\hat{\boldsymbol{\rho}}$ is regressed on all directions.

When the joint coefficient equals the effect on the activation. $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\gamma}}$ whenever $W^\top \hat{\boldsymbol{\rho}} = \mathbf{0}$. This holds if the SAE reconstructs every text exactly, if the reconstruction error does not change with the treatment (then $\hat{\boldsymbol{\rho}} = \mathbf{0}$ even if each $\boldsymbol{\xi}_i$ is large), or if it changes only orthogonally to every direction. Unlike $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\gamma}}$, it does not require orthogonal directions.

The remainder is informative. The reconstruction moves along $\mathbf{w}_j$ only through the activations, so whatever movement of the texts the activations do not follow stays in $\boldsymbol{\xi}_i$. This happens in the two cases of Section 3.4: an activation censored at zero, and a feature that loses its slot under TopK. If more of the movement is missed in one arm than in the other, it shows up in $\hat{\boldsymbol{\rho}}$, and $\big(G^{-1}W^\top \hat{\boldsymbol{\rho}}\big)_j$ adds it back to $\hat{\delta}_j$, while $\hat{\gamma}_j$ misses it. For example, suppose feature j is off in every text of both arms, the directions are orthogonal, and the treatment moves treated texts 0.2 further away from $\mathbf{w}_j$. The activations stay at zero, so $\hat{\gamma}_j = 0$, while the movement stays in the treated texts' reconstruction error, so $\mathbf{w}_j^\top \hat{\boldsymbol{\rho}} = -0.2$ and $\hat{\delta}_j = -0.2$, the movement of the texts. If instead another feature whose direction points away from $\mathbf{w}_j$ switched on more strongly, the SAE would reconstruct part of the movement through that feature's activation. $\hat{\boldsymbol{\delta}}$ would still attribute the movement to $\mathbf{w}_j$, because it

depends on $\hat{\boldsymbol{\beta}}$ alone, and $\hat{\boldsymbol{\rho}}$ would offset the other feature's $\hat{\gamma}_k$.

B Labeling and Fidelity Prompts

Labeling. For each feature j, the labeler receives the prompt below. The training texts are split into three bins of their score s_{ij} (percentiles 0–10, 45–55 and 90–100), and the prompt shows seven randomly drawn texts from each bin, each cut to at most 256 words. The model is GPT-4.1 at temperature zero. Line breaks in the prompts are added for display, and angle brackets mark what is filled in.

```
You are a machine learning researcher studying one direction in a text-embedding
space. Every text has a score on this direction.
Below, texts are grouped by their score, from the lowest scores (Group 1) to the
highest (Group 3). Each group shows randomly chosen texts from its range.
Your goal is to describe what the texts with the highest scores (Group 3) do that
the other texts do not. Use the lower groups to tell which features of Group 3 are
distinctive and which are common to all texts.

GROUP 1 (score percentiles 0-10; 7 of <n> texts shown):
- <text>
- <text>
  [7 texts in all, each cut to at most 256 words]

GROUP 2 (score percentiles 45-55; 7 of <n> texts shown):
- <text>
- <text>
  [7 texts in all, each cut to at most 256 words]

GROUP 3 (score percentiles 90-100; 7 of <n> texts shown):
- <text>
- <text>
  [7 texts in all, each cut to at most 256 words]

Rules for your label:
- Describe a concrete, objective feature of the texts (what they say, argue or
  propose, or how they say it), not an abstract concept.
- The feature must fit most texts in Group 3, not only one or two of them, and be
  rarer in the lower groups. Do not describe something that all groups share.
- Use at most 15 words.

Example labels from other studies:
- "praises the taste of the food"
- "gives specific numbers and dates"

Answer in exactly this format and nothing else:
LABEL: "<label>"
```

Label fidelity. To check whether a label picks out its feature's texts among texts the labeler did not see, a second call reads one text at a time together with the label and

answers YES or NO, following Movva et al. (2025, Appendix C.1) with the bins of s_{ij} in place of their deciles of activation. For each feature, it reads 50 training texts from the top bin and 50 from the two lower bins (25 from each), excluding the texts shown to the labeler, with the same model and temperature. Precision is the share of YES answers that fall on top-bin texts, recall is the share of top-bin texts that receive a YES, and the F-score is their harmonic mean (Table A1).

```
You will read a text and a description of a feature that texts can have.

Feature: "<label>"

Text: "<text>"

Does the text have this feature? Answer with exactly one word: YES or NO.
```

Table A1: Label fidelity, UK House of Commons written questions (47,448 distinct questions tabled in 2023, at most 255 characters each; $M = 20$, $K = 4$). For each feature, a second LLM reads 100 training questions the labeler did not see (50 from the top 10% of scores, 50 from the lower bins) with the label only and answers YES or NO. Precision: share of YES answers in the top bin. Recall: share of top-bin questions with YES.

Feature	Label	Precision	Recall	F
F12	explicitly references previous official documents, reports, or answers in the question text	0.98	1.00	0.99
F17	asks for an assessment or evaluation of adequacy, impact, or implications of a policy or event	0.98	0.98	0.98
F0	references foreign countries, international relations, or UK actions involving other nations	1.00	0.92	0.96
F10	asks what steps are being taken to address or improve a specific issue	0.89	1.00	0.94
F8	requests a policy change, action, or assessment from a government official or body	0.84	0.98	0.91
F4	asks about future plans, policies, or publication of reports or responses	0.80	0.94	0.86
F19	mentions discussions or consultations with specific people or organizations	0.95	0.72	0.82
F18	requests specific quantitative data or statistics about government programs or expenditures	0.97	0.70	0.81
F14	requests quantitative data or statistics about specific issues or departmental activities	0.86	0.74	0.80
F9	requests specific quantitative data or statistics (e.g., numbers, proportions, sizes) about a subject	0.84	0.74	0.79
F16	asks for specific quantitative data, such as counts, proportions, or measurements	0.76	0.78	0.77
F5	mentions specific locations, facilities, or infrastructure in the context of operational or logistical details	0.88	0.58	0.70
F11	requests specific factual information about past events, actions, or statistics	0.56	0.88	0.69
F13	requests quantitative data or statistics about specific outcomes, populations, or timeframes	0.78	0.58	0.67
F6	requests recent or future actions, estimates, or statistics rather than past assessments or consultations	0.50	0.48	0.49
F15	requests specific quantitative data such as costs, numbers, proportions, or sizes	0.61	0.40	0.48
F7	asks for specific quantitative data or timelines about healthcare services or medical issues	0.81	0.34	0.48
F2	requests data or statistics broken down by categories such as region, school type, or condition	0.79	0.30	0.43
F1	requests specific monetary amounts or financial figures	1.00	0.22	0.36
F3	asks about assessments or plans regarding health, social care, or LGBT-related policies and services	0.36	0.16	0.22
	Mean	0.81	0.67	0.71

C Additional Results on Applications

Table A2: Decomposition onto principal components instead of SAE directions, Egami et al. Experiment 3. The first 7 principal components of the training-half embeddings, both ends labeled with the same labeler. Each component is oriented so that treated answers move towards the first end; effect with 95% jackknife interval. The components are orthogonal, so their shares r_k add up to the joint share (0.90).

	Treated answers move towards	... and away from	Effect	r_k
PC2	proposes both prison time and deportation, often emphasizing harsher or longer sentences	expresses support for legal pathways or leniency for non-criminal, non-violent immigration	0.198 [0.172, 0.224]	0.606
PC1	emphasizes prior violent offenses or repeat offending as justification for harsher punishment	argues against imprisonment and proposes deportation as the preferred response	0.134 [0.104, 0.165]	0.280
PC6	requests or considers additional information or circumstances before deciding on punishment or action	states simply that breaking the law or being illegal justifies punishment or deportation	0.017 [-0.001, 0.034]	0.004
PC5	argues that illegal entry alone should not be punished with imprisonment	emphasizes sending the person back to their home country for punishment or justice	0.014 [-0.006, 0.035]	0.003
PC4	explicitly argues against prison time for illegal entry, favoring deportation instead	proposes considering personal circumstances or contributions before deciding on deportation or amnesty	0.011 [-0.011, 0.032]	0.002
PC7	expresses uncertainty or requests more information before making a judgment or recommendation	mentions whether the person has prior offenses or a criminal record when recommending consequences	0.009 [-0.010, 0.028]	0.001
PC3	emphasizes following legal immigration processes or suggests specific legal pathways to citizenship	opposes prison for illegal entry unless there are prior or violent offenses	0.006 [-0.020, 0.031]	0.000

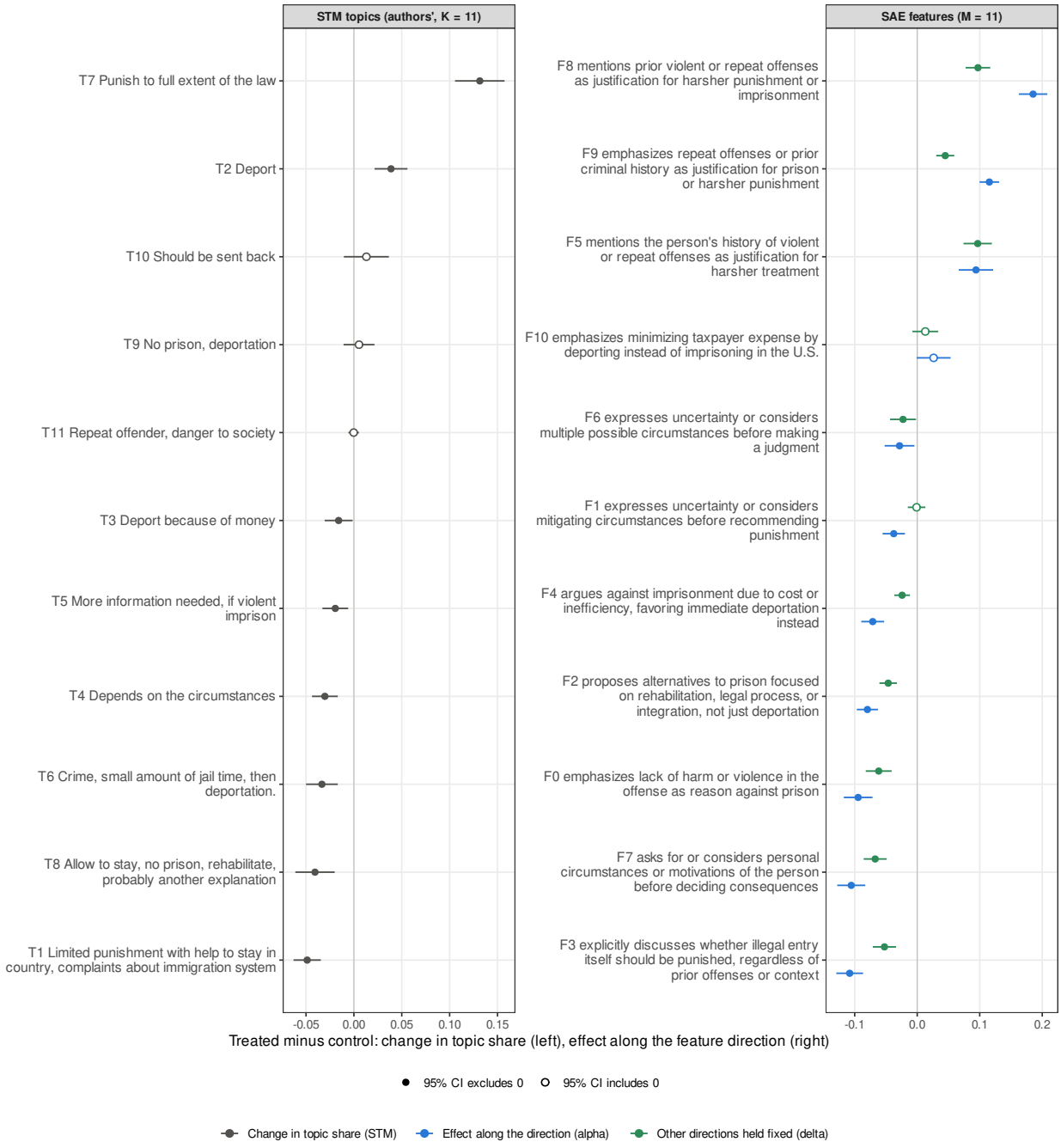


Figure A1: Variation in M. The application of Figure 3 with $M = 11$ (same embeddings and split).

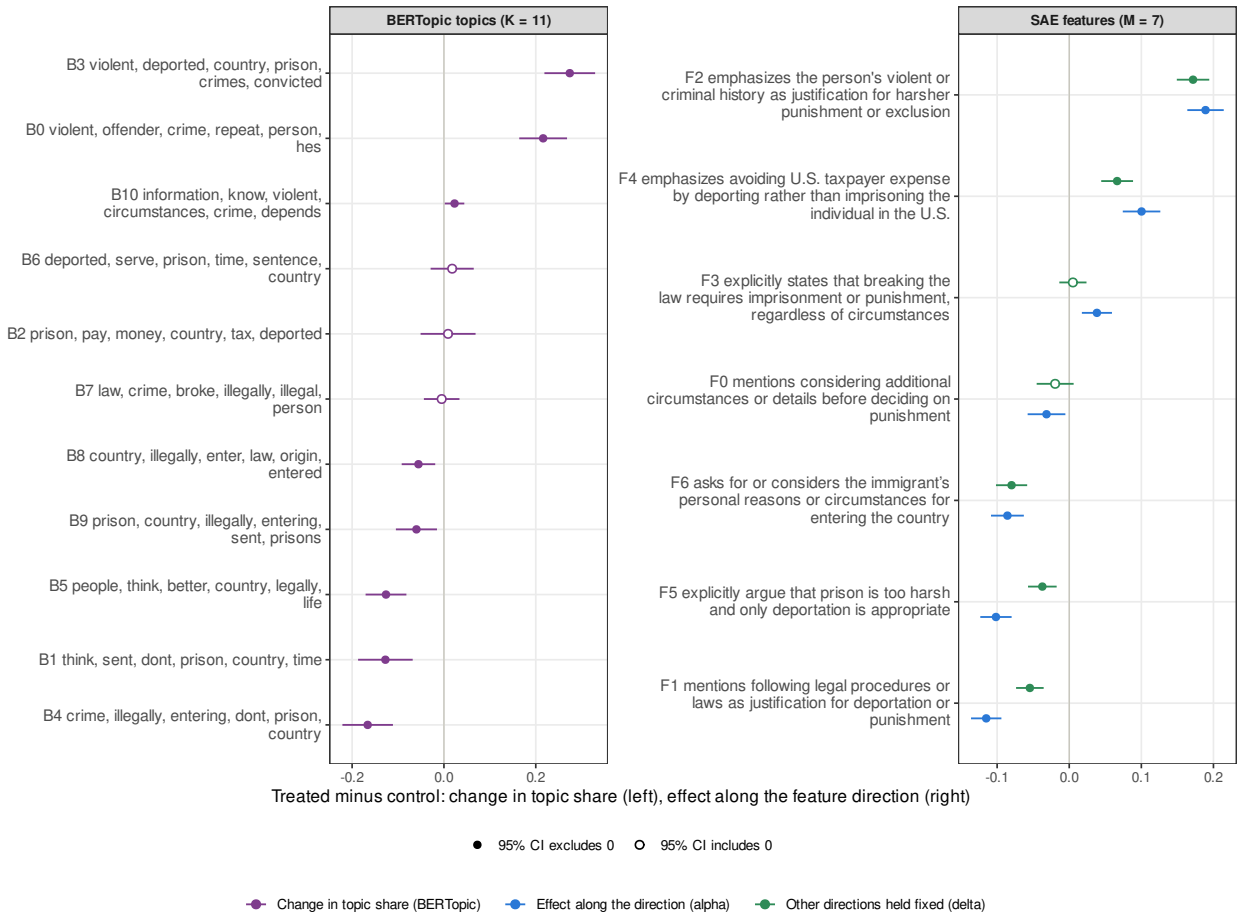


Figure A2: BERTopic in place of the STM, Egami et al. Experiment 3 (test half). Left: the treated minus control share of responses assigned to each BERTopic topic, with 95% intervals for a difference in proportions. BERTopic was fit on the training half with the SAE's embeddings, using UMAP with BERTopic's default settings and, in place of its default HDBSCAN, K-means with 11 topics, as many as the authors' STM, so that no response falls into an outlier topic. The test half is assigned with the fitted model, and each topic is labeled by its six highest-weighted words (c-TF-IDF). Right: the SAE panel of Figure 3.

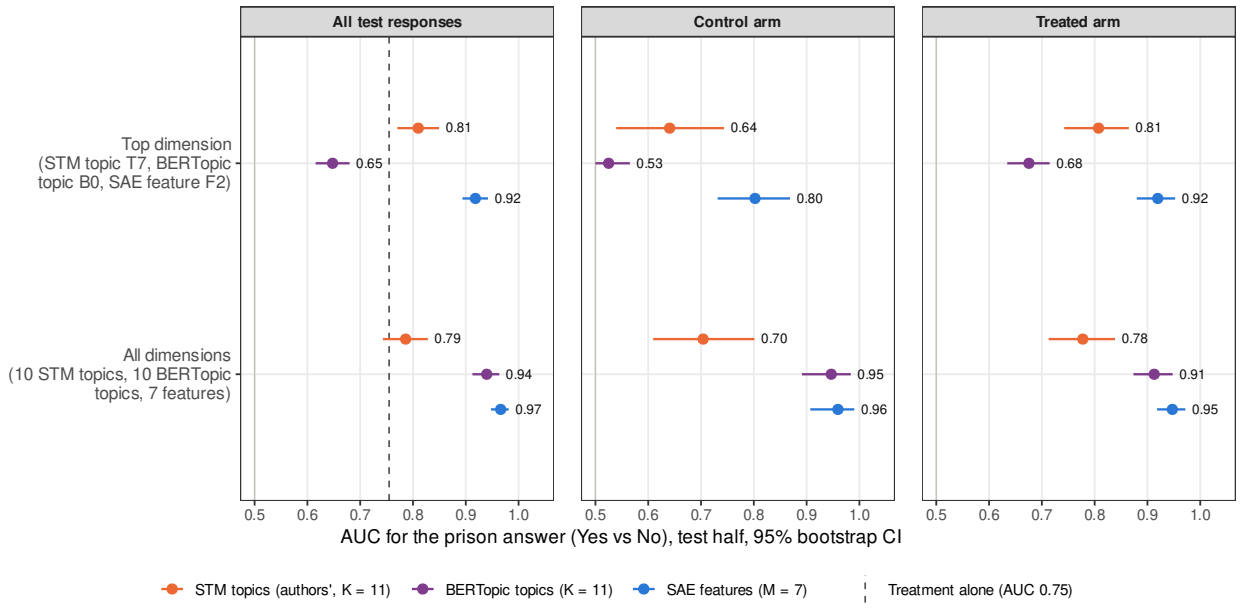


Figure A3: The prison-answer check of Figure 5 with BERTopic. A response's BERTopic topic enters as indicators: all topics but one, or the single topic with the largest effect on the training half. Weights are learned on the training half and the AUC is computed on the test half; 95% bootstrap intervals.

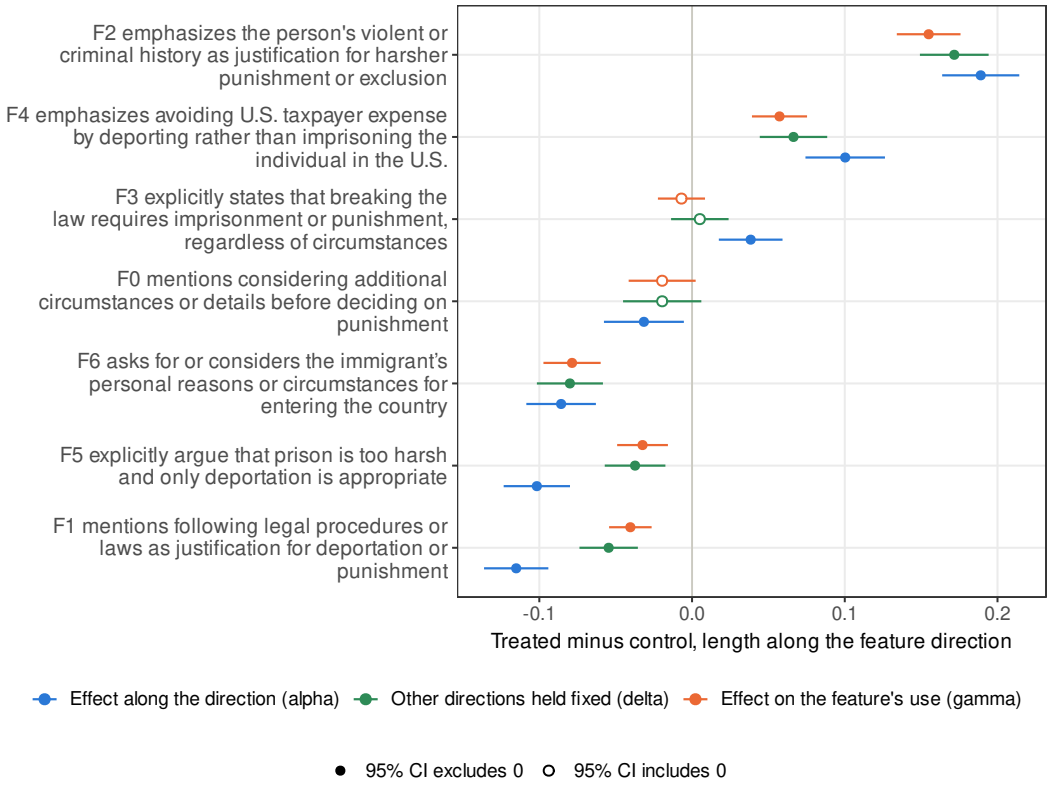


Figure A4: Three readings of each feature's effect, Egami et al. Experiment 3. The simple projection $\hat{\alpha}_j$, the joint coefficient $\hat{\delta}_j$ (other directions held fixed) and the effect on the SAE's activations $\hat{\gamma}_j$; 95% jackknife intervals.

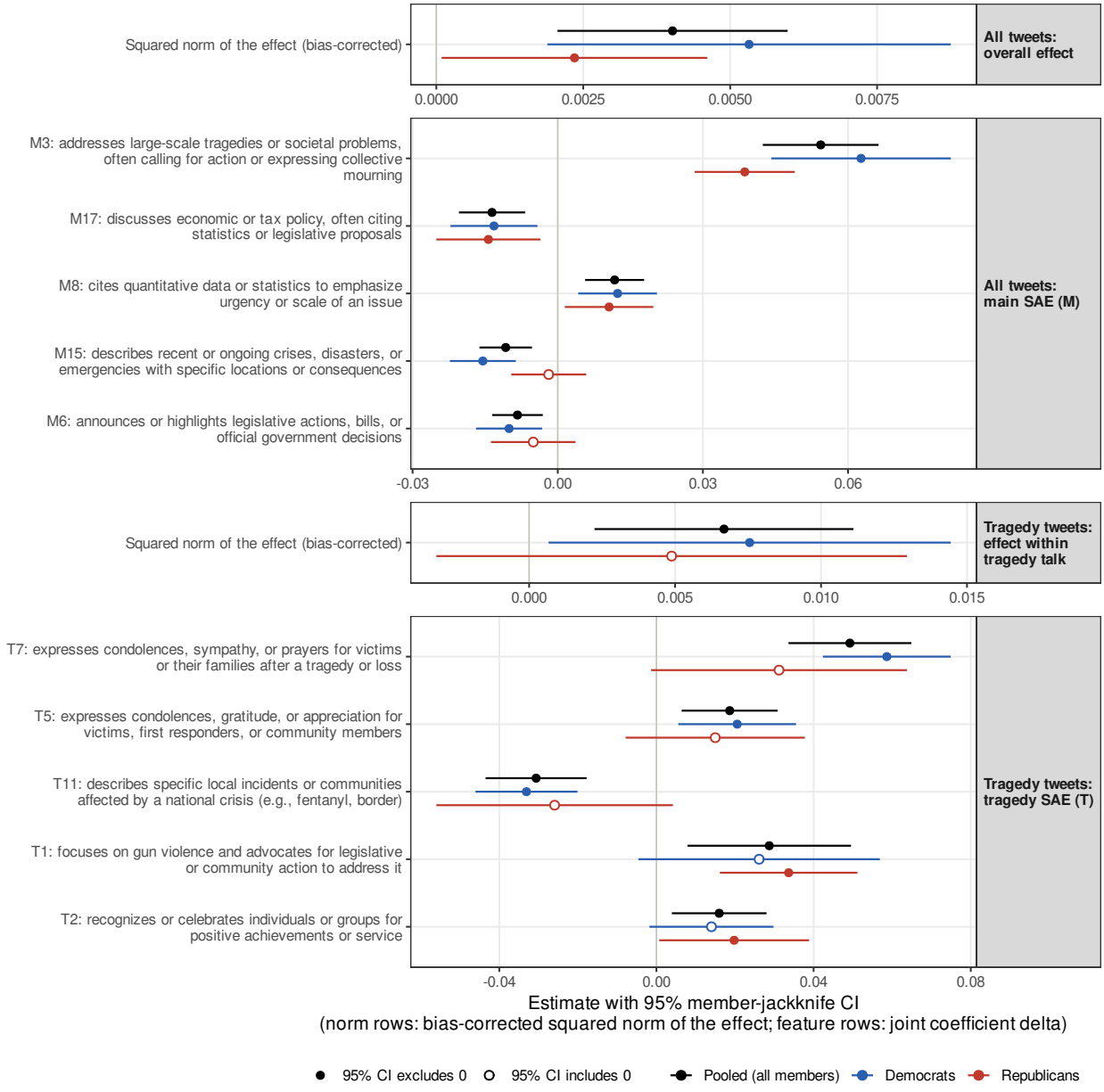


Figure A5: As Figure 6, with day × stack × party fixed effects. Treated members are compared with their party colleagues on the same day, so each party's national reaction to the shooting drops out. The party estimates equal separate regressions on each party's members, and their difference is the triple-difference coefficient. Features as in Figure 6.

Table A3: All features of the two SAEs in the pooled design of Figure 6. Panel A: the main SAE ($M = 20$, $K = 4$), all tweets. Panel B: the tragedy SAE ($M = 12$, $K = 3$), tragedy tweets. Rows are ordered by the bias-corrected share q_j, so the first five rows of each panel are the features shown in Figure 6. $\hat{\delta}_j$ is the joint coefficient, with its 95% leave-one-member-out jackknife interval.

Feature	Label	$\hat{\delta}_j$	95% CI	q_j
A. All tweets, main SAE				
M3	addresses large-scale tragedies or societal problems, often calling for action or expressing collective mourning	0.054	[0.042, 0.067]	0.532
M17	discusses economic or tax policy, often citing statistics or legislative proposals	−0.013	[−0.020, −0.007]	0.048
M8	cites quantitative data or statistics to emphasize urgency or scale of an issue	0.013	[0.007, 0.019]	0.033
M15	describes recent or ongoing crises, disasters, or emergencies with specific locations or consequences	−0.011	[−0.017, −0.006]	0.031
M6	announces or highlights legislative actions, bills, or official government decisions	−0.007	[−0.013, −0.002]	0.023
M7	advocates for expanding or protecting healthcare access or specific health-related policies	−0.011	[−0.017, −0.005]	0.023
M0	criticizes or challenges foreign governments or U.S. national policy decisions	−0.011	[−0.017, −0.005]	0.021
M11	commemorates or honors a group, individual, or historical event or expresses collective remembrance	0.011	[0.004, 0.018]	0.021
M1	states or argues a strong policy position on a controversial national issue	−0.012	[−0.018, −0.005]	0.017
M5	expresses gratitude, congratulations, or invitations to participate in events or celebrations	−0.007	[−0.012, −0.001]	0.014
M10	references the military, veterans, or national security issues	0.006	[0.001, 0.011]	0.011
M16	directly criticizes or blames specific political leaders or officials for negative outcomes	−0.009	[−0.017, −0.001]	0.010
M2	criticizes government actions or policies and calls for specific changes or accountability	−0.010	[−0.016, −0.005]	0.008
M18	comments on or reacts to recent news, hearings, or political events as they happen	−0.005	[−0.011, 0.001]	0.003
M4	announces or invites to live or current events, often specifying time or location	0.005	[0.001, 0.009]	0.002
M13	issues direct advice, alerts, or calls to action to the public	0.003	[−0.004, 0.011]	0.000
M19	mentions holidays, local events, or expresses greetings and well-wishes	0.000	[−0.002, 0.003]	0.000
M12	announces events, opportunities, or services with instructions or details for public participation	0.001	[−0.004, 0.006]	−0.002
M9	directly urges or calls on the reader to take action or get involved	0.001	[−0.004, 0.007]	−0.002
M14	describes attending or participating in local events or meetings in specific places	0.000	[−0.006, 0.005]	−0.002
B. Tragedy tweets, tragedy SAE				
T7	expresses condolences, sympathy, or prayers for victims or their families after a tragedy or loss	0.054	[0.038, 0.069]	0.418
T5	expresses condolences, gratitude, or appreciation for victims, first responders, or community members	0.017	[0.004, 0.030]	0.086
T11	describes specific local incidents or communities affected by a national crisis (e.g., fentanyl, border)	−0.030	[−0.043, −0.016]	0.082
T1	focuses on gun violence and advocates for legislative or community action to address it	0.022	[0.000, 0.043]	0.058
T2	recognizes or celebrates individuals or groups for positive achievements or service	0.016	[0.004, 0.028]	0.041
T6	commemorates or honors victims, communities, or individuals in response to tragedy or achievement	0.010	[−0.002, 0.022]	0.031
T9	offers congratulations or well-wishes to individuals or groups for achievements or events	0.011	[−0.002, 0.023]	0.030
T4	criticizes Democrats, liberals, or Biden by name or refers to their actions negatively	−0.025	[−0.037, −0.013]	0.025
T0	mentions introducing, passing, or supporting specific legislation or government action	−0.018	[−0.035, −0.002]	0.017
T3	mentions specific individuals or groups by name or title	−0.014	[−0.024, −0.004]	0.004
T8	focuses on children or students as primary subjects or beneficiaries	0.000	[−0.012, 0.011]	−0.005
T10	expresses condolences or honors individuals for their service, sacrifice, or passing	−0.009	[−0.020, 0.002]	−0.021

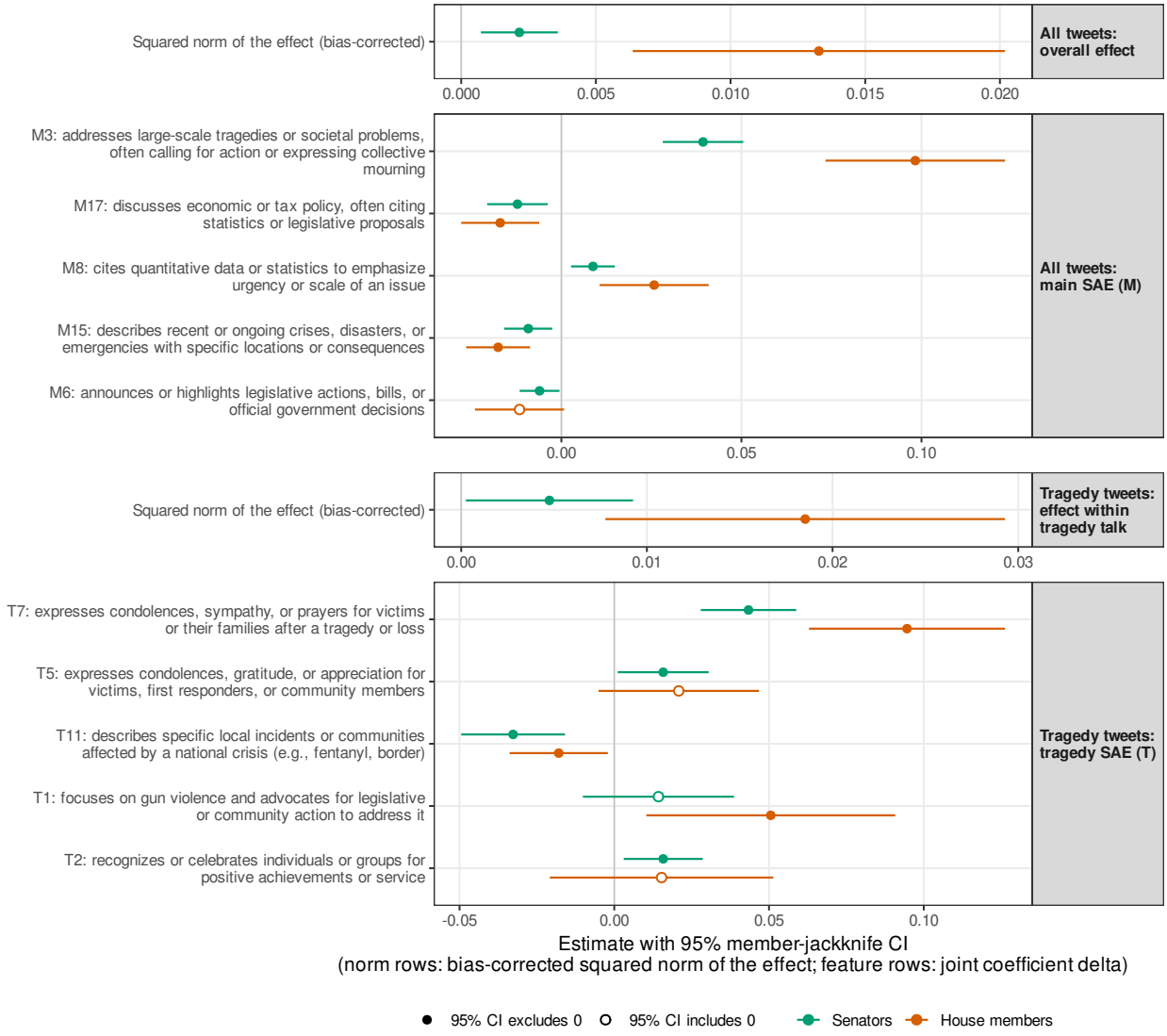


Figure A6: As the party estimates of Figure 6, for senators and House members. One regression with treatment × post × chamber against the common day effects. The House member is treated by a shooting in the district, the senators by a shooting anywhere in the state. Features as in Figure 6.

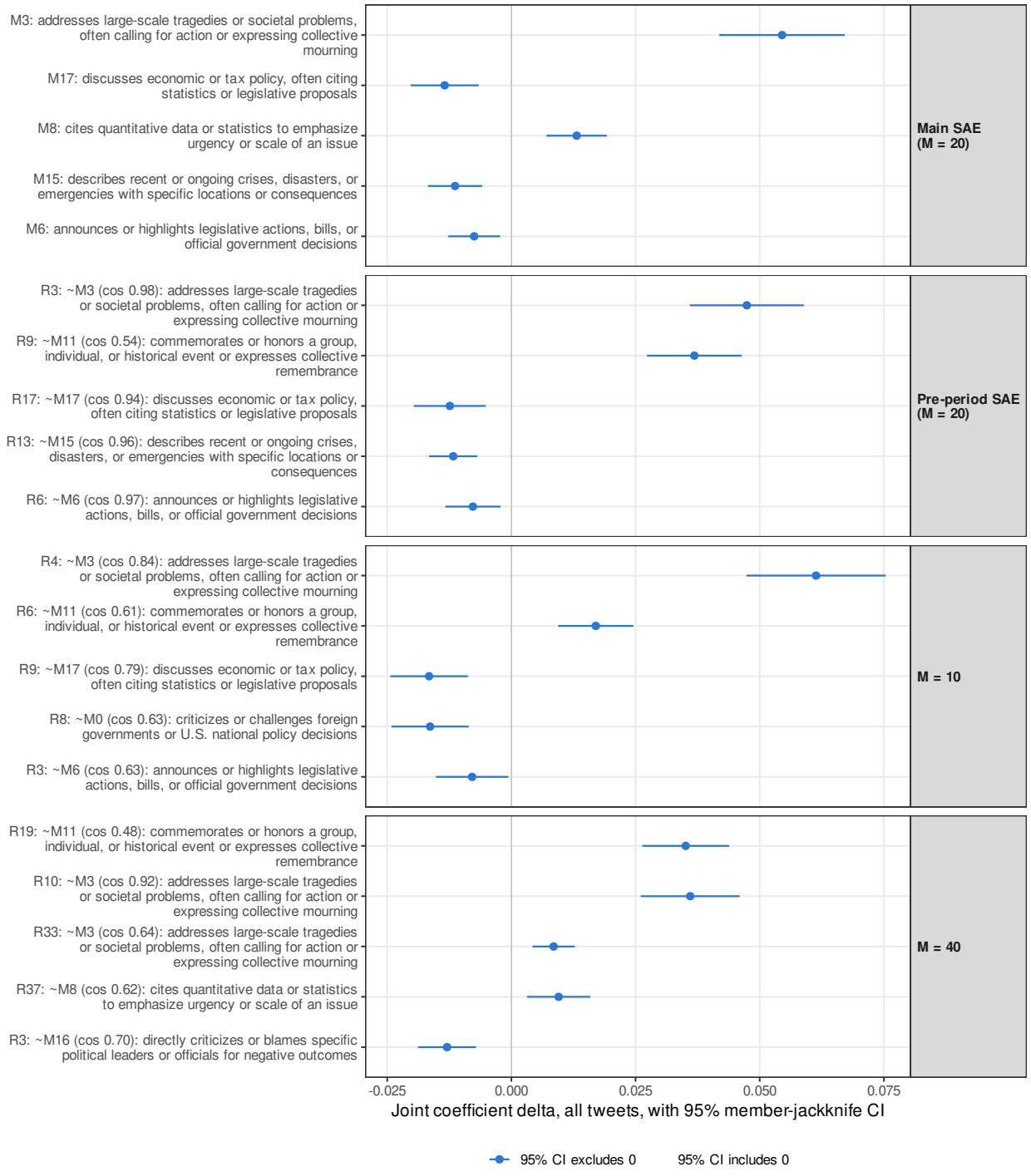


Figure A7: The all-tweets effect of Figure 6 on other SAEs. The main SAE ($M = 20$, $K = 4$), an SAE trained only on tweets from the two weeks before each shooting ($M = 20$, $K = 4$), and SAEs with $M = 10$ and $M = 40$ ($K = 4$), all trained on members left out of the estimate. Each row shows the five features with the largest bias-corrected shares q_j. The other SAEs are not labeled; their features (R) are named by the closest main-SAE feature and the cosine between the two decoder directions. Joint coefficients $\hat{\delta}_j$ with 95% member-jackknife intervals.

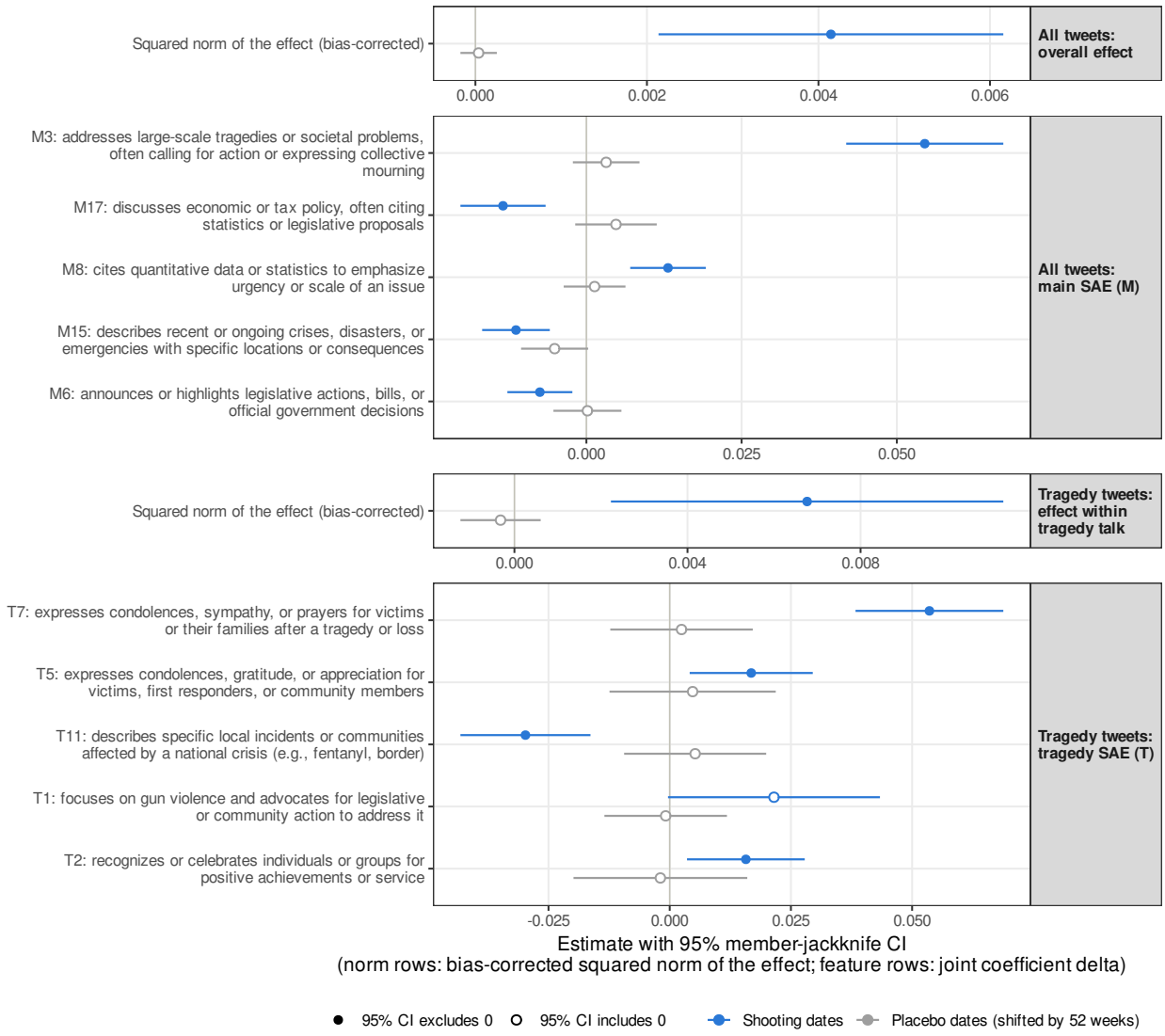


Figure A8: Placebo. The design of Figure 6 on windows shifted by 52 weeks (forward where the shifted window would start before the archive; placebo windows that overlap a shooting in the same state dropped; 42 placebo shootings). The members holding the shooting's House and Senate seats on the placebo date are treated. Gray: placebo dates (global tests $p = 0.31$ for all tweets and 0.55 for tragedy tweets); blue: the estimates of Figure 6. Features as in Figure 6; 95% member-jackknife intervals.

D Additional Results on Validations

Human-label design (blind sheet, scoring rule)

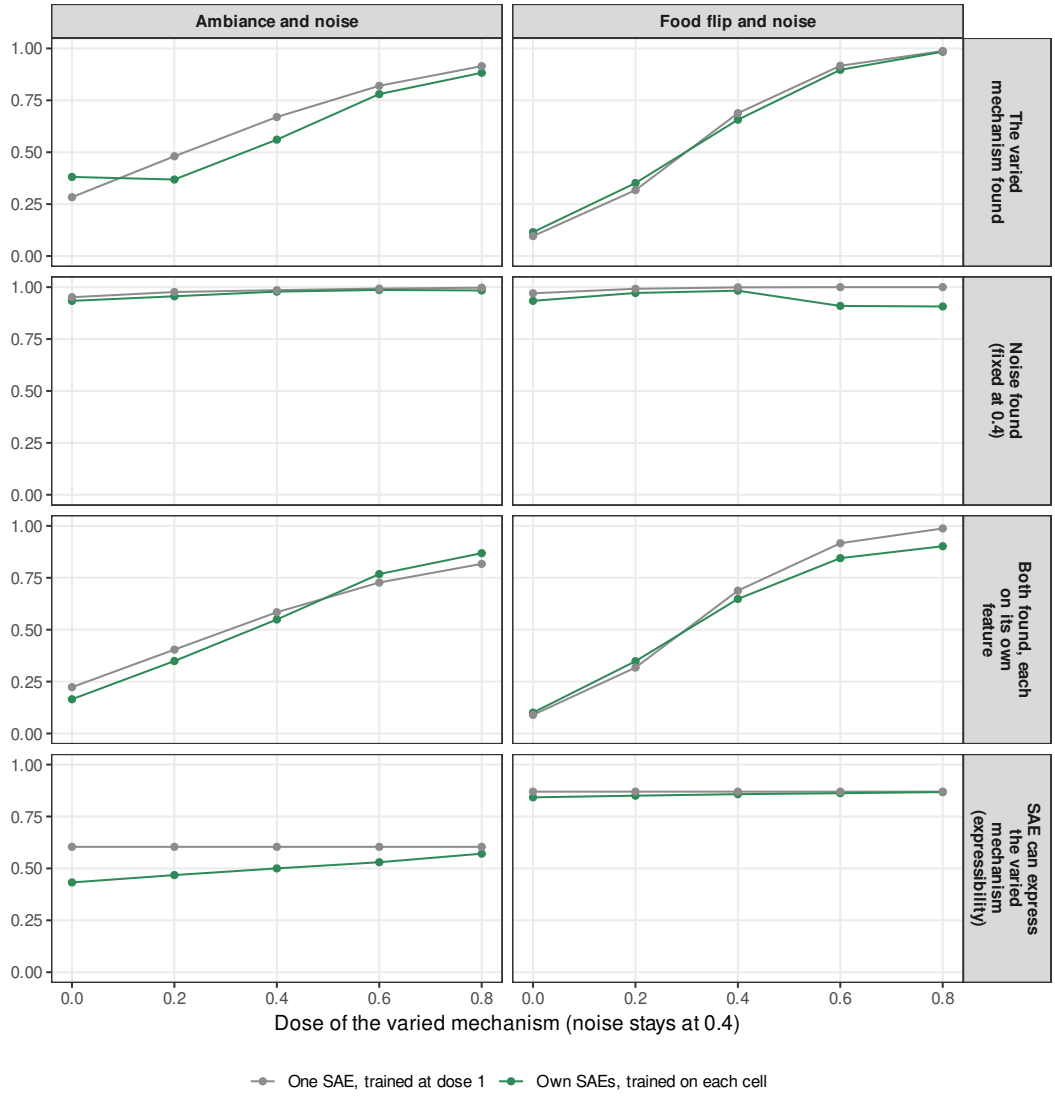


Figure A9: Selectivity with each condition's own SAE, CEBaB ground truth (2,000 texts, $M = 10$, $K = 4$). As in Figure 10, noise is held at a dose of 0.4 while the second mechanism's dose rises. Gray: one SAE, trained at a dose of one. Green: SAEs trained on each condition's own assignment, noise at 0.4 and the second mechanism at its dose. Rows: how often the second mechanism and noise are found (main feature with a significant $\hat{\delta}_j$), how often both are found on features of their own, and the share of the second mechanism's effect the SAE can express.

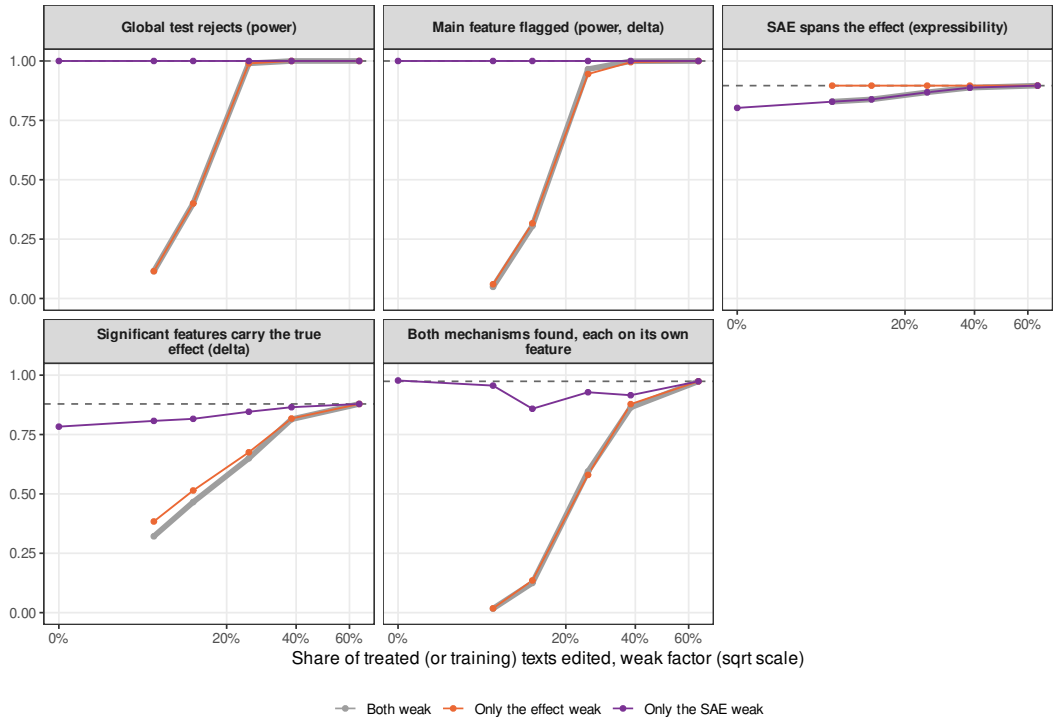


Figure A10: Weak SAE against weak effect, CEBaB ground truth, turning praise of the food into a complaint and adding a complaint about noise (2,000 texts, $M = 10$, $K = 4$). Both weak: the SAE's training texts and the experiment have the same share of edited texts. Only the effect weak: the SAE is trained at a dose of one. Only the SAE weak: the experiment is at a dose of one and the SAE is trained at a lower dose (zero means no edited text). Dashed lines: both strong. Horizontal axis: share of treated (or training) texts edited for the weak factor, square-root scale.

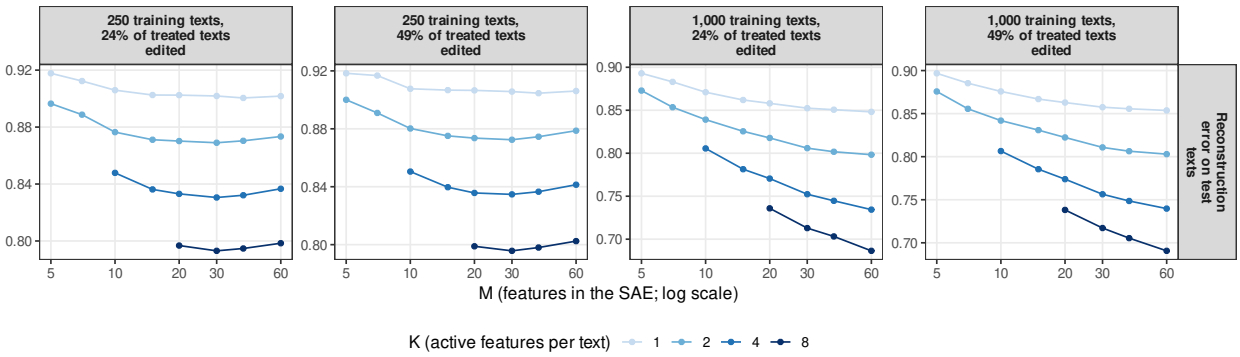


Figure A11: Reconstruction error on the test texts, for the SAEs of Figure 12. The SAE's error on the experiments' test samples, measured as the fraction of variance unexplained (the mean squared error divided by the texts' variance). A researcher can compute it without knowing the effect. Each panel has its own vertical axis.

E Data Sources

Experiment. The answers of Experiment 3 in Egami et al. (2022) come from their replication data.

Mass shootings. The shootings come from the Mother Jones database (Follman et al., 2026), in its version of 2 August 2026, and the tweets from the congresstweets archive (Litel, 2026). Members' seats and parties on each event date come from the congress-legislators data (@unitedstates Project, 2026a). Shootings are assigned to House districts with the Census Bureau's cartographic boundary files for the 115th, 116th and 118th Congresses (U.S. Census Bureau, 2026). For the 117th Congress, we use the 116th districts with North Carolina's 2020 map (@unitedstates Project, 2026b), the only state that redrew its districts between the two. Mother Jones gives no coordinates for the 18 shootings from May 2022 on. We locate 16 of them with the Census geocoder, one (Michigan State University) with OpenStreetMap (© OpenStreetMap contributors, ODbL) and one (Half Moon Bay) from the English Wikipedia article on the shooting (CC BY-SA).

Ground truth. The restaurant reviews and their edits are the CEBaB data (Abraham et al., 2022). The original reviews come from OpenTable.

Labels. The written questions come from the UK Parliament's Written Questions API (UK Parliament, 2026), downloaded on 6 October 2026. Contains Parliamentary information licensed under the Open Parliament Licence v3.0.