SUPPLEMENTARY MATERIAL FOR "THE MECHANISTIC FIXED-EFFECTS TOPIC MODEL: RECOVERING CORPORATE CULTURE FROM EMPLOYEE REVIEWS"

BY DAN BIDERMAN[1,a], DAVID BLEI[2,b], WEI CAI[2,c], ANDREA CICCARONE[3,e], AMIR FEDER[4,f] AND ANDREA PRAT[2,d]

[1]*Stanford University,* [a]*biderman@stanford.edu*

[2]*Columbia University,* [b]*david.blei@columbia.edu;* [c]*wc2419@columbia.edu;* [d]*andrea.prat@columbia.edu*

[3]*ETH Zurich,* [e]*aciccarone@ethz.ch*

[4]*The Hebrew University of Jerusalem,* [f]*afeder@cs.huji.ac.il*

This supplement supplies implementation details and additional analyses for the main paper.

APPENDIX A: ADDITIONAL INTRODUCTION TABLES

This appendix documents the three steps behind the introductory Amazon–Aflac example. Table A1 reports the reviews' leading topics under the historical bag-of-words models. These historical estimates motivate the use of fixed effects but are not directly comparable to the current $K = 50$ models.

TABLE A1
Leading Topic Shares under Historical Bag-of-Words Models

Review	Model	Top Words	Share
Amazon	**VTM**	*managers, industry, products, communication, organization, awesome, highly, cost, provide*	0.24
		told, generous, firm, executives, timings, left, directors, saying, defined	0.15
		fun, days, companies, flexibility, associates, plenty, clients, daily, treat	0.08
		lack, home, cons, family, bonus, issues, processes, cool, quickly	0.08
	FETM	*overall, looking, free, left, chance, workplace, reason, lunch, food*	0.38
		focus, major, follow, groups, encouraged, teams, feeling, executives, credit	0.12
		based, department, job, given, performance, managers, supervisors, truly, help	0.12
		jobs, customer, person, issues, problems, treat, hiring, possible, drive	0.08
Aflac	**VTM**	*training, high, changes, performance, stressful, constant, turnover, layoffs, lunch*	0.16
		poor, decent, free, workers, insurance, shift, skills, world, helpful	0.12
		told, generous, firm, executives, timings, left, directors, saying, defined	0.08
		bad, focus, pressure, especially, discounts, ibm, services, phone, pays	0.07
	FETM	*overall, looking, free, left, chance, workplace, reason, lunch, food*	0.29
		jobs, customer, person, issues, problems, treat, hiring, possible, drive	0.14
		opportunities, advancement, career, growth, development, professional, willing, technology, systems	0.10
		learning, diversity, stock, takes, constantly, options, promote, tools, point	0.07

Notes. Historical $K = 30$ bag-of-words illustration archived in the earlier FETM draft. VTM denotes the historical vanilla topic model and FETM its fixed-effects extension; neither is the current mechanistic mVTM or mFETM. These estimates are retained to document the motivating example and are not directly comparable to the current $K = 50$ results.

Table A2 contrasts the historical word representation with the current bag of concepts. Rows are paired by rank rather than aligned as latent objects.

TABLE A2
From Words to Concepts in the Introductory Example

Bag of words: top words	Bag of concepts: current topic
Amazon	
overall, looking, free, left, chance, workplace, reason, lunch	Warehouse Operations and Equipment Handling: technical terms related to online support services; phrases and terms related to automated processes and claims management; terms related to home appliance service and repair
focus, major, follow, groups, encouraged, teams, feeling, executives	Attendance and time-off policies: terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); phrases indicating tasks and responsibilities; mentions of job performance and employment situations
based, department, job, given, performance, managers, supervisors, truly	Wages and Working Hours: terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); mentions of job performance and employment situations; quantitative data and time metrics related to events or activities
Aflac	
overall, looking, free, left, chance, workplace, reason, lunch	Attendance and time-off policies: terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); phrases indicating tasks and responsibilities; mentions of job performance and employment situations
jobs, customer, person, issues, problems, treat, hiring, possible	Warehouse Operations and Equipment Handling: technical terms related to online support services; phrases and terms related to automated processes and claims management; terms related to home appliance service and repair
opportunities, advancement, career, growth, development, professional, willing, technology	Variable work schedules and hours: mentions of scheduling appointments and time management; terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); mentions of job performance and employment situations

Notes. For each review, the rows are its three leading topics, paired by rank for exposition rather than aligned as latent objects. Left: top words from the historical bag-of-words FETM in Table A1. Right: the current mFETM topic name followed by its three highest-ranked retained concept descriptions. The current mFETM estimates give the two reviews different leading topics.

Table A3 displays the concept side of the naming procedure, which also uses representative reviews. Names label estimated topics and do not enter estimation.

TABLE A3
From Concepts to Names: the Interpretation Step

Top concepts	Topic name
technical terms related to online support services; phrases and terms related to automated processes and claims management; terms related to home appliance service and repair; references to order-picking machinery and their operational details	Warehouse Operations and Equipment Handling
terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); phrases indicating tasks and responsibilities; mentions of job performance and employment situations; references to past events or experiences	Attendance and time-off policies
terms related to employee rights and leaves of absence under the Family Medical Leave Act (FMLA); mentions of job performance and employment situations; quantitative data and time metrics related to events or activities; various forms of currency and mentions of monetary values	Wages and Working Hours

Notes. Three topics from the example. Left: the highest-weighted concepts under $\mathrm{softmax}(\beta_k^\star)$, given by the features' automatic descriptions. Right: the generated name, which also uses representative high-share reviews as described in Section D. Names label the output; estimation never uses them.

APPENDIX B: ESTIMATED TOPIC NAMES

Table A4 lists the language-model-generated names of all $K = 50$ topics. Candidate concepts are drawn from $\mathrm{softmax}(\beta_k^\star)$ for mFETM and $\mathrm{softmax}(\beta_k)$ for mVTM; both models then use the same refinement and review-grounding procedure described in Section D. Topic numbers are model-specific, so rows are not aligned across columns.

TABLE A4
Estimated topic names at $K = 50$.

	mFETM	mVTM
1	Employee compensation and career development	Employee benefits and career growth opportunities
2	Abusive management and poor working conditions	Poor management practices
3	Enterprise project management and integration	Supportive management and coworkers
4	Attendance and time-off policies	Employee autonomy and growth opportunities
5	Management care for employees	Employee food and beverage benefits
6	Leadership and organizational culture	Enterprise technology strategy and innovation
7	Entry-level hiring and career opportunities	Positive workplace experience
8	Workplace Job Experience Evaluation	Overwork and workplace stress
9	Supportive Coworker Relationships	Workplace management and support
10	Work-life balance and flexibility	Flexible work arrangements and remote work
11	Workplace management fairness and transparency	Employee compensation and career advancement
12	Employee discounts and sales incentives	Manual material handling strain
13	Organizational change adaptability	Workplace communication and cooperation
14	Warehouse Operations and Equipment Handling	Call center workload and performance monitoring
15	Employee perks and workplace amenities	Market downturns and workforce reductions
16	Competitive pay and steady work hours	Positive fast-paced work environment
17	Variable work schedules and hours	Employee salary competitiveness
18	Positive workplace culture and environment	Workplace culture and employee experience
19	Wages and Working Hours	Employee well-being and safety culture
20	Workplace environment and culture	Company reputation and industry leadership
21	Retail staffing and management conditions	Workload and management practices
22	Organizational values alignment and integrity	Organizational culture and leadership development
23	Unethical management and unfair advancement	Politicized, exclusionary workplace culture
24	Employee benefits packages	Organizational culture and values
25	Workplace workload and work-life balance	Forced-ranking performance evaluations
26	Employee compensation and benefits	Employee compensation and workload conditions
27	Company market positioning and competitiveness	Entry-level hiring and advancement criteria
28	Coworker quality and collaboration	Abusive and retaliatory management culture
29	Organizational culture and management practices	Organizational bureaucracy and decision-making delays
30	Telecommunications company offerings and employee benefits	Retail staffing and management problems
31	Work Effort and Advancement	Workload and work-life balance
32	Employee pay and salary levels	Employee benefits package
33	Career advancement opportunities	Career advancement opportunities
34	Career advancement opportunities	Job security and advancement pressure
35	Coworker friendliness and collegiality	High-pressure sales quotas and product pushing
36	Positive employee experience and workplace relationships	Variable work schedules and limited flexibility
37	Shareholder-driven restructuring and layoffs	Workplace fairness and advancement opportunities
38	Workplace location and physical facilities	Absence of reported workplace drawbacks
39	Workplace management quality and conditions	Positive coworker relationships and caliber
40	Workplace management and organizational culture	Low pay and compensation competitiveness
41	Employee compensation and benefits	Workplace stress and workload pressure
42	Positive workplace sentiment	Workplace location and commuting conditions
43	Workplace location and relocation tradeoffs	Workplace compensation and scheduling conditions
44	Industry downturn and market volatility	Employee training and skill development
45	Performance evaluation and career advancement	Management quality and leadership effectiveness
46	Fast-paced workplace culture	Abundance of workplace opportunities
47	Poor management and unfair treatment	Positive workplace culture
48	Poor management and workplace support	Pay, hours, and scheduling conditions
49	No reported workplace issues	Workplace facilities and amenities
50	Employee training and professional development	Compensation and senior management priorities

APPENDIX C: BAG-OF-CONCEPTS CONSTRUCTION DETAILS

Each document is truncated at 256 tokens and passed through the base Gemma-2-9B model. We read the residual stream at layer 39 (3,584 dimensions) and encode each token's activation with the width-16,384 JumpReLU autoencoder (Rajamanoharan et al., 2024) from the Gemma Scope suite (Lieberum et al., 2024). The SAE writes each activation as a sparse, nonnegative vector of feature firings, and we keep the nonzero (document, token, feature, magnitude) entries.

Three pruning rules follow:

1. before applying the strength threshold q_v, drop features with any positive activation in fewer than 0.1% of documents, which are too rare to learn from, or more than 90%, which are too common to discriminate;
2. drop features active on more than 1% of tokens in the SAE's training data, excluding ubiquitous low-level features that would otherwise create "syntax topics";
3. show each remaining feature's automatic label to GPT-4o-mini and drop descriptions of programming, mathematics, grammar, text formatting, or stop words (prompt in Section D).

The three rules reduce the dictionary from $16{,}384$ candidate features to $6{,}889$. For each remaining feature, the firing threshold q_v is the 80th percentile of its nonzero activation distribution in the SAE's training data, computed from Neuronpedia's per-feature histogram.[1] The implementation uses the corresponding corpus nonzero-activation percentile when a histogram is unavailable, although every feature in the final run had a histogram. After thresholding, we again require positive counts in at least 0.1% of documents, leaving $4{,}735$ features.

The supplied Neuronpedia feature labels describe activations in general web text, which may not match their meaning in workplace reviews; we note the divergences where they matter. The resulting count matrix is $906{,}938 \times 4{,}735$ and about 2.4% dense. Both models use this same matrix, fit/validation/test partition, topic count, and encoder architecture; their separately selected optimization schedules are reported in the main text.

APPENDIX D: LLM PROMPTS AND DETAILS

Three steps use a language model: filtering spurious SAE features (Main-text section 2.3), removing out-of-place features from a topic's representation, and naming topics (Main-text section 4.1). We use fixed prompts and model versions and hold the resulting outputs fixed across analyses. This appendix reports the models and substantive prompt instructions.

D.1. Content Filter. We use `gpt-4o-mini-2024-07-18`, the version pinned by Zheng et al. (2025) (Appendix A.1). Zheng et al. do not publish their template; ours mirrors the paper's wording and asks for a binary decision:

```
Below is a textual description of a feature learned by a
sparse autoencoder.
Decide whether this description is about programming, math,
grammar, text formatting, or stop words.
Reply with ONLY 'yes' or 'no'.
Description: {label}
Answer:
```

Features answered "yes" are dropped.

D.2. Topic Refinement. For each mFETM topic, the top $n + m = 12$ features by $\mathrm{softmax}(\beta_k^\star)$ are shown to the model; mVTM uses $\mathrm{softmax}(\beta_k)$. The model flags up to $m = 2$ features that do not belong, and the surviving top $n = 10$ form the topic's textual representation. We use `gpt-5.6-luna`. Relative to the F.1 prompt of Zheng et al. (2025), we add context explaining that the labels are automated interpretations of SAE features estimated on general web text, so that a label's surface domain is not read literally in workplace reviews:

[1] Neuronpedia, `https://www.neuronpedia.org/gemma-2-9b`, source `39-gemmascope-res-16k`; accessed May 2026.

```
The goal of this task is to evaluate a list of features
produced by an automatic method. We call this list a
"topic". Given a topic, you'll be answering the question:
"Which feature(s) don't belong in this list?" For each
topic, choose the feature(s) whose meaning does not match
with what the list seems to be about.
Reply with a brief reasoning for your choice, and up to two
numbers corresponding to the features that don't belong (or
-1 if there are less than two).
Each feature is an automated interpretability description
(from Neuronpedia) of a sparse-autoencoder feature,
describing the contexts in which it typically activates in
general web text. The documents here are employee reviews
of companies, where the same feature often activates on
workplace analogues of the described context. Judge
belonging by what the features would jointly indicate
within an employee review, not by the surface domain of any
single description. The list is ranked: earlier features
have greater topic weight and should count as stronger
evidence.
Important: Prioritize identifying feature(s) that are oddly
specific and/or clearly out of place. Use your world
knowledge in considering whether a feature belongs. If you
are not sure, do not choose the feature.
```

D.3. Naming. The feature descriptions alone can be misleading because they summarize where an SAE feature activates in general web text, not how it is used in this corpus. We therefore ground the ten surviving labels with representative reviews. For each topic, we rank documents by the deterministic share $\mathrm{softmax}(\mu_{\theta_d})_k$ and scan the 1,000 highest-share documents in descending order until obtaining 20 unique excerpts containing 5–45 words. The naming model receives only the ranked feature descriptions and excerpt text, never firm, industry, polarity, or document identifiers. We use the same selection rule for mFETM and mVTM, and use `gpt-5.6-luna` for naming as well. We then adapt the F.2 prompt of Zheng et al. (2025) as follows:

```
You name latent topics from a topic model fitted to
employee reviews. You receive two ranked evidence sources:
automated sparse-autoencoder feature descriptions and
representative employee-review excerpts with the highest
posterior topic shares. SAE descriptions come from general
web text and may be noisy, polysemantic, or expressed
through unrelated surface domains. Review excerpts are
direct in-corpus evidence, but any single excerpt may
contain incidental details.
Infer the narrowest workplace construct supported by
repeated patterns across the excerpts and broadly
compatible with the features. Prioritize repeated in-corpus
meaning when translating a feature description into its
workplace analogue. Treat earlier items within each list as
stronger evidence.
Rules:
1. Return only a 3-8 word noun phrase, with no preamble or
period.
```

```
2. Do not name restaurants, shopping, products, customers,
family, games, or another surface domain unless that
literal subject recurs in the employee-review excerpts.
3. Preserve genuine industry, occupation, policy, or
sentiment content only when several high-ranked excerpts
support it. Do not name sentiment when it is merely
evaluative language around another workplace construct.
4. Do not invent specificity. If the evidence is diffuse,
use a transparent broad workplace label rather than joining
unrelated fragments.
5. Prefer concrete workplace nouns. Avoid vague filler such
as "elements," "aspects," "various," "references,"
"personal recommendations," or "experiences" unless that
meaning is explicitly central.
6. Name one central construct.
```

Names label the output only; estimation never sees them.

APPENDIX E: INFERENCE ALGORITHM

Algorithm 1 states the mFETM estimation procedure. The likelihood uses reparameterized draws of θ, $\beta^\star$, and the auxiliary coefficients u, which are projected to centered deviations γ. The fixed coefficient-specific precision determines the scale of each u perturbation before centering. The displayed multinomial log likelihood omits the document-specific combinatorial constant. For a minibatch $\mathcal{B}$ from N training documents, we minimize the Monte Carlo per-document negative ELBO

$$-\frac{1}{|\mathcal{B}|}\sum_{d\in\mathcal{B}}\Big[\log p(\mathbf{c}_d \mid \tilde{\theta}_d, \tilde{\beta}^\star, \tilde{\gamma}) - w\,\mathrm{KL}_{\theta_d}\Big] + \frac{w}{N}\left(\mathrm{KL}_{\beta^\star} + \mathrm{KL}_u\right),$$

where w is the annealed KL weight. This separates the document-level KL from the two global KL terms. Setting $\gamma \equiv 0$ yields the mVTM likelihood.

mFETM is trained for 80 epochs with minibatches of 256 and an exponentially decaying learning rate from 0.01 to 0.001; its KL weight rises linearly from zero to one over the first 40 epochs. mVTM is trained for 60 epochs with the same minibatch size and a learning rate from 0.005 to 0.001; its KL warm-up lasts 30 epochs. Both models use the same validation-based selection rule, with performance assessed every five epochs. We reserve 10% of the original training split for validation and select the full-KL epoch with the lowest document-completion perplexity. We then evaluate the selected checkpoint on the test split without further tuning. Downstream topic shares use $\mathrm{softmax}(\mu_{\theta_i})$; held-out probabilities also use the posterior means of $\beta^\star$ and the centered γ.

Algorithm 1 Mechanistic Fixed-Effects Topic Model (mFETM)

Input: concept counts $\{\mathbf{c}_i\}$, environment indicators $\{\mathbf{x}_i\}$; topics K, concepts V, environments E
Output: topic shares $\hat{\pi}_d = \mathrm{softmax}(\mu_{\theta_d})$, global topics $\hat{\beta}^\star$, centered fixed-effect shifts $\hat{\gamma}$
Compute: fit-sample frequencies w_e^I, w_e^P
Draw once: $\tau_{e,k,v} \sim \mathrm{Gamma}(a_\tau, b_\tau)$
Initialize: encoder weights; $\mu_{\beta^\star}, \sigma_{\beta^\star}, \mu_u$ with $\sum_{e\in g} w_e^g \mu_{u_e} = 0$
for each of 80 training epochs **do**
    sample a minibatch of documents d
    sample $\epsilon_\theta, \epsilon_\beta, \epsilon_u \sim \mathcal{N}(0, I)$
    $(\mu_\theta, \sigma_\theta) \leftarrow \mathrm{encoder}(\mathbf{c}_d)$
    $\tilde{\theta}_d \leftarrow \mu_\theta + \sigma_\theta \odot \epsilon_\theta$ ▷ draw topic logits
    $\tilde{\pi}_d \leftarrow \mathrm{softmax}(\tilde{\theta}_d)$ ▷ topic shares
    $\tilde{\beta}^\star \leftarrow \mu_{\beta^\star} + \sigma_{\beta^\star} \odot \epsilon_\beta$ ▷ reparameterize
    $\tilde{u}_{e,k,v} \leftarrow \mu_{u_{e,k,v}} + \tau_{e,k,v}^{-1/2} \epsilon_{u_{e,k,v}}$ ▷ auxiliary Gaussian draw
    $\tilde{\gamma}_e^g \leftarrow \tilde{u}_e^g - \sum_{e'\in g} w_{e'}^g \tilde{u}_{e'}^g$ ▷ identify each block
    $\eta_d \leftarrow \sum_k \tilde{\pi}_{d,k} \mathrm{softmax}\big(\tilde{\beta}_k^\star + \sum_e x_{d,e} \tilde{\gamma}_{e,k}\big)$
    $\log p(\mathbf{c}_d \mid \tilde{\theta}, \tilde{\beta}^\star, \tilde{\gamma}) \leftarrow \sum_v c_{d,v} \log \eta_{d,v}$
    compute the Gaussian KL terms for θ, $\beta^\star$, and u
    average the likelihood and θ KL over the minibatch; divide the global $\beta^\star$ and u KLs by N
    clip the total gradient norm at 1; update trainable parameters with Adam; project μ_u back to weighted mean zero
    if the epoch is a multiple of 5 **then**
        compute validation completion perplexity using the fixed Bernoulli half-split
        once $w = 1$, retain the parameter state whenever validation perplexity improves
    end if
end for
restore the retained full-KL parameter state
return $\hat{\pi}_d = \mathrm{softmax}(\mu_{\theta_d})$, $\hat{\beta}^\star$, and $\hat{\gamma}^g = P_g \mu_u^g$

APPENDIX F: TOPIC GEOMETRY

We compare the models along two geometric dimensions: separation between topic centroids and dispersion among the leading concepts within a topic. Each SAE feature i has a decoder direction $\mathbf{W}_{\mathrm{dec}}[i] \in \mathbb{R}^{3584}$ in Gemma's internal representation. Write $b_k = \beta_k^\star$ for mFETM and $b_k = \beta_k$ for mVTM. We construct each topic embedding as the feature-weighted average of these directions:

$$\mathbf{e}_k \;=\; \sum_{i=1}^{V_{\mathrm{kept}}} \mathrm{softmax}(b_k)_i \cdot \mathbf{W}_{\mathrm{dec}}[i] \;\in\; \mathbb{R}^{3584}.$$

All metrics use this space and the same $V_{\mathrm{kept}} = 4{,}735$ features, constructed once independently of K. This provides a model-native geometric comparison using the fitted topic weights and SAE decoder space, without external embeddings or additional language-model inference.

F.1. Pairwise Topic Distance. For each model, we compute pairwise cosine distance between topic embeddings:

$$d(k, k') \;=\; 1 - \cos\big(\mathbf{e}_k, \mathbf{e}_{k'}\big),$$

and average over all $\binom{K}{2} = 1{,}225$ unordered pairs. A larger mean indicates greater separation among topic centroids.

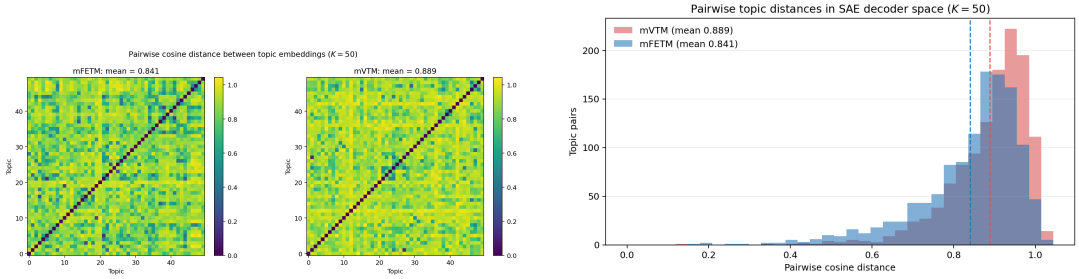


FIG A1. *Pairwise cosine distances between topic embeddings at $K = 50$. mFETM topic centroids are closer on average than mVTM topic centroids.*

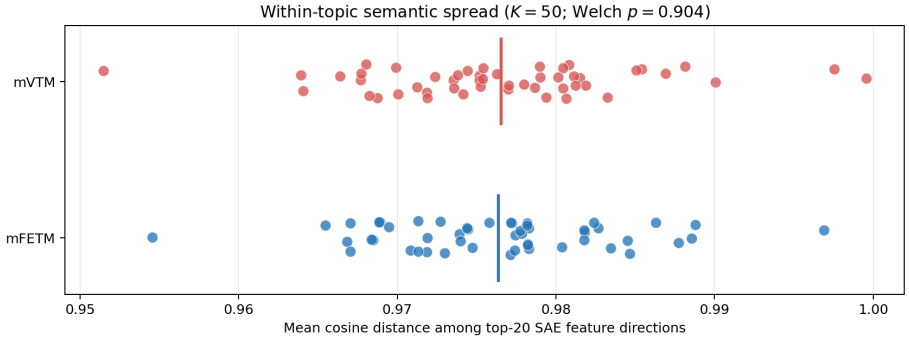


FIG A2. *Within-topic semantic spread at $K = 50$. Each dot is one topic; vertical lines mark each model's mean. The two distributions nearly coincide.*

The mean distance is 0.841 under mFETM (sd 0.127) and 0.889 under mVTM (sd 0.099), a gap of -0.048. mFETM topic centroids are therefore closer on average (Figure A1). Because pairs share topics, the 1,225 distances are not independent; we therefore report their distribution without treating the pair count as the inferential sample size.

F.2. Within-Topic Semantic Spread. For each topic, we take the 20 concepts with the largest weights under $\mathrm{softmax}(b_k)$. Writing $\mathbf{w}_i$ for concept i's decoder direction, topic k's spread is the average cosine distance over all pairs of its leading concepts,

$$S_k \;=\; \mathrm{mean}_{i<j}\,\big[\,1-\cos(\mathbf{w}_i,\mathbf{w}_j)\,\big],$$

with i and j ranging over 190 pairs. A low S_k indicates a tightly clustered topic; a high value indicates greater geometric dispersion. The distance is bounded between 0 and 2 and lies between 0 and 1 in these data.

Mean spread is 0.9764 under mFETM and 0.9766 under mVTM, a gap of -0.0002 (Welch $p = 0.904$). The within-topic geometry is effectively unchanged across models (Figure A2).

On these measures, mFETM changes the geometry between topics rather than broadening individual topics: within-topic spread is virtually unchanged, while topic centroids are closer together. The centroid shift is consistent with less environment-specific extremity, although greater overlap and redundancy remain observationally equivalent explanations.

APPENDIX G: STEERING FLUENCY CHECK

The faithfulness result is informative only if stronger steering does not simply make the continuation less plausible. We therefore generate a continuation for each of Topics 1–10 at

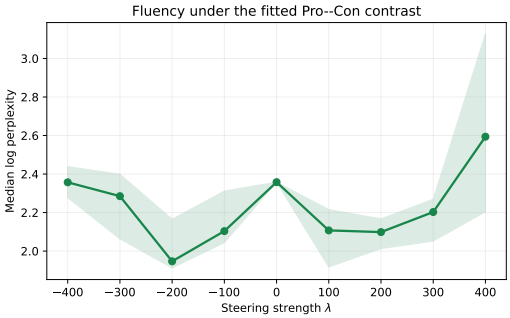


FIG A3. *Continuation fluency along the fitted Pro-minus-Con direction. The line reports median log perplexity under unsteered Gemma-2-9B across ten topic-specific generations; the band is the interquartile range. Generation settings match the main steering exercise.*

every $\lambda \in \{-400, -300, \ldots, 400\}$ using the same prompt and generation procedure as the main exercise. Each continuation is then scored under unsteered Gemma-2-9B. We report the median natural log perplexity across topics with an interquartile band; using the median limits the influence of occasional high-perplexity generations.

Median log perplexity remains between 1.95 and 2.59 across the full path. It equals 2.36 at the direction-specific ablation, falls over most intermediate positive strengths, and reaches 2.59 at $\lambda = 400$. The intervention therefore changes the Pro–Con likelihood gap without producing a general collapse in continuation fit, although the dispersion widens at the strongest positive setting.

APPENDIX H: REMOVAL-SWEEP ROBUSTNESS AND DECOMPOSITION

The main text reports normalized fit retention because it is directly comparable across models. For the topic-level tables, we estimate one normalization-invariant contrast per topic: the row topic relative to the equally weighted average of the other 49 topics. Re-estimating the logits with the second-largest-share topic omitted instead of the largest-share topic leaves both models' BH-selected sets unchanged; coefficient and standard-error differences are below 2×10^{-5}. This verifies that the reported topic associations do not depend on the fitting baseline.

With all topics included, mFETM has a pseudo-R^2 of 0.158, compared with 0.136 for mVTM. The firm-clustered, BIC-corrected Vuong test (Vuong, 1989) favors mFETM but does not reject equal relative fit at the 5% level ($z = 1.60$, two-sided $p = 0.109$). In a combined logit, each model's 49-topic block remains jointly significant conditional on the other (Wald $p = 2.5 \times 10^{-20}$ for mFETM and $p = 2.4 \times 10^{-15}$ for mVTM), indicating that both contain incremental firm-level information.

For the removal exercise, topics are ranked separately within each model by the larger of their absolute pro/con Cohen's d and standardized cross-industry range. Scores use the full document corpus, with industry ranges restricted to cells containing at least 1,000 documents. The nominal five-point cutoffs remove 2, 5, 8, 10, 12, and 15 of the 50 topics. After each removal, we renormalize the remaining shares and refit the year-fixed-effects logit. At the nominal 5% cutoff, mFETM retains 97.1% of its full-model fit and mVTM 99.5%; from 10% through 30%, mFETM retains substantially more.

Figure A4 reports the companion firm-clustered, BIC-corrected Vuong statistic against a random-removal benchmark. At the nominal 5% cutoff, the targeted statistic is $z = 1.32$ and remains inside the random range. From 10% through 30%, it ranges from 2.68 to 3.38 and exceeds the pointwise 95% random-removal range at every cutoff. From 10% onward,

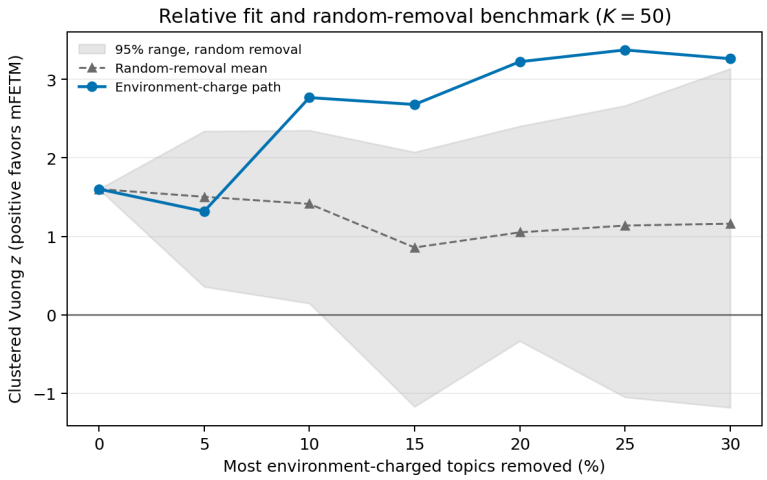


FIG A4. *Vuong comparison after removing environment-charged topics, $K = 50$. The targeted line is the firm-clustered, BIC-corrected Vuong z; positive values favor mFETM. The dashed line and band show the mean and pointwise 2.5th–97.5th percentile range from 20 random-removal draws at each positive cutoff; the 0% point is deterministic. Cutoffs are nominal and use the rounded topic counts reported in the text.*

targeted removal shifts relative fit toward mFETM more than removal of the same number of randomly selected topics.

The complementary polarity-only sweep removes one separately ranked topic from each model at every step. The path can be nonmonotone because either removed topic may also carry firm-level signal. Figure A5 pairs the Vuong statistic with both fit paths, allowing movements in relative fit to be traced to losses in mFETM, mVTM, or both. The environment-charge sweep is the primary specification because it evaluates both fixed-effect blocks; the polarity-only sweep isolates the contribution of review-side association.

REFERENCES

LIEBERUM, T., RAJAMANOHARAN, S., CONMY, A., SMITH, L., SONNERAT, N., VARMA, V., KRAMÁR, J., DRAGAN, A., SHAH, R. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *arXiv preprint arXiv:2408.05147*.

RAJAMANOHARAN, S., LIEBERUM, T., SONNERAT, N., CONMY, A., VARMA, V., KRAMÁR, J. and NANDA, N. (2024). Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. *arXiv preprint arXiv:2407.14435*.

VUONG, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57** 307–333. https://doi.org/10.2307/1912557

ZHENG, C., BELTRAN-VELEZ, N., KARLEKAR, S., SHI, C., NAZARET, A., MALLIK, A., FEDER, A. and BLEI, D. M. (2025). Model directions, not words: Mechanistic topic models using sparse autoencoders. *arXiv preprint arXiv:2507.23220*.

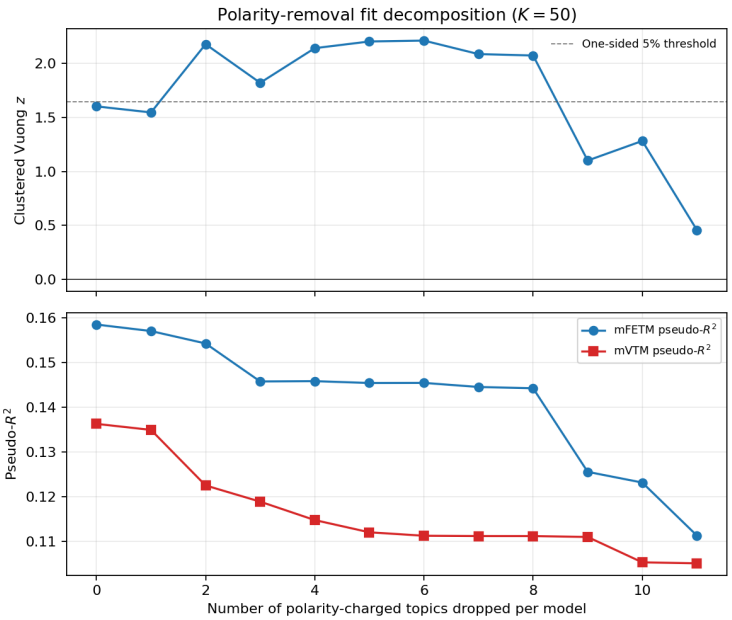


FIG A5. *Polarity-removal fit decomposition at $K = 50$. Top: Vuong z as the most polarity-charged topics are removed; the dashed line marks the one-sided 5% threshold, $z = 1.645$. Bottom: each model's pseudo-R^2 over the same sweep. At each step, each model loses its own next most polarity-charged topic, so movements in z are interpreted jointly with both fit paths.*